UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO

JEAN FELIPE MARTINS DA COSTA

Word Association Retrieval (WAR): Um Método probabilístico para recuperação de termos associados em textos multissegmentados

Dissertação apresentada como requisito parcial para a obtenção do grau de Mestre em Ciência da Computação

Orientador: Prof. Dra. Renata Galante

Porto Alegre
2022

CIP — CATALOGAÇÃO NA PUBLICAÇÃO

Costa, Jean Felipe Martins da

Word Association Retrieval (WAR): Um Método probabilístico para recuperação de termos associados em textos multissegmentados / Jean Felipe Martins da Costa. – Porto Alegre: PPGC da UFRGS, 2022.

158 f.: il.

Dissertação (mestrado) – Universidade Federal do Rio Grande do Sul. Programa de Pós-Graduação em Computação, Porto Alegre, BR–RS, 2022. Orientador: Renata Galante.

1. Recuperação de Informação. 2. Regras de Associação. 3. Base de Dados. 4. Mineração de Dados. 5. Mineração de Processos. 6. Ranking de Termos. I. Galante, Renata. II. Título.

“The roots of education are bitter,
but the fruit is sweet.”
— ARISTOTLE

RESUMO

Esta dissertação apresenta o WAR - Word Association Retrieval, um novo método probabilístico para recuperação de termos associados em textos multissegmentados. O método WAR trabalha com o cenário de recuperação de palavras em um contexto único, permitindo quantificar a correlação dos termos mesmo estando em segmentos distintos. Este cenário é quando um evento ou processo possui várias etapas de descrições textuais, podendo assim ser representado de forma tabular, onde cada coluna representa uma etapa (segmento) e o processo total (contexto) é representado em uma linha de uma tabela, ou seja, vários segmentos de um mesmo contexto. Como exemplo de dois segmentos é a capacidade de buscar associações como nos segmentos de texto de descrição inicial com a descrição final, de uma pergunta e a resposta, da descrição de uma consulta médica e a conduta do médico descrita etc. Como em recuperação de informações o método Bag Of Words busca os documentos associados apenas contando as ocorrências. Já o método TF/IDF e suas variações aplicam pesos ponderados nas ocorrências o que por sua vez apresentam resultados melhores. Nas regras de associação temos o algoritmo clássico Apriori que também apenas contabiliza as ocorrências, mas não aplicada pesos ponderados de associação. Assim o WAR apresenta como solução de pesos ponderados de associação. Este método permite buscar as associações dos termos entre os segmentos de um texto, evitando o *overfitting* das técnicas modernas e a visão limitada do Apriori. Desta forma, usando lógica de pesos ponderados já aplicado na recuperação de informação nas regras de associação, o método WAR propõe duas matrizes de associação multidimensionais para termos de todas as fontes para apresentar uma classificação em forma de ranque dos termos em resposta às palavras de pesquisa. O método WAR foi aplicado em uma base de dados artificial como análise prévia e posteriormente na coleção de dados de teleconsulta médica real, e permitiu identificar resultados de associação relevantes pesquisados em vários estágios do processo de atendimento primário à saúde. Assim, com uso da base de prontuários médicos, foi possível retornar uma classificação por meio de uma lista ordenada de termos associados em relação à busca.

Palavras-chave: Recuperação de Informação. Regras de Associação. Base de Dados. Mineração de Dados. Mineração de Processos. Ranking de Termos.

Word Association Retrieval (WAR): Probabilistic method for associated term retrieval in multi-segmented texts

ABSTRACT

This thesis presents the WAR - Word Association Retrieval, a new probabilistic method for retrieving associated terms in multi-segmented texts. The WAR method works with the scenario of word retrieval in a single context that encompasses categorical data or free text, allowing quantification of the correlation of terms even when in distinct segments. This scenario is when an event or process has several steps of textual descriptions, so it can be represented in a tabular form, where each column represents a step (segment). The full process (context) is described in a row of a table, resulting in several segments of the same context. An example of two segments is the ability to search for associations as: in the text segments of initial description with the final description, of a question and answer, of the description of a medical appointment and the described doctor's conduct, and etc. As in information retrieval, the Bag Of Words method searches the associated documents just by counting the occurrences. On the other hand, the TF/IDF method and its variations apply weighted occurrences, which in turn present better results. In association rules, we have the classic Apriori algorithm that also only counts occurrences, but does not apply association weights. So the WAR presents as a weighted association weights solution. This method allows searching for the associations of terms between the segments of a text, avoiding the overfitting of modern techniques and the limited vision of Apriori. Thus, using weighted logic already applied in the retrieval of information in association rules, the WAR method proposes two multidimensional association matrices for terms to present a ranking of terms in response to the search words. The WAR was applied to an artificial database for prior analysis and later using a medical teleconsultation database and allowed to identify relevant association results searched at various stages of the primary health care process. Thus, using the medical records database, it was possible to return a classification through an ordered list of terms associated with the search.

Keywords: Information Retrieval, Term Ranking, Association Rules, Process Mining, Data Mining, Database.

LISTA DE ABREVIATURAS E SIGLAS

APS	Atenção Primária à Saúde
CF	Combinatory Frequency
CID	Classificação Internacional de Doenças
CIAP	Classificação Internacional de Assistência Primária
CSV	Comma-Separated Values
ED	Entailment Distribution
EP	Entailment Probability
EF	Entailment Frequency
IDF	Inverse Document Frequency
IR	Information Recover
JSON	JavaScript Object Notation
MMRT	Multidimensional Matrix of Related Terms Matriz Multidimensional de Termos Relacionados
MMRI	Multidimensional Matrix of Reciprocal Implication Matriz Multidimensional de Implicação Recíproca
NLP	Natural Language Processing
PSV	Pipe-Separated Values
TD	Term Distribution
TF	Term frequency
WAR	Word Association Recovery
XML	eXtensible Markup Language

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

A área de computação conhecida por Recuperação de Informação ou, em Inglês, *Information Retrieval* (IR) se apresenta como uma ciência que lida com a procura de informações em bases de dados textuais. Tais fontes de dados podem ser caracterizadas por texto ou metadados aplicados em documentos de texto.

A estrutura de um método de IR consiste na utilização de uma função de ranqueamento, ou seja, relacionar os documentos da base de dados com os termos de uma consulta por meio de um escore de relevância, o ranque. Este processo é dividido em duas frentes: a concepção de um arcabouço lógico para representar os documentos, geralmente adotando termos de indexação como base de apontamento para os documentos da base pesquisada; assim como a definição da função de ranqueamento, a qual computa a relação entre os documentos com a consulta efetuada (BAEZA-YATES; RIBEIRO-NETO, 2013).

Conforme a natureza dos documentos a serem recuperados, diversas técnicas vêm se aperfeiçoando para melhor se adequarem na interpretação dos conteúdos e assim retornarem ranques mais pertinentes a busca. Hoje contamos com estratégias de recuperação para textos não estruturados, semiestruturados, oriundos da rede de hipertexto (internet) e documentos multimídia como som, vídeo, imagem entre outros.

A recuperação clássica, processamento em textos não estruturados, interpreta o conteúdo textual simplesmente como uma sequência de palavras, enquanto o modelo semiestruturado identifica componentes estruturais do texto, segmentos de uma parte integral do documento (MELLO, 2002).

Os registros textuais são ricas fontes de informações, entretanto, dois desafios são necessários para a extração de valor. Primeiramente, é fundamental compreender o formato dos registros para estruturar os dados, seguido da lógica necessária para a recuperação as informações de forma mais eficiente ao contexto da busca do usuário (MELLO, 2000).

1.1 Desafio da pesquisa

Na definição desta pesquisa, a função de IR uma estrutura multissegmentada, ela carece de uma análise diferenciada a qual considere as associações entre as subdivisões de texto dos documentos e sua interação com os termos da busca.

Para muitas implementações, os modelos que evoluíram a partir dos métodos clássicos são aplicados analisando os documentos em seu contexto único, mas não permitindo a busca por relação entre os termos dos segmentos.

Registros podem ser fracionados em segmentos, entretanto é necessário que os segmentos possam ser analisados de forma individual, mas respeitando que eles fazem parte do contexto do registro. Como exemplo de dois segmentos é a capacidade de buscar associações como nos segmentos de texto de descrição inicial com a descrição final, de uma pergunta e a resposta, da descrição de uma consulta médica e a conduta do médico descrita etc. Quando visualizamos o registro como um processo em que os segmentos representam as atividades, analisar as associações permite a mineração do processo, e para isso o método WAR viabiliza dentro do campo de pesquisas textuais, a busca de termos associados.

Seguindo os modelos atuais, a função de recuperação utiliza os termos que compõem a busca cruzando com o arcabouço lógico que associa termos com os documentos, independente se são estruturados ou semiestruturados, ou seja, um indexador de termos para documentos. Tal abordagem resulta em um ranque de documentos relacionados aos termos de busca, entretanto uma análise a nível de associação de palavras de segmentos distintos não pode ser realizada de maneira satisfatória por meio das atuais estratégias.

Na área da saúde, a teledermatologia vem tornando-se um elemento importante no cuidado à saúde ao redor do mundo nas últimas duas décadas, em linha com o crescimento em tecnologia da informação (HS, 2002).No Brasil, com suporte do Ministério da Saúde, em 2005 foram criados nove grupos de telemedicina ligados a universidades públicas, entre elas a Universidade Federal do Rio Grande do Sul (UFRGS) com o projeto TelessaúdeRS (2007, 2007).

Neste contexto, foi implementado em 2017 no Rio Grande do Sul, pelo projeto TelessaúdeRS, o aplicativo DermatoNET, uma aplicação específica para o telediagnóstico em dermatologia. As informações da solicitação contêm dados textuais descritivos sobre o caso clínico. A descrição da solicitação é lida e avaliada pelo médico teleconsultor. Se as informações adquiridas forem suficientes, o dermatologista prossegue com o telediagnóstico, elaborando um texto com descrição das lesões com base na inspeção das fotos, sugerindo uma hipótese diagnóstica principal, classificando o caso de acordo com o Código Internacional de Doenças (CID 10) e elaborando um texto livre com a conduta sugerida e finalmente decidindo por sugerir encaminhamento para avaliação presencial com dermatologista ou por manutenção na Atenção Primária à saúde.

Essa base de dados contém registrado todo o processo de avaliação de teledermatologia, desde o início com a dúvida gerada pelo médico da APS até à emissão de um diagnóstico mais provável e à sugestão de conduta inicial. Essas informações contém um potencial de revelação e de geração de conhecimento diversificado e abrangente. O principal problema identificado é conseguir associar os termos ocorridos em diferentes etapas do processo de forma a entender a associação entre um termo de uma atividade inicial em relação a uma atividade intermediária ou final, ou vice-versa.

Desta forma, esta pesquisa tem como objetivo apresentar um método recuperação de termos associados em registros multissegmentados a fim de, responder se há a possibilidade de associar termos considerando peso ponderado ao invés da ocorrência de associação como no caso do Apriori. Para isso, a pesquisa se orientou pelos seguintes objetivos específicos: recuperar os termos de maior relevância com base nos termos da busca; ranquear associações dos termos considerando associações para múltiplos registros; e considerar o registro como uma unidade que não pode ser dissociada, entretanto considerando que os dados estão em múltiplos segmentos.

1.2 Solução da proposta

Para permitir identificar o quanto um termo está associado com os termos de outro segmento, é necessário mudar a estrutura da função de recuperação, implementando um indexador que faz o apontamento de termos de busca com termos de segmentos.

O método *Word Association Recovery* (WAR) proposto nesta dissertação, introduz uma abordagem de regras de associação e recuperação de informações que permite a análise em documentos de um ou mais segmentos. Tendo como entrada um vetor de consulta, o processamento da função resulta na implicação dos termos de todos os segmentos presentes por uma ordem de relevância, para isso, utilizando um indexador multidimensional que implementa cálculos de regras de associação.

A estruturação do um ranque que ordena o retorno conforme seu valor se dá pela proeminência do termo em comparação aos demais. Assim, termos que são vulgarmente encontrados no universo de documentos analisados apresentam uma relevância menor aos demais, pois sua associação é caracterizada de forma menos determinista nas outras ocorrências. Em registros onde cada segmento representa uma etapa de um processo, se torna possível recuperar a informação do quanto os termos de uma determinada etapa estão associados com os termos das outras.

O método WAR foi avaliado através de dois blocos de experimentos: o primeiro com dados sintéticos para testar todas as características e peculiaridades do método e o segundo com dados reais do TelessaúdeRS.

Esta dissertação foi desenvolvida em parceria entre dois programas de pós-graduação da UFRGS (Universidade Federal do Rio Grande do Sul), sendo os programadas PPGC (Programa de Pós-graduação em Computação) e PPGCM (Programa de Pós-graduação em Ciências Médicas). Este trabalho apresenta a análise sobre o viés computacional da técnica elaborada de recuperação de termos associados.

Ao executar os experimentos, foi possível validar que o método proposto possui um comportamento de regras de associação, não apresentando as limitações do Apriori pois apresenta pesos ponderados como das técnicas de recuperação de informações. Assim, o método WAR fornece uma forma diferente aos termos de associação clássicos.

1.3 Estrutura da dissertação

Os próximos capítulos se dividem da seguinte forma: O capítulo 2 introduz detalhadamente os conceitos utilizados ao longo da dissertação, sendo eles referentes ao uso de técnicas e métricas de recuperação de informações e regras de associação. O capítulo 3 traz estudos já realizados na área da computação como também na parte de recuperação de informações e regras de associação em bases do mesmo segmento. O capítulo 4 apresenta em detalhes a base utilizada e o processo a qual ela representa. O capítulo 5 descreve o método proposto nesta dissertação, bem como as atividades e métricas que o compõe. O capítulo 6 avalia o método proposto com dados sintéticos e reais, apresentando os experimentos e os seus resultados. Por fim, o capítulo 7 traz as conclusões obtidas, além de apresentar as possibilidades de melhoria, possíveis implementações e de trabalhos futuros voltados a estender o tópico da pesquisa.

2 FUNDAMENTAÇÃO TEÓRICA

Neste Capítulo são apresentados os conceitos base para a compreensão do método proposto nesta dissertação. Inicialmente, apresenta-se uma revisão dos conceitos nas áreas de pesquisa relacionadas, seguido da formulação para a implementação de técnicas para recuperação e associação de informações textuais. Por fim, são apresentados os benefícios e a aplicabilidade dessas técnicas em bases de dados para a elaboração de um ranque de associação de elementos de texto.

2.1 Análise de dados

Como decorrência da crescente digitalização, conjuntos de dados estão disponíveis cada vez em maior quantidade e consequentemente fornecendo maior número de coleções de registros textuais em diferentes formatos e necessidades de análise. Contando apenas com palavras, os métodos de mineração de texto carecem de uma atividade de pré-processamento que interprete o texto e forneça uma representação numérica que apoie a recuperação dessas informações (WEISS et al., 2005).

2.1.1 Quantificação de valores textuais

Os métodos de mineração de dados precisam que os dados sejam fornecidos seguindo uma estruturação ou se torna necessário adicionar uma atividade prévia de transformação dos dados originais.

Com os dados estruturados, os registros iniciam um processo que permite uma perspectiva quantitativa dos dados. Esta atividade é um dos principais temas de suporte a mineração de texto, pois é responsável por aplicar métodos de quantificação dos registros, seja por meio da utilização de informações como a frequência de um termo na coleção de documentos ou de algoritmos de aprendizado de máquina (WEISS et al., 2005).

2.2 Mineração de texto

A Mineração de Texto, *Text Mining* em inglês, busca a produção de conhecimento pela aplicação de técnicas sobre registros textuais. Tais técnicas extraem informações

pela busca de padrões, associações, categorização, agrupamento, produção de taxonomias, análise de sentimentos e modelagem de relações entre entidades (BERRY; CASTELLANOS, 2008).

A análise de um conjunto de documentos textuais escritos em linguagem natural, tem como aplicação: categorizar o texto, associando seu conteúdo a rótulos; ou desenvolver um índice de pesquisa com a utilização das informações extraídas. Modelos de Processamento de Linguagem Natural, em inglês *Natural Language Processing* (NLP), utilizam métodos analíticos para a classificação do texto por meio do reconhecimento de padrões (MEADOW; BOYCE; KRAFT, 2000).

Outra aplicabilidade está no estudo do vocabulário dos documentos e conjunto de termos. Esta abordagem é composta de variadas técnicas de recuperação e análise lexical a fim de estruturar a frequência da distribuição de palavras, coletando as informações nos textos para a aplicação de um método baseado no desenvolvimento de um índice de termos (BERRY; CASTELLANOS, 2008).

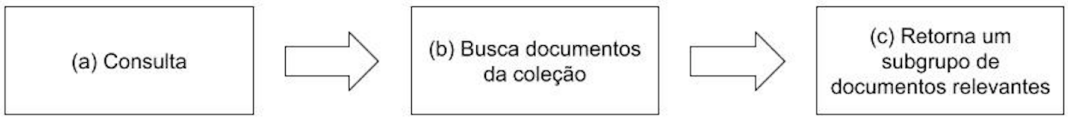
Nos modelos de mineração textual, é possível aplicar técnicas de recuperação de informações e análise preditiva de documentos. Para a implementação destes modelos, são aplicados práticas como: classificação de documentos; clusterização; e extração de informações (WEISS et al., 2005).

2.2.1 Modelos de mineração textual

Na mineração de texto, conta-se com dois modelos, um para a recuperação de informações e outro voltado para a categorização em um modelo de mineração preditiva. Modelos de melhor precisão e performance são elaborados para que as informações textuais possam ser interpretadas e o benefício desse conhecimento seja aplicado no documento (WEISS et al., 2005).

Hoje em dia, a aplicação da recuperação de informações não se restringe a utilização de técnicas clássicas de recuperação de documentos armazenados em bancos de dados, mas também a documentos como páginas Web. Assim, quando o objetivo é a recuperação de documentos por ordem de relevância, as etapas principais deste modelo são: (a) é fornecida uma descrição geral da consulta, (b) a coleção de documentos é pesquisada e (c) subconjuntos de documentos relevantes são retornados, conforme Figura 2.1.

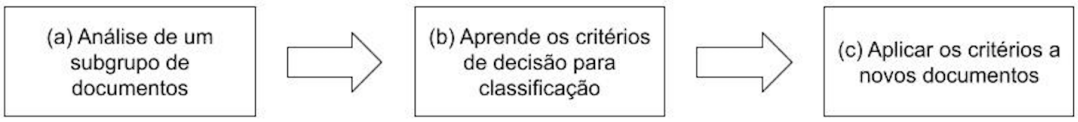
Figura 2.1: Etapas da recuperação de informações



Fonte: Adaptado de (WEISS et al., 2005).

Quando aplicado um modelo de mineração preditiva de texto, o objetivo passa pela aprendizagem dos critérios de uma parte da coleção para a classificação de outros documentos. Este modelo de mineração permite a categorização de documentos com base em uma amostragem, um exemplo são regras de *spam* para e-mail. A implementação deste modelo segue as etapas: (a) examinar uma coleção de documentos; (b) aprender os critérios de decisão para classificação; e (c) aplicar esses critérios a novos documentos. Tal fluxo é apresentado na Figura 2.2.

Figura 2.2: Etapas da mineração de texto preditiva



Fonte: Adaptado de (WEISS et al., 2005).

Em ambos os modelos, dificilmente os valores de entrada correspondem exatamente com os documentos em comparação e, para isso, é aplicado um processamento de similaridade, o qual realiza a associação por meio de um ranque de proximidade.

2.3 Recuperação de informação

A área de computação conhecida por Recuperação de Informações, em inglês *Information Retrieval* (IR), se apresenta como uma ciência que lida com a procura de informações em bases textuais. Algoritmos deste campo de pesquisa buscam encontrar os registros mais pertinentes aos itens de uma consulta (MEADOW; BOYCE; KRAFT, 2000).

Os modelos mais comuns na recuperação de informações funcionam por meio de um mecanismo de pesquisa, onde têm como entrada uma lista de palavras-chave e como retorno é obtida uma listagem de documentos relevantes.

Quando os termos são utilizados em uma consulta, a busca pode ser vista como um pequeno registro no qual tem a sua semelhança medida com os demais documentos da coleção, retornando seus resultados de valores associados com base em uma ordem de relevância elaborada pela função de ranqueamento (WEISS et al., 2005) e (BAEZA-YATES; RIBEIRO-NETO, 2013).

2.3.1 Modelos de Recuperação de Informação

Um modelo de ranqueamento para recuperação de informações é composto por quatro elementos principais: um conjunto de representações lógicas de documentos; um conjunto de representações lógicas das necessidades de informação do usuário, a consulta; o arcabouço ou *framework* responsável pela implementação da abordagem de instrução de ranqueamento; e a função de ranqueamento (BAEZA-YATES; RIBEIRO-NETO, 2013).

A ordenação do retorno é apresentada na forma de um ranque que aproxima os elementos da consulta e os documentos por meio do seu grau de similaridade. Este fator é representado pela pontuação que representa o fator de implicação dos termos da consulta sobre os documentos com base nos pesos gerados pelo *framework*.

Para um documento, o *framework* de ponderação que mapeia o número de ocorrência dos termos no documento para um valor real positivo pode ser visto como uma sumarização quantitativa deste documento. Na abordagem conhecida como Saco de Palavras, *Bag of Words* (BoW) em inglês, a ordem dos termos do documento é ignorada, considerando apenas o número de ocorrências de cada termo. Por reter apenas a informação da quantidade de ocorrências de cada termo, os documentos de conteúdo “Maria é mais veloz que João” são considerado idêntico a “João é mais veloz que Maria” (MANNING; RAGHAVAN; SCHUTZE, 2008).

2.3.2 Modelos de visão lógica dos documentos

Os termos que compõem os textos dos documentos são utilizados como base para a associação entre a consulta e os documentos. Entretanto, é possível que o conteúdo do documento siga uma abordagem de representação lógica no qual um conjunto de palavras representem o seu conteúdo, um exemplo é a representação dos termos por meio de *tokens*.

Quando uma representação lógica de um documento conta com a utilização integral do conjunto de palavras do documento, esta representação é nomeada como *full-text*. A adoção desta técnica resulta em um alto custo computacional e, quando empenhado sobre grandes conjuntos de dados, a opção de uma representação lógica mais sucinta é uma alternativa a ser considerada (BAEZA-YATES; RIBEIRO-NETO, 2013).

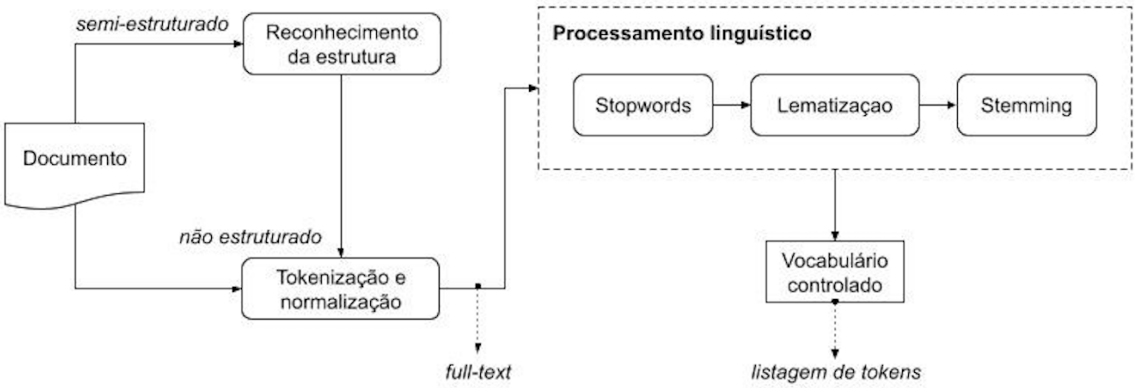
Técnicas de controle da variação do vocabulário tem como objetivo reduzir os efeitos de ambiguidade das representações lógicas, seja no momento da criação de um registro, da pesquisa ou de ambas. Controlar o vocabulário significa reduzir o número de valores possíveis usados como atributos de referência (MEADOW; BOYCE; KRAFT, 2000).

Desta forma, a utilização de técnicas de pré-processamento permite alterar a representação do texto de um documento para uma visão lógica. Esta visualização pode contar com a utilização integral dos termos ou com uma listagem sucinta, a qual pode ser utilizada para a elaboração de um índice.

2.3.3 Transformação do texto de entrada

A definição do vocabulário de termos utilizados pelo sistema é resultado dos processos da transformação do texto de entrada. Tais práticas estruturam a visão lógica do conteúdo de um documento e passam pela alteração e diversas técnicas que buscam normalizar, padronizar e otimizar o desempenho da recuperação com base nos termos nele contido (MANNING; RAGHAVAN; SCHUTZE, 2008). Este fluxo segue ilustrado na Figura 2.3.

Figura 2.3: Visão lógica da transformação do texto de um documento



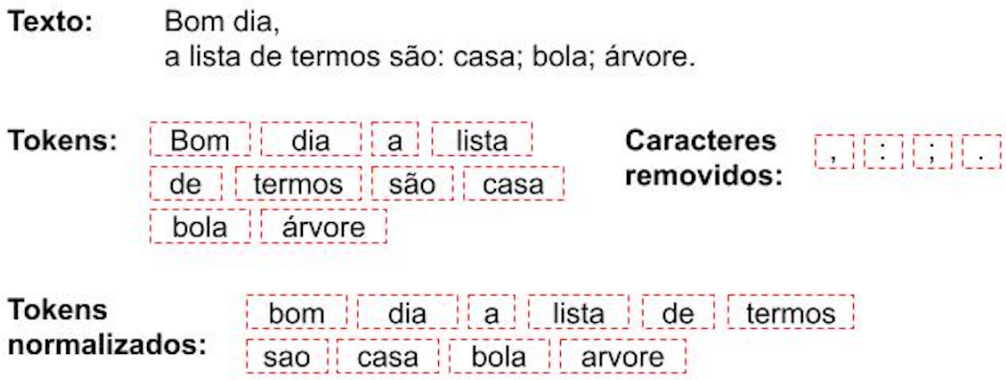
Fonte: Adaptado de (WEISS et al., 2005).

Tendo em vista o conteúdo do documento, é necessário que seja identificado o formato de entrada, sendo não estruturado ou semiestruturado. Caso o documento possua uma estrutura pré-definida de texto, é necessário que um modelo siga a estrutura para que o texto seja interpretado corretamente e assim siga para a etapa de tokenização e normalização.

2.3.3.1 Tokenização e Normalização

A etapa de tokenização tem como responsabilidade dividir o texto em pedaços chamados *tokens* e, neste processo, algumas alterações podem ocorrer, como a remoção e alteração de caracteres, conforme exemplificado na Figura 2.4, (MANNING; RAGHAVAN; SCHUTZE, 2008).

Figura 2.4: Processo de tokenização e normalização



Fonte: Dos Autores.

A criação dos *tokens* geralmente é derivada do resultado de normalização baseado nos termos encontrados no texto. Tais atividades de normalização contam com a remoção de acentuação e a alteração do formato dos caracteres (maiúsculo ou minúsculo) para fins de padronização. Desta forma, um *token* é composto de uma sequência de caracteres que representa um termo de um ou mais documentos da coleção e, quando compreendidos, representa uma unidade semântica útil para o processamento.

2.3.3.2 Processamento linguístico

A etapa de processamento linguístico é composta por múltiplas atividades que resultam na otimização da listagem final de *tokens*. Tais ações são complementares e são baseadas nas particularidades de cada idioma ou coleção de documentos. O objetivo está em remover elementos não relevantes e agrupar múltiplas representações em menos *tokens* (MANNING; RAGHAVAN; SCHUTZE, 2008).

Em um vocabulário existem palavras que são comumente encontradas, pois são base na estruturação das sentenças ou estão vulgarizadas no domínio. Para métodos de mineração de texto, tais termos podem ser ignorados, pois não apresentam nenhuma melhoria relevante na recuperação dos documentos e agregam custo ao processamento (MEADOW; BOYCE; KRAFT, 2000).

Para que os sistemas removam esses termos do grupo de elementos a serem considerados, o processamento conta com a utilização de uma listagem de palavras denominadas *stopwords*. Assim, durante a atividade de processamento linguístico, os termos desta lista são removidos.

Outra aplicação realizada na lista de termos de entrada é a lematização, tal processo conta com um modelo que deflexionam palavras na sua forma reduzida, o lexema. Esta ação tem como objetivo minimizar a variação de termos por significado, removendo as variações como nos casos das conjugações (BAEZA-YATES; RIBEIRO-NETO, 2013).

Como alternativa a utilização de vastas listas para a lematização, a abordagem de *Stemming* realiza a redução das variações de termos com base em regras pré-estipuladas. Após este processamento, cada *Stemmer* resultante representa um *token* e não necessariamente um termo associado a um significado (MANNING; RAGHAVAN; SCHUTZE, 2008).

As regras de *Stemming* aplicam remoções de caracteres nas extremidades das palavras com base nos prefixos ou sufixos, tamanho do radical, caracteres que antecedem os sufixos, dentre outras abordagens. Dependendo da regra a ser aplicada por idioma, esta abordagem permite a redução de variações nas conjugações, flexões gramaticais de quantidade como termos no plural e alterações morfológicas como aumentativos e diminutivos.

A transformação do texto de entrada conta com técnicas de tokenização e normalização, os quais moldam os termos aplicando regras para uniformizar os registros. Em adição, técnicas de processamento linguístico contam com modelos que buscam remover termos irrelevantes e reduzir as variações gramaticais oriundas de: termos compostos; polissemia; sinonímia; flexões gramaticais, plurais; negações; variação do termo (erros e gírias); e alterações morfológicas, aumentativos e diminutivos. Desta forma, é obtido como resultado um vocabulário controlado compondo um universo de *tokens* reduzido e assim facilitando a estruturação de índices.

2.3.4 Processo de indexação

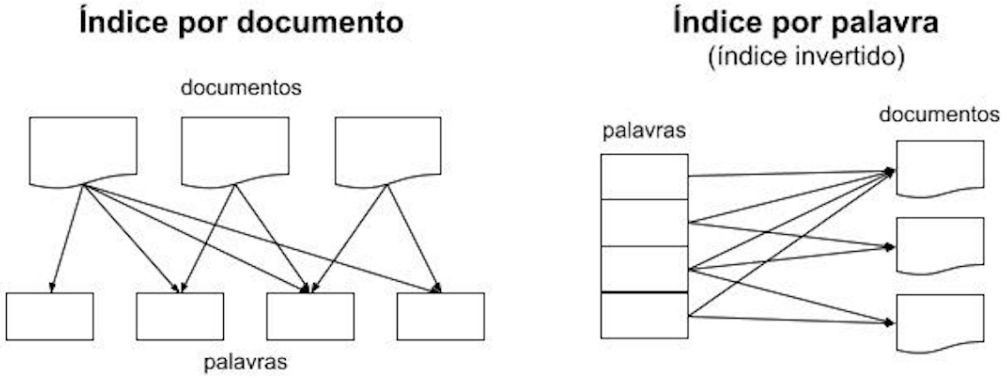
A comparação de todos os documentos de uma coleção é uma operação de alto custo computacional. Contudo, técnicas de IR fazem com que esta operação funcione em um tempo reduzido e são projetadas para reduzir a complexidade da comparação sequencial de todos os documentos (WEISS et al., 2005).

À medida que aumenta a quantidade de informações a serem acessadas, a utilização de mecanismos otimizados de pesquisa se torna cada vez mais importante para a recuperação de documentos relevantes. Algoritmos com o processamento linguístico amenizam os problemas da linguagem natural, tais como: homógrafos, sinônimos, menções passadas, menções negativas e outros. Como resultado deste processamento, os termos da coleção de documentos são resumidos em um vocabulário controlado composto por uma listagem de tokens e por fim são disponíveis para o processo de indexação (MEADOW; BOYCE; KRAFT, 2000).

Com foco na otimização da recuperação das informações, os sistemas de busca de IR aplicam os modelos de indexação de palavras apontando para documentos ao invés de documentos apontando para as palavras (WEISS et al., 2005). Este processo é chamado de lista invertida ou índice invertido e a diferença é ilustrada na Figura 2.5.

O índice invertido é a chave para a eficiência dos sistemas de recuperação de informações, pois a base do desempenho está em identificar os termos da consulta com os termos associados a listagem de documentos.

Figura 2.5: Comparação entre índice de documentos e índice por palavras



Fonte: Dos Autores.

Para o desenvolvimento de um índice invertido, é necessário que sejam realizadas quatro etapas: seleção dos documentos a serem indexados; tokenização e normalização; processamento linguístico; e indexação de cada documento com cada ocorrência do termo (MANNING; RAGHAVAN; SCHUTZE, 2008).

Com um índice invertido estruturado sobre um vocabulário controlado, é possível executar os processos de recuperação e ranqueamento dos documentos com base nos termos da consulta e sua similaridade calculada.

2.3.5 Processo de Recuperação e Ranqueamento

Cada documento da coleção é representado pelo montante de palavras que consta em seu conteúdo. Para otimizar o processo de IR, os documentos contam com uma visão lógica do documento estruturado sobre a construção de um índice invertido de *tokens* resultantes dos métodos de transformação do texto.

Para a recuperação de informações, é fundamental que sejam estabelecidas as técnicas de indexação, recuperação e ranqueamento para que a consulta apresente melhor desempenho na recuperação dos documentos da coleção. Tais etapas trabalham de forma complementar, conforme apresentado na Figura 2.6.

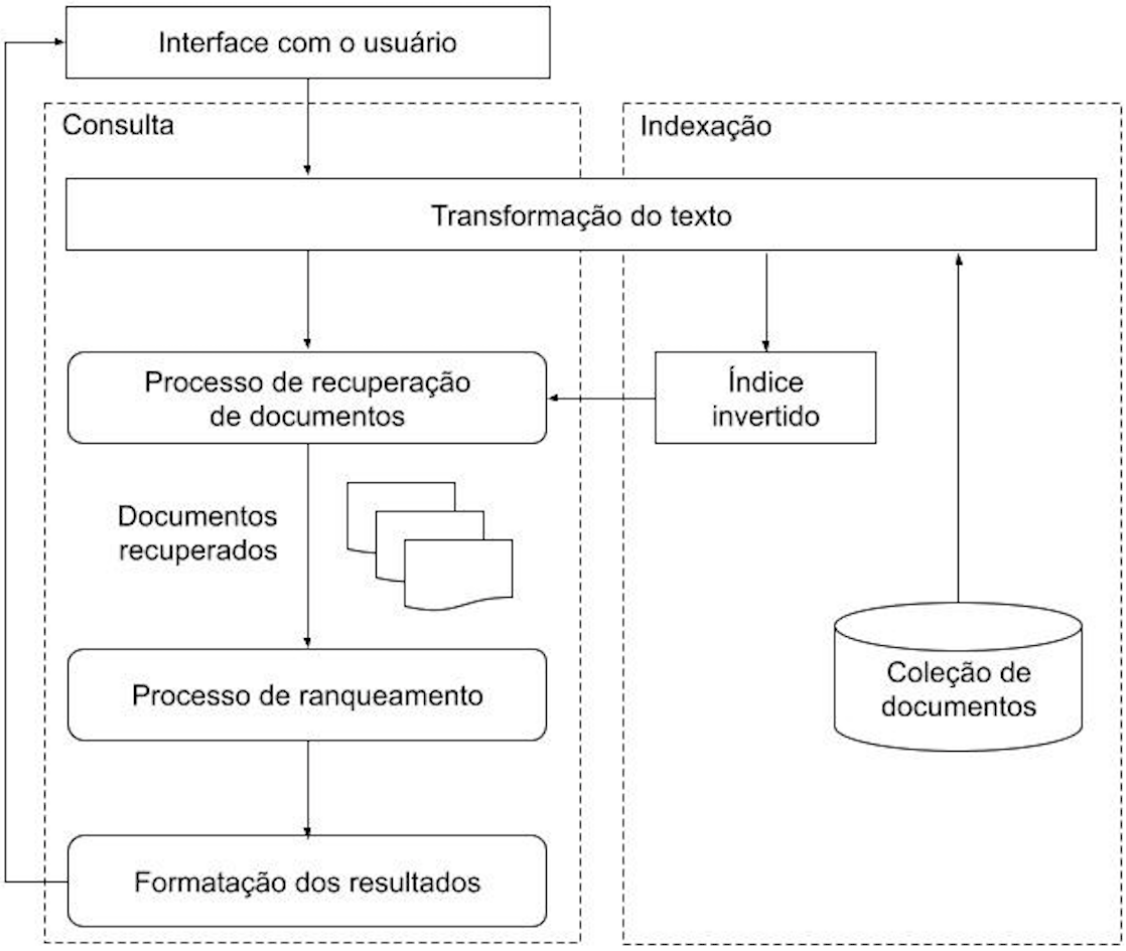
Quando uma consulta é realizada, o processo de recuperação analisa cada uma das palavras e, conforme cada termo é processado, a medida de similaridade com o documento é alterada pela contabilização dos elementos da consulta e documentos onde o termo aparece.

Mesmo contando com poucas palavras de entrada no mecanismo de pesquisa, estes termos são considerados chave na representação dos documentos e assim a semelhança pode ser facilmente calculada (WEISS et al., 2005).

Para a apresentação dos documentos recuperados, apenas a informação binária de relevância pode não ser suficiente, por isso, é necessário a elaboração de um ranque de similaridade. Tais técnicas contam com cálculos que buscam identificar o quão relacionados estão a consulta com os documentos recuperados. Portanto, se uma palavra estiver ausente da consulta, ela não terá contribuição para a medida de similaridade.

A estruturação do um ranque que ordena o retorno conforme seu valor se dá pela proeminência do termo em comparação aos demais. Assim, termos que são vulgarmente encontrados no universo de documentos analisados apresentam uma relevância menor em relação aos outros, pois sua associação é caracterizada de forma menos determinista nas outras ocorrências (WEISS et al., 2005).

Figura 2.6: Processo de indexação, recuperação e ranqueamento de documentos



Fonte: Adaptado de (BAEZA-YATES; RIBEIRO-NETO, 2013).

2.3.6 Ponderação pela Frequência dos Termos - TF/IDF

Proposta por Luhn, a hipótese de Luhn baseia-se na suposição de que o peso de relevância de um documento é aplicado de forma proporcional a frequência do termo (TF) nele contido. Isto significa que, quanto mais frequente o termo de consulta estiver no documento, maior será o peso da sua relevância (BAEZA-YATES; RIBEIRO-NETO, 2013).

Certos termos têm pouco ou nenhum poder discriminador em determinar a relevância. Dessa forma, é introduzido um mecanismo de redução de peso conforme o número de registros aumenta na coleção (MANNING; RAGHAVAN; SCHUTZE, 2008). Termos que são vulgarmente encontrados no universo de documentos analisados apresentam uma relevância menor aos demais, pois a sua associação é caracterizada de forma menos determinista nas outras ocorrências.

Esta afirmação sustenta a interpretação estatística de Sparck Jones chamada de Frequência Inversa de Documentos, em inglês *Inverse Document Frequency* (IDF), (BAEZA-YATES; RIBEIRO-NETO, 2013). Sendo a especificidade de um termo quantificada pela função inversa do número de documentos em que ele ocorre, os termos mais raros têm pesos mais altos porque são mais seletivos, IDF, entretanto os termos mais frequentes dentro de um documento possuem frequências relativamente altas, TF.

Com base nestes conceitos, a atribuição de pesos a termos de uma coleção de documentos genérica, onde não se possui nenhuma informação prévia, apresentam resultados eficazes. Para a implantação destes conceitos, são disponibilizadas equações variantes para o cálculo dos pesos, os quais apresentam esquemas de ponderação para TF, Tabela 2.1, e IDF, na Tabela 2.2.

Tabela 2.1: Variantes dos pesos TF

Esquema de ponderação	*Peso TF*
binário	$\{0, 1\}$
frequência bruta	$f_{i,j}$
normalização logarítmica	$1 + \log f_{i,j}$
normalização dupla 0,5	$0,5 + 0,5\frac{f_{i,j}}{\max_i f_{i,j}}$
normalização dupla K	$K + (1 - K)\frac{f_{i,j}}{max_i f_{i,j}}$

Fonte: Os Autores

Como produto da multiplicação dos pesos ponderados para TF e IDF, é obtido o valor de relevância quando considerados os conceitos supracitados. Desta forma, é possível aprimorar a relação entre os termos da pesquisa com os termos dos documentos, o que resulta em uma melhor ordenação no ranqueamento.

Tabela 2.2: Variantes dos pesos IDF

Esquema de ponderação	*Peso IDF*
unitário	1
frequência inversa	$\log \frac{N}{n_i}$
frequência inversa suave	$\log \left(1 + \frac{N}{n_i}\right)$
frequência inversa máxima	$\log \left(1 + \frac{\max_i n_i}{n_i}\right)$
frequência inversa probabilística	$\log \frac{N-n_i}{n_i}$

Fonte: Os Autores

2.3.7 Métricas de Avaliação

Um passo muito importante no desenvolvimento de um método de IR são as métricas de avaliação (BAEZA-YATES; RIBEIRO-NETO, 2013). Existem diversos métodos que permitem a avaliação comparativa dos resultados a ponto de julgar a qualidade do modelo de classificação de texto. A seguir, são descritos os principais métodos e que são utilizados nos experimentos desta dissertação.

2.3.7.1 Matriz de Confusão

A tabela de contingência realiza o cruzamento de múltiplas variáveis categóricas, as quais são contabilizadas. Para a implementação da recuperação de informações, há um tipo de implementação desta tabela que é a tabela de confusão, ou também chamada de matriz de confusão.

Para os modelos de IR, é possível submeter um conjunto de documentos a um método de classificação e com isso obter a pertinência do documento quando comparado a documentos já observados.

O resultado categorizado, tem sua pertinência apresentada de forma *booleana* e quando tabulado com os documentos já observados, pode ser apresentado em uma matriz de duas linhas e duas colunas, contabilizando: verdadeiros positivos; falsos positivos; falsos negativos; e verdadeiros negativos, permitindo assim uma análise detalhada da precisão do método de classificação, conforme ilustrado na Figura 2.7.

Figura 2.7: Matriz de confusão

		Documentos Observados		
		1	0	
Resultado Categorizado	1	VP (verdadeiro positivo)	FP (falso positivo)	RP (resultado positivo)
	0	FN (falso negativo)	VN (verdadeiro negativo)	RN (resultado negativo)
		R (relevantes) VP + FN	NR (não relevantes) FP + VN	

Fonte: Dos Autores.

2.3.7.2 Acurácia e Erro

A métrica de acurácia, ou *accuracy*, é a representação da proporção de resultados corretos resultante do método de categorização. Esta medida não faz distinção entre positivos e negativos, mas considera o acerto total em relação a todos os positivos e negativos existentes. A equação da acurácia se dá pela seguinte equação:

$$ACC = (VP + VN)/(R + NR)$$

Em oposição a acurácia, a medida de erro, ou *error*, se dá pela fração de documentos resultantes que foram atribuídos a classes incorretas e seu cálculo resulta pela seguinte equação:

$$ERR = (FP + FN)/(R + NR)$$

Uma das desvantagens dessas métricas está no impacto quando uma das classes representa um conjunto muito pequeno de dados, pois uma alteração considerável dentro desta classe pode ser diluída quando analisada em relação aos acertos totais.

2.3.7.3 Precisão e Revocação

Buscando solucionar as implicações negativas das métricas previamente apresentadas, as métricas de precisão e revocação, respectivamente, conhecidas por *precision* e *recall*, também buscam quantificar a qualidade de um classificador textual.

A precisão apresenta a medida de documentos recuperados que são realmente relevantes em relação a todos os documentos recuperados identificados como relevantes, incluindo falsos positivos. Este valor é expresso pela equação:

$$precision = VP/RP$$

O cálculo de revocação apresenta a relação entre o número de documentos relevantes recuperados em relação ao total de documentos relevantes, tendo a sua fórmula representada pela equação:

$$recall = VP/R$$

Para a melhor utilização dessas métricas, ambas podem ser combinadas para elaborar outra métrica de qualidade sobre os documentos recuperados.

2.3.7.4 Medida-F e F1

A medida-F, ou em inglês *F-measure*, combina as métricas de precisão e revocação para quantificar a qualidade geral do modelo, mesmo quando uma das classes possui um conjunto muito pequeno de documentos. Tal medida trabalha com uma variável de ponderação a qual atribui o peso entre as métricas. Representado por α (alpha), onde quando igual a 0, apenas a precisão é considerada, já quanto maior o valor, maior o peso associado a revocação. Esta medida é expressa pelo seguinte cálculo:

$$F_\alpha = \frac{\left(\alpha^2+1\right)\times precision \times recall}{\alpha^2 \times precision + recall}$$

Uma variação desta medida considera pesos iguais para precisão e revocação e, para isso, considera a variável de ponderação igual a 1. Esta medida é chamada de Medida-F1, ou *F1-Score*, e é obtida pela equação:

$$F_1 = \frac{2 \times precision \times recall}{precision + recall}$$

2.4 Mineração de regras de associação

A Mineração de Regras de Associação, ou *Association Rule Mining* (ARM) em inglês, tem como objetivo examinar o conteúdo de uma base de dados e encontrar associações entre dados. Tendo como premissa principal encontrar elementos que implicam na ocorrência de outros itens, ou padrões frequentes em uma coleção analisada (BRAMER, 2013).

Os modelos populares de ARM utilizam a frequência de conjuntos de itens para quantificar o nível de associação, (AGGARWAL, 2015). Sendo inicialmente proposto no contexto de análise dos dados da cesta de supermercado, esta aplicação se propôs a identificar itens de compras que estão relacionados a conjuntos de itens já comprados. Entretanto, com a representação de dados textuais geralmente é provida em uma listagem de termos, a mineração de padrões também pode ajudar a identificar palavras-chave co-ocorrentes.

Alicerçado no primeiro caso de aplicação, a terminologia postulada para este segmento de estudo segue a analogia da cesta de mercado (AGGARWAL, 2015). Dessa forma: um item (*item*) referência um elemento do universo analisado; o conjunto de *itens* representam um grupo ordenado, os quais geralmente seguem uma ordenação lexicográfica e formam os grupos de itens (*itemsets*). Quando analisado o grupo de ocorrências registradas, cada transação (*transaction*) registra os *itens* de compra em um *itemsets*.

Na análise da cesta de mercado, os elementos *"açúcar"*, *"leite"*, *"pão"*, *"peixe"*, *"queijo"* são exemplos de *items*. Os mesmos quando analisados em conjunto, {*"leite", "pão", "queijo"*} são considerados como um grupo, ou seja, um *itemset*. Por fim, as *transaction* correspondem a um grupo, *itemset*, de compras feitas, por exemplo {*"leite", "pão", "queijo"*} e {*"açúcar", "leite", "pão", "peixe", "queijo"*}.

2.4.1 Conceitos de associação

Para exemplificar o funcionamento das regras de associação por meio de exemplos e implementações, é adotada as seguintes definições: D para representar todo o *dataset*; N para o total de *transactions* existentes; J para o conjunto de *itens* distintos existentes no *dataset*; e M o total dos *itens* representados no conjunto J. Dessa forma, considerando a base D expressa na Tabela 2.3, é compreendido um total de 8 transações, com 5 diferentes tipos de *itens*, assumindo assim N = 5, M = 4 e J = { "a", "b", "c", "d"}.

Outra variável utilizada na recuperação das associações é o limite do suporte, valor que aponta o mínimo necessário para considerar os conjuntos de itens, descartando assim quando o valor está abaixo desse limite. Esta variável pode ser identificada pela representação nos algoritmos pelo sinal de sigma, sendo S, Σ ou σ.

Tabela 2.3: Exemplo de transações

Número da transação	*Transação*
1	{ "d", "c", "b", "a"}
2	{ "d", "c"}
3	{ "d", "c", "b", "a"}
4	{ "c"}
5	{ "b", "a"}

Fonte: Os Autores

Analisando a base de exemplo, é possível identificar que quando o *item* {"c"} é comprado, o item d também é frequentemente comprado e esta regra é expressa por {"c"} $\rightarrow$ {"d"}. Essa representação significa implicação, onde o conjunto {"c"} implica em {"d"}. Quando explorada a implicação, é identificada que a regra é satisfeita em 75% dos casos, ou seja, nas transações de número 1, 2 e 3, mas não na transação 4, podendo assim criar uma escala de predição para as ocorrências futuras.

O formalismo da notação de implicação é expressa por dois conjuntos de itens, o lado esquerdo e outro do lado direito da seta, respectivamente, representados pelas letras L e R, mas também conhecido por antecedente (*antecedent*) e consequente (*consequent*), assim como corpo (*body*) e cabeça (*head*). Sendo L $\rightarrow$ R, ambos os conjuntos devem possuir no mínimo 1 *item*, e não possuir *itens* em comum.

2.4.2 Índices estatísticos de associação

Alguns índices são computados para avaliar a importância das regras em relação a base de dados estudada (BRAMER, 2013). Dentre eles, o cálculo do suporte, o qual é utilizado para calcular a proporção de um determinado *itemset* em relação a base. Para o cálculo é utilizado o total de *transactions* do *itemset* X dividido pelo total de *transactions* N da base, conforme equação a seguir.

$$sup(X) = count(X)/N$$

Outro índice muito utilizado pelas regras de associação é o cálculo da confiança. Esta métrica representa a probabilidade condicional de ambos *itemsets* ocorrerem dado o total do *itemset* L. Sendo expressa pela equação a seguir.

$$conf(L \rightarrow R) = \frac{sup(L \cap R)}{sup(L)}$$

Como medida estatística comumente utilizada para indicar a relação de frequência entre *itemsets*, é utilizado o valor de *lift*. O cálculo dessa métrica resulta no impacto da ocorrência de L na frequência de R. Quando o resultado é maior que 1, é possível saber o grau de dependência. Fórmula apresentada na equação a seguir.

$$lift(L \rightarrow R) = \frac{sup(L \cap R)}{sup(L) \times sup(R)}$$

Também utilizada como medida de análise em alguns casos, a métrica que retorna a diferença entre o suporte de ambos os *itemsets* da implicação e o suporte esperado caso cada *itemset* fosse independente é apresentada como *leverage*. Quanto mais baixo o valor obtido, em comparação ao suporte da união de L e R, menor é a quantidade de *transactions* em que os *itemsets* se relacionam. Resultado é obtido por meio da equação a seguir.

$$leverage(L \rightarrow R) = sup(L \cap R) - sup(L) \times sup(R)$$

Conforme visto, diversas métricas estão disponíveis para a melhor compreensão das regras de implicação. Para exemplificar a aplicação das regras apresentadas, quando analisado a implicação da regra $\text{c} \rightarrow \text{d}$ são obtidos os dados da Tabela 2.4.

Tabela 2.4: Valores das métricas de associação sobre regra de exemplo

Métrica	*Valor*	*Descrição*
suporte L	0,8	O *itemset* L está em 80% das *transactions*
suporte R	0,6	O *itemset* R está em 60% das *transactions*
suporte L ∩ R	0,6	União dos *itemsets* L e R está em 60% das *transactions*
confiança	0,75	Em todas as *transactions* com L, 75% tem contém R
lift	1,25	A ocorrência de R está relacionada às ocorrências do grupo L
leverage	0,12	As ocorrências de ambos os subgrupos possuem alta relação

Fonte: Os Autores

Tendo exemplificado apenas algumas das diversas métricas disponíveis, outros índices estatísticos podem ser implementados para a compreensão das implicações de grupos de itens. À vista disto, algoritmos utilizam essas métricas para proporcionar consultas de associações.

2.5 Algoritmos de Associação

Com a grande capacidade de armazenamento de informações, o número de regras de implicação que podem ser geradas a partir de uma base de dados tende a ser potencialmente muito grande (BRAMER, 2013) e, consequentemente, necessitando de uma grande capacidade computacional para o processamento das regras de associação. Para isso, há algoritmos especializados em proporcionar uma maneira de decidir quais regras descartar e quais manter e assim executar os resultados.

Dentre os que se destacam como alternativa a implementação de busca simples, está o Apriori, Eclat e FP-Growth, todos apresentando diferentes modelos para extração de *itemsets* frequentemente associados.

2.5.1 Implementação de Busca Simples

O levantamento da frequência dos *itemsets* de um *dataset* não é uma atividade difícil, entretanto, o desafio segue na desempenho dessa ação (GARCIA-MOLINA; ULLMAN; WIDOM, 2008). Uma abordagem de busca simples utiliza todos os conjuntos de itens possíveis e, após contabilizar o seu suporte, descarta todos os itens abaixo de um determinado limite.

Mesmo possuindo uma complexidade e processamento simples, a demanda por recursos de memória é alto no momento em que é necessário gerar todas as possibilidades de *itemsets* possíveis e armazenar as contagens da frequência de cada um, conforme apresentado na Tabela 2.5. Contudo, quando considerado um ambiente real, não se torna uma prática viável (Heaton, 2016).

Tabela 2.5: Tabela de frequência de todos os *itemsets* possíveis

Itemset possível	*Frequência*
{ a }	3
{ b }	3
{ c }	4
{ d }	3
{ a, b }	3
{ a, c }	2
{ b, c }	2
{ a, b, c }	2
{ a, d }	2
{ b, d }	2
{ a, b, d }	2
{ c , d }	3
{ a, c , d }	2
{ b, c , d }	2
{ a, b, c , d }	2

Fonte: Os Autores

2.5.2 Apriori

Introduzido inicialmente para fornecer melhorias de desempenho em relação a implementação de busca simples, o algoritmo Apriori é reconhecido como uma implementação clássica de um modelo de mineração de regras de associação (Heaton, 2016).

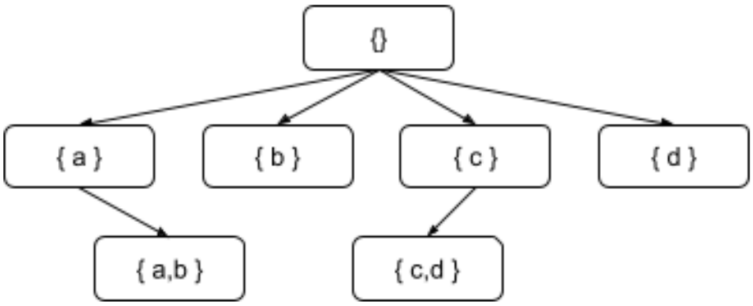
Iniciando pela análise de frequência de *items* de um *dataset*, ele emprega a busca em profundidade criando *itemsets* candidatos. Tais conjuntos que não seguem os padrões de frequência calculados pelo suporte mínimo são eliminados (AGGARWAL, 2015).

Na primeira etapa do processo, é realizada a análise de frequência em *itemsets* individuais (k=1). Conjuntos que possuem a frequência superior a configurada são combinadas a ocorrências de outros *items* (k=k+1) para assim repetir a análise de frequência do novo conjunto, processo que é repetido até o comprimento máximo de *items*.

Para seguir essa implementação, é necessário executar uma varredura do *dataset* para identificar todos os candidatos, o que consequentemente resulta em muitos subconjuntos. Logo, a requisição por recursos de memória se torna significativo e um problema na execução (AGGARWAL, 2015).

Como exemplo, a Figura 2.8 mostra a estruturação da árvore de conjuntos do Apriori na busca pelo conjunto com suporte maior que 0,5. Nessa ordem, um subconjunto de um conjunto de *items* frequentes também deve ser frequente, assim como um superconjunto de um conjunto de itens pouco frequentes acompanha o índice baixo da frequência (GARCIA-MOLINA; ULLMAN; WIDOM, 2008).

Figura 2.8: Árvore de conjuntos gerada pelo Apriori



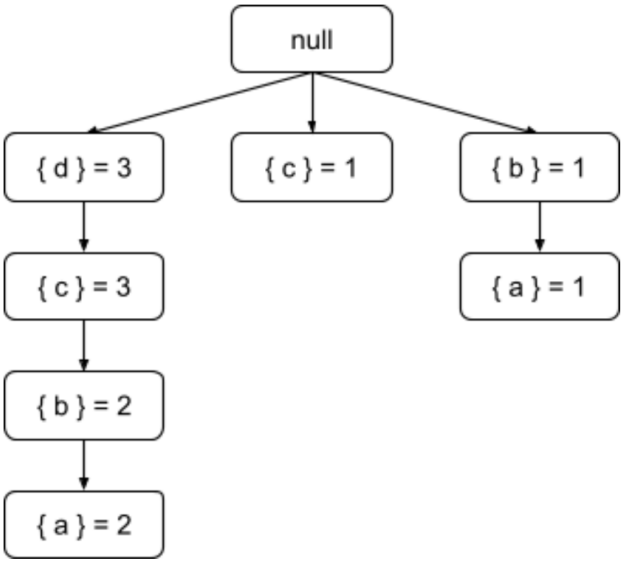
Fonte: Dos Autores.

2.5.3 Eclat

As deficiências do algoritmo Apriori levaram a necessidade de outras abordagens pela busca de eficiência, sendo um deles o Eclat, (Heaton, 2016). Significando *Equivalence Class Clustering and bottom up Lattice Traversal*, o Eclat se diferencia do Apriori pela substituição da primeira pesquisa por uma busca recursiva em profundidade.

A estruturação dos dados dessa técnica utiliza uma árvore de recursão, Trie, a qual permite o processamento em profundidade. Assim sendo, inicialmente um item é definido como um prefixo, sendo ele um padrão que deve estar presente em demais conjuntos de itens encontrados, permitindo uma construção recursiva profunda dos conjuntos de itens. À medida que os conjuntos de itens são encontrados, eles são adicionados em novos nós da árvore. O item de um segundo nível corresponde a um filho do primeiro nível, e nesse formato, nenhum pai tem mais de um filho com o mesmo nome, entretanto, o mesmo item pode aparecer em diversos locais da árvore conforme apresentado na Figura 2.9 (Heaton, 2016).

Figura 2.9: Árvore Trie gerada pelo Eclat



Fonte: Dos Autores.

Durante a montagem da Trie, os nós são criados para representar todos os conjuntos de itens justaposto dos índices de suporte. Caso um item já possua um nó na estrutura, o nó do item mais a direita tem a contagem do suporte aumentado e, caso não existam, são criados com o suporte igual a um. Assim, fazendo com que o Eclat use menos memória em comparação ao Apriori, pois *itemsets* frequentes são representadas apenas uma vez em ramificações principais da árvore (Heaton, 2016).

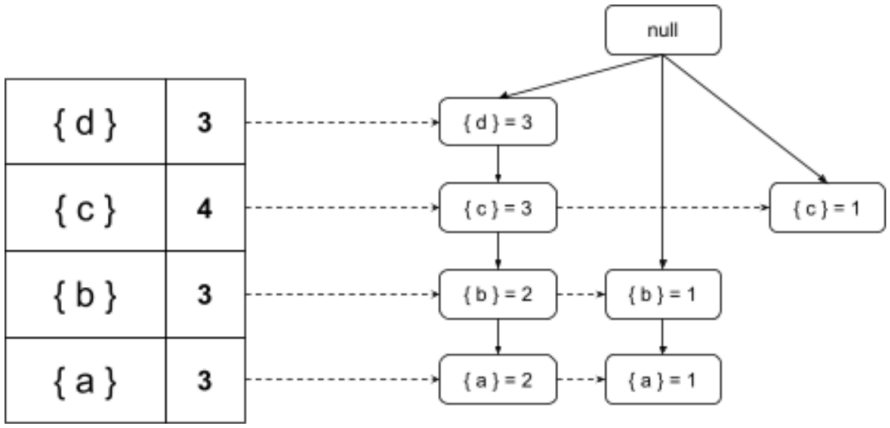
2.5.4 FP-Growth

Como alternativa as demais implementações, foi desenvolvido o FP-Growth, *Frequent Pattern Growth*. Assim como a técnica Eclat, o FP-Growth também utiliza uma Trie, ou seja, implementa uma estrutura vertical dos dados para o armazenamento (Heaton, 2016).

Além da estrutura em árvore, este algoritmo também conta com a utilização de uma tabela auxiliar que contém em sua listagem uma ocorrência para cada item, vinculando todos os nós do mesmo tipo (GARCIA-MOLINA; ULLMAN; WIDOM, 2008). Essa prática fornece uma visão horizontal dos dados assim como uma visão vertical da árvore conforme representado na Figura 2.9.

Contudo, quando comparado a implementação do FP-Growth com Eclat, é declarado a similaridade em questões de processamento, estruturação e requisitos de memória, (Heaton, 2016).

Figura 2.10: Visão horizontal e vertical dos dados gerada pelo FP-Growth



Fonte: Dos Autores.

2.5.5 Implementações de algoritmos de associação

Sendo um algoritmo de mineração de conjuntos no qual tem o objetivo de recuperar conjuntos de itens frequentes, o Apriori é facilmente compreensível e por isso é um ponto de partida para estudos na área. No entanto, como foi apresentado, esta técnica apresenta deficiências de escalabilidade pelo problema da gestão dos recursos de memória, inviabilizando o uso em grandes conjuntos de dados (Heaton, 2016).

Em resposta, o Eclat e o FP-Growth apresentam resultados melhores que o Apriori em relação ao desempenho, sendo o FP-Growth ligeiramente melhor que o Eclat. Entretanto, segue sendo necessário acompanhar novas implementações, bem como modificações de modelos existentes para a avaliação de alternativas que apresentem melhores desempenhos ou novas funcionalidades.

2.6 Considerações

Processar grandes conjuntos de dados pode ser uma atividade desafiadora quando deparado com alguns requisitos de estruturação dos dados. Quando a base de dados conta com registros textuais, a necessidade de recursos que buscam estruturar, formatar e quantificar a base são fundamentais.

A disciplina de mineração de texto guia o desenvolvimento de modelos responsáveis por processar e quantificar bases textuais a fim de viabilizar a extração das informações. Hoje em dia é possível contar com linhas de pesquisas que focam na solução destes desafios, tais como: *word embeddings*, que quantifica a correlação entre os termos com base no contexto empregado; recuperação de informações, que busca recuperar documentos textuais com base na comparação de termos de um domínio de busca; análise de associação para técnicas preditivas; entre outras.

Neste Capítulo, foram descritas algumas técnicas de descoberta de conhecimento em bases de dados. Conforme apresentado, a recuperação das informações conta com índices e métricas de recuperação e validação entre a correlação de busca e resultado. Contudo, as atividades de transformação do texto de entrada é um ponto de atenção independente da técnica utilizada. Outra abordagem apresentada foi a mineração de regras de associação, a qual mostrou implementações que buscam a correlação entre termos.

Conforme apresentado pela base de dados dos registros do teleatendimento do sistema DermatoNET da plataforma TelessaúdeRS utilizada nesta dissertação, cada *transaction* registrada na base estudada é composta por múltiplos segmentos de dados, as quais são representados por colunas de uma tabela. O tipo do dado textual contido em cada coluna varia, podendo ser representado por um dado categórico ou texto livre, linguagem natural. Com este panorama, a necessidade de recuperação de informações, identificando as associações entre as descrições de cada segmento do processo, é um desafio que necessita da combinação dos benefícios de ambas as técnicas apresentadas.

3 TRABALHOS RELACIONADOS

Este Capítulo tem como objetivo descrever os trabalhos relacionados que apresentam relação com a pesquisa. Para isso, este capítulo está organizado na seguinte forma: como prólogo é explicado na seção 3.1 a motivação e os principais desafios da pesquisa; na seção 3.2 são apresentados os trabalhos que serviram de *baseline* para este trabalho assim como o projeto desenvolvido em paralelo pelo núcleo de pesquisa da medicina; e por fim, na seção 3.3 são introduzidas as considerações da relação deste trabalho a luz das referências levantadas.

3.1 Motivação e desafios da pesquisa

Esta dissertação foi desenvolvida em parceria entre dois programas de pós-graduação da UFRGS (Universidade Federal do Rio Grande do Sul), sendo os programadas PPGC (Programa de Pós-Graduação em Computação) e PPGCM (Programa de Pós-Graduação em Ciências Médicas). Este trabalho apresenta a análise sobre o viés computacional da técnica elaborada de recuperação de informação.

Como motivação, esta dissertação buscou solucionar o problema de mineração dos dados na característica dos dados gerados durante o processo de atendimento do clínico via teleconsultoria, o qual é detalhado no Capítulo 4. Os problemas identificados estavam em conseguir associar os termos ocorridos em diferentes etapas do processo de forma a entender a associação entre um termo de uma atividade inicial em relação a uma atividade intermediária ou final, ou vice-versa.

O desenvolvimento de um método que permita buscar os termos associados com base na relevância de associação, abre possibilidades para:

- análise dos dados gerados para identificar comportamentos na descrição da dúvida clínica do atendimento médico e a relação com a conduta ou medicamentos e tratamentos relacionados;
- identificar a correlação de termos que resultam em atendimentos prioritários. Por exemplo, buscar características que informem o diagnóstico de câncer ou necessidade de encaminhamento especializado; e
- possibilitar a otimização do processo para identificar os itens supracitados com maior antecedência automaticamente no processo.

Como desafio técnico, esta pesquisa buscou a implementação conjunta de dois princípios distintos para melhorar a precisão da recuperação dos termos associados. Para a busca de termos em um ranque de relevância, é considerada a frequência do termo em relação a ocorrência do corpus, o qual reduz o peso dos termos vulgarmente repetidos e remove ocorrências de *outliers*. Justaposto a esta técnica, foi utilizado a base das regras de associação que agregam peso ao maior número de ocorrências.

Com esta característica distinta de ambos os conceitos, esta pesquisa não teve facilidade em encontrar pesquisas que se propusessem a fazer algo na mesma linha de pesquisa, resultando na utilização de trabalhos parcialmente similares e o trabalho que já utilizou o modelo desenvolvido.

3.2 Descrição dos Trabalhos Relacionados

Os resultados obtidos com o método de recuperação de termos associados em documentos médicos podem gerar inovação na saúde pública por meio da transparência e análise dos dados gerados durante a execução do processo. Entender os laudos dermatológicos permite um melhor planejamento e ações assertivas orientadas a dados. Com o método aplicado aos registros da teledermatologia, foi possível extrair informações dos laudos de telediagnóstico dermatológico, o que por fim possibilita estudos para planejamento em saúde pública, melhorando a tecnologia utilizada pelo TelessaúdeRS reduzindo o tempo no processo, auxiliando na elaboração dos desfechos de telediagnóstico e gerando economia em saúde pública Costa, Gonçalves and Bakos (2021).

Na pesquisa de Kulkarni, Tokekar and Kulkarni (2012), a busca de termos associados a um contexto foi uma necessidade base para atender a crescente demanda de categorização devido ao aumento de textos produzidos. Para isso, a abordagem foi utilizar como base o algoritmo Apriori onde as regras de associação das palavras são utilizadas para derivar conjuntos de recursos de texto pré-classificados documentos. Esta classificação necessitou de um treinamento devido ao esforço humano custoso.

Com base nisso, Madani, Boussaid and Zegour (2013) apresentam um levantamento sobre métodos para mineração de documentos semiestruturados o qual conclui que a maioria dos até então algoritmos tradicionais de mineração de dados não são adequados para documentos semiestruturados e outras abordagens precisam de adaptações para esses dados.

Alguns pesquisadores focam na Web Semântica para indexar documentos semiestruturados e propõem uma recuperação focada em tempo real e interação seletiva para a web semiestruturada Lipczak et al. (2013). Uma abordagem adaptativa de recuperação de documentos Goswami and Kundu (2013) usa a ontologia e as técnicas de processamento de álgebra vetorial para recuperar documentos semiestruturados.

Ainda na linha da busca de associação de termos, Singh and Sethi (2015) apresentam uma nova técnica de implementação cruzada de algoritmos buscando uma melhor eficiência em relação aos mesmos níveis de confiança e suporte. Contudo, a pesquisa evidenciou os ganhos na performance quando não apenas criando uma variação do Apriori, como o FP-growth, mas também juntado com técnicas como Reverse-Apriori.

Com a utilização de técnicas mais recentes de análise de texto como *Word Embeddings*, Ai et al. (2016) analisam o uso do modelo de vetor de parágrafos (PV-DBOW) para recuperação de informações, a qual utiliza o treinamento para prever cada palavra observada nele. Contudo, na pesquisa foram discutidos três problemas que restringem a eficácia nos cenários IR, são eles: *overfitting* em documentos curtos, estratégia de amostragem negativa imprópria e falta de modelagem de substituição de palavras.

Mesmo os pesquisadores iniciados há alguns anos, o domínio da informação ainda é um desafio devido à dificuldade de compreensão dos textos livres. Atualmente, há muita pesquisa na área de processamento de texto em informações semiestruturadas. Nadig, Braschler and Stockinger (2020) apresentam um estudo sobre consulta de linguagem natural de dados semiestruturados comparando métodos de busca em banco de dados e recuperação de informação.

O contexto de recuperação de informações de bancos de dados médicos vem ganhando destaque principalmente em um mundo pós pandemia e a compreensão dos registros pode auxiliar na ampliação dos serviços de saúde. Para isso, são aplicadas desde técnicas de IR a redes neurais, conforme apresentado na pesquisa de Herrera et al. (2020), no qual redes neurais convolucionais são utilizadas para detectar alterações clínicas em registros de múltiplas origens referente ao histórico do paciente a fim de identificar possíveis padrões futuros. Entretanto, este modelo não pode ser replicado a bases processuais como a analisada nesta dissertação.

3.3 Considerações sobre os trabalhos

Esta pesquisa teve como dificuldade inicial encontrar uma técnica que possa ser definida como um *baseline* para o modelo proposto, pois a pesquisa buscou o melhor das abordagens de mineração de regras de associação e os princípios que evitam os vieses dos modelos de recuperação de informação. Para isso, foi necessário definir um algoritmo que permita controle do comportamento e modificações para buscar a associação entre segmentos diferentes. Para evitar *overfitting* devido aos textos de poucas palavras e de viés médico, foi escolhida a utilização do modelo Apriori com ajuste para identificar os termos de diferentes segmentos.

4 BASE DE DADOS TEXTUAL - REGISTROS DE TELEATENDIMENTO

Este Capítulo tem como objetivo a apresentação da base de dados relacionado ao processo de atendimento médico que utilizam a plataforma de atendimento do TelessaúdeRS. Inicialmente é apresentado a plataforma e um dos sistemas que o compõe e que é base para esta pesquisa, o DermatoNET.

Em seguida, é exposto o conceitual que sustenta o entendimento de teleatendimento e as implementações nacionais deste modelo. Com este entendimento, é revisada a aplicação e o processo que origina estes registros e quais dados são esperados na base desta aplicação. Por fim, são listadas as definições de coleta e processamento do grupo de registros a fim de estabelecer a coleção de dados para a pesquisa.

4.1 Teleconsultoria

Sendo definida por um canal de diálogo entre profissionais da área da saúde, a teleconsultoria faz uso de instrumentos de telecomunicações para o esclarecimento de dúvidas sobre procedimentos clínicos (HADDAD, 2012), segundo a portaria GM/MS 2546, (BRASIL, 2011), definida como:

> [...] consulta registrada e realizada entre trabalhadores, profissionais e gestores da área de saúde, por meio de instrumentos de telecomunicação bidirecional, com o fim de esclarecer dúvidas sobre procedimentos clínicos, ações de saúde e questões relativas ao processo de trabalho [...]

Sendo o provedor de apoio assistencial as teleconsultorias, os Núcleos do Telessaúde (NT) auxiliam nos procedimentos de diagnósticos dos médicos da atenção primária. Com a utilização da plataforma digital, este recurso age de maneira distribuída, permitindo que a atenção primária à saúde (APS) utilize suporte remoto dos teleconsultores.

Ainda conforme a portaria GM/MS 2546 (BRASIL, 2011), o modelo de assistência à APS via teleconsultorias ocorre em duas modalidades: teleconsultoria síncrona, sendo realizada em tempo real, geralmente por chat, web ou videoconferência; ou teleconsultoria assíncrona, realizada por meio de mensagens *off-line*.

No Brasil, com suporte do Ministério da Saúde, em 2007 foi criado o Projeto Telessaúde Brasil Redes (BRASIL, 2007). Iniciou-se como um projeto piloto que foi implantado em nove diferentes estados, sendo eles: Amazonas, Ceará, Goiás, Minas Gerais, Pernambuco, Rio de Janeiro, Rio Grande do Sul, Santa Catarina e São Paulo.

4.2 TelessaúdeRS e DermatoNET

Dentre as nove implantações piloto, a implantação do estado do Rio Grande do Sul contou com a parceria com a Universidade Federal do Rio Grande do Sul (UFRGS) com o projeto TelessaúdeRS (BRASIL, 2007). Em 2010, o projeto piloto da UFRGS foi expandido e incorporado a um programa governamental, o Telessaúde Brasil Redes, um sistema de suporte a equipes de APS, com o foco principal em aumentar a resolutividade e otimizar o fluxo entre os níveis de cuidado.

Atualmente, o TelessaúdeRS conta com mais de 200 colaboradores, os quais atuam no atendimento das teleconsultorias dos sistemas da plataforma de Telessaúde do ministério da saúde.

Em 2017 foi adicionado à plataforma do Telessaúde uma aplicação específica para o telediagnóstico de dermatologia, o DermatoNET. A teledermatologia consiste no uso de tecnologia no campo da telecomunicação para aproximar o atendimento do paciente junto a equipe de APS com os especialistas em dermatologia. Desta forma, resultando na redução das limitações de tempo e de espaço, o qual auxilia na melhora sobre os cuidados dermatológicos e na eficiência no processo de atendimento com ações mais bem embasadas e retornos mais precisos.

4.3 Processo de teleatendimento

Sendo oriundo de uma dúvida médica do profissional da saúde, o processo de teleatendimento do Telessaúde Brasil ocorre no ambiente virtual da plataforma e tal fluxo pode ser formalizado em três etapas: a solicitação; a tele regulação; e a resposta.

As atividades de solicitação iniciam quando o profissional da saúde que possui a dúvida clínica entra com suas credenciais na plataforma de teleatendimento e seleciona a seção relacionada a solicitação de auxílio que está sendo criada. O médico de APS deve prover algumas informações básicas para a próxima etapa, como o tipo de teleconsultoria, sendo síncrono ou assíncrono. Quando o tipo é assíncrono, o questionamento passa por uma triagem e é encaminhada para o teleconsultor mais adequado, o qual tem o prazo de 72 horas para responder (BRASIL, 2012b).

Como etapa intermediária, o tele regulador tem como objetivo receber, analisar, classificar e orientar o fluxo das requisições de teleatendimento provendo as informações dos códigos internacionais de CID e CIAP.

Nessa atividade, é realizada a triagem de encaminhamento de dúvidas ao profissional. Para isso, é triangulada a categoria da dúvida do solicitante, com o formato de atendimento solicitado e a área profissional do teleconsultor justaposto do seu histórico de atendimento em relação ao tema. De forma complementar, o tele regulador também é incumbido da auditoria do resultado do atendimento: avaliando, revisando e complementando com eventuais considerações e informações complementares (BRASIL, 2012c).

Como última etapa, a partir do tipo selecionado, o profissional teleconsultor atende as solicitações respondendo às questões com base na melhor evidência clínica disponível. Para isso, são analisadas as descrições e informações fornecidas e formalizadas sobre a dúvida clínica, elaborando assim a hipótese diagnóstica, categorização do caso e abordagem de conduta e, como referencial, é realizada a busca por solicitações similares já respondidas na plataforma (BRASIL, 2012a).

4.4 Base de dados da teleconsultoria assíncrona

Como resultado do processo de teleatendimento, a plataforma do Telessaúde armazena os registros de cada solicitação em uma base de dados, a qual é sistematicamente enriquecida por novos atendimentos. Cada atendimento assíncrono resulta em um registro composto por vários campos de dados textuais, os quais fornecem informações de ponta-a-ponta do fluxo de atendimento.

Com base na procura do paciente por atendimento, o processo de um registro é iniciado pela situação de dúvida do médico da APS e finalizado pela emissão de um diagnóstico mais provável junto da sugestão de conduta inicial pelo médico especializado. Provendo um canal de conexão entre o médico da APS e um especialista dermatológico, todos os registros de atendimento integram um banco de dados com os registros dos laudos com o processo completo de avaliação dermatológica, o qual tem suas etapas apresentadas na Tabela 4.1.

A teledermatologia vem tornando-se um elemento importante no cuidado à saúde ao redor do mundo nas últimas duas décadas em proveito do desenvolvimento da tecnologia da informação (PAKHS, 2002). Documentos contendo informações clínicas que descrevem fenótipos e tratamentos de pacientes representam uma fonte de dados as quais, com uso da computação, podem revelar novos princípios de correlações nas doenças (JENSEN; JENSEN; BRUNAK, 2012).

Tabela 4.1: Etapas do processo do Telessaúde

Etapa	*Elemento*	*Descrição*
Etapa 1	Atividade	Dúvida clínica
	Agente	Médico da APS
	Descrição	Descreve situação clínica podendo adicionar fotos como evidências
	Dados de entrada	Imagem, texto e dados categóricos
Etapa 2	Atividade	Elaboração de hipótese diagnóstica
	Agente	Especialista teleconsultor do NT
	Descrição	Apresenta a hipótese diagnóstica à luz das informações da consulta
	Dados de entrada	Texto
Etapa 3	Atividade	Categoriza o caso clínico
	Agente	Especialista teleconsultor do NT
	Descrição	Seleciona os códigos clínicos referente ao hipótese diagnóstica
	Dados de entrada	Categórico
Etapa 4	Atividade	Sugestão de abordagem clínica
	Agente	Especialista teleconsultor do NT
	Descrição	Elabora a descrição da conduta clínica sugerida ao APS
	Dados de entrada	Texto

Fonte: Os Autores

4.5 Mineração de textos de laudos médicos do DermatoNET

Com a digitalização dos meios de atendimento, muitos dados são gerados e salvos como registros históricos. O domínio desses dados pode resultar em benefícios quando aplicadas técnicas de análise de dados, seja no campo de recuperação de informação, aprendizagem de máquina ou análise preditiva.

A utilização de mineração textual favorece no reconhecimento de padrões, nas dúvidas dos médicos da atenção primária, assim como o perfil clínico dos pacientes. Tais técnicas de descoberta de conhecimento permitem compreender o conteúdo das consultas de forma quantitativa, propiciando a análise sobre as categorias e textos de hipóteses diagnósticas e condutas que estão sendo aventadas pelos teledermatologistas.

Para esta dissertação, foi delimitada a utilização dos atendimentos registrados até abril de 2020. Considerando apenas registros que completaram o processo de atendimento via DermatoNET, a base utilizada totalizou 12.199 laudos para serem utilizados.

4.5.1 Estrutura dos laudos utilizados

Durante todas as etapas do processo de atendimento, dados são criados e associados de forma segmentada ao mesmo registro. Todos os registros possuem múltiplos dados onde cada um segue um tipo de estruturação.

Mesmo fazendo parte de um modelo estruturado de comunicação, dados descritivos são necessários para melhorar a comunicação entre os médicos. Desta forma, mesmo possuindo informações que se pode resumir em dados categóricos já estabelecidos, campos de texto com linguagem natural se tornam fundamentais na personalização da descrição e incremento de maiores e melhores detalhes.

Conforme o andamento das etapas do processo de atendimento, são gerados registros médicos com dados categóricos, sendo tais informações em texto livre, linguagem natural e apenas alguns registros categóricos. Para esta dissertação, nem todos os campos foram disponibilizados, assim, apenas os dados listados na Tabela 4.2 foram disponibilizados para utilização.

Tabela 4.2: Estruturação dos registros obtidos da base do DermatoNET

Identificador	*Criação*	*Descrição*	*Tipo de dado*
dconsdados	Etapa 1	Conjunto de informações categóricas que compõe um texto sobre hipótese diagnóstica ou descrição do quadro clínico	Texto livre
dconsdesc	Etapa 1	Descrição da consulta	Texto livre
dsolpacsexocod	Etapa 1	Sexo do paciente	Categórico
dconshipdiag	Etapa 2	Descrição da hipótese diagnóstica	Texto livre
dconsciap1cod	Etapa 3	Código CIAP	Categórico
dconscid1cod	Etapa 3	Código CID	Categórico
dconssugenccod	Etapa 4	Sugestão de encaminhamento com base na análise clínica	Categórico
dconscondsug	Etapa 4	Descrição da conduta sugerida	Texto livre

Fonte: Os Autores

Com uma composição híbrida entre os tipos de dados, as técnicas de mineração de texto se apresentam fundamentais para extração de informações da base de dados. Usufruindo dos dados históricos para a aprendizagem, benefícios como automação, recomendação e associação de ações durante o atendimento podem ser recursos para a melhoria do processo.

4.5.2 Pré-Processamento dos Textos

Para esta dissertação, optou-se pela realização do pré-processamento da base de dados, estruturando as informações para que permitam melhores resultados nos algoritmos utilizados na etapa de experimentação.

Contando com dados categóricos e como texto livre, o pré-processamento foi implementado em todos os campos de dados para todos os registros. Para isso foram definidas as regras de transformação de texto, as quais implementam as etapas de tokenização, normalização, seguidas de técnicas de processamento linguístico tanto nos registros quanto nos termos da consulta. O detalhamento destas regras é descrita na experimentação no Capítulo 6.

5 PROPOSTA DO MÉTODO WAR

O objetivo deste Capítulo é apresentar os detalhes do modelo proposto para a recuperação de termos considerando a associação entre múltiplos segmentos. O método é denominado *Word Association Retrieval* (WAR), um acrônimo para Modelo Probabilístico para Recuperação de Termos Associados em Textos Multissegmentados. Utilizando o conceito de recuperação de informações acompanhado das técnicas de associação, o modelo WAR apresenta como contribuição a capacidade de recuperar a lista de termos relevantes com base em uma consulta, mas levando em consideração a relação com termos de diferentes segmentos.

5.1 Definição preliminar - Segmentação dos documentos

Cada documento representa um registro composto por texto, o qual pode ser analisado de forma atômica ou por suas subdivisões. Com base no tipo de estrutura do documento que compõe o *corpus*, é possível que cada subdivisão do documento seja vista como uma parte a ser analisada.

Em sistemas de busca de arquivos ou buscadores WEB, o documento da base de dados (*dataset*) é analisado de forma única, identificando se possui os termos de busca, ou de forma segmentada, atribuindo pesos diferentes para segmentos distintos. Mesmo possuindo uma ou mais subdivisões, os modelos de recuperação de informações consideram o registro de forma única, pois os termos da consulta são aplicados apenas a fim de recuperar o documento.

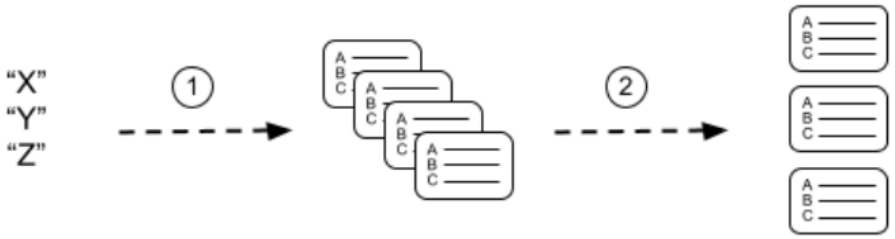
Para esta dissertação, quando um método considera todos os possíveis segmentos de um registro e os analisa de forma desmembrada, mas correlacionada, eles são categorizados como documentos multissegmentados. Por conseguinte, os valores do ranque também são apresentados de forma distinta.

Quando se analisa os documentos de um único segmento, o resultado apresenta como retorno os documentos relevantes aos termos da consulta. Em contrapartida, ao analisar documentos multissegmentados, a consulta é sobre os segmentos e apresenta o retorno não por documento, mas por termos dos segmentos.

A seguir, são apresentados exemplos de um processo de duas etapas, onde a primeira representa a busca por termos em uma base de registros e a segunda o retorno ranqueado.

No caso da Figura 5.1 de uma consulta de termos “X”, “Y” e “Z”, quando o método não considera as segmentações de texto dos documentos, todos os termos são aplicados a todos os segmentos e o ranque ordena os documentos mais relevantes.

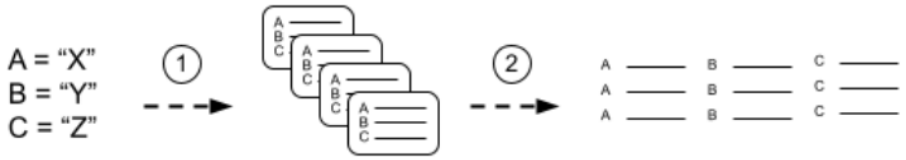
Figura 5.1: Etapas de consulta e ranque para documentos de segmento único



Fonte: Dos Autores.

Quando o método considera de maneira multissegmentada os documentos do *corpus*, os termos de entrada da busca podem ser direcionados para segmentos específicos. Conforme exemplo da Figura 5.2, os segmentos “A”, “B”, e “C” recebem respectivamente os termos “X”, “Y” e “Z”, o que por sua vez resulta em um ranque também segregado por segmento.

Figura 5.2: Etapas de consulta e ranque para documentos de multissegmentos



Fonte: Dos Autores.

A apresentação dos resultados de forma segmentada, permite identificar o impacto dos termos pesquisados no próprio segmento e nos demais. Quando aplicado este método, é possível identificar o nexo entre as etapas por meio dos termos utilizados, assim como o impacto de um termo em todos os demais segmentos.

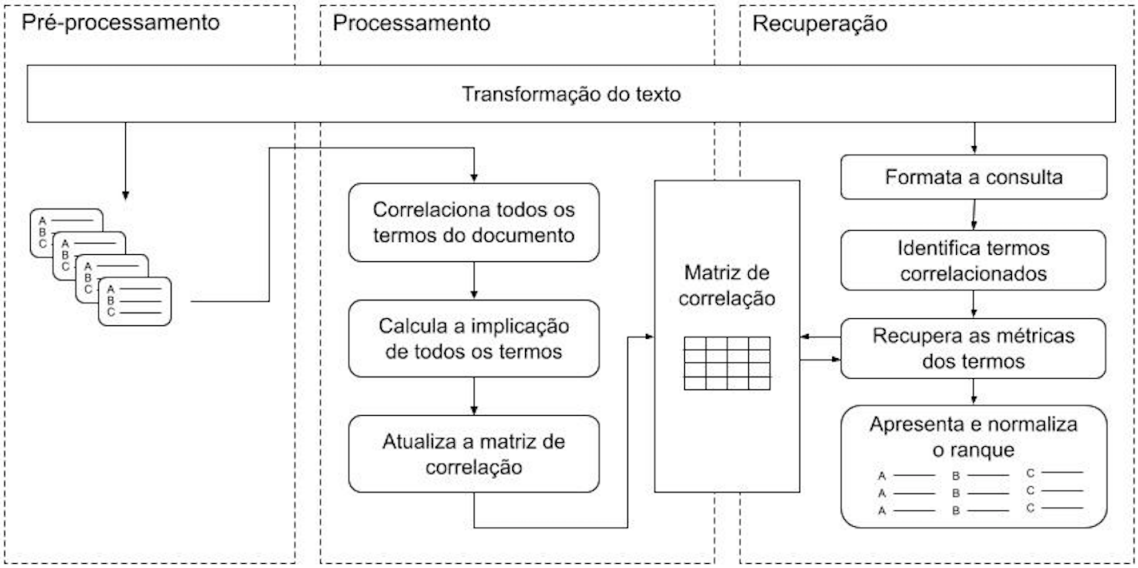
Como exemplo de aplicabilidade desse método, a compreensão da correlação dos termos da descrição de uma postagem com os encontrados nos comentários se torna viável, assim como na análise de um processo, em que cada etapa representa um segmento e é constituída por registros de texto.

São documentos elegíveis a característica multissegmentada, registros com dados estruturados como XML, planilhas, bases de dados, documento que apresente seus dados de forma tabular e qualquer outro formato que seja possível segmentar conteúdo.

5.2 Visão Geral

O método proposto nesta dissertação conta com um fluxo de tratamento de dados que conta com três etapas, sendo elas: pré-processamento, processamento e recuperação. Mesmo tendo suas atividades com processamento independente, técnicas de transformação do texto e a matriz de correlações são compartilhadas entre etapas, conforme Figura 5.3.

Figura 5.3: Fluxo do método de três etapas



Fonte: Dos Autores.

Por tratar de entradas de texto livre, é necessário que o método conte com uma etapa responsável por reduzir a variação do significado de um termo. Contando com regras pré-estabelecidas, a atividade de transformação do texto as aplica tanto quando um documento serve de entrada na atividade de pré-processamento, quanto nos termos da consulta na etapa de recuperação.

Cada etapa realiza um papel importante no processamento como um todo. A atividade de pré-processamento formata os registros de entrada com base nas regras definidas para transformação do texto. O processamento interpreta os registros de entrada realizando a correlação entre os termos dos textos, calcula a implicação entre eles e atualiza a matriz de correlação. Por fim, a recuperação das informações da matriz conta com valores registrados em uma consulta que usa os dados da matriz para elaborar a saída em formato de um ranque.

5.2.1 Etapa de Pré-Processamento

Esta etapa tem como objetivo aplicar as regras de transformação do texto em todos os registros da base de dados utilizada. Tendo sido já mencionados como atividades importantes na recuperação de informação, estes métodos compreendem a análise léxica dos textos do *corpus* e podem ser aplicados de forma concomitante para obter uma melhor estruturação no pré-processamento.

Buscando aumentar o desempenho da análise dos dados, é aplicado um conjunto de regras que buscam higienizar, normalizar e padronizar a base de dados. Tais métodos de análise léxica moldam os dados, aplicando as seguintes regras:

- Normalização, que conta com um pré-processamento linguístico, uniformizando os registros entre letras maiúsculas e minúsculas, variações gramaticais, termos compostos, polissemia, sinonímia, plurais, negações, variação do termo (erros e gírias) e alterações morfológicas como aumentativos e diminutivos;
- Higienização, que remove *stop-words* e estruturas não relevantes; e
- Padronização, que realiza a tokenização dos termos além de poder contar com técnicas como *stemming*.

5.2.2 Etapa de Processamento

Com o registro pré-processado, cada termo é a representação do resultado da etapa de tokenização. Após esta etapa, é realizada a análise dos termos visando identificar como eles se correlacionam com os demais elementos dos documentos.

O levantamento de correlação busca identificar a implicação da ocorrência de um termo com os elementos que estão diretamente relacionados a ele. O resultado dos cálculos realizados resulta em uma matriz que contém os pesos relacionados às implicações.

5.2.3 Etapa de Recuperação

A etapa de recuperação, além de formatar os registros de entrada aos moldes das regras de transformação do texto usadas no pré-processamento, os usa na consulta aos valores da matriz de correlação gerada na etapa de processamento.

Desta forma, utilizando o processamento de correlação, os valores de relação dos termos são utilizados para identificar os pesos de implicação. Com base em uma consulta composta de termos associados aos segmentos, é estruturado um ranque que ordena o retorno conforme a proeminência do termo em comparação aos demais.

Termos que são vulgarmente encontrados no universo do segmento apresentam uma relevância menor aos outros quando não associado ao termo de busca, pois sua associação é caracterizada de forma menos determinista nas outras ocorrências. Em contrapartida, os termos que estão fortemente ligados apresentam uma pontuação de maior destaque sendo assim apresentados no topo do ranque.

5.2.4 Método Proposto

O método proposto neste trabalho tem como foco prover a criação e recuperação dos valores da matriz de correlação. É apresentada uma abordagem para as etapas de processamento e recuperação dos valores de implicação da matriz de correlação. Assim sendo, mesmo aplicando regras de transformação de texto, a explanação do método não se desdobra sobre as atividades de análise léxica do pré-processamento.

5.3 Processamento

A implementação do método de recuperação de associação conta com duas etapas no nível de processamento. Nesta etapa, são desenvolvidas duas matrizes multidimensionais sendo elas a matriz multidimensional de termos relacionados (MMRT) e matriz multidimensional de implicação recíproca (MMRI).

Para a primeira etapa, o desenvolvimento da MMRT considera cada segmento como um estado e relaciona todos os pares de estados possíveis, sendo eles distintos ou iguais. Como resultado, é obtido um conjunto de matrizes que representam a implicação dos termos por meio do produto cartesiano, resultando da multiplicação de todos os termos combinatórios.

Para a segunda etapa, é utilizada como base a MMRT, pois é iniciada a estruturação de outra matriz, a MMRI. Isso se dá por meio de uma sequência de cálculos que consideram a frequência das relações calculando a implicação dos termos dos segmentos representados em cada par de dimensões da matriz.

5.3.1 MMRT: Matriz Multidimensional de Termos Relacionados

O desenvolvimento da MMRT realiza o levantamento de todas as relações existentes entre termos e, para isso, são adotadas algumas definições e representações: cada segmento dentro de todos os documentos é considerado um estado e é representado pela letra "E". Para a representação de uma relação entre dois estados, sendo eles distintos ou não, é utilizada a letra R. A contabilização de todos os estados é representada por TE. Para representar o número total de relações entre estados, é utilizado TR.

O número total de relações em estados de documentos heterogêneos é obtido quando aplicada a seguinte equação:

$$TR = TE + (TE \times (TE - 1) \div 2)$$

As relações possuem validação bidirecional, tornando irrelevante a ordem da associação dos estados. Logo, a ordem em que os segmentos são relacionados não os distingue, desta forma, quando encontradas relações com os mesmos segmentos, com ordens iguais ou diferentes, apenas uma é considerada, desconsiderando as duplicidades.

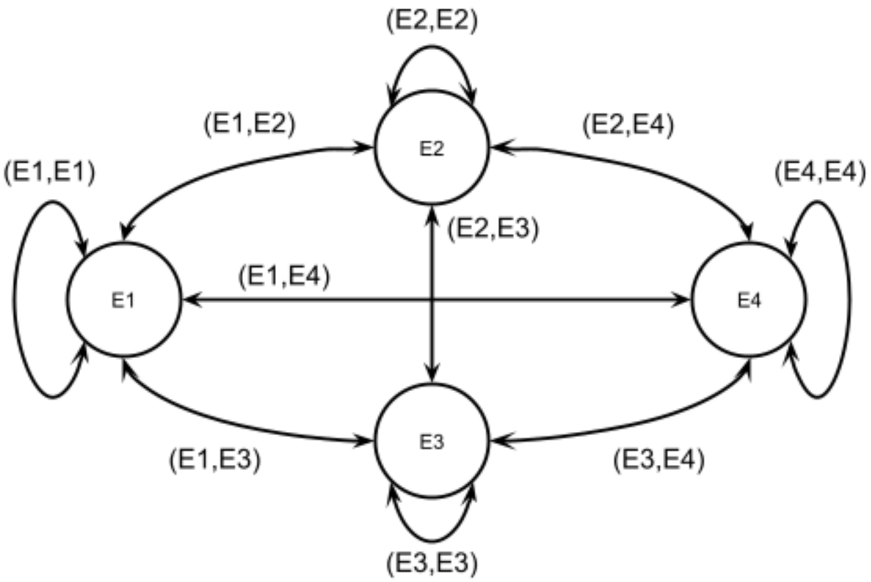
Por buscar a comutação entre os estados, em um cenário de uma base composta por 4 segmentos, o total de estados reflete a mesma quantidade, entretanto, o total de relações é igual a 10. Assumindo assim os estados E1, E2, E3, E4 , TE = 4, e quando relacionados TR = 10 representando por (E1,E1), (E1,E2), (E1,E3), (E1,E4), (E2,E2), (E2,E3), (E2,E4), (E3,E3), (E3,E4), (E4,E4) conforme apresentado na Figura 5.4.

Baseado nesta definição, é parte do método a interação entre todas as relações considerando cada estado uma dimensão, sendo elas representadas por x e y. Os termos são multiplicados entre as dimensões, ou seja, todos os termos da dimensão x serão multiplicados pela dimensão y e isso se repete para todas as relações conforme equação a seguir:

$$\sum_{i=1}^{n} Xi \times Yi = \{(x, y) | x \in Xi \wedge y \in Yi\}$$

Sendo as dimensões representadas pela letra D, quando multiplicadas é obtida uma lista de implicações. Contudo, por não tratar de elementos aritméticos e sim *tokens* que são analisados como registros categóricos, a multiplicação resulta em um produto cartesiano de termos.

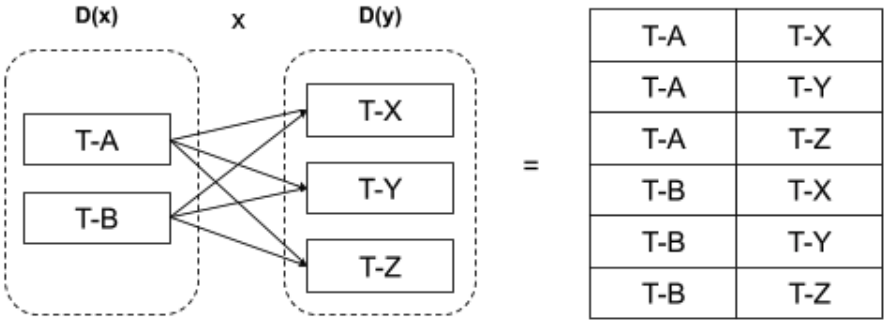
Figura 5.4: Relações entre estados para documentos de 4 segmentos



Fonte: Dos Autores.

No exemplo da Figura 5.5, a relação possui dois termos na primeira dimensão, D(x), e três na segunda, D(y). Quando realizada a multiplicação, cada termo se associa com todos os termos do outro grupo, assim gerando uma lista com seis implicações.

Figura 5.5: Produto cartesiano entre as dimensões X e Y



Fonte: Dos Autores.

Este processo é repetido até que o levantamento de implicação identifique a relação de todos os termos de todos os estados contra todos os outros mais ele mesmo, desta forma obtendo uma matriz multidimensional composta pela geração dos produtos cartesianos.

Esta matriz resultante é definida como MMRT, uma matriz multidimensional que considera todos os estados de segmentos. Como próxima etapa, é realizada a navegação entre todas as relações de associação bidimensional, realizando os cálculos que servem de base para a geração da MMRI.

5.3.2 MMRI: Matriz Multidimensional de Implicação Recíproca

Utilizando a lista de implicações resultantes do cálculo do produto cartesiano de termos, a MMRI atribui a quantificação dos termos e pesos das implicações entre eles de forma recíproca para cada dimensão da relação. Isto significa que, para cada relação entre segmentos identificados, esta etapa executa os cálculos de associação, mas diferentemente dos métodos clássicos de associação que analisam entre os *items* do mesmo seguimento, os cálculos comparam os termos entre segmentos distintos.

Desta forma, são aplicados para todas as relações bidimensionais de implicação os cálculos das métricas a seguir, as quais são denominadas como x e y as respectivas dimensões da relação analisada. Para trazer a reciprocidade nas implicações, todos os cálculos são repetidos considerando x uma dimensão e depois a outra, exceto a métrica EF.

O primeiro valor a ser gerado é a contabilização da frequência da implicação, assim, caso um ou mais registros apresentem a mesma associação entre termos para os mesmos estados, o valor deste campo da matriz deve acompanhar o total de ocorrências (EF - *entailment frequency*).

O próximo passo conta o número total de ocorrências do termo em determinado estado considerando todo o *corpus*, contabilizando a frequência do termo (TF - *term frequency*) para as duas dimensões. Em outras palavras, para cada associação, cada termo é contabilizado em sua respectiva dimensão, trazendo o levantamento de quantas vezes ele aparece no *corpus* da sua própria dimensão.

Em ambas as métricas apresentadas, EF e TF, a lógica foi similar ao cálculo do suporte, entretanto, apresentando os dados de forma discreta e não percentual.

A próxima métrica é calculada pela distribuição da associação (ED - *entailment distribution*) e apresenta a relação percentual da associação sobre todos as implicações com o termo analisado. Similar ao cálculo da confiança, esta métrica calcula a relação da associação em relação a ocorrência do termo da dimensão analisada. A fórmula a seguir apresenta a divisão da frequência da associação pela frequência total do termo no *corpus*.

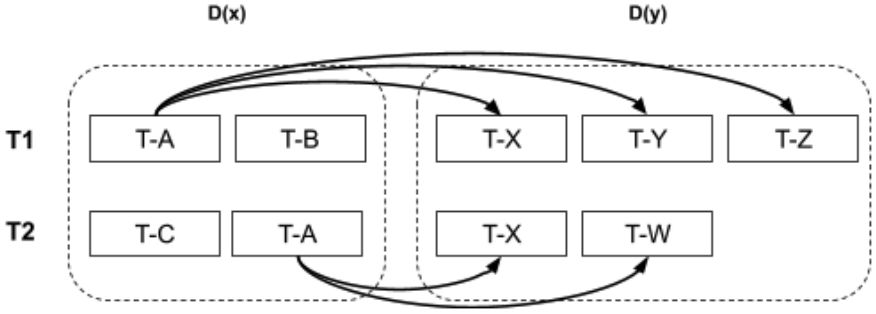
$$EDx = EF \div TFx$$

A etapa seguinte analisa a frequência combinatória (CF - *combinatory frequency*) das implicações entre os termos de ambas dimensões da relação, aplicando a seguinte fórmula e considerando Z o termo analisado.

$$CFx = \sum_{i=1}^{n} Xi \times Yi = \{(x, y)|x \in Xi \wedge y \in Yi|x = Z\}$$

Para todos os documentos da relação, o cálculo contabiliza, para cada termo, os termos do estado oposto para todo o *corpus*. No exemplo da Figura 5.6, o *item* "T-A" tem a métrica CF é igual a 5 quando analisados os registros das *transactions* T1 e T2. Esta contabilização não descarta repetidas associações de termos da dimensão oposta quando encontradas em diferentes documentos, como no caso do *item* "T-A" de D(x) que se relaciona com o "T-X" em diferentes *transactions*.

Figura 5.6: Levantamento da métrica CF para o item T-A do D(x)



Fonte: Dos Autores.

A próxima métrica calcula o peso da associação referente ao termo de cada dimensão, (TD - *term distribution*). A representação percentual obtida representa o resto da divisão entre o total de ocorrências da associação pela frequência combinatória de cada dimensão, conforme fórmula a seguir.

$$TDx = EF \div CFx$$

O cálculo da probabilidade da implicação (EP - *entailment probability*) considera o percentual da distribuição da implicação, ED, sobre a distribuição do termo analisado, TD, conforme próxima fórmula. A utilização do TD regula a taxa do ED caso o termo seja utilizado em muitos documentos, sendo assim considerado de pouca relevância. Este cálculo segue o mesmo conceito do IDF (*inverse document frequency*).

$$EPx = (EDx \times TDx) \div 100$$

Com a elaboração dos cálculos utilizados pelo método WAR, é possível relacionar todos os termos de todos os estados entre si. O levantamento destes valores resulta em uma matriz de pesos de implicação entre todos os elementos de texto para todos os segmentos existentes. Todas as métricas e suas respectivas descrições estão sumarizadas na Tabela 5.1. Com estas métricas na matriz MMIR, é possível iniciar o mecanismo de recuperação dos termos.

Tabela 5.1: Métricas utilizadas no método WAR

Métrica	*Descrição*
EF	Número de vezes que essa combinação ocorre
TF_(X|Y)	Total de vezes que o termo ocorre na respectiva dimensão
ED_(X|Y)	Peso da implicação sobre todas as ocorrências do termo
CF_(X|Y)	Total de relações com termos da dimensão oposta
TD_(X|Y)	Probabilidade de implicação dos termos
EP_(X|Y)	Probabilidade ponderada da implicação

Fonte: Os Autores

5.3.3 Lógica do processamento

Com os segmentos definidos, a elaboração da lista dos estados se da por meio do método matemático de combinações únicas das dimensões mais a auto-relação.

Cada relação encontrada é composta por duas dimensões, sendo cada uma a representação de um estado de dados a ser analisado. Assim, é gerado o conteúdo da MMRT quando realizando o levantamento de relações entre registros de ambas as dimensões para cada relação. Para este feito, para todas as relações, é repetida uma iteração sobre todos os registros textuais onde são multiplicado os termos da primeira dimensão, D(x), pelos termos da segunda, D(y).

Recebendo dois dados de entrada, o $dataSet$ recebe os dados processados e tabulados, enquanto a variável $stateList$ recebe o nome dos segmentos do $dataSet$ a serem considerados. Como próximo passo, é elaborada uma matriz bidimensional de todas as associações de estados possíveis baseada na listagem $stateList$. Dentro da iteração das associações, cada estado é vinculado a uma dimensão que percorre todo o $dataSet$ considerando apenas as dimensões da associação. Para cada *transaction*, ele separa os *tokens* de palavras de ambas as dimensões e as multiplica. Entretanto, por ser um dado categórico, o resultado obtido é um produto cartesiano de termos, vide Figura 5.5, o qual, por fim, são adicionados na associação da MMRT.

Algorithm 1: Construção da MMRT

Input: $dataSet$
Input: $stateList$
Result: MMRT

```
MMRT ← []
stateList_combinations ← combinations(stateList, 2)
foreach association in stateList_combinations do
    dimensionX ← association.state1
    dimensionY ← association.state2
    foreach dataRow in dataSet do
        listItemSetsDx ← dataRow[dimensionX].splitEachWord()
        listItemSetsDy ← dataRow[dimensionY].splitEachWord()
        listItemSets = listItemSetsDx × listItemSetsDy
        foreach entailment in listItemSets do
            MMRT[association][x] ← entailment[x]
            MMRT[association][y] ← entailment[y]
```

Com a utilização da MMRT, é iniciado o processo de desenvolvimento da matriz multidimensional de implicação recíproca. Para cada relação, o desenvolvimento da MMRI utiliza os valores MMRT como base de cálculo. Inicialmente, é contabilizada a frequência das associações, métrica EF e, subsequentemente, são calculadas as demais métricas sobre a perspectiva de cada dimensão, são elas: TF; ED; CF; TD; e EP. Este processo é descrito pelo algoritmo a seguir.

5.4 Recuperação

A etapa de recuperação conta com os cálculos já realizados na MMRI para a consulta e elaboração do ranque. Por conseguinte, com base em termos de consulta associados a segmentos, a etapa de recuperação considera a implicação dos termos e seus pesos para apresentar quais termos estão diretamente relacionados e, assim, os apresenta por ordem do peso da implicação.

A recuperação dos termos correlacionados é executada em quatro etapas: formatação dos termos de consulta; levantamento dos termos correlacionados aos da consulta; recuperação das métricas dos termos de busca e correlacionados; e por fim, normalização e apresentação dos resultados em forma de ranque.

Algorithm 2: Construção da MMRT

Input: $MMRT$
Result: MMRI

$MMRI \leftarrow []$
foreach $association$ *in* $MMRT$ **do**
  foreach $entailment$ *in* $MMRT[association]$ **do**
    $MMRI[association][x] < MMRT[association][x]$
    $MMRI[association][y] < MMRT[association][y]$
    $MMRT[association][EF] \leftarrow metricEF()$
    $dimensionX \leftarrow association.entailment[x]$
    $dimensionY \leftarrow association.entailment[y]$
    $MMRT[association][TF_x] \leftarrow metricTF(dimensionX)$
    $MMRT[association][TF_y] \leftarrow metricTF(dimensionY)$
    $MMRT[association][ED_x] \leftarrow metricED(dimensionX)$
    $MMRT[association][ED_y] \leftarrow metricED(dimensionY)$
    $MMRT[association][CF_x] \leftarrow metricCF(dimensionX)$
    $MMRT[association][CF_y] \leftarrow metricCF(dimensionY)$
    $MMRT[association][TD_x] \leftarrow metricTD(dimensionX)$
    $MMRT[association][TD_y] \leftarrow metricTD(dimensionY)$
    $MMRT[association][EP_x] \leftarrow metricEP(dimensionX)$
    $MMRT[association][EP_y] \leftarrow metricEP(dimensionY)$

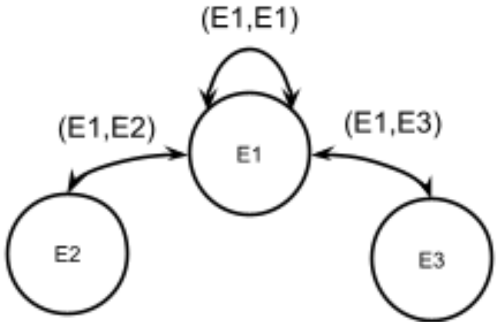
5.4.1 Formatação dos Termos de Consulta

A formatação dos termos de entrada deve ser realizada a luz do método também definido e aplicado de transformação do texto para os registros, permitindo assim a padronização entre base de dados e termos da busca. Como entrada, a consulta deve receber uma lista de um ou mais termos para um ou mais estados.

5.4.2 Levantamento dos Termos Correlacionados com os da Consulta

Com os termos de entrada na consulta já formatados, o algoritmo gera uma lista e os define como “Termos de busca”. Para cada estado consultado, o algoritmo busca a relação de todos os termos na busca com os termos relacionados no mesmo estado e em estados distintos. No exemplo de uma busca em um corpus de 3 estados, conforme a Figura 5.7, quando consultado os termos do estado “E1”, o algoritmo busca nas relações os termos implicados no próprio estado e nos estados distintos, sendo eles os estados “E2”, “E3” e o próprio “E1”.

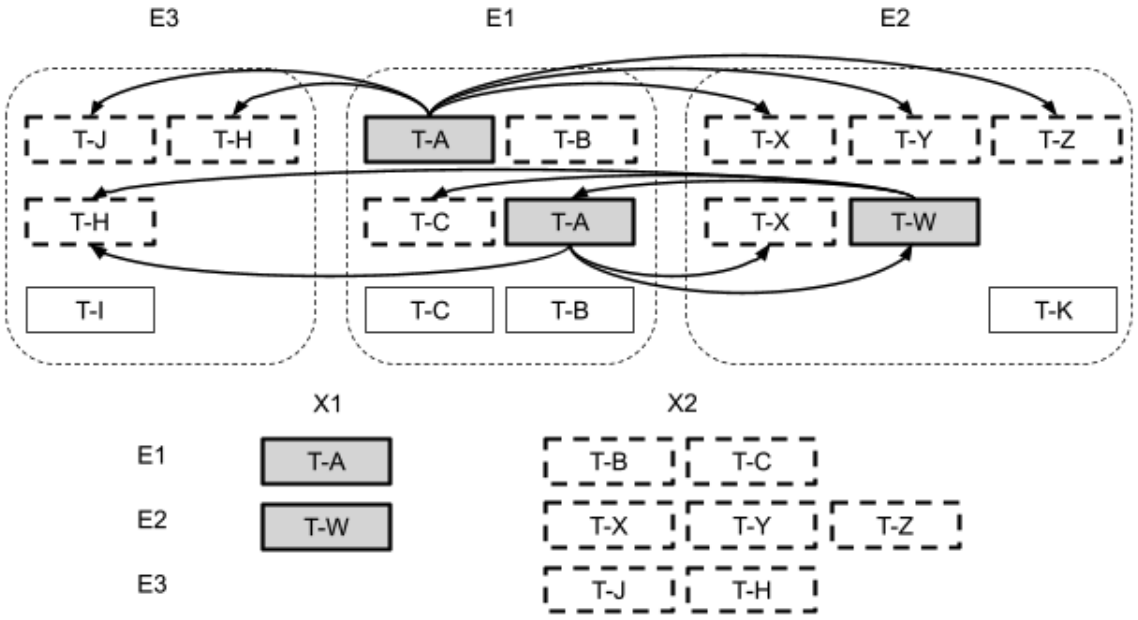
Figura 5.7: Busca de relações na consulta no estado E1 para método de 3 estados



Fonte: Dos Autores.

Após o levantamento de correlação, os termos resultantes consideram apenas os itens distintos entre todas as relações e, assim, retornando a lista de "Termos de implicações" por estado. No exemplo da Figura 5.8, uma busca do "T-A" no estado "E1" e "T-W" no estado "E2", quando analisadas as implicações no próprio estado, o "T-A" encontra "T-B" e "T-C" e o "T-W" encontra o "T-X". Quando analisadas as implicações em estados distintos, os itens "T-X", "T-Y", "T-Z", "T-J" e "T-H" são adicionados à lista de implicações já que este processo é repetido para todas as *transactions*.

Figura 5.8: Levantamento de implicações de termos na busca



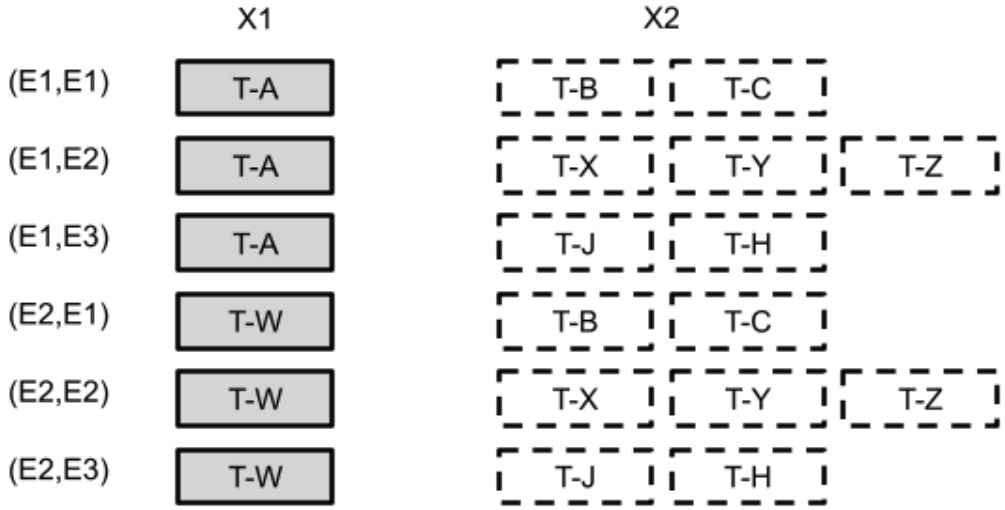
Fonte: Dos Autores.

5.4.3 Recuperação das Métricas dos Termos

Utilizando as listas de “Termos de busca” e “Termos de implicações”, o levantamento das métricas busca as relações entre os estados de ambas as listas. O resultado do levantamento destas implicações retorna as métricas de todos os termos implicados com base na perspectiva da relação do termo de busca.

Na Figura 5.9, o item “T-A” do estado “E1”, por exemplo, busca todas as métricas de relações possíveis com os termos da lista de implicações, sendo eles: “T-B”, “T-C”, “T-X”, “T-Y”, “T-Z”, “T-J”, e “T-H”. O mesmo é realizado para todos os demais termos da busca.

Figura 5.9: Exemplo do levantamento de implicações de termos na busca



Fonte: Dos Autores.

5.4.4 Cálculo do Ranque

Para cada termo implicado é calculado o valor do peso do termo em relação aos termos da busca. Este cálculo é realizado seguindo a fórmula apresentada abaixo que utiliza as métricas ED, distribuição da implicação, em relação a métrica EP, probabilidade da implicação. Com todos os valores calculados, é realizada uma média aritmética dos pesos encontrados resultando assim no peso final de implicação do termo.

$$\frac{1}{n}\sum_{i=1}^{n}(EPx_i \times EDy_i + EPy_i \times EDx_i) \div 4$$

Como última etapa da recuperação dos termos relacionados, a apresentação dos pesos separa por estado a lista de elementos implicados pelos itens da consulta. Neste resultado, cada termo é apresentado com o seu respectivo peso final de implicação, seguindo a ordenação do maior para o menor, o qual o maior valor representa a maior associação.

5.4.5 Lógica do Processamento

Com as entradas de texto formatadas, o próximo objetivo é realizar o levantamento dos termos correlacionados com os da consulta. De maneira inicial, o algoritmo define os termos de entrada como "Termos de busca" e, com base neles, gera uma lista de termos relacionados e os define como "Termos de implicações". Para cada relação de segmento existente, o sistema busca na MMRI as métricas de cada associação que estão relacionadas aos termos da busca e aos termos relacionados.

Para permitir a definição de termos de busca específicos por segmento de texto, o algoritmo recebe como entrada uma matriz de busca que associa para cada segmento os termos. Como próxima etapa, os termos são relacionados e as métricas são recuperadas e calculadas na matriz MMRI. Por fim, o ranque é ordenado e estruturado em uma matriz que separa por segmento de texto conforme apresentado no seguinte algoritmo.

Algorithm 3: Levantamento das relações entre estados

Input: $MatrixSearch$
Result: MatrixRanking

```
1 MatrixRanking ← getAllSegments()
2 MatrixAssociation ← getAllAssociation(MatrixSearch)
3 foreach searchTerm in MatrixSearch do
4     foreach associationTerm in MatrixAssociation do
5         metrics ← getMetric(MMRI, searchTerm, associationTerm)
          MatrixRanking[state] ← calcRanking(metrics)
6 MatrixRanking ← sortRanking(MatrixRanking)
```

6 EXPERIMENTOS E RESULTADOS

Este Capítulo descreve os experimentos realizados para avaliar o método proposto. O objetivo geral da avaliação é identificar a contribuição técnica avaliando de duas maneiras: a primeira em um experimento preliminar baseado em uma base de dados gerada artificialmente e, em um segundo momento, com a utilização de uma base de dados real de teleconsultoria médica.

Para o desenvolvimento com uso da base de dados médica, foi necessário o domínio do tema da base de dados e devido a esta necessidade, esta pesquisa foi realizada em parceria entre as áreas de computação e medicina da Universidade Federal do Rio Grande do Sul. O desenvolvimento do pré-processamento da base de dados, execução dos cálculos das matrizes propostas e a recuperação das associações, seguem os métodos e fórmulas apresentadas no Capítulo 5 e exemplificadas nas seções seguintes.

Desde a autorização para o uso da base de dados à validação dos resultados intermediários, passando pela elaboração das técnicas de pré-processamento, todas as etapas foram validadas em conjunto com uma médica e seu orientador especializados em dermatologia, no contexto do seu trabalho de mestrado. Assim, mantendo o desenvolvimento e a melhoria da lógica de processamento e recuperação do método proposto com atenção exclusiva desta pesquisa.

6.1 Experimento Preliminar - Base de Dados Sintética

A fim de identificar o resultado do desempenho do método proposto que permita analisar sem a necessidade de um especialista no assunto, foi elaborada uma base de dados artificial com menos registros. Possuindo apenas dois segmentos de texto representados nas colunas "Segmento 1"e "Segmento 2"com cinco termos diferentes cada, totalizando nove registros, a base de dados é apresentada na Tabela 6.1.

Esta base de dados foi utilizada na execução de três cenários de comparação entre os algoritmos Apriori (considerado baseline) e WAR. O primeiro cenário retrata a análise de associação dos termos em um segmento único de texto. O segundo cenário buscou identificar a relação entre os termos dos segmentos diferentes. Por fim, o terceiro com busca de termos em ambos os seguimentos.

Tabela 6.1: Base de dados gerada artificialmente para análise preliminar

	Segmento 1	*Segmento 2*
1	neoplasia dermatite sifilis ceratose	sabonete acido perfume
2	neoplasia dermatite	sabonete sintoma perfume hidratante
3	neoplasia dermatite sifilis eczema	sintoma hidratante acido
4	neoplasia dermatite ceratose eczema	sabonete perfume
5	dermatite ceratose	sabonete acido perfume
6	dermatite	sintoma acido perfume hidratante
7	dermatite	perfume
8	ceratose eczema	sabonete perfume
9	ceratose eczema	sabonete

Fonte: Os Autores

A implementação clássica do Apriori busca encontrar a associação das ocorrências em cada transação, traduzindo para os termos de busca textual, ou seja, buscando a associação de cada termo com os demais termos da transação.

Quando o algoritmo WAR é implementado para analisar somente um segmento de texto, ele tem um comportamento similar ao do Apriori, entretanto, processa as associações de maneira distinta, pois considera a lógica de frequência dos termos de forma geral e não somente das ocorrências de associação.

6.1.1 Comparação Apriori e WAR - Buscas na base de dados de um seguimento

Utilizando a base de dados gerada artificialmente que foi apresentada na Tabela 6.1, foi definido como objetivo identificar a associação dos termos de um segmento de texto com os demais termos do mesmo segmento, neste caso, o segmento escolhido foi a coluna "Segmento 1". Como primeira etapa de comparação, a coluna "Segmento 1"da base de dados foi submetida ao processamento dos métodos WAR e Apriori. Nesta etapa, o objetivo é listar por meio de um ranque de peso os termos do mesmo segmento que são associados com um termo de busca.

Submetendo a base ao processamento do método WAR, a Tabela 6.2 traz de maneira desmembrada todas as métricas para exemplificação do passo a passo dos cálculos necessários para a elaboração do ranque. Esta tabela apresenta todas as combinações possíveis entre termos distintos encontrados no mesmo segmento.

Para fins de comparação, a mesma base de dados foi submetida ao processamento do Apriori e para realizar uma análise geral, o suporte foi definido em 0.05, permitindo a análise de todas as ocorrências de associação.

Tabela 6.2: Processamento WAR: Associação entre os termos do Segmento 1

x	y	EF	TF_X	TF_Y	ED_X	ED_Y	CF_X	CF_Y	TD_X	TD_Y	EP_X	EP_Y
dermatite	ceratose	3	7	5	42.86	60.0	4	4	0.75	0.75	0.32	0.45
dermatite	eczema	2	7	4	28.57	50.0	4	4	0.5	0.5	0.14	0.25
dermatite	neoplasia	4	7	4	57.14	100.0	4	4	1.0	1.0	0.57	1.0
dermatite	sifilis	2	7	2	28.57	100.0	4	4	0.5	0.5	0.14	0.5
eczema	ceratose	3	4	5	75.0	60.0	4	4	0.75	0.75	0.56	0.45
eczema	neoplasia	2	4	4	50.0	50.0	4	4	0.5	0.5	0.25	0.25
eczema	sifilis	1	4	2	25.0	50.0	4	4	0.25	0.25	0.06	0.12
neoplasia	ceratose	2	4	5	50.0	40.0	4	4	0.5	0.5	0.25	0.2
neoplasia	sifilis	2	4	2	50.0	100.0	4	4	0.5	0.5	0.25	0.5
sifilis	ceratose	1	2	5	50.0	20.0	4	4	0.25	0.25	0.12	0.05

Fonte: Os Autores

Com o intuito de identificar os termos associados em relação aos termos de busca, foram filtradas as associações entre dois termos somente e os resultados obtidos são mostrados na Tabela 6.3.

Tabela 6.3: Processamento Apriori: Associação entre os termos do Segmento 1

Suporte	*Termo 1*	*Termo 2*
0.444444	dermatite	neoplasia
0.333333	dermatite	ceratose
0.333333	eczema	ceratose
0.222222	dermatite	sifilis
0.222222	dermatite	eczema
0.222222	neoplasia	sifilis
0.222222	ceratose	neoplasia
0.222222	eczema	neoplasia
0.111111	eczema	sifilis
0.111111	ceratose	sifilis

Fonte: Os Autores

6.1.1.1 Busca 1: termo neoplasia

Ao buscar por **neoplasia**, foi possível identificar que, para ambos os algoritmos, o termo **dermatite** foi considerado de alta associação. Entretanto, o comportamento das demais associações foi diferente, conforme pode ser observado na Tabela 6.4.

Para exemplificar a diferença de análise dos algoritmos, as ocorrências dos termos foram estruturadas em forma de grade na Figura 6.1. As ocorrências do termo de busca estão destacadas em amarelo, enquanto os termos associados são representados na cor azul, por fim, os termos não associados em laranja.

Tabela 6.4: Resultado da comparação entre Apriori e WAR: cenário 1 e busca 1

Apriori			WAR		
Posição	Suporte	Associado	Posição	ER	Associado
1	0.444444	dermatite	1	28,535	dermatite
2	0.222222	sifilis	2	12,500	sifilis
2	0.222222	eczema	3	6,250	eczema
2	0.222222	ceratose	4	5,000	ceratose

Fonte: Os Autores

Dessa forma, é possível observar que o Apriori usa as ocorrências em azul para definir o grau de associação, enquanto o método WAR usa as ocorrências em azul para aumentar o peso no ranking e as não associações em laranja para reduzir. Trazendo assim o critério de desempate para termos com o mesmo número de termos associadas.

Figura 6.1: Grade de associação dos termos na busca por neoplasia

neoplasia	dermatite	sifilis	ceratose	
neoplasia	dermatite			
neoplasia	dermatite	sifilis		eczema
neoplasia	dermatite		ceratose	eczema
	dermatite		ceratose	
	dermatite			
	dermatite			
			ceratose	eczema
			ceratose	eczema

Fonte: Dos Autores.

Outra forma de visualizar os dados da Figura 6.1 é tabular as ocorrências distribuindo a frequência em termos **associados** e **não associados** e esta distribuição pode ser observada na Tabela 6.5. Quando analisadas as associações, observa-se que o termo que lidera o ranque é o que mais possui ocorrências em relação ao termo de busca. Todos os demais termos possuem o mesmo número de ocorrência de termos **associados**, entretanto, as ocorrências de **não associados** ajustam o peso e a ordem do ranque.

Tabela 6.5: Tabela de associação dos termos na busca por neoplasia

Termo	*Frequência*	*Associados*	*Não Associados*
dermatite	7	4	3
sifilis	2	2	0
eczema	4	2	2
ceratose	5	2	3

Fonte: Os Autores

6.1.1.2 Busca 2: termo dermatite

Ao realizar a busca pelo termo **dermatite**, as associações são distribuídas conforme a Tabela 6.6. Com base nesta distribuição, quando analisada a execução dos algoritmos Apriori e WAR, o resultado é apresentado na Tabela 6.7.

Tabela 6.6: Tabela de associação dos termos na busca por dermatite

Termo	*Frequência*	*Associados*	*Não Associados*
neoplasia	4	4	0
ceratose	5	3	2
sifilis	2	2	0
eczema	4	2	2

Fonte: Os Autores

Tabela 6.7: Resultado da comparação entre Apriori e WAR: cenário 1 e busca 2

Apriori			WAR		
Posição	Suporte	Associado	Posição	ER	Associado
1	0.444444	neoplasia	1	28,535000	neoplasia
2	0.333333	ceratose	2	9,621750	ceratose
3	0.222222	sifilis	3	7,071250	sifilis
4	0.222222	eczema	4	3,535625	eczema

Fonte: Os Autores

6.1.1.3 Busca 3: termo sifilis

Ao realizar a busca pelo termo **sifilis**, as associações são distribuídas conforme a Tabela 6.8 e, quando executados os algoritmos Apriori e WAR, o resultado é o apresentado na Tabela 6.9.

Tabela 6.8: Tabela de associação dos termos na busca por sifilis

Termo	*Frequência*	*Associados*	*Não Associados*
neoplasia	5	2	2
dermatite	7	2	5
eczema	4	1	3
ceratose	5	1	4

Fonte: Os Autores

Tabela 6.9: Resultado da comparação entre Apriori e WAR: cenário 1 e busca 3

Apriori			WAR		
Posição	Suporte	Associado	Posição	ER	Associado
1	0.222222	neoplasia	1	12,50000	neoplasia
1	0.222222	dermatite	2	7,07125	dermatite
2	0.111111	eczema	3	1,50000	eczema
2	0.111111	ceratose	4	1,22500	ceratose

Fonte: Os Autores

6.1.1.4 Busca 4: termo ceratose

Ao realizar a busca pelo termo **ceratose**, as associações são distribuídas conforme a Tabela 6.10 e os algoritmos tem os resultados apresentados na Tabela 6.11.

Tabela 6.10: Tabela de associação dos termos na busca por ceratose

Termo	*Frequência*	*Associados*	*Não Associados*
eczema	4	3	1
dermatite	7	3	4
neoplasia	4	2	2
sifilis	2	1	1

Fonte: Os Autores

6.1.1.5 Busca 5: termo eczema

Ao realizar a busca pelo termo **eczema**, as associações são distribuídas conforme a Tabela 6.12 e, quando executados os algoritmos Apriori e WAR, o resultado é o apresentado na Tabela 6.13.

Tabela 6.11: Resultado da comparação entre Apriori e WAR: cenário 1 e busca 4

Apriori			WAR		
Posição	Suporte	Associado	Posição	ER	Associado
1	0.333333	eczema	1	16,83750	eczema
1	0.333333	dermatite	2	9,62175	dermatite
2	0.222222	neoplasia	3	5,00000	neoplasia
3	0.111111	sifilis	4	1,22500	sifilis

Fonte: Os Autores

Tabela 6.12: Tabela de associação dos termos na busca por eczema

Termo	*Frequência*	*Associados*	*Não Associados*
ceratose	5	3	2
neoplasia	4	2	2
dermatite	7	2	5
sifilis	2	1	1

Fonte: Os Autores

6.1.1.6 Considerações sobre a Consulta em um Seguimento de Texto

Em todos os casos de comparação, foi possível identificar que o algoritmo de associação Apriori e o método WAR apresentam os resultados em ordens equivalentes. Entretanto, para buscar maior precisão no ranque, o método WAR também considera os métodos **não associados**, ou seja, o cálculo considera a frequência dos termos **associados** e **não associados**.

6.1.2 Comparação Apriori e WAR - Buscas na Base de Dados de Dois Seguimentos

Assim como na primeira rodada de pesquisa, a qual comparou o método WAR com o Apriori, este método repetirá a comparação, entretanto, comparando com dois segmentos de texto, ou seja, considerando as colunas "Segmento 1"e "Segmento 2"apresentadas na Tabela 6.1.

Nesta nova rodada de testes, foram utilizado os mesmos termos já utilizados, entretanto, desta vez julgando o impacto do termo de busca, "Segmento 1", nos termos do outro seguimento, o "Segmento 2".

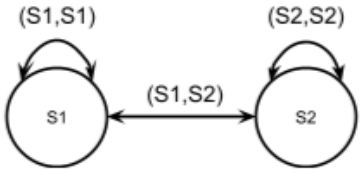
Quando submetido ao processamento do método WAR uma base de dados com dois segmentos S1 e S2, são encontradas 3 associações, sendo elas: (S1,S1); (S1,S2); e (S2,S2). Esta representação pode ser vista na Figura 6.2.

Tabela 6.13: Resultado da comparação entre Apriori e WAR: cenário 1 e busca 5

Apriori			WAR		
Posição	Suporte	Associado	Posição	ER	Associado
1	0.333333	ceratose	1	16,837500	ceratose
2	0.222222	neoplasia	2	6,250000	neoplasia
2	0.222222	dermatite	3	3,535625	dermatite
3	0.111111	sifilis	4	1,500000	sifilis

Fonte: Os Autores

Figura 6.2: Grafo de associações mapeadas pelo WAR



Fonte: Dos Autores.

Como resultado desse processamento, é obtida a matriz MMRI apresentada na Tabela 6.14, a qual é apresentada na integra. Entretanto, para identificar a implicação entre os segmentos S1 e S2 é necessário a utilização apenas da implicação (S1, S2).

Devido ao processamento com lógicas diferentes, enquanto o método WAR permite por padrão a entrada e o processamento de duas colunas de texto e valida a associação entre elas, o Apriori precisa que todos os termos sejam unidos em uma coluna única. Para isso, foi realizada uma etapa de pré-processamento que adiciona o prefixo "s1_"e "s2_"nos termos das respectivas colunas "Segmento 1"e "Segmento 2". Este ajuste pode ser visto na Tabela 6.15.

Ao submeter esta base de dados a lógica de associação do Apriori, foi definido um *threshold* para o suporte em 0,05. Com o objetivo de identificar a associação entre segmentos diferentes, o resultado foi filtrado considerando apenas associações entre segmentos diferentes e descartando segmentos iguais, gerando o resultado apresentado na Tabela 6.16.

Com base nos resultados de processamento obtidos, foram realizadas cinco buscas com os termos já utilizados, entretanto, buscando no ranque a ordem de associação dos termos do outro segmento.

6.1.2.1 Busca 1: termo neoplasia

Ao realizar a busca pelo termo **neoplasia**, os algoritmos Apriori e WAR apresentaram o resultado conforme apresentado na Tabela 6.17.

Tabela 6.14: Processamento WAR: MMRI dos Segmentos 1 e 2

x	y	EF	TF_X	TF_Y	ED_X	ED_Y	CF_X	CF_Y	TD_X	TD_Y	EP_X	EP_Y
					(S1,S1)							
dermatite	ceratose	3	7	5	42.86	60.0	4	4	0.75	0.75	0.32	0.45
dermatite	eczema	2	7	4	28.57	50.0	4	4	0.5	0.5	0.14	0.25
dermatite	neoplasia	4	7	4	57.14	100.0	4	4	1.0	1.0	0.57	1.0
dermatite	sifilis	2	7	2	28.57	100.0	4	4	0.5	0.5	0.14	0.5
eczema	ceratose	3	4	5	75.0	60.0	4	4	0.75	0.75	0.56	0.45
eczema	neoplasia	2	4	4	50.0	50.0	4	4	0.5	0.5	0.25	0.25
eczema	sifilis	1	4	2	25.0	50.0	4	4	0.25	0.25	0.06	0.12
neoplasia	ceratose	2	4	5	50.0	40.0	4	4	0.5	0.5	0.25	0.2
neoplasia	sifilis	2	4	2	50.0	100.0	4	4	0.5	0.5	0.25	0.5
sifilis	ceratose	1	2	5	50.0	20.0	4	4	0.25	0.25	0.12	0.05
					(S2,S2)							
acido	hidratante	2	4	3	50.00	66.67	4	4	0.50	0.50	0.25	0.33
acido	perfume	3	4	7	75.00	42.86	4	4	0.75	0.75	0.56	0.32
acido	sabonete	2	4	6	50.00	33.33	4	4	0.50	0.50	0.25	0.17
acido	sintoma	2	4	3	50.00	66.67	4	4	0.50	0.50	0.25	0.33
hidratante	perfume	2	3	7	66.67	28.57	4	4	0.50	0.50	0.33	0.14
hidratante	sabonete	1	3	6	33.33	16.67	4	4	0.25	0.25	0.08	0.04
hidratante	sintoma	3	3	3	100.00	100.00	4	4	0.75	0.75	0.75	0.75
perfume	sabonete	5	7	6	71.43	83.33	4	4	1.25	1.25	0.89	1.04
perfume	sintoma	2	7	3	28.57	66.67	4	4	0.50	0.50	0.14	0.33
sabonete	sintoma	1	6	3	16.67	33.33	4	4	0.25	0.25	0.04	0.08
					(S1,S2)							
ceratose	acido	2	5	4	40.00	50.00	3	5	0.67	0.40	0.27	0.20
ceratose	perfume	4	5	7	80.00	57.14	3	5	1.33	0.80	1.06	0.46
ceratose	sabonete	5	5	6	100.00	83.33	3	5	1.67	1.00	1.67	0.83
dermatite	acido	4	7	4	57.14	100.00	5	5	0.80	0.80	0.46	0.80
dermatite	hidratante	3	7	3	42.86	100.00	5	4	0.60	0.75	0.26	0.75
dermatite	perfume	6	7	7	85.71	85.71	5	5	1.20	1.20	1.03	1.03
dermatite	sabonete	4	7	6	57.14	66.67	5	5	0.80	0.80	0.46	0.53
dermatite	sintoma	3	7	3	42.86	100.00	5	4	0.60	0.75	0.26	0.75
eczema	acido	1	4	4	25.00	25.00	5	5	0.20	0.20	0.05	0.05
eczema	hidratante	1	4	3	25.00	33.33	5	4	0.20	0.25	0.05	0.08
eczema	perfume	2	4	7	50.00	28.57	5	5	0.40	0.40	0.20	0.11
eczema	sabonete	3	4	6	75.00	50.00	5	5	0.60	0.60	0.45	0.30
eczema	sintoma	1	4	3	25.00	33.33	5	4	0.20	0.25	0.05	0.08
neoplasia	acido	2	4	4	50.00	50.00	5	5	0.40	0.40	0.20	0.20
neoplasia	hidratante	2	4	3	50.00	66.67	5	4	0.40	0.50	0.20	0.33
neoplasia	perfume	3	4	7	75.00	42.86	5	5	0.60	0.60	0.45	0.26
neoplasia	sabonete	3	4	6	75.00	50.00	5	5	0.60	0.60	0.45	0.30
neoplasia	sintoma	2	4	3	50.00	66.67	5	4	0.40	0.50	0.20	0.33
sifilis	acido	2	2	4	100.00	50.00	5	5	0.40	0.40	0.40	0.20
sifilis	hidratante	1	2	3	50.00	33.33	5	4	0.20	0.25	0.10	0.08
sifilis	perfume	1	2	7	50.00	14.29	5	5	0.20	0.20	0.10	0.03
sifilis	sabonete	1	2	6	50.00	16.67	5	5	0.20	0.20	0.10	0.03
sifilis	sintoma	1	2	3	50.00	33.33	5	4	0.20	0.25	0.10	0.08

Fonte: Os Autores

Tabela 6.15: Base de dados gerada artificialmente ajustada para o Apriori

	Segmento 1 e 2
1	s1_neoplasia s1_dermatite s1_sifilis s1_ceratose s2_sabonete s2_acido s2_perfume
2	s1_neoplasia s1_dermatite s2_sabonete s2_sintoma s2_perfume s2_hidratante
3	s1_neoplasia s1_dermatite s1_sifilis s1_eczema s2_sintoma s2_hidratante s2_acido
4	s1_neoplasia s1_dermatite s1_ceratose s1_eczema s2_sabonete s2_perfume
5	s1_dermatite s1_ceratose s2_sabonete s2_acido s2_perfume
6	s1_dermatite s2_sintoma s2_acido s2_perfume s2_hidratante
7	s1_dermatite s2_perfume
8	s1_ceratose s1_eczema s2_sabonete s2_perfume
9	s1_ceratose s1_eczema s2_sabonete

Fonte: Os Autores

Tabela 6.16: Processamento Apriori: Associação entre os termos do Segmento 1

Suporte	*Termo 1*	*Termo 2*
0.555556	s1_dermatite	s2_acido
0.555556	s1_dermatite	s2_perfume
0.555556	s1_ceratose	s2_sabonete
0.444444	s1_dermatite	s2_sabonete
0.333333	s1_neoplasia	s2_sabonete
0.333333	s1_dermatite	s2_sintoma
0.333333	s1_dermatite	s2_hidratante
0.333333	s1_ceratose	s2_perfume
0.333333	s1_eczema	s2_sabonete
0.333333	s1_eczema	s2_perfume
0.222222	s1_neoplasia	s2_acido
0.222222	s1_neoplasia	s2_perfume
0.222222	s1_neoplasia	s2_sintoma
0.222222	s1_neoplasia	s2_hidratante
0.222222	s1_sifilis	s2_acido
0.222222	s1_ceratose	s2_acido
0.111111	s1_sifilis	s2_sabonete
0.111111	s1_sifilis	s2_perfume
0.111111	s1_sifilis	s2_sintoma
0.111111	s1_sifilis	s2_hidratante
0.111111	s1_eczema	s2_acido
0.111111	s1_eczema	s2_sintoma
0.111111	s1_eczema	s2_hidratante

Fonte: Os Autores

6.1.2.2 Busca 2: termo dermatite

Ao realizar a busca pelo termo **dermatite**, os algoritmos Apriori e WAR apresentaram o resultado conforme apresentado na Tabela 6.18.

Tabela 6.17: Resultado da comparação entre Apriori e WAR: cenário 2 e busca 1

Apriori			WAR		
Posição	Suporte	Associado	Posição	ER	Associado
1	0.333333	sabonete	1	11,25000	sabonete
2	0.222222	perfume	2	9,69675	perfume
2	0.222222	hidratante	3	7,45850	hidratante
2	0.222222	sintoma	3	7,45850	sintoma
2	0.222222	acido	4	5,00000	acido

Fonte: Os Autores

Tabela 6.18: Resultado da comparação entre Apriori e WAR: cenário 2 e busca 2

Apriori			WAR		
Posição	Suporte	Associado	Posição	ER	Associado
1	0.555556	perfume	1	44,14065	perfume
1	0.555556	acido	2	22,92800	acido
2	0.444444	sabonete	3	15,23810	sabonete
2	0.333333	hidratante	4	14,53625	hidratante
2	0.333333	sintoma	4	14,53625	sintoma

Fonte: Os Autores

6.1.2.3 Busca 3: termo sifilis

Ao realizar a busca pelo termo **sifilis**, os algoritmos Apriori e WAR apresentaram o resultado conforme apresentado na Tabela 6.19.

Tabela 6.19: Resultado da comparação entre Apriori e WAR: cenário 2 e busca 3

Apriori			WAR		
Posição	Suporte	Associado	Posição	ER	Associado
1	0.222222	acido	1	10,00000	acido
2	0.111111	hidratante	2	1,83325	hidratante
2	0.111111	sintoma	2	1,83325	sintoma
2	0.111111	perfume	3	0,79175	perfume
2	0.111111	sabonete	4	0,73225	sabonete

Fonte: Os Autores

6.1.2.4 Busca 4: termo ceratose

Ao realizar a busca pelo termo **ceratose**, os algoritmos Apriori e WAR apresentaram o resultado conforme apresentado na Tabela 6.20.

Tabela 6.20: Resultado da comparação entre Apriori e WAR: cenário 2 e busca 4

Apriori			WAR		
Posição	Suporte	Associado	Posição	ER	Associado
1	0.555556	sabonete	1	55,540275	sabonete
2	0.333333	perfume	2	24,342100	perfume
2	0.222222	acido	3	5,375000	acido

Fonte: Os Autores

6.1.2.5 Busca 5: termo eczema

Ao realizar a busca pelo termo **eczema**, os algoritmos Apriori e WAR apresentaram o resultado conforme apresentado na Tabela 6.21.

Tabela 6.21: Resultado da comparação entre Apriori e WAR: cenário 2 e busca 5

Apriori			WAR		
Posição	Suporte	Associado	Posição	ER	Associado
1	0.333333	sabonete	1	11,250000	sabonete
1	0.333333	perfume	2	2,803500	perfume
2	0.111111	hidratante	3	0,916625	hidratante
2	0.111111	sintoma	3	0,916625	sintoma
2	0.111111	acido	4	0,625000	acido

Fonte: Os Autores

6.1.2.6 Considerações sobre a pesquisa em base com dois seguimentos de texto

Após a execução das buscas em termos associados em outro segmento de texto, foi possível buscar a coerência de resultados quando comparado o método proposto com o Apriori. Ambos os métodos resultaram em uma listagem de termos priorizados pela associação, contudo o método WAR apresentou uma precisão diferente nos resultados, além da ausência da necessidade de pré-processamento.

6.1.3 Comparação Apriori e WAR - Buscas com termos nos dois seguimentos

As buscas desta seção contam com termos compostos em ambos os segmentos, ou seja, dois termos no "Segmento 1"e dois termos no "Segmento 2". Para definir o par de termos, foram escolhidos os de maior associação de cada segmento e os de menor peso de associação que não contivessem termos do par de maior associação. As buscas realizadas relacionam todas as relações de par de termos apresentadas na Tabela 6.22.

Tabela 6.22: Relação de par de termos de maior e menor associação

Associação	**Segmento 1**	**Segmento 2**
Maior	(dermatite, neoplasia)	(sintoma, hidratante)
Menor	(ceratose, sifilis)	(acido, sabonete)

Fonte: Os Autores

Para este cenário de busca, não foi necessário reprocessar a base de dados. Para o método WAR, foi utilizada a MMRI já apresentada na Tabela 6.14. Para o processamento do Apriori foi necessário manter a definição do *threshold* para o suporte em 0,05, mas filtrar as associações com 5 ocorrências sendo no mínimo duas de cada segmento, gerando o resultado apresentado na Tabela 6.23 para impacto no "Segmento 1"e na Tabela 6.24 para impacto no "Segmento 2".

Tabela 6.23: Processamento Apriori: Associação do Apriori entre 3 termos

Suporte	*Termo 1*	*Termo 2*	*Termo 3*	*Termo 4*	*Termo 5*
0.111111	s1_sifilis	s1_eczema	s1_dermatite	s2_perfume	s2_sintoma
0.111111	s1_sifilis	s1_eczema	s1_dermatite	s2_hidratante	s2_perfume
0.111111	s1_sifilis	s1_eczema	s1_dermatite	s2_hidratante	s2_sintoma
0.111111	s1_dermatite	s1_neoplasia	s1_eczema	s2_acido	s2_perfume
0.111111	s1_dermatite	s1_neoplasia	s1_eczema	s2_acido	s2_sintoma
0.111111	s1_dermatite	s1_neoplasia	s1_eczema	s2_acido	s2_hidratante
0.111111	s1_dermatite	s1_neoplasia	s1_eczema	s2_perfume	s2_sintoma
0.111111	s1_dermatite	s1_neoplasia	s1_eczema	s2_hidratante	s2_perfume
0.111111	s1_dermatite	s1_neoplasia	s1_eczema	s2_hidratante	s2_sintoma
0.111111	s1_sifilis	s1_neoplasia	s1_eczema	s2_acido	s2_perfume
0.111111	s1_sifilis	s1_neoplasia	s1_eczema	s2_acido	s2_sintoma
0.111111	s1_sifilis	s1_neoplasia	s1_eczema	s2_acido	s2_hidratante
0.111111	s1_dermatite	s1_sifilis	s1_eczema	s2_acido	s2_perfume
0.111111	s1_dermatite	s1_sifilis	s1_eczema	s2_acido	s2_sintoma
0.111111	s1_dermatite	s1_sifilis	s1_eczema	s2_acido	s2_hidratante
0.111111	s1_dermatite	s1_sifilis	s1_neoplasia	s2_acido	s2_sabonete
0.111111	s1_dermatite	s1_sifilis	s1_neoplasia	s2_acido	s2_perfume
0.111111	s1_dermatite	s1_sifilis	s1_neoplasia	s2_acido	s2_sintoma
0.111111	s1_dermatite	s1_sifilis	s1_neoplasia	s2_acido	s2_hidratante
0.111111	s1_dermatite	s1_sifilis	s1_neoplasia	s2_perfume	s2_sintoma
0.111111	s1_dermatite	s1_sifilis	s1_neoplasia	s2_hidratante	s2_perfume
0.111111	s1_dermatite	s1_sifilis	s1_neoplasia	s2_hidratante	s2_sintoma
0.111111	s1_ceratose	s1_dermatite	s1_neoplasia	s2_acido	s2_sabonete
0.111111	s1_ceratose	s1_sifilis	s1_neoplasia	s2_acido	s2_sabonete
0.111111	s1_neoplasia	s1_eczema	s1_sifilis	s2_perfume	s2_sintoma
0.111111	s1_neoplasia	s1_eczema	s1_sifilis	s2_hidratante	s2_perfume
0.111111	s1_neoplasia	s1_eczema	s1_sifilis	s2_hidratante	s2_sintoma
0.111111	s1_dermatite	s1_ceratose	s1_sifilis	s2_acido	s2_sabonete

Fonte: Os Autores

Tabela 6.24: Processamento Apriori: Associação do Apriori entre 3 termos

Suporte	*Termo 1*	*Termo 2*	*Termo 3*	*Termo 4*	*Termo 5*
0.111111	s1_dermatite	s1_neoplasia	s2_acido	s2_perfume	s2_sintoma
0.111111	s1_dermatite	s1_neoplasia	s2_acido	s2_hidratante	s2_perfume
0.111111	s1_dermatite	s1_neoplasia	s2_acido	s2_hidratante	s2_sintoma
0.222222	s1_dermatite	s1_neoplasia	s2_hidratante	s2_perfume	s2_sintoma
0.111111	s1_dermatite	s1_neoplasia	s2_hidratante	s2_perfume	s2_sabonete
0.111111	s1_dermatite	s1_neoplasia	s2_hidratante	s2_sintoma	s2_sabonete
0.111111	s1_sifilis	s1_neoplasia	s2_hidratante	s2_acido	s2_perfume
0.111111	s1_sifilis	s1_neoplasia	s2_hidratante	s2_acido	s2_sintoma
0.111111	s1_neoplasia	s1_eczema	s2_hidratante	s2_acido	s2_perfume
0.111111	s1_neoplasia	s1_eczema	s2_hidratante	s2_acido	s2_sintoma
0.111111	s1_dermatite	s1_sifilis	s2_hidratante	s2_acido	s2_perfume
0.111111	s1_dermatite	s1_sifilis	s2_hidratante	s2_acido	s2_sintoma
0.111111	s1_dermatite	s1_eczema	s2_hidratante	s2_acido	s2_perfume
0.111111	s1_dermatite	s1_eczema	s2_hidratante	s2_acido	s2_sintoma
0.111111	s1_sifilis	s1_eczema	s2_hidratante	s2_acido	s2_perfume
0.111111	s1_sifilis	s1_eczema	s2_hidratante	s2_acido	s2_sintoma
0.111111	s1_dermatite	s1_neoplasia	s2_perfume	s2_sintoma	s2_sabonete
0.111111	s1_sifilis	s1_neoplasia	s2_perfume	s2_acido	s2_sintoma
0.111111	s1_neoplasia	s1_sifilis	s2_perfume	s2_hidratante	s2_sintoma
0.111111	s1_neoplasia	s1_eczema	s2_perfume	s2_acido	s2_sintoma
0.111111	s1_eczema	s1_neoplasia	s2_perfume	s2_hidratante	s2_sintoma
0.111111	s1_dermatite	s1_sifilis	s2_perfume	s2_acido	s2_sintoma
0.111111	s1_sifilis	s1_dermatite	s2_perfume	s2_hidratante	s2_sintoma
0.111111	s1_dermatite	s1_ceratose	s2_perfume	s2_acido	s2_sabonete
0.111111	s1_dermatite	s1_eczema	s2_perfume	s2_acido	s2_sintoma
0.111111	s1_eczema	s1_dermatite	s2_perfume	s2_hidratante	s2_sintoma
0.111111	s1_sifilis	s1_eczema	s2_perfume	s2_acido	s2_sintoma
0.111111	s1_eczema	s1_sifilis	s2_perfume	s2_hidratante	s2_sintoma

Fonte: Os Autores

6.1.3.1 Busca 1: S1 = (dermatite, neoplasia) e S2 = (sintoma, hidratante)

Ao realizar a busca composta por dois termos tanto no "Segmento 1"quanto no "Segmento 2", foi obtido o resultado apresentado na Tabela 6.25. Contando com ambos os pares de termos de maior associação, sendo S1 = (dermatite, neoplasia) e S2 = (sintoma, hidratante).

Tabela 6.25: Resultado da comparação entre Apriori e WAR: cenário 3 e busca 1

Apriori				WAR			
Segmento	Posição	Suporte	Associado	Segmento	Posição	ER	Associado
Seg 1	1	0.111111	sifilis	Seg 1	1	7.310875	ceratose
	1	0.111111	eczema		2	5.809437	sifilis
					3	2.904719	eczema
Seg 2	1	0.111111	perfume	Seg 2	1	15.804588	perfume
	1	0.111111	acido		2	11.127938	acido
					3	6.955375	sabonete

Fonte: Os Autores

6.1.3.2 Busca 2: S1 = (dermatite, neoplasia) e S2 = (acido, sabonete)

Ao realizar a busca composta por dois termos tanto no "Segmento 1"quanto no "Segmento 2", foi obtido o resultado apresentado na Tabela 6.26. Contando com o par do "Segmento 1"com maior associação, sendo S1 = (dermatite, neoplasia), e o par do "Segmento 2"com menor associação, sendo S2 = (acido, sabonete).

Tabela 6.26: Resultado da comparação entre Apriori e WAR: cenário 3 e busca 2

Apriori				WAR			
Segmento	Posição	Suporte	Associado	Segmento	Posição	ER	Associado
Seg 1	1	0.111111	ceratose	Seg 1	1	18.884256	ceratose
					2	7.590750	sifilis
					3	5.415156	eczema
Seg 2	1	0.111111	perfume	Seg 2	1	25.737631	perfume
					2	7.738331	hidratante
					2	7.738331	sintoma

Fonte: Os Autores

6.1.3.3 Busca 3: S1 = (ceratose, sifilis) e S2 = (sintoma, hidratante)

Ao realizar a busca composta por dois termos tanto no "Segmento 1"quanto no "Segmento 2", foi obtido o resultado apresentado na Tabela 6.27. Contando com o par do "Segmento 1"com menor associação, sendo S1 = (ceratose, sifilis), e o par do "Segmento 2"com maior associação, sendo S2 = (sintoma, hidratante).

Tabela 6.27: Resultado da comparação entre Apriori e WAR: cenário 3 e busca 3

Apriori				WAR			
Segmento	Posição	Suporte	Associado	Segmento	Posição	ER	Associado
Seg 1				Seg 1	1	11.441375	dermatite
					2	8.104250	neoplasia
					3	5.042687	eczema
Seg 2				Seg 2	1	14.416356	sabonete
					2	8.613825	perfume
					3	7.989688	acido

Fonte: Os Autores

6.1.3.4 Busca 4: S1 = (ceratose, sifilis) e S2 = (acido, sabonete)

Ao realizar a busca composta por dois termos tanto no "Segmento 1"quanto no "Segmento 2", foi obtido o resultado apresentado na Tabela 6.28. Contando com ambos os pares de termos de menor associação, sendo S1 = (ceratose, sifilis) e S2 = (acido, sabonete).

Tabela 6.28: Resultado da comparação entre Apriori e WAR: cenário 3 e busca 4

Apriori				WAR			
Segmento	Posição	Suporte	Associado	Segmento	Posição	ER	Associado
Seg 1	1	0.111111	dermatite	Seg 1	1	13.714775	dermatite
	1	0.111111	neoplasia		2	8.437500	neoplasia
					3	7.553125	eczema
Seg 2	1	0.111111	perfume	Seg 2	1	18.546869	perfume
					2	3.597275	hidratante
					2	3.597275	sintoma

Fonte: Os Autores

6.1.3.5 Considerações sobre a pesquisa em dois seguimentos de texto

Em um cenário que busca par de termos em cada segmento de texto, foi identificado que pela obrigatoriedade de ocorrências a abordagem apresentada pelo Apriori vai perdendo precisão e retornando ranques sem termos associados. Para o WAR, a visão categórica e booleana que se tem todas as associações não existe, trazendo sempre a visão em um ranque de escala quantitativa, superando assim a limitação apresentada pelo Apriori.

6.2 Experimento 2: base de dados do DermatoNET

A base de dados contou com um intervalo amostral que conta com todos os registros de teleatendimento do DermatoNET do TelessaúdeRS até a data de 20 de Abril de 2020. Cada registro na base significa um laudo e, para esta dissertação, foram considerados apenas laudos completos, ou seja, os que descrevem a conduta sugerida pelo teleconsultor. Esta definição resultou em uma base de 12.219 *transactions*.

Para definir um escopo de validação para a dissertação, foi definido o total de 3 campos para o processamento, sendo eles 2 campos de texto livre e um de dados categóricos:

- **dconsdados:** referente a consulta, sendo um texto livre com o conjunto de informações sobre a hipótese diagnóstica ou descrição do quadro clínico;
- **dconscondsug:** referente a conduta, sendo um texto livre com a descrição da conduta sugerida pelo médico teleconsultor; e
- **dconscid1cod:** referente ao código CID-10, o qual é um dado categórico do código utilizado na classificação de doenças.

6.2.1 Pré-processamento

Com foco na estabilização dos dados textuais, o arquivo de entrada é submetido a uma série de atividades de normalização e higienização para reduzir a variação dos termos de mesmo significado e remover palavras irrelevantes.

Para esta dissertação, os textos foram simplificados com a execução de técnicas de conversão de texto. Todas as palavras foram convertidas em letras minúsculas, seguido da remoção de tags de marcação de texto, assim como a remoção da acentuação e de qualquer caractere especial, ou seja, tudo que não seja letra ou número.

Por se tratar de um *corpus* de palavras em português, todos os registros foram analisados a fim de identificar a desinência verbal de cada termo e, quando necessário, alterar o sufixo de plural para singular. Este processo realiza a substituição dos sufixos ‘oes’, ‘aos’, ‘aes’, ‘oes’, ‘ais’, ‘eis’, ‘eis’, ‘ois’, ‘uis’, ‘res’, ‘zes’, ‘ses’, ‘is’ e ‘ns’, respectivamente por ‘ao’, ‘ao’, ‘ao’, ‘ao’, ‘al’, ‘el’, ‘il’, ‘ol’, ‘ul’, ‘r’, ‘z’, ‘s’, ‘il’ e ‘m’. Em todas as outras palavras onde os sufixos anteriores não foram encontrados, quando encontrado, a letra ‘s’ como última letra, a mesma é removida.

Um dos pontos a serem considerados na análise textual está na identificação de contextos negativos. Sendo uma área distinta de pesquisa, esta dissertação aplicou apenas uma lógica simples de considerar palavras anteriores para identificar um contexto de negação. Quando identificadas as palavras anteriores **"nega"** e **"nao possui"**, o algoritmo engloba a próxima palavra e substitui os espaços em branco por ‘_’. No exemplo da descrição médica **"nega prurido"**, o token resultante será **"nega_prurido"**.

Como etapa seguinte, é realizado a validação de termos concatenados indevidamente na base de dados original. Foi realizado um levantamento das possíveis anomalias sistêmicas que podem ocorrer em termos armazenados indevidamente sem espaço. Utilizando os termos apresentados no apêndice A, o *dataSet* foi validado separando estes termos.

Outra etapa de processamento utilizado foi o processo de tokenização. Este processo busca ocorrências textuais e as substitui por uma entidade de texto única, um *token*, ou simplesmente a remove. Para a busca e a alteração de registros, foi utilizada a tabela do apêndice B para consulta e apêndice C para o segmento conduta.

Por fim, é realizada a filtragem da base de dados com o uso de *word lists* permissivas. Em cada segmento de texto já processado, é aplicada uma filtragem de *tokens* que devem ser considerados e, para o segmento de consulta, foi aplicada a *word list* apresen-

tada no apêndice D e para conduta foram os *tokens* listados no apêndice E.

6.2.2 Processamento

Com a base processada e os cálculos validados, toda a lógica foi desenvolvida utilizando Python, homologado na versão 3.8.8. Para a estruturação do base de dados, foi utilizada a biblioteca pandas na versão 1.2.4.

Nesta implementação, após o processamento as métricas são fechadas para consumo dos experimentos e não mais alimentadas com novos registros. Cada relação da matriz MMRI é exportada de maneira tabular em arquivos CSV com as colunas: Termo$_x$; Termo$_y$; EF; TF$_x$; TF$_y$; ED$_x$; ED$_y$; CF$_x$; CF$_y$; TD$_x$; TD$_y$; EP$_x$ e EP$_y$.

6.3 Experimentos

Com a base pré-processada e as etapas que estruturam as matrizes MMRT e MMRI já realizadas, é possível realizar as consultas e validar o desempenho do método WAR. Para isso, é necessário definir os termos que serão utilizados nos modelos de validação.

A fim de definir os termos de busca, foi distribuído em uma curva ABC a contabilização dos termos, os quais foram ordenados conforme o valor acumulado das ocorrências e assim criando as seguintes classes: Classe A representando 20% do total máximo acumulado; Classe B com demais valores máximos acumulados até o limite de 50% acumulado; e Classe C para os demais 50%.

Com o valor acumulado dividido em 3 classes, foram selecionados 3 termos com ocorrência máxima dentro da faixa da classe. Esta definição foi aplicada para cada um dos 3 segmentos de texto e o resultado dos termos obtidos é mostrado na Tabela 6.29.

Para a análise dos resultados, é necessário considerar o total de ocorrências de termos em cada segmento assim como a natureza de cada um. O termo utilizado no **cid** é um termo categórico único por registro que representa como o teleconsultor categoriza a doença. Em contrapartida, a estrutura do campo **conduta** é texto livre e permitiu maior preservação da maioria dos termos. Já o texto de **consulta** é preenchido com base em um template da ferramenta, onde os termos adicionados no template foram desconsiderados no pré-processamento.

Tabela 6.29: Levantamento de três termos por segmento e classe na curva ABC

Segmento	Classe	Ocorrências	Termo	Significado
cid	A	864	L82	Ceratose seborréica
		626	C44.9	Neoplasia maligna da pele
		488	L57.0	Ceratose actínica
	B	440	D22	Nevos melanocíticos
		427	C44	Outras neoplasias malignas da pele
		417	L20	Dermatite atópica
	C	150	L01	Impetigo
		145	C43	Melanoma maligno da pele
		143	L81.4	Outras formas de hiperpigmentação
conduta	A	1952	sintoma	Sintoma
		1744	sabonete	Sabonete
		1741	roupa	Roupa
	B	891	girassol	Girassol
		869	afastar	Afastar
		854	lavagem	Lavagem
	C	423	manutencao	Manutenção
		420	biopsia	Biópsia
		417	edema	Edema
consulta	A	7103	mal	Mal
		6188	prurido	Prurido
		2690	sinal	Sinal
	B	1776	pruriginosa	Pruriginosa
		1333	descamativa	Descamativa
		1330	mancha	Mancha
	C	339	cervical	Cervical
		337	dedo	Dedo
		335	aumentado	Aumentado

Fonte: Os Autores

O valor total de **conduta** e **consulta** são maiores em relação ao total de registros da base devido ao fato de o campo ser texto livre e possuir mais de um termo por registro, resultando respectivamente em 162.041 e 96.498 como o total de ocorrências.

Na distribuição nas classes ABC foram obtidos os seguintes intervalos de registros acumulados para **conduta**: **Classe A**, de 0 à 32.408; **Classe B**, de 32.409 à 81.021; e **Classe C**, de 81.022 à 162.041. Para **consulta**, os intervalos das classes foram: **Classe A**, de 0 à 19.300; **Classe B**, de 19.301 à 48.249; e **Classe C**, de 48.350 à 96.498.

As ocorrências do **cid** apresentam 12.219 ocorrências, sendo o mesmo valor do total de registros, pois se trata de dados categóricos único por registro. Os valores do intervalo das classes ABC são: **Classe A**, de 0 à 2.444; **Classe B**, 2.445 à 6.110; e **Classe C**, de 6.111 à 12.219.

Com os entendimentos necessários sobre os dados e a forma de validação, foram realizados os seguintes modelos de validação, sendo eles: Cenário 1 que faz a busca de um termo validando o resultado em todos os segmentos; e Cenário 2 que faz a busca em múltiplos segmentos e valida a precisão do ranque.

6.3.1 Cenário 1: busca de um termo validando o resultado em todos os segmentos

Em cada busca deste cenário, além da listagem com os termos ranqueados, também são apresentadas as seguintes informações: **EF:** frequência da associação entre o termo do ranque e da busca; **TF:** frequência total do termo do ranque; **Ordem EF:** ordem do TF, onde 1 é a maior frequência; e **EF / TF:** relação da frequência da associação em relação a frequência total.

6.3.1.1 Buscas pelo CID

Por ser um dado categórico único por registro, os termos de **cid** não possuem relações deste segmento com ele mesmo, desta forma havendo somente termos associados nos segmentos de **conduta** e **consulta**.

Para análise do resultado, foi submetido os termos de CID da Tabela 6.29 na busca e obtendo os resultados dos termos da **Classe A** nas Tabelas 6.30, 6.31 e 6.32, os termos da **Classe B** nas Tabelas 6.33, 6.34, 6.35 e por fim os termos da **Classe C** nas tabelas 6.36, 6.37 e 6.38.

6.3.1.2 Buscas pelo conduta

Nesta nova rodada de buscas, foi submetido os termos de conduta da tabela 6.29 na busca, sendo obtido os resultados dos termos da **Classe A** nas tabelas 6.39, 6.40 e 6.41, os termos da **Classe B** nas tabelas 6.42, 6.43, 6.44 e por fim os termos da **Classe C** nas tabelas 6.45, 6.46 e 6.47.

Diferentemente dos resultados da busca por **cid**, a busca por termos da **conduta** permitiu obter uma lista de termos associados para todos os três segmentos de texto.

Tabela 6.30: Busca na classe A pelo cid L82

Termo	Ranque	Ordem EF	EF	EF / TF	TF
Conduta					
benigna	28.18200	1	414	43%	959
modificacao	25.96700	3	398	41%	960
cor	24.90465	2	401	40%	1006
cancer	24.67980	8	328	48%	681
especialidade	22.04325	5	331	48%	684
priorizar	15.60180	7	329	44%	743
exerese	13.09850	12	281	39%	728
excesso	12.06460	10	315	32%	993
benignidade	11.87420	11	301	38%	793
acompanhamento	10.95355	4	346	35%	999
suspeita	10.79680	6	331	34%	966
cronica	8.65925	9	326	27%	1189
teledermatologia	5.82060	19	105	43%	244
diametro	5.41525	15	135	43%	316
dermatoscopia	2.32190	13	183	15%	1261
Consulta					
mal	3.971300	1	638	9%	7103
nevo	1.768350	8	107	26%	418
hipercromica	1.124700	3	163	18%	892
marrom	1.056975	26	54	42%	130
crescimento	1.004275	5	130	19%	702
irregular	0.866200	7	109	22%	500
acastanhada	0.759375	19	69	32%	215
prurido	0.553700	2	307	5%	6188
coloracao	0.542450	17	76	23%	325
cor	0.511400	16	79	22%	363
motoboy	0.485700	14	82	14%	587
laboral	0.447425	13	84	14%	612
pedreiro	0.413700	15	82	14%	601
agricultor	0.397950	12	84	13%	628
borda	0.375950	10	89	16%	540

Fonte: Os Autores

Tabela 6.31: Busca na classe A pelo cid C44.9

Termo	Ranque	Ordem EF	EF	EF / TF	TF
Conduta					
dermoscopy	155.231400	2	505	56%	899
dermatoscopia	76.101800	1	511	41%	1261
neoplasica	43.100400	4	335	43%	780
brevidade	36.947100	3	454	41%	1101
aferir	17.052900	6	330	31%	1054
limitacao	14.859900	5	331	28%	1169
afastar	13.612700	7	284	33%	869
biopsia	0.984450	8	95	23%	420
oncologia	0.429975	11	22	27%	83
tumor	0.155975	10	23	20%	115
hospital	0.070400	19	11	28%	39
autorizado	0.019200	27	8	18%	45
neoplasia	0.016800	17	14	11%	132
prioridade	0.008375	12	21	5%	405
ligar	0.007975	13	20	5%	376
Consulta					
motoboy	6.320125	5	173	29%	587
laboral	6.054375	2	178	29%	612
pedreiro	5.479025	6	173	29%	601
agricultor	5.298825	3	177	28%	628
mal	1.638200	1	419	6%	7103
cbc	0.829400	54	20	42%	48
ulcerada	0.696450	13	58	28%	205
nariz	0.573050	12	60	18%	325
cicatriza	0.325550	41	24	32%	74
nasal	0.285750	20	47	17%	282
crescimento	0.276975	8	78	11%	702
nevo	0.249625	18	50	12%	418
sangrante	0.197125	39	26	23%	111
sangramento	0.179450	7	86	8%	1022
nasolabial	0.172800	150	6	60%	10

Fonte: Os Autores

Tabela 6.32: Busca na classe A pelo cid L57.0

Termo	Ranque	Ordem EF	EF	EF / TF	TF
Conduta					
efurix	236.022075	3	323	88%	366
5fluoruracila	166.056075	13	231	88%	263
2012protocolo	111.489000	18	154	90%	172
queimadura	105.725800	8	288	78%	369
rigorosa	52.723100	10	287	60%	481
rediscutido	51.317050	17	191	79%	241
comprida	48.942450	14	231	71%	325
bone	40.005500	6	301	49%	616
5fluorouracil	31.525325	19	84	85%	99
edema	31.357700	11	242	58%	417
chapeu	29.728375	2	326	40%	808
ardencia	26.415325	7	291	40%	727
camada	24.443050	4	302	32%	936
fina	23.445700	5	302	32%	954
fotoprotecao	20.465900	1	334	35%	955
Consulta					
actinica	2.263025	20	44	39%	112
mal	1.708650	1	383	5%	7103
nariz	1.079700	10	68	21%	325
agricultor	0.947925	4	91	14%	628
motoboy	0.914025	6	83	14%	587
laboral	0.834400	5	84	14%	612
pedreiro	0.826500	7	83	14%	601
crioterapia	0.748250	38	20	43%	46
nasal	0.656800	13	59	21%	282
prurido	0.346100	2	207	3%	6188
descamativa	0.277200	3	99	7%	1333
aspera	0.196800	24	32	19%	167
cbc	0.180000	57	11	23%	48
efurix	0.170200	74	9	41%	22
mancha	0.092400	11	66	5%	1330

Fonte: Os Autores

Tabela 6.33: Busca na classe B pelo cid D22

Termo	Ranque	Ordem EF	EF	EF / TF	TF
Conduta					
cor	2.716350	1	153	15%	1006
modificacao	2.194150	2	140	15%	960
diametro	2.117125	8	79	25%	316
benigna	1.320850	3	121	13%	959
dermatoscopia	1.268000	4	121	10%	1261
melanoma	1.175450	14	42	31%	136
exerese	1.148750	5	101	14%	728
benignidade	0.743475	6	95	12%	793
teledermatologia	0.735400	15	42	17%	244
artificial	0.508875	43	13	48%	27
simetria	0.488925	75	7	70%	10
bronzeamento	0.390875	46	13	45%	29
irregular	0.333125	61	9	50%	18
limitacao	0.314200	7	85	7%	1169
anotando	0.302575	101	4	67%	6
Consulta					
nevo	2.414300	4	94	22%	418
mal	1.003975	1	319	4%	7103
cor	0.356825	9	56	15%	363
crescimento	0.280200	6	68	10%	702
melanocitico	0.273000	32	20	25%	79
nevos	0.225000	271	2	100%	2
sinal	0.198200	3	117	4%	2690
raspar	0.190400	219	3	75%	4
prurido	0.183175	2	174	3%	6188
dorso	0.161350	5	72	8%	932
verruga	0.131250	15	33	11%	302
aumento	0.105000	8	61	7%	854
irregular	0.088175	11	40	8%	500
nascimento	0.081750	24	24	15%	164
atrito	0.073775	65	10	26%	38

Fonte: Os Autores

Tabela 6.34: Busca na classe B pelo cid C44

Termo	Ranque	Ordem EF	EF	EF / TF	TF
Conduta					
reavaliacao	2.052375	4	108	21%	514
discussao	1.672875	3	108	19%	559
autorizacao	1.100525	15	32	43%	74
brevidade	0.965250	1	113	10%	1101
limitacao	0.891600	2	110	9%	1169
aferir	0.884650	5	103	10%	1054
dermatoscopia	0.782150	6	95	8%	1261
neoplasica	0.719550	7	73	9%	780
oncologia	0.697150	17	22	27%	83
ligar	0.669600	9	68	18%	376
prioridade	0.583200	10	65	16%	405
afastar	0.350850	8	71	8%	869
tumor	0.279075	16	23	20%	115
secretaria	0.265950	12	38	18%	216
dermoscopy	0.213075	11	49	5%	899
Consulta					
motoboy	1.327075	6	90	15%	587
laboral	1.244100	4	92	15%	612
pedreiro	1.179050	5	91	15%	601
agricultor	1.079625	3	92	15%	628
mal	0.589775	1	269	4%	7103
nariz	0.350825	11	45	14%	325
basocelular	0.228850	39	17	27%	63
cbc	0.206400	60	11	23%	48
prurido	0.129925	2	143	2%	6188
cicatriza	0.114075	45	15	20%	74
nasal	0.112350	18	32	11%	282
carcinoma	0.105300	44	15	21%	73
crescimento	0.101400	9	48	7%	702
ulcerada	0.098400	24	28	14%	205
irregular	0.081225	15	36	7%	500

Fonte: Os Autores

Tabela 6.35: Busca na classe B pelo cid L20

Termo	Ranque	Ordem EF	EF	EF / TF	TF
Conduta					
amaciante	50.510425	5	281	54%	521
exacerbacao	36.381875	17	214	53%	406
fragrancia	25.407200	10	248	41%	608
recidivante	21.935400	11	244	39%	627
cronicidade	19.732925	16	222	38%	590
abrasivo	18.225500	19	213	38%	557
pequena	18.194850	6	281	34%	834
cremosa	16.858650	20	210	41%	518
sabonete	15.237000	1	368	21%	1744
reforcar	13.788075	9	249	32%	787
perfume	13.511850	4	330	22%	1488
roupa	12.115925	2	346	20%	1741
diario	12.037425	12	243	29%	826
girassol	11.878525	8	249	28%	891
mometasona	11.197225	14	237	30%	785
Consulta					
prurido	1.731625	1	345	6%	6188
dermatite	0.862350	7	78	21%	364
mal	0.449950	2	228	3%	7103
descamativa	0.417125	4	106	8%	1333
asma	0.330450	10	53	18%	298
pruriginosa	0.262725	5	105	6%	1776
poplitea	0.259200	37	18	35%	52
postero	0.240000	283	2	100%	2
cubital	0.240000	41	16	36%	44
rinite	0.222900	21	31	22%	140
sinal	0.205550	3	117	4%	2690
fossa	0.171600	30	22	26%	84
placa	0.151825	8	63	8%	818
piora	0.125675	9	59	8%	754
cotovelo	0.118650	17	33	13%	259

Fonte: Os Autores

Tabela 6.36: Busca na classe C pelo cid L01

Termo	Ranque	Ordem EF	EF	EF / TF	TF
Conduta					
contaminar	88.540300	11	33	97%	34
cefalexina	5.105150	1	111	21%	523
higiene	3.046525	2	94	17%	540
remocao	2.251000	3	85	16%	540
creche	1.500000	46	6	50%	12
mupirocina	1.401700	8	43	24%	176
toalha	1.065700	6	48	11%	454
delicada	0.829175	12	29	21%	139
proximo	0.742200	9	36	16%	219
suspensao	0.472400	7	45	10%	462
escola	0.399750	35	8	33%	24
lavagem	0.323100	5	50	6%	854
gentil	0.283475	26	10	22%	46
antibioticoterapia	0.216750	21	13	19%	67
antibiotico	0.207100	10	33	8%	392
Consulta					
impetigo	0.458125	23	11	37%	30
drenagem	0.179275	3	39	7%	594
infectadas	0.167500	272	1	100%	1
pilosos	0.167500	314	1	100%	1
sujo	0.113050	167	2	50%	4
pus	0.109400	7	27	8%	348
mosquito	0.100125	71	4	27%	15
bolhosa	0.100000	19	12	15%	79
prurido	0.090825	1	107	2%	6188
purulento	0.058275	57	5	20%	25
crosta	0.057875	8	26	6%	447
bolha	0.050000	13	15	9%	160
secrecao	0.046700	10	22	4%	549
amarelada	0.045000	27	9	12%	78
romperam	0.043225	166	2	33%	6

Fonte: Os Autores

Tabela 6.37: Busca na classe C pelo cid C43

Termo	Ranque	Ordem EF	EF	EF / TF	TF
Conduta					
urgencia	1.580375	3	56	12%	468
dermatoscopia	1.155550	1	69	5%	1261
brevidade	0.653550	2	61	6%	1101
dermoscopy	0.626900	6	46	5%	899
neoplasica	0.517000	7	43	6%	780
melanoma	0.455075	11	20	15%	136
limitacao	0.435175	4	52	4%	1169
aferir	0.421250	5	50	5%	1054
urgente	0.310500	33	3	60%	5
afastar	0.306600	8	41	5%	869
secretaria	0.252075	10	23	11%	216
prioridade	0.225300	9	29	7%	405
ligar	0.070625	12	19	5%	376
oncologia	0.060375	17	7	8%	83
tumor	0.045000	16	8	7%	115
Consulta					
irregular	0.246900	2	37	7%	500
nevo	0.171950	5	27	6%	418
coloracao	0.154475	6	26	8%	325
enegrecida	0.120750	23	10	17%	60
miiase	0.117300	126	2	50%	4
mal	0.098800	1	108	2%	7103
crescimento	0.077975	3	32	5%	702
borda	0.054675	7	25	5%	540
agricultor	0.050925	8	24	4%	628
cor	0.050675	12	21	6%	363
pedreiro	0.049225	9	23	4%	601
irregulares	0.048300	62	4	19%	21
bordo	0.047725	13	19	6%	317
motoboy	0.047300	11	22	4%	587
laboral	0.046900	10	22	4%	612

Fonte: Os Autores

Tabela 6.38: Busca na classe C pelo cid L81.4

Termo	Ranque	Ordem EF	EF	EF / TF	TF
Conduta					
despigmentante	13.231825	7	47	49%	95
fluocinolona	7.927975	10	40	41%	98
acetonida	7.606850	22	28	43%	65
amplo	4.641700	20	33	40%	83
hidroquinona	4.104150	8	47	32%	145
tretinoina	2.714700	11	40	26%	152
bone	2.425425	3	78	13%	616
azelaico	2.012350	18	35	23%	152
verao	1.987100	30	21	46%	46
fotoprotecao	1.568475	1	94	10%	955
rigorosa	1.555675	4	59	12%	481
chapeu	1.536300	2	79	10%	808
intervalo	1.295050	32	19	40%	48
protecao	0.708300	13	38	15%	258
reforcar	0.498400	5	56	7%	787
Consulta					
melasma	2.271775	7	21	42%	50
hipercromica	0.259000	3	50	6%	892
mancha	0.220125	2	64	5%	1330
cloasma	0.199500	65	3	38%	8
mal	0.098800	1	108	2%	7103
gestacao	0.057675	10	11	12%	89
zigomatica	0.043750	45	5	16%	32
recidivar	0.043750	288	1	50%	2
angustiada	0.043750	140	1	50%	2
cresceram	0.038500	101	2	29%	7
hiperpigmentacao	0.031500	33	6	15%	40
macula	0.024475	9	14	5%	304
cafe	0.021000	28	6	9%	67
medo	0.021000	78	3	16%	19
marrao	0.019250	250	1	33%	3

Fonte: Os Autores

Tabela 6.39: Busca na classe A pela conduta sintoma

Termo	Ranque	Ordem EF	EF	EF / TF	TF
		CID			
L57.0	4.204650	1	288	59%	488
L40	1.059500	2	165	48%	343
B86	0.654050	3	84	77%	109
L50	0.371600	4	68	76%	90
L63	0.285625	7	62	60%	103
L29	0.164925	5	66	42%	156
L42	0.106875	10	44	60%	73
L21	0.102975	6	64	28%	228
B35	0.067025	8	53	21%	248
B07	0.047175	9	48	14%	344
L91.0	0.033900	19	22	40%	55
B35.4	0.022000	11	43	28%	154
L71	0.020500	12	40	32%	124
B09	0.016200	27	14	78%	18
L81.1	0.012125	21	19	35%	54
		Conduta			
efurix	3.673650	15	314	86%	366
queimadura	3.265100	13	322	87%	369
roupa	2.748150	1	599	34%	1741
bone	2.745300	8	376	61%	616
rigorosa	2.666375	12	329	68%	481
chapeu	2.399000	4	408	50%	808
ardencia	2.316975	5	395	54%	727
questionar	2.316675	2	450	49%	915
comprida	2.025050	23	255	78%	325
fotoprotecao	1.934350	3	416	44%	955
5fluoruracila	1.931050	29	219	83%	263
2012protocolo	1.884000	50	161	94%	172
edema	1.796850	18	288	69%	417
rediscutido	1.724325	28	220	91%	241
fina	1.718825	6	391	41%	954
		Consulta			
prurido	5.060775	1	1163	19%	6188
mal	3.058275	2	1034	15%	7103
pruriginosa	0.585600	4	363	20%	1776
ardencia	0.447875	5	353	18%	1989
sinal	0.410900	3	384	14%	2690
descamativa	0.379350	6	285	21%	1333
eritematosa	0.259825	7	222	24%	943
placa	0.241575	11	193	24%	818
descamacao	0.204600	10	194	21%	922
mao	0.204150	8	217	19%	1167
membro	0.178275	12	180	25%	714
mancha	0.163425	9	215	16%	1330
cabeludo	0.143100	14	167	23%	726
piora	0.141800	13	168	22%	754
corpo	0.139125	17	149	25%	593

Fonte: Os Autores

Tabela 6.40: Busca na classe A pela conduta sabonete

Termo	Ranque	Ordem EF	EF	EF / TF	TF
		CID			
L20	15.237000	1	368	88%	417
L71	1.553000	3	108	87%	124
L29	1.267475	2	115	74%	156
L70	0.634900	4	100	46%	216
L30.0	0.430150	8	55	90%	61
L30.1	0.405450	6	66	58%	113
L11.0	0.367950	12	38	93%	41
L20.9	0.311200	10	39	89%	44
L28.1	0.291975	11	39	76%	51
L30	0.235150	7	58	57%	101
L28.0	0.222600	9	48	62%	78
L23	0.168500	5	72	25%	283
L28	0.053625	17	25	83%	30
L85.3	0.027250	22	19	83%	23
L70.0	0.026700	13	31	39%	80
		Conduta			
perfume	17.942900	1	986	66%	1488
hidratante	13.122025	2	940	57%	1660
neutro	10.459100	10	616	91%	678
girassol	10.300200	5	678	76%	891
abrasivo	10.280500	11	555	100%	557
cuidado	9.053700	4	737	61%	1213
diario	8.846275	7	634	77%	826
roupa	8.822650	3	831	48%	1741
amaciante	8.801550	13	508	98%	521
fragrancia	8.345925	12	534	88%	608
pequena	8.311100	9	624	75%	834
antihistaminico	7.521900	6	663	59%	1127
hidroxizine	7.065700	8	628	61%	1032
cremosa	5.950175	16	451	87%	518
exacerbacao	5.249650	18	384	95%	406
		Consulta			
prurido	7.721000	1	1279	21%	6188
mal	3.485150	2	1023	14%	7103
pruriginosa	0.934200	5	411	23%	1776
ardencia	0.885850	4	415	21%	1989
sinal	0.719500	3	437	16%	2690
descamativa	0.393225	6	268	20%	1333
piora	0.278700	7	202	27%	754
eritematosa	0.213050	8	193	20%	943
acne	0.208475	24	98	38%	255
corpo	0.186650	12	164	28%	593
dermatite	0.179650	20	128	35%	364
mao	0.157975	9	184	16%	1167
perna	0.157025	10	167	24%	682
membro	0.151425	13	164	23%	714
asma	0.139850	23	99	33%	298

Fonte: Os Autores

Tabela 6.41: Busca na classe A pela conduta roupa

Termo	Ranque	Ordem EF	EF	EF / TF	TF
		CID			
L20	12.115925	1	346	83%	417
L57.0	3.502625	2	237	49%	488
B86	1.463200	3	102	94%	109
L70	0.635200	4	100	46%	216
B35.4	0.323050	5	69	45%	154
B35	0.156650	6	64	26%	248
L30	0.136725	10	41	41%	101
L80	0.124075	7	54	31%	174
L01	0.121625	8	51	34%	150
L29	0.091475	9	45	29%	156
L20.9	0.067275	11	36	82%	44
L73.2	0.042900	15	23	79%	29
L30.0	0.028175	13	28	46%	61
L70.0	0.014900	14	26	33%	80
T69.1	0.014375	57	4	100%	4
		Conduta			
sabonete	8.822650	1	831	48%	1744
amaciante	8.793050	6	507	97%	521
toalha	6.658475	11	432	95%	454
fragrancia	6.116550	7	482	79%	608
locao	5.252825	5	511	68%	748
exacerbacao	5.064225	20	378	93%	406
perfume	4.531450	2	624	42%	1488
abrasivo	4.371000	13	422	76%	557
comprida	3.957475	27	307	94%	325
cremosa	3.947200	18	399	77%	518
pequena	3.618100	10	475	57%	834
girassol	3.600350	8	478	54%	891
hidratante	3.303625	4	591	36%	1660
sintoma	2.748150	3	599	31%	1952
antihistaminico	2.696700	9	475	42%	1127
		Consulta			
prurido	5.934550	1	1167	19%	6188
mal	3.059350	2	979	14%	7103
pruriginosa	0.831200	4	390	22%	1776
sinal	0.646475	3	416	15%	2690
ardencia	0.510875	5	347	17%	1989
descamativa	0.392050	6	267	20%	1333
corpo	0.276375	10	177	30%	593
membro	0.223050	11	175	25%	714
eritematosa	0.175875	9	183	19%	943
dermatite	0.174175	19	124	34%	364
mao	0.158125	8	184	16%	1167
mancha	0.154925	7	192	14%	1330
piora	0.153125	12	169	22%	754
placa	0.148750	13	169	21%	818
drenagem	0.111525	16	131	22%	594

Fonte: Os Autores

Tabela 6.42: Busca na classe B pela conduta girassol

Termo	Ranque	Ordem EF	EF	EF / TF	TF
CID					
L20	11.878525	1	249	60%	417
L29	0.922725	2	76	49%	156
L30	0.809500	3	61	60%	101
L30.0	0.596100	4	45	74%	61
L28.1	0.594200	5	40	78%	51
L20.9	0.263125	8	31	70%	44
I87.2	0.137875	6	36	35%	103
L23.9	0.081975	9	22	28%	79
L23	0.038400	7	33	12%	283
L28.2	0.016800	10	20	36%	55
L85.3	0.009225	16	11	48%	23
L30.5	0.008400	11	15	23%	64
L90.8	0.004675	101	1	100%	1
B86	0.003650	12	13	12%	109
L28.0	0.003650	13	13	17%	78
Conduta					
cremosa	16.538775	5	494	95%	518
fragrancia	12.294700	7	471	77%	608
locao	11.965625	4	522	70%	748
sabonete	10.300200	1	678	39%	1744
perfume	10.219050	2	633	43%	1488
amaciante	9.559725	13	406	78%	521
abrasivo	8.966450	11	414	74%	557
exacerbacao	7.742025	16	341	84%	406
hidratante	7.192075	3	594	36%	1660
reforcar	5.639900	12	411	52%	787
hidroxizine	5.562600	9	446	43%	1032
antihistaminico	5.359325	8	462	41%	1127
diario	4.718325	14	395	48%	826
pequena	4.573000	15	394	47%	834
cuidado	4.290725	10	445	37%	1213
Consulta					
prurido	3.252375	1	741	12%	6188
mal	0.734850	2	475	7%	7103
pruriginosa	0.650000	3	272	15%	1776
descamativa	0.183175	6	163	12%	1333
ardencia	0.174500	5	192	10%	1989
membro	0.169050	7	134	19%	714
sinal	0.154325	4	207	8%	2690
mmii	0.103050	22	66	26%	250
perna	0.103000	8	111	16%	682
corpo	0.102200	10	104	18%	593
braco	0.086225	13	86	15%	566
dermatite	0.065775	20	68	19%	364
piora	0.064275	9	105	14%	754
placa	0.058600	11	100	12%	818
tronco	0.054425	21	68	14%	481

Fonte: Os Autores

Tabela 6.43: Busca na classe B pela conduta afastar

Termo	Ranque	Ordem EF	EF	EF / TF	TF
		CID			
C44.9	13.612700	1	284	45%	626
L82	0.613900	2	118	14%	864
D04	0.605150	7	29	62%	47
C44	0.350850	3	71	17%	427
C43	0.306600	6	41	28%	145
D04.0	0.207000	16	6	100%	6
D22	0.061225	5	43	10%	440
L57.0	0.057750	4	44	9%	488
C49.9	0.033250	27	3	100%	3
C43.0	0.021875	25	3	75%	4
D09.7	0.007500	75	1	100%	1
D09.9	0.007500	76	1	100%	1
C00	0.007500	71	1	100%	1
L44.0	0.006000	101	1	100%	1
H02.8	0.004200	84	1	100%	1
		Conduta			
neoplasica	74.465800	1	697	89%	780
aferir	37.176000	3	679	64%	1054
limitacao	29.982800	2	695	59%	1169
dermatoscopia	16.095950	4	583	46%	1261
dermoscopy	13.606450	5	451	50%	899
brevidade	2.864625	6	332	30%	1101
reavaliacao	0.047975	8	62	12%	514
discussao	0.047025	7	64	11%	559
koeber	0.036400	128	9	100%	9
favorecam	0.026000	125	9	90%	10
completamente	0.025875	15	30	38%	80
melanoma	0.013800	26	24	18%	136
tireoglobulina	0.013050	230	5	71%	7
antihbc	0.011500	234	4	80%	5
urgencia	0.010650	10	37	8%	468
		Consulta			
mal	1.427025	1	567	8%	7103
laboral	0.572175	3	150	25%	612
motoboy	0.560775	5	145	25%	587
pedreiro	0.514775	6	145	24%	601
agricultor	0.471675	4	147	23%	628
prurido	0.310300	2	332	5%	6188
crescimento	0.181125	8	115	16%	702
nevo	0.102325	19	70	17%	418
aumento	0.089050	10	102	12%	854
sangramento	0.085450	9	110	11%	1022
borda	0.079950	16	77	14%	540
irregular	0.076350	18	71	14%	500
progressivo	0.075800	21	66	15%	436
hipercromica	0.053950	14	95	11%	892
surgimento	0.048550	15	81	10%	802

Fonte: Os Autores

Tabela 6.44: Busca na classe B pela conduta lavagem

Termo	Ranque	Ordem EF	EF	EF / TF	TF
		CID			
B36.0	4.506650	1	147	58%	255
L30.1	0.913975	4	64	57%	113
L23	0.863925	2	89	31%	283
L21	0.852800	3	81	36%	228
L01	0.323100	5	50	33%	150
L25	0.115000	7	29	36%	81
L20	0.059650	6	40	10%	417
L24.0	0.025850	20	8	89%	9
L02	0.022800	15	13	52%	25
L24.9	0.022575	17	11	73%	15
L40	0.018225	8	25	7%	343
L24	0.013200	13	15	37%	41
L23.9	0.011700	9	20	25%	79
L70.0	0.009950	11	17	21%	80
L60	0.005550	10	19	17%	110
		Conduta			
dimeticone	4.636500	10	295	71%	414
alimento	4.547700	8	309	67%	460
excessiva	4.325350	4	344	54%	642
borracha	4.220700	9	302	61%	494
luva	4.115300	5	330	54%	610
recorrente	3.875100	7	310	57%	548
algodao	3.860500	3	345	49%	703
rotina	3.485025	17	260	64%	409
encontrado	3.318275	16	261	63%	417
cronico	2.980275	11	293	52%	561
enxague	2.765800	23	167	91%	184
mantendo	2.428425	22	170	89%	192
xampu	2.389800	13	279	43%	644
metal	2.204750	18	254	47%	540
desencadeante	2.087875	15	272	42%	646
		Consulta			
prurido	1.616950	1	564	9%	6188
mao	0.881800	3	260	22%	1167
mal	0.741750	2	462	7%	7103
ardencia	0.289350	4	216	11%	1989
descamativa	0.183700	7	160	12%	1333
sinal	0.183675	5	216	8%	2690
dedo	0.179825	10	94	28%	337
luva	0.124200	21	53	56%	94
pruriginosa	0.115825	6	162	9%	1776
descamacao	0.093525	8	112	12%	922
mancha	0.072225	9	108	8%	1330
piora	0.058700	11	94	12%	754
drenagem	0.054225	13	76	13%	594
hipocromica	0.049150	17	64	12%	526
sensibilidade	0.043250	15	72	9%	812

Fonte: Os Autores

Tabela 6.45: Busca na classe C pela conduta manutencao

Termo	Ranque	Ordem EF	EF	EF / TF	TF
		CID			
L21	7.259400	1	126	55%	228
L40	1.588525	2	87	25%	343
L20	0.031200	3	26	6%	417
L40.3	0.029500	5	10	53%	19
L80	0.020125	4	14	8%	174
L75.0	0.015975	25	3	100%	3
L21.0	0.011875	15	4	67%	6
L21.9	0.003550	22	3	43%	7
L40.8	0.002375	17	4	25%	16
R61	0.001800	84	1	50%	2
L90.0	0.001775	26	3	20%	15
H01	0.001200	51	1	50%	2
I89.0	0.001200	53	1	100%	1
L91.9	0.001200	83	1	50%	2
R23.4	0.001175	38	2	29%	7
		Conduta			
regressivo	2.473650	20	90	96%	94
xampu	1.839700	1	196	30%	644
inverno	1.451200	21	88	82%	107
hormonal	1.109125	22	86	69%	124
emocional	0.955000	11	111	43%	258
cetoconazol	0.950450	2	175	21%	815
couro	0.838150	3	164	20%	819
estresse	0.831700	12	110	38%	292
temperatura	0.531375	17	100	31%	318
lavagem	0.501225	4	142	17%	854
cronico	0.416050	7	117	21%	561
lcd	0.411600	27	71	32%	222
recidivante	0.394050	8	117	19%	627
informar	0.356175	18	99	24%	411
cronicidade	0.307500	16	101	17%	590
		Consulta			
prurido	0.545400	1	296	5%	6188
mal	0.406575	2	270	4%	7103
cabeludo	0.271300	4	106	15%	726
descamativa	0.139475	5	104	8%	1333
psoriase	0.108125	12	50	20%	255
descamacao	0.075950	8	67	7%	922
sinal	0.070200	3	126	5%	2690
pruriginosa	0.036300	7	86	5%	1776
ardencia	0.034725	6	92	5%	1989
dermatite	0.024725	14	36	10%	364
piora	0.016575	11	50	7%	754
placa	0.016500	9	54	7%	818
seborreica	0.013600	22	23	19%	123
eritematosa	0.012450	13	47	5%	943
mao	0.011350	10	53	5%	1167

Fonte: Os Autores

Tabela 6.46: Busca na classe C pela conduta biopsia

Termo	Ranque	Ordem EF	EF	EF / TF	TF
		CID			
C44.9	0.984450	1	95	15%	626
C44	0.080300	2	34	8%	427
B55	0.054000	36	2	67%	3
D48	0.030000	88	1	100%	1
L93.2	0.029750	15	5	71%	7
L98	0.021650	3	20	9%	231
B42	0.021300	21	3	60%	5
C44.2	0.020825	11	5	42%	12
L41	0.020400	44	2	100%	2
L90.9	0.019800	120	1	100%	1
L90.3	0.015000	119	1	100%	1
C46	0.015000	77	1	100%	1
L41.1	0.012425	29	3	43%	7
L12.0	0.012425	25	3	75%	4
B43.0	0.012000	70	1	100%	1
		Conduta			
ligar	0.306225	5	81	22%	376
prioridade	0.294675	6	81	20%	405
brevidade	0.255825	1	110	10%	1101
elucidacao	0.226600	11	41	42%	98
dermatoscopia	0.200350	2	101	8%	1261
dermoscopy	0.162000	7	80	9%	899
aferir	0.155950	4	82	8%	1054
esclarecimento	0.128475	19	24	67%	36
confirmacao	0.113575	8	78	13%	581
limitacao	0.102650	3	83	7%	1169
reavaliacao	0.051925	10	48	9%	514
discussao	0.051100	9	49	9%	559
neoplasica	0.020250	12	34	4%	780
secretaria	0.019050	13	32	15%	216
histologica	0.019000	159	4	50%	8
		Consulta			
mal	0.303400	1	240	3%	7103
prurido	0.052000	2	161	3%	6188
motoboy	0.041850	10	41	7%	587
pedreiro	0.041450	11	41	7%	601
laboral	0.041150	9	41	7%	612
agricultor	0.040725	7	41	7%	628
sinal	0.026175	3	94	3%	2690
ardencia	0.018600	4	74	4%	1989
pulsos	0.015000	662	1	100%	1
assintomaticos	0.012000	432	1	100%	1
ceratoticas	0.012000	450	1	100%	1
sangramento	0.011000	6	45	4%	1022
eritematosa	0.010875	8	41	4%	943
drenado	0.008875	208	3	43%	7
pruriginosa	0.007750	5	55	3%	1776

Fonte: Os Autores

Tabela 6.47: Busca na classe C pela conduta edema

Termo	Ranque	Ordem EF	EF	EF / TF	TF
		CID			
L57.0	31.35770	1	242	50%	488
I83.1	0.49865	2	22	73%	30
X32	0.00600	35	2	67%	3
L03	0.00540	44	1	100%	1
I83.2	0.00480	12	4	40%	10
L27.0	0.00360	7	6	22%	27
I83.9	0.00360	43	1	100%	1
I83	0.00360	17	3	50%	6
M34	0.00300	63	1	100%	1
C44.1	0.00120	25	2	29%	7
T78.4	0.00120	34	2	33%	6
L23.3	0.00120	47	1	50%	2
L56.8	0.00060	57	1	50%	2
T69.1	0.00060	64	1	25%	4
L56	0.00000	54	1	20%	5
		Conduta			
5fluoruracila	14.939150	5	246	94%	263
efurix	9.790000	6	244	67%	366
2012protocolo	8.904150	17	155	90%	172
rediscutido	8.688500	16	211	88%	241
comprida	7.633700	12	223	69%	325
queimadura	6.122825	13	221	60%	369
bone	4.979375	3	255	41%	616
telefonico	4.832550	15	216	59%	367
rigorosa	4.781650	11	224	47%	481
chapeu	4.292275	2	271	34%	808
ardencia	3.511475	4	248	34%	727
camada	2.392300	8	236	25%	936
fotoprotecao	2.244725	7	238	25%	955
fina	2.227325	9	236	25%	954
sintoma	1.796850	1	288	15%	1952
		Consulta			
mal	0.533200	1	291	4%	7103
prurido	0.141525	2	206	3%	6188
nasal	0.063900	15	43	15%	282
agricultor	0.059850	7	60	10%	628
laboral	0.056450	8	56	9%	612
motoboy	0.056400	9	55	9%	587
pedreiro	0.054850	10	54	9%	601
nariz	0.049275	19	36	11%	325
actinica	0.046800	28	26	23%	112
crioterapia	0.033600	45	14	30%	46
descamativa	0.028900	5	77	6%	1333
mancha	0.023700	6	63	5%	1330
ardencia	0.019350	4	77	4%	1989
cbc	0.018000	65	10	21%	48
membro	0.015400	14	44	6%	714

Fonte: Os Autores

6.3.1.3 Buscas pela consulta

Por fim, os termos da consulta da Tabela 6.29 foram submetidos à busca, e assim obtendo os resultados dos termos da **Classe A** apresentados nas Tabelas 6.48, 6.49 e 6.50, os termos da **Classe B** nas Tabelas 6.51, 6.52, 6.53 e por fim os termos da **Classe C** nas Tabelas 6.54, 6.55 e 6.56.

Assim como a busca por termos da **conduta**, os termos do segmento **consulta** permitiu obter uma lista de termos associados para todos os três segmentos de texto.

Tabela 6.48: Busca na classe A pela consulta mal

Termo	Ranque	Ordem EF	EF	EF / TF	TF
		CID			
L82	3.971300	1	638	74%	864
L57.0	1.708650	3	383	78%	488
C44.9	1.638200	2	419	67%	626
D22	1.003975	4	319	73%	440
C44	0.589775	5	269	63%	427
L40	0.562900	7	227	66%	343
L20	0.449950	6	228	55%	417
L70	0.426900	9	162	75%	216
B07	0.245700	8	176	51%	344
L21	0.224900	12	130	57%	228
B36.0	0.217850	11	135	53%	255
L98	0.203300	13	129	56%	231
L23	0.177450	10	136	48%	283
C43	0.098800	14	108	74%	145
L81.4	0.098800	15	108	76%	143
		Conduta			
dermatoscopia	3.523550	5	870	69%	1261
sabonete	3.485150	2	1023	59%	1744
roupa	3.059350	3	979	56%	1741
sintoma	3.058275	1	1034	53%	1952
cronica	2.973900	6	838	70%	1189
hidratante	2.802225	4	931	56%	1660
limitacao	2.496150	9	767	66%	1169
acido	2.466400	7	817	62%	1316
fina	2.340750	14	684	72%	954
camada	2.323625	16	674	72%	936
benigna	2.221250	17	671	70%	959
excesso	2.219625	11	695	70%	993
brevidade	2.161425	10	729	66%	1101
suspeita	2.154500	15	682	71%	966
aferir	2.084400	12	691	66%	1054
		Consulta			
prurido	29.144400	1	3447	56%	6188
sinal	4.677125	2	1386	52%	2690
ardencia	3.964050	3	1163	58%	1989
pruriginosa	2.504800	4	949	53%	1776
mancha	2.385400	5	826	62%	1330
descamativa	2.254450	6	811	61%	1333
sangramento	1.816025	7	690	68%	1022
hipercromica	1.422825	9	587	66%	892
aumento	1.360800	10	569	67%	854
crescimento	1.176500	15	497	71%	702
mao	1.140450	8	632	54%	1167
agricultor	1.132075	18	437	70%	628
dorso	1.124600	11	562	60%	932
laboral	1.109650	19	425	69%	612
pedreiro	1.086100	21	419	70%	601

Fonte: Os Autores

Tabela 6.49: Busca na classe A pela consulta prurido

Termo	Ranque	Ordem EF	EF	EF / TF	TF
CID					
L20	1.731625	1	345	83%	417
L40	0.996750	3	265	77%	343
L23	0.792650	4	240	85%	283
L82	0.553700	2	307	36%	864
L29	0.393950	12	136	87%	156
L21	0.347825	8	159	70%	228
L57.0	0.346100	5	207	42%	488
B35.4	0.340875	13	122	79%	154
B35	0.285275	9	157	63%	248
B36.0	0.243150	11	139	55%	255
D22	0.183175	7	174	40%	440
C44	0.129925	10	143	33%	427
C44.9	0.126500	6	175	28%	626
B86	0.109900	17	97	89%	109
L30.1	0.108500	18	96	85%	113
Conduta					
hidratante	7.950925	2	1263	76%	1660
sabonete	7.721000	1	1279	73%	1744
perfume	7.521200	3	1188	80%	1488
antihistaminico	6.238275	6	997	88%	1127
roupa	5.934550	4	1167	67%	1741
hidroxizine	5.467400	7	921	89%	1032
sintoma	5.060775	5	1163	60%	1952
girassol	3.252375	10	741	83%	891
cuidado	2.990325	8	806	66%	1213
locao	2.610000	12	625	84%	748
dexclorfeniramina	2.444950	18	568	91%	626
corticoide	2.275400	11	672	71%	943
acido	2.228225	9	746	57%	1316
diario	2.190750	13	620	75%	826
mometasona	2.065900	15	598	76%	785
Consulta					
mal	29.144400	1	3447	49%	7103
ardencia	9.173450	3	1470	74%	1989
pruriginosa	7.931850	4	1329	75%	1776
sinal	6.873650	2	1507	56%	2690
descamativa	3.742575	5	910	68%	1333
eritematosa	1.909050	7	654	69%	943
placa	1.854850	10	602	74%	818
mao	1.810000	6	684	59%	1167
descamacao	1.806400	8	631	68%	922
sangramento	1.547950	9	616	60%	1022
piora	1.412700	12	533	71%	754
membro	1.343425	14	504	71%	714
mancha	1.035850	11	588	44%	1330
dorso	1.029700	13	521	56%	932
corpo	0.945675	18	419	71%	593

Fonte: Os Autores

Tabela 6.50: Busca na classe A pela consulta sinal

Termo	Ranque	Ordem EF	EF	EF / TF	TF
		CID			
L20	0.205550	3	117	28%	417
D22	0.198200	2	117	27%	440
L23	0.120000	4	89	31%	283
L98	0.117625	6	76	33%	231
L82	0.110650	1	129	15%	864
L40	0.096850	5	88	26%	343
B35	0.095250	7	69	28%	248
L57.0	0.048200	8	69	14%	488
L70	0.027000	12	58	27%	216
B35.1	0.026700	15	48	35%	138
B36.0	0.024900	9	67	26%	255
B07	0.017850	11	64	19%	344
L21	0.015300	13	55	24%	228
I87.2	0.013000	19	35	34%	103
L80	0.010875	17	39	22%	174
		Conduta			
perfume	0.722000	4	408	27%	1488
sabonete	0.719500	1	437	25%	1744
hidratante	0.715000	2	425	26%	1660
roupa	0.646475	3	416	24%	1741
cuidado	0.428975	7	315	26%	1213
sintoma	0.410900	5	384	20%	1952
acido	0.388050	6	318	24%	1316
antihistaminico	0.355050	8	299	27%	1127
corticoide	0.336200	10	264	28%	943
hidroxizine	0.300775	9	265	26%	1032
unha	0.257650	12	253	23%	1086
cronica	0.248400	11	254	21%	1189
algodao	0.236250	16	215	31%	703
desencadeante	0.227175	23	201	31%	646
micologico	0.217975	13	229	28%	829
		Consulta			
prurido	6.873650	1	1507	24%	6188
mal	4.677125	2	1386	20%	7103
ardencia	1.149500	3	542	27%	1989
pruriginosa	0.517400	4	396	22%	1776
mao	0.314050	6	291	25%	1167
flogistico	0.300750	31	108	97%	111
descamativa	0.274250	5	294	22%	1333
mancha	0.257650	7	276	21%	1330
sensibilidade	0.190425	10	220	27%	812
sangramento	0.190125	8	237	23%	1022
eritematosa	0.189225	9	230	24%	943
placa	0.183000	12	212	26%	818
piora	0.180550	14	203	27%	754
descamacao	0.159725	11	217	24%	922
dorso	0.152425	13	208	22%	932

Fonte: Os Autores

Tabela 6.51: Busca na classe B pela consulta pruriginosa

Termo	Ranque	Ordem EF	EF	EF / TF	TF
CID					
L20	0.262725	1	105	25%	417
L40	0.169750	2	84	24%	343
B86	0.154175	7	47	43%	109
B35	0.084375	6	59	24%	248
L23	0.080950	4	62	22%	283
L57.0	0.040475	5	62	13%	488
L42	0.038200	16	34	47%	73
L43	0.033000	14	39	42%	93
L82	0.031825	3	74	9%	864
L28.1	0.025550	22	26	51%	51
C44.9	0.025400	8	47	8%	626
B35.4	0.024800	10	44	29%	154
L20.9	0.020650	27	21	48%	44
L30	0.018600	17	33	33%	101
L29	0.017700	12	42	27%	156
Conduta					
antihistaminico	1.490550	4	398	35%	1127
hidroxizine	1.270400	6	371	36%	1032
hidratante	1.182550	1	428	26%	1660
perfume	1.117100	3	406	27%	1488
sabonete	0.934200	2	411	24%	1744
roupa	0.831200	5	390	22%	1741
girassol	0.650000	8	272	31%	891
sintoma	0.585600	7	363	19%	1952
locao	0.551475	10	243	32%	748
dexclorfeniramina	0.435850	13	211	34%	626
cuidado	0.389175	9	266	22%	1213
mometasona	0.340600	14	210	27%	785
diario	0.317500	12	219	27%	826
corticoide	0.306650	11	223	24%	943
reforcar	0.295950	17	200	25%	787
Consulta					
prurido	7.931850	1	1329	21%	6188
mal	2.504800	2	949	13%	7103
descamativa	1.620575	3	443	33%	1333
eritematosa	0.925450	5	322	34%	943
sinal	0.517400	4	396	15%	2690
placa	0.320300	7	220	27%	818
membro	0.310550	9	201	28%	714
ardencia	0.264550	6	286	14%	1989
piora	0.244700	12	179	24%	754
surgimento	0.242850	11	183	23%	802
dolorosa	0.213925	21	115	40%	286
mao	0.213200	8	215	18%	1167
dorso	0.207800	10	189	20%	932
tronco	0.189700	16	139	29%	481
corpo	0.171075	13	152	26%	593

Fonte: Os Autores

Tabela 6.52: Busca na classe B pela consulta descamativa

Termo	Ranque	Ordem EF	EF	EF / TF	TF
		CID			
L40	0.841875	1	124	36%	343
L20	0.417125	2	106	25%	417
L57.0	0.277200	3	99	20%	488
L21	0.222325	4	71	31%	228
B35	0.159875	6	62	25%	248
L23	0.149050	5	64	23%	283
B35.4	0.069975	10	35	23%	154
C44.9	0.031700	8	54	9%	626
C44	0.028625	9	37	9%	427
L82	0.026700	7	56	6%	864
B35.3	0.020300	12	27	30%	90
L30.1	0.009000	13	24	21%	113
L30	0.008650	14	23	23%	101
L42	0.006750	16	18	25%	73
D04.8	0.006600	113	1	100%	1
		Conduta			
hidratante	0.973925	1	365	22%	1660
perfume	0.579950	2	298	20%	1488
couro	0.428550	7	210	26%	819
sabonete	0.393225	4	268	15%	1744
roupa	0.392050	5	267	15%	1741
sintoma	0.379350	3	285	15%	1952
cronicidade	0.305800	19	155	26%	590
hidroxizine	0.277650	8	188	18%	1032
cetoconazol	0.273000	10	176	22%	815
clobetasol	0.270300	13	165	23%	714
xampu	0.267350	18	156	24%	644
acido	0.242325	6	214	16%	1316
mometasona	0.227450	14	164	21%	785
micologico	0.220600	16	163	20%	829
miconazol	0.214650	21	145	21%	685
		Consulta			
prurido	3.742575	1	910	15%	6188
mal	2.254450	2	811	11%	7103
pruriginosa	1.620575	3	443	25%	1776
placa	0.775250	6	256	31%	818
eritematosa	0.522350	7	239	25%	943
ardencia	0.448300	4	298	15%	1989
sinal	0.274250	5	294	11%	2690
mao	0.243450	8	202	17%	1167
cabeludo	0.211550	9	147	20%	726
hiperemiada	0.149150	13	113	26%	439
membro	0.087775	11	121	17%	714
sangramento	0.086150	10	139	14%	1022
eritemato	0.082500	54	40	67%	60
piora	0.076300	14	108	14%	754
cotovelo	0.076000	21	81	31%	259

Fonte: Os Autores

Tabela 6.53: Busca na classe B pela consulta mancha

Termo	Ranque	Ordem EF	EF	EF / TF	TF
		CID			
B36.0	0.472475	2	98	38%	255
L82	0.303600	1	132	15%	864
L81.4	0.220125	5	64	45%	143
L80	0.185325	6	63	36%	174
L81.1	0.138075	7	35	65%	54
L57.0	0.092400	3	66	14%	488
D22	0.061375	4	65	15%	440
L30.5	0.054500	10	29	45%	64
L81.3	0.041650	25	13	81%	16
L81	0.033900	19	15	65%	23
L81.0	0.021450	15	19	44%	43
B36.1	0.010925	65	3	100%	3
L56.2	0.010200	32	9	45%	20
I78.1	0.010125	17	18	34%	53
L59.9	0.010000	181	1	100%	1
		Conduta			
fotoprotecao	0.576625	1	244	26%	955
chapeu	0.413250	3	201	25%	808
protecao	0.276725	30	104	40%	258
hidroquinona	0.262225	44	82	57%	145
fluocinolona	0.225500	74	60	61%	98
benignidade	0.212900	7	154	19%	793
bone	0.197775	8	147	24%	616
cor	0.183575	6	173	17%	1006
rigorosa	0.172200	25	118	25%	481
cronica	0.171250	5	174	15%	1189
sintoma	0.163425	2	215	11%	1952
despigmentante	0.162400	81	54	57%	95
roupa	0.154925	4	192	11%	1741
repigmentacao	0.150500	46	80	42%	189
enxague	0.146500	49	78	42%	184
		Consulta			
mal	2.385400	1	826	12%	7103
hipercromica	1.593850	3	327	37%	892
prurido	1.035850	2	588	10%	6188
hipocromica	1.024350	5	226	43%	526
sinal	0.257650	4	276	10%	2690
escura	0.107975	18	82	39%	209
dorso	0.080525	8	125	13%	932
braco	0.080125	13	98	17%	566
descamacao	0.078300	9	121	13%	922
aumento	0.050950	10	106	12%	854
perna	0.049350	15	89	13%	682
sensibilidade	0.048100	14	97	12%	812
ardencia	0.044550	6	142	7%	1989
acastanhada	0.044400	28	59	27%	215
aumentando	0.043950	21	78	23%	335

Fonte: Os Autores

Tabela 6.54: Busca na classe C pela consulta cervical

Termo	Ranque	Ordem EF	EF	EF / TF	TF
CID					
D22	0.010800	2	19	4%	440
B45	0.010150	64	1	100%	1
E85.4	0.010150	73	1	100%	1
L20	0.009600	4	16	4%	417
Q18	0.007975	107	1	100%	1
D21.0	0.007350	20	5	33%	15
C44.9	0.006800	3	17	3%	626
L82	0.006650	1	23	3%	864
L74.2	0.005075	97	1	100%	1
D21.9	0.004425	13	6	15%	39
L23.2	0.004425	49	2	40%	5
R61	0.003625	109	1	50%	2
L10	0.002950	46	2	40%	5
B00.0	0.002900	61	1	50%	2
A49.8	0.002900	60	1	50%	2
Conduta					
tesoura	0.017700	107	12	26%	47
anestesico	0.014750	118	10	24%	41
hemostasia	0.014750	124	10	28%	36
mosaicismo	0.013275	477	2	67%	3
lidocaina	0.013250	147	9	24%	38
assepsia	0.013250	136	9	26%	34
esteril	0.013250	142	9	27%	33
acompanhamento	0.010250	8	41	4%	999
hidroxizine	0.010175	7	42	4%	1032
perfume	0.009925	2	59	4%	1488
girassol	0.009825	11	35	4%	891
antihistaminico	0.009550	6	43	4%	1127
benigna	0.009125	9	35	4%	959
roupa	0.009050	1	63	4%	1741
sabonete	0.008450	3	59	3%	1744
Consulta					
prurido	0.183225	1	216	3%	6188
mal	0.106000	2	188	3%	7103
acrocordons	0.023925	383	1	100%	1
sinal	0.021450	3	77	3%	2690
pruriginosa	0.018600	4	66	4%	1776
ardencia	0.015100	5	60	3%	1989
hipercromica	0.009525	7	34	4%	892
linfonodo	0.007350	125	5	38%	13
descamativa	0.006575	6	35	3%	1333
linfonodomegalia	0.005150	82	7	24%	29
escoriacoes	0.005075	471	1	100%	1
sudorese	0.004425	113	6	21%	29
rash	0.003675	134	5	28%	18
pediculada	0.003675	129	5	21%	24
obesa	0.002950	159	4	19%	21

Fonte: Os Autores

Tabela 6.55: Busca na classe C pela consulta dedo

Termo	Ranque	Ordem EF	EF	EF / TF	TF
CID					
B07	0.204125	1	40	12%	344
L60	0.155700	5	23	21%	110
L30.1	0.135850	4	23	20%	113
B35.1	0.100400	3	23	17%	138
L25	0.049775	8	13	16%	81
L23	0.045950	2	26	9%	283
G59.0	0.024750	53	1	100%	1
L60.8	0.022275	9	10	26%	38
B35	0.016125	6	16	6%	248
M20.0	0.015000	81	1	100%	1
L24.9	0.011100	17	5	33%	15
I73.8	0.010500	54	1	100%	1
L20	0.007800	7	13	3%	417
F06.4	0.006000	52	1	100%	1
L24	0.005925	10	8	20%	41
Conduta					
luva	0.580950	2	124	20%	610
excessiva	0.529200	3	116	18%	642
dimeticone	0.516800	10	96	23%	414
alimento	0.431100	9	98	21%	460
borracha	0.429275	8	101	20%	494
algodao	0.349000	5	107	15%	703
acido	0.303000	1	134	10%	1316
metal	0.232700	12	81	15%	540
rotina	0.192150	16	71	17%	409
encontrado	0.190500	15	71	17%	417
lavagem	0.179825	11	94	11%	854
unha	0.171200	7	103	9%	1086
hidratante	0.167450	4	112	7%	1660
desencadeante	0.146525	13	77	12%	646
perfume	0.145600	6	103	7%	1488
Consulta					
mao	1.567575	1	210	18%	1167
unha	0.195850	5	69	19%	369
falange	0.083125	25	16	52%	31
mal	0.075400	2	165	2%	7103
prurido	0.067925	3	153	2%	6188
sinal	0.046500	4	100	4%	2690
anelar	0.041600	63	7	64%	11
joanete	0.020025	188	3	100%	3
tunel	0.020025	211	3	100%	3
detergente	0.020025	50	9	35%	26
fissura	0.017800	15	24	20%	120
ardencia	0.017100	6	68	3%	1989
fungo	0.014100	19	19	20%	95
luva	0.014100	20	19	20%	94
sensibilidade	0.013550	8	44	5%	812

Fonte: Os Autores

Tabela 6.56: Busca na classe C pela consulta aumentado

Termo	Ranque	Ordem EF	EF	EF / TF	TF
CID					
L82	0.034475	1	17	2%	864
L02.4	0.026550	30	1	100%	1
D21.2	0.005900	21	1	33%	3
L70.1	0.005875	13	2	15%	13
R23.8	0.002950	49	1	17%	6
A30	0.000000	14	1	3%	39
L66	0.000000	40	1	3%	37
L23	0.000000	34	1	0%	283
L24.7	0.000000	35	1	13%	8
L28	0.000000	36	1	3%	30
L28.2	0.000000	37	1	2%	55
L40	0.000000	11	2	1%	343
L50	0.000000	38	1	1%	90
L57.0	0.000000	12	2	0%	488
L57.3	0.000000	39	1	8%	13
Conduta					
continuam	0.005900	207	1	33%	3
benignidade	0.004725	1	15	2%	793
despigmentoso	0.002950	222	1	20%	5
ota	0.002950	308	1	20%	5
pulsada	0.002950	324	1	20%	5
5fluorouracil	0.000000	175	1	1%	99
oxido	0.000000	149	2	2%	91
otimizar	0.000000	309	1	3%	35
oncologico	0.000000	307	1	20%	5
oncologia	0.000000	306	1	1%	83
oclusao	0.000000	305	1	0%	275
observar	0.000000	148	2	0%	538
nivea	0.000000	304	1	1%	128
neutro	0.000000	36	5	1%	678
nitidez	0.000000	303	1	14%	7
Consulta					
mal	0.031125	1	59	1%	7103
nacimento	0.011800	223	1	50%	2
prurido	0.008700	2	36	1%	6188
mancha	0.008250	3	22	2%	1330
parkinsonismo	0.005900	234	1	33%	3
infiltracao	0.005875	98	2	18%	11
hemorragia	0.005875	96	2	18%	11
sinal	0.003700	4	20	1%	2690
hiperglicemia	0.002950	197	1	33%	3
microscopica	0.002950	221	1	33%	3
sublingual	0.002950	268	1	25%	4
ressecao	0.002950	257	1	33%	3
parotida	0.002950	235	1	33%	3
queima	0.002950	249	1	25%	4
nervosa	0.002950	226	1	33%	3

Fonte: Os Autores

6.3.1.4 Modelo de avaliação dos resultados das buscas

Ao analisar as tabelas de resultados, é possível identificar que a ordem dos ranques além de respeitar a quantidade de associações, coluna **EF**, também considera o percentual desta associação para o universo de associações do termo do ranque, coluna **TF**. Em outras palavras, mesmo palavras muito associadas com o termo de busca podem possuir posições inferiores no ranque devido a vulgaridade de outras associações do termo, dado que pode ser visto no percentual da coluna **EF / TF**.

A fim de identificar o comportamento do ranque em relação a quantidade de termos encontrados e variação da quantidade de ocorrência dos termos da busca, conforme levantado pelas classes da curva ABC de ocorrência da Tabela 6.29, os quadros de análise foram segmentados entre os 5, 10 e 15 primeiros resultados do ranque assim como para cada uma das classes ABC. Cada quadro de análise é desenvolvido a luz das métricas apresentadas a seguir:

- **MA:** A fim de analisar a precisão do ranque com base na métrica de associação, usando 5, 10 ou 15 itens analisados no ranque, quantos resultados possuem a "Order EF"superior a este limite. Por exemplo, nos 5 primeiros resultados, quando apresentadas as "Order EF": 1, 3, 4, 2 e 6; há 1 resultado maior, ou seja, quanto mais próximo de 0, mais preciso é. Valor representa a média dos 3 termos de busca;
- **MB:** Para identificar a variação da participação do termo da busca nos termos do ranque, foi análisado o percentual da associação em relação a frequência do termos, valor apresentado na coluna **EF / TF**. Esta métrica é calculada pela média da variação dos 3 termos de busca em relação a quantidade total de itens analisados;
- **MC:** Assim como a métrica **MB**, esta métrica utiliza o valor apresentado na coluna **EF / TF** e aponta a ocorrência de *outliers* por meio da diferença entre a maior e menor variação do percentual. Quanto maior o valor, maior a atuação da lógica do método para trazer termos com base na probabilidade da ocorrência.

6.3.1.5 Análise dos resultados das buscas

Para identificar o desempenho do ranque em relação ao número de termos retornados e termos com menor frequência, foram estruturadas as tabelas a seguir.

Quando utilizado o **cid** na busca, foram obtidas as tabelas de análise 6.57 para **conduta** e 6.58 para **consulta**, baseadas nos resultados apresentados nas Tabelas 6.30, 6.31, 6.32, 6.33, 6.34, 6.35, 6.36, 6.37 e 6.38. Analisando os valores obtidos na métrica **MA**, é possível identificar que os dois fatores analisados apresentam um impacto suave na precisão do ranque por meio da métrica de associação.

Outra consideração a ser feita está no impacto da necessidade de usar a probabilidade de associação em relação a ocorrências de associação. Na Tabela 6.57, o número de ocorrências fora do grupo de associação é menor que os apresentados na Tabela 6.58. Um fator que pode justificar este comportamento é que **conduta** possui 68% mais associações que a **consulta**.

Tabela 6.57: Análise dos termos de conduta para busca por CID

	MA			**MB**			**MC**		
	5	**10**	**15**	**5**	**10**	**15**	**5**	**10**	**15**
A	2	2,6	2,3	19,3	29,3	47,3	22	25	25
B	2	3,3	3,6	21,6	31,6	45	19	16	29
C	3	3,6	4	35	60	63	74	48	49

Fonte: Os Autores

Tabela 6.58: Análise dos termos de consulta para busca por CID

	MA			**MB**			**MC**		
	5	**10**	**15**	**5**	**10**	**15**	**5**	**10**	**15**
A	2	4	5	30	37,6	43,6	11	4	17
B	2	3,6	5,6	16,6	73	73	10	72	72
C	3	5	7	59	64,6	64,6	53	50	50

Fonte: Os Autores

Os valores da métrica **MB** mostram o impacto da atuação da lógica do método WAR em trazer maior relevância a termos de menor posição no ranque assim como os termos de menor número de associação, já a métrica **MC** mostra o peso desta atuação.

É possível identificar que os termos com menor associação, ou seja, os da classe C, que pela métrica **MB** houve uma atuação maior do método WAR em relação as outras classes e, quando observada a **MC**, é possível ver que o método considerou a probabilidade de ocorrência já nos 5 primeiros resultados.

Assim sendo, quando utilizado um dado categórico como busca, o uso da métrica de associação tem maior relevância nos primeiros itens do ranque. Quanto menor for o peso do termo no ranque ou o termo possuir menor associação com os termos mapeados, o método WAR atua para equilibrar estes fatores e os trazer na listagem.

Quando realizada a busca pela **conduta**, foram obtidas as tabelas de análise 6.59 para **cid**, 6.60 para **conduta** e 6.61 para **consulta**. Este resultado foi compilado com base nos resultados apresentados nas Tabelas 6.39, 6.40, 6.41, 6.42, 6.43, 6.44, 6.45, 6.46 e 6.47.

Tabela 6.59: Análise dos termos de CID para busca por conduta

	MA			**MB**			**MC**		
	5	**10**	**15**	**5**	**10**	**15**	**5**	**10**	**15**
A	0,6	0,6	2	40,6	58,3	67,6	20	24	10
B	0,6	2	4	34,6	78,6	87	21	25	9
C	2	6,3	10	67	88	88,6	43	16	14

Fonte: Os Autores

Tabela 6.60: Análise dos termos de conduta para busca por conduta

	MA			**MB**			**MC**		
	5	**10**	**15**	**5**	**10**	**15**	**5**	**10**	**15**
A	3	4	3,3	48,3	53,3	59,3	10	3	14
B	1,3	3	3,6	38,6	56,6	66,6	39	67	43
C	3,3	4,6	5	42,3	66,3	73,6	39	19	16

Fonte: Os Autores

Tabela 6.61: Análise dos termos de consulta para busca por conduta

	MA			**MB**			**MC**		
	5	**10**	**15**	**5**	**10**	**15**	**5**	**10**	**15**
A	0	1,6	2	7	18	18,3	2	14	13
B	1	1	3	13,3	29,3	29,3	9	30	30
C	2,3	3	3	26,6	63,6	64,3	56	81	81

Fonte: Os Autores

Na utilização dos termos de busca da **consulta**, foram obtidas as Tabelas de análise 6.62 para **cid**, 6.63 para **conduta** e 6.64 para **consulta**. Este resultado foi compilado com base nos resultados apresentados nas Tabelas 6.48, 6.49, 6.50, 6.51, 6.52, 6.53, 6.54, 6.55 e 6.56.

Analisando a métrica **MA** das Tabelas 6.59 e 6.62, é possível identificar um desempenho fortemente baseado nas associações para os primeiros cinco itens do ranque para os termos da Classe A.

Outro fator a ser analisado é que os termos de **conduta** estendem o desempenho baseado em associação para os 10 primeiros itens da Classe A e cinco primeiros da Classe B, pois o segmento possui mais registros de associações.

Tabela 6.62: Análise dos termos de CID para busca por consulta

	MA			MB			MC		
	5	**10**	**15**	**5**	**10**	**15**	**5**	**10**	**15**
A	0,6	2	1,3	28	33	35,6	36	30	35
B	1,3	1,6	3,3	29	46,3	74,3	34	37	51
C	3	6	8,3	67,6	97	98	89	6	3

Fonte: Os Autores

Tabela 6.63: Análise dos termos de conduta para busca por consulta

	MA			MB			MC		
	5	**10**	**15**	**5**	**10**	**15**	**5**	**10**	**15**
A	1	1	1,6	13,6	16,6	21,3	18	15	23
B	1,6	2,3	4	19,6	24,6	26	25	35	39
C	4	6,3	7,3	26,3	36	37,6	38	50	48

Fonte: Os Autores

Tabela 6.64: Análise dos termos de consulta para busca por consulta

	MA			MB			MC		
	5	**10**	**15**	**5**	**10**	**15**	**5**	**10**	**15**
A	0,3	0,6	1,6	14,3	38,3	42,3	19	58	58
B	0,6	1,6	2,3	24,6	24,6	39	13	13	27
C	1,3	4,6	8,6	65,3	81,3	81,3	48	49	49

Fonte: Os Autores

Também é possível considerar que quanto mais termos são exigidos no ranque, maior é a ação do método WAR para identificar mais associações. Entretanto, quando analisados os termos da Classe C, é possível identificar que os termos de busca de pouca associação com o **cid** fazem com que o ranque seja populado com 40% à 66,6% de itens identificados pela probabilidade de associação do método WAR.

A variação da métrica **MB** sinaliza que quanto menos registros de associação, maior será a utilização da lógica de probabilidade de associação. Já a métrica **MC** mostrou que quanto menos associações, mais antecipada é a variação dos termos no ranque.

Tal indicador pode ser comprovado observando o comportamento entre a classe C em relação as classes A e B, que apresentam a variação muito mais acentuada já nos primeiros cinco registros do índice e depois oscilações baixas, diferentemente do aumento gradual das classes com mais associações.

Ambos os comportamentos podem ser justificados por meio dos valores do ranque do **cid**, nos quais os valores têm uma redução acentuada no peso ao listar dos termos associados. Isto aponta que há termos fortemente ligados aos termos de **cid**. Entretanto, também há associações raras que possuem relevância média ao ranque e, por fim, despriorizando ocorrências que tem maior peso com outros termos de **cid**.

Antes de proceder com a análise da **conduta** e da **consulta**, é necessário levantar o comportamento da distribuição das auto associações dos termos do mesmo segmento. Para isso, foi desenvolvida a Tabela 6.65 com as informações de total de associações ocorridas, maior ocorrência de associação entre termos e a média para os 25, 50, 100 e 150 valores de maior ocorrência de associações.

Tabela 6.65: Comportamento das distribuições de associação de consulta e conduta

Métrica	**Conduta**	**Consulta**
Total de termos	162.041	96.498
Total de termos associados	196.066	116.268
Total de associações	1.611.047	434.999
Total de associações únicas	83.133	67.781
Média de associações por termo	9,94	4,50
Maior associação	1045	3447
Média 25 maiores ocorrências	767,48	919,92
Média 50 maiores ocorrências	684,1	694,44
Média 100 maiores ocorrências	604,52	485,94
Média 150 maiores ocorrências	552,25	399,16

Fonte: Os Autores

Analisando a Tabela 6.65 é possível identificar que a **conduta** gerou uma média de associação por termo maior que o dobro em relação a **consulta**. Quando analisadas as associações de ocorrência única, o percentual da **consulta** representa 15,58% das associações, enquanto da **conduta** representa apenas 5,16%. Quando analisado o total de associações únicas sobre o total de termos associados em relação as associações únicas, é obtido um total de 42,4% de associações únicas para **conduta** e 58,30% para **consulta**. Com estas informações, é possível cruzar estes dados com os apresentados nas tabelas de análise para explicar e justificar o comportamento do método neste cenário.

Os termos da **consulta** apresentados do ranque sobre os termos de busca para **conduta** e **consulta**, respectivamente, geram as Tabelas de análise 6.61 e 6.64, permitindo que quando analisada a métrica **MA** seja possível identificar que o ranque prioriza entre 100% à 80% dos resultados os termos pela ordem de peso das associações para as classes A e B. Contudo, para a classe C os termos listados com base na associação reduzem para 80% à 42,6%.

Quando analisadas as métricas **MB** e **MC**, é visto uma maior variação nos termos da classe C, associada a baixa associação de termos devido a queda acentuada das médias de associações do segmento de **consulta**.

Quando analisado os termos da **conduta** apresentados no ranque sobre os termos de busca para **conduta** e **consulta**, respectivamente, geram as Tabelas de análise 6.60 e 6.63. É possível identificar pela métrica **MA** que por buscar em uma base de dados que possui menor número de associações e distribuição menos uniforme de termos associados, é possível identificar que o método WAR atuou para elencar mais associações identificadas como associações prováveis em relação aos termos com alto registro de associações.

Ao analisar as métricas **MB** e **MC**, é enfatizado que as buscas da base do segmento **conduta** apresentam o número reduzido de termos que utilizam o total de associações em relação do segmento **consulta**. Mesmo apresentando um maior número de associações, as mesmas são apresentadas com distribuição uniforme, permitindo com o método WAR atue buscando outros termos com base na probabilidade de ocorrer.

Por fim, quando comparado as auto associações de **conduta** na Tabela 6.60 e **consulta** na Tabela 6.64, é possível identificar via métrica **MA** que a **consulta** teve um resultado muito mais relacionado ao peso das associações. Ao utilizar termos com menos associações, classe C, os termos listados com base na probabilidade de ocorrência tiveram maior peso.

Observando a distribuição da **conduta**, é possível identificar que a variação de termos relacionada a probabilidade de ocorrência é muito mais estável. Este comportamento vai ao encontro com o identificado na média de maiores ocorrências de associação apresentada na Tabela 6.65, onde para a **conduta** a distribuição dos termos é mais uniforme, já para a **consulta** é mais acentuada.

6.3.1.6 Considerações das análises

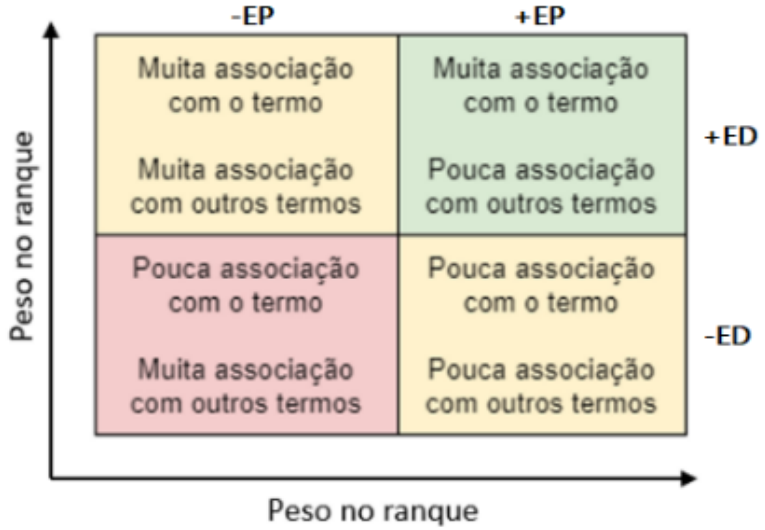
Com estes experimentos foi possível identificar que o ranque apresentado pelo método WAR não se baseia unicamente pelo peso de associações, evitando que associações vulgares tenham destaque.

Distribuições mais acentuadas tiveram maior peso para as associações nos primeiros termos do ranque e associados com maiores ocorrências de associação, mas também, de forma acentuada, apresentam termos por meio da probabilidade de ocorrência quando necessário listar mais termos ou termos associados a buscas de menor registro de associações.

Em distribuições uniformes e de maior número de associações, a abordagem do método foi mais presente para trazer o critério com base na probabilidade de ocorrência.

Assim sendo, este cenário de validação permitiu comprovar o comportamento proposto pelo método WAR em equilibrar o peso dos termos do ranque com base na exclusividade e vulgaridade da associação, informações apresentadas, respectivamente, pelas métrica ED e EP, conforme apresentadas na Figura 6.3.

Figura 6.3: Grade dos critérios de peso no ranque do método WAR



Fonte: Dos Autores.

6.3.2 Cenário 2: busca em múltiplos segmentos validando a precisão do ranque

Neste cenário, são realizadas buscas para identificar a efetividade do ranque. Para isso, também foram utilizados os termos levantados na Tabela 6.29. Para isso, foram utilizados os três termos da classe A para **conduta** e **consulta**.

Em paralelo, foi realizado o levantamento da base de dados de origem de todos os registros de laudos que possuem todos os três termos para cada segmento. Como resultado, foi obtido o total de 15 registros que possuem todos os registros da classe A, ou seja **conduta** com "sintoma", "sabonete"e "roupa"e **consulta** com "mal", "prurido"e "sinal".

Com a utilização destes 15 registros levantados na filtragem, foi possível cruzar com os termos listados no ranque a fim de validar a efetividade do método, o que resultou na Tabela 6.66 que apresenta o termo, o peso do ranque e o número de registros que possuem o termo no levantamento realizado.

No andamento, a análise foi dividida em 2 pontos: identificar o impacto do método nos dados categóricos de **cid**; e o impacto nos termos de texto livre de **conduta** e **consulta**.

Tabela 6.66: Resultado do WAR para a busca de consulta e conduta da classe A

CID			Conduta			Consulta		
Termo	ER	#	Termo	ER	#	Termo	ER	#
L20	4.956.675	7	perfume	5.490.704	7	ardencia	2.688.600	5
L57.0	1.635.038	0	hidratante	4.690.267	5	pruriginosa	2.217.508	2
L82	0.772608	0	amaciante	3.247.354	4	descamativa	1.239.317	2
L40	0.452667	0	antihistaminico	3.104.358	10	mancha	0.673554	0
B86	0.372900	0	girassol	3.018.912	7	mao	0.630792	2
L29	0.328871	3	fragrancia	2.788.929	5	sangramento	0.620379	2
C44.9	0.295154	0	hidroxizine	2.739.888	8	eritematosa	0.563000	2
L70	0.292667	0	abrasivo	2.733.837	5	placa	0.553692	0
L71	0.281767	0	cuidado	2.635.058	5	piora	0.525625	1
D22	0.230892	0	diario	2.462.633	8	descamacao	0.520729	1
L23	0.211200	0	pequena	2.442.512	3	dorso	0.428046	1
C44	0.120725	0	locao	2.239.079	6	membro	0.411817	0
B35.4	0.120300	0	neutro	2.207.996	7	hipercromica	0.362554	0
L21	0.115787	0	cremosa	1.933.938	4	sensibilidade	0.346471	1
B35	0.106288	1	exacerbacao	1.899.379	4	corpo	0.344271	3

Fonte: Os Autores

6.3.2.1 Análise do resultado do CID

No levantamento de laudos, foram obtidos 7 diferentes códigos **cid**, sendo: L20 com 7 ocorrências; L29 com 3 ocorrências; e B35, L90.0, I83.1, L30.0 e L60.9 com 1 ocorrência cada. No ranque o **cid** de código L20 foi o mais bem colocado, enquanto os **cid** L29 e B35 também aparecem no ranque, entretanto 57% dos códigos **cid** não foram apresentados.

A fim de exemplificar a relação das associações, foi estruturada na Tabela 6.67 a relação da associação com os termos da busca com os 7 primeiros termos encontrados no ranque, L20, L57.0, L82, L40, B86, L29 e C44.9 .

No levantamento da Tabela 6.67, é possível identificar o impacto de múltiplos critérios na definição do peso do ranque. Um dos principais critérios é o total de associações, entretanto, ele não é determinístico para a ordem do ranque. Como pode ser visto no compilado da Tabela 6.68, o total de associações não é determinístico para o peso, pois não segue uma ordem como do método Apriori.

Tabela 6.67: Total de associações entre termos de CID com os da busca

	Conduta				**Consulta**			
	sintoma	sabonete	roupa	Outros	mal	prurido	sinal	Outros
L20	33	368	346	8.862	228	345	117	2.900
L57.0	288	5	237	6.245	383	207	69	3.220
L82	34	4	16	6.996	638	307	129	5.316
L40	165	12	7	7.217	227	265	88	2.514
B86	84	26	102	2.888	21	97	109	802
L29	66	115	45	3.747	94	136	28	1.138
C44.9	10	4	8	3.593	419	175	67	4.440
Outros	1.528	1.210	980		5.093	4.656	2.083	
Total	2.208	1.744	1.741		7.103	6.188	2.690	

Fonte: Os Autores

Tabela 6.68: Total de associações entre termos de CID com outros segmentos

	Total de associações		
	Conduta	Consulta	Ambos
L20	747	690	1437
L57.0	530	659	913
L82	54	1074	1128
L40	184	580	764
B86	212	227	439
L29	226	258	484
C44.9	22	661	683

Fonte: Os Autores

Um dos fatores que fazem com que no ranque, Tabela 6.66, priorize outros 12 termos de **cid** no lugar dos outros cinco termos do levantamento de laudo que não foram listados nos primeiros 15 registros, está no fato do peso da probabilidade da associação.

Ao analisar o resultado de associação do **cid** L20, é visto que os termos que mais se destacam em associação da **conduta** são sabonete e roupa, com 368 e 346 associações respectivamente. Na **consulta**, o termo prurido também possui um número elevado de associações. Entretanto, ao analisar a proporcionalidade com o total do uso do termo, vemos que na **conduta** das associações, sabonete representou 21,1% e roupa 19,8% do total de associações e, no segmento **consulta**, prurido representou apenas 5,6% do total do termo da busca.

Para visualizar o impacto desta alteração, foi realizada novamente a busca, entretanto, removendo os termos sabonete e roupa da **conduta**. Na Tabela 6.69 é visto que o termo do **cid** caiu para a quarta posição.

Tabela 6.69: Total de associações entre termos de CID com os da busca

Ranque		**Conduta**		**Consulta**			
Termo	ER	sintoma	Outros	mal	prurido	sinal	Outros
L57.0	1.576900	288	6.487	383	207	69	3.220
L82	1.158913	34	6.716	638	307	129	5.316
L40	0.679000	165	7.236	227	265	88	2.514
L20	0.596781	33	9.576	228	345	117	2.900

Fonte: Os Autores

Outra maneira de visualizar a outra forma que o método WAR atua, está em ver a especificidade da associação pelo termo que foi apresentado no ranque. Para isso, é realizada uma outra busca com apenas o segmento da **consulta** pelos termos prurido e sinal, o qual resultou nos 6 primeiros resultados nos itens da Tabela 6.70.

Tabela 6.70: Total de associações entre termos de CID com os da busca

Ranque		**Consulta**		
Termo	ER	prurido	sinal	Outros
L20	0.968588	345	117	3.128
L40	0.546800	265	88	2.741
L23	0.456325	204	89	2.221
L82	0.332175	307	129	5.954
L29	0.198275	136	28	1.232
L57.0	0.197150	207	69	3.403

Fonte: Os Autores

Neste resultado, ao comparar os valores dos termos de **cid** L23 e L82, é possível ver que L82 está abaixo de L23 mesmo possuindo mais associações para ambos os termos. Isso se da pelo fato que L82 tem mais associações com outros termos quando comparado com o termo L23, caracterizando suas associações como vulgares e, consequentemente, reduzindo seu impacto no peso do ranque.

Triangulando as informações sobre o comportamento do método WAR, resultados obtidos no ranque e os dados obtidos no levantamento dos registros dos laudos filtrados por todos os termos da classe A, foi possível identificar um comportamento coerente com as expectativas. Os termos de **cid** que foram encontrados no levantamento da base se marcaram presente no resultado do ranque quando possuíam maior número de ocorrências.

Ocorrências únicas de **cid** não se fizeram presente no ranque pois, como foi visto, os fatores de probabilidade de associação despriorizam termos vulgares fazendo com que resultados que ocorram poucas vezes com todos os termos de busca tenha peso menor quando identificado um peso de exclusividade em parte dos termos.

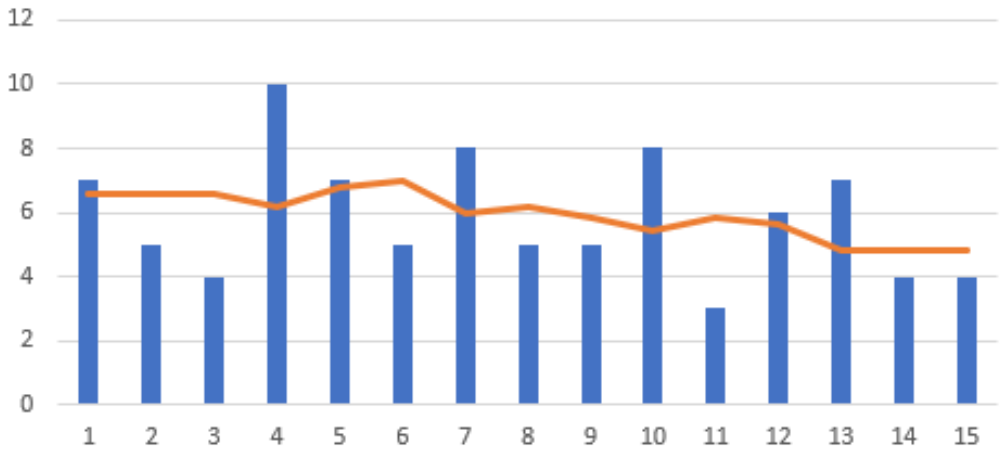
6.3.2.2 Análise do resultado de consulta e conduta

A fim de identificar a precisão do ranque, foi utilizado novamente o levantamento com 15 laudos que possuem obrigatoriamente todos os termos da busca, entretanto, avaliando a precisão do ranque sobre os segmentos de texto **conduta** e **consulta** dos laudos.

Ao analisar o resultado da busca apresentado na Tabela 6.66, são observados os termos da **conduta** que possuem registro no levantamento dos laudos para todos os 15 itens do ranque. Para identificar a primeira ocorrência de zero, o ranque foi estendido para mais unidades. O primeiro item apresentando no raque que não conta no levantamento foi o termo de posição 20. Na **consulta**, o resultado apresentou um maior número de termos no ranque que não constam no levantamento, sendo 4 dos 15 registros.

Ao somar o total de associações para cada grupo de 5 termos no ranque da **conduta**, é contabilizado que os cinco primeiros tiveram uma média de 6,6 ocorrências, já os próximos cinco resultados tiveram a média de 6,2 e no último a média foi de 4,8. Ao realizar a mesma avaliação para **consulta**, é obtida a média de 2,2, 1,2 e 1. Quando representado com a média móvel para **conduta** e **consulta**, são obtidos os respectivos gráficos das Figuras 6.4 e 6.5. Em ambas as perspectivas, é possível ver que a resposta do método WAR ordenou os termos de maior ocorrência e aos poucos isso vai reduzindo ao crescer do número de itens no ranque, o que significou maior precisão.

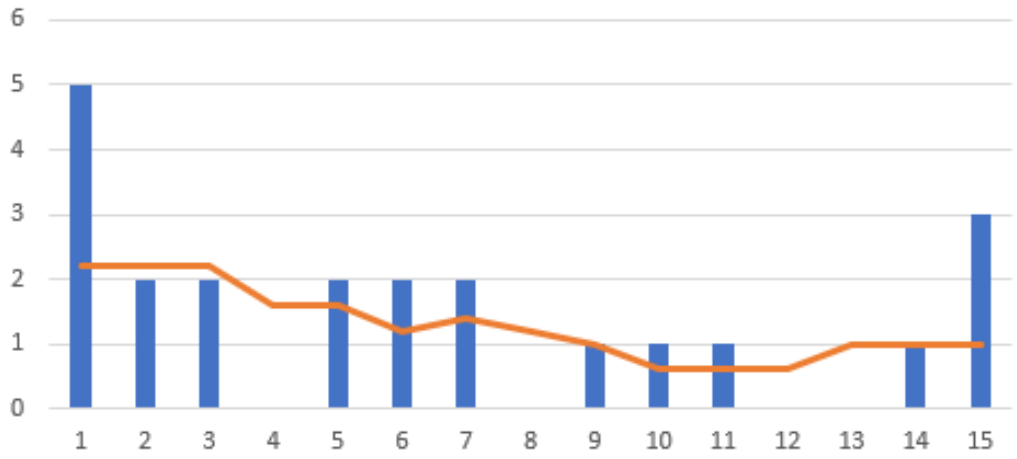
Figura 6.4: Media móvel das ocorrências da conduta do ranque nos laudos



Fonte: Dos Autores.

Este comportamento é explicado pela distribuição não uniforme dos termos usados na busca pelo segmento da **consulta** e vulgaridade dos mesmos em relação aos termos da **conduta**.

Figura 6.5: Media móvel das ocorrências da consulta do ranque nos laudos



Fonte: Dos Autores.

Como último modelo de análise deste cenário, foi realizada a contabilização e ordenação dos termos do levantamento para cada segmento. O resultado geral para **conduta** é apresentado no anexo F e para **consulta** no anexo G.

O segmento de texto **conduta** possui 227 termos diferentes, sendo: 2 com 15; 1 com 12; 1 com 10; 3 com 8; 4 com 7; 4 com 6; 7 com 5; 14 com 4; 26 com 3; 56 com 2; e 109 com 1 ocorrência. Já para o segmento de texto **consulta** há 84 termos diferentes, sendo: 2 com 15; 1 com 12; 1 com 5; 3 com 3; 15 com 2; e 62 com 1 ocorrência.

Ao relacionar os termos do ranque com as ocorrências do levantamento, é obtido a tabela 6.71 para o segmento de **conduta** e a Tabela 6.72 para **consulta**. Nesse formato, fica claro que o modelo de priorização do método WAR evita que extremos definam a prioridade do ranque fazendo com que ocorrências mais centralizadas tenham maior destaque no ranque.

Quando o número de ocorrências é contabilizado sobre o eixo das ocorrências do levantamento, é obtido a figura 6.6 pra **conduta** e 6.7 para **consulta**. Nestas ilustrações é confirmado o comportamento de evitar os extremos formando uma curva *gaussiana* sobre o total de ocorrências.

Mesmo um dos gráficos estando mais deslocado para o extremo de menos ocorrências, o método tende a centralizar a curva conforme mais ocorrências de associação vão surgindo.

Analisando o gráfico da Figura 6.6 dos termos de **conduta**, resultou em um melhor equilíbrio entre número de associações e probabilidade de associação sobre as 227 ocorrências, o que resultou em uma curva centralizada.

Tabela 6.71: Relação do ranque de conduta com as ocorrências do levantamento

Ranque		**Levantamento**	
Termo	ER	Ordem	Ocorrências
perfume	1	5	7
hidratante	2	7	5
amaciante	3	8	4
antihistaminico	4	3	10
girassol	5	5	7
fragrancia	6	7	5
hidroxizine	7	4	8
abrasivo	8	7	5
cuidado	9	7	5
diario	10	4	8
pequena	11	9	3
locao	12	6	6
neutro	13	5	7
cremosa	14	8	4
exacerbacao	15	8	4

Fonte: Os Autores

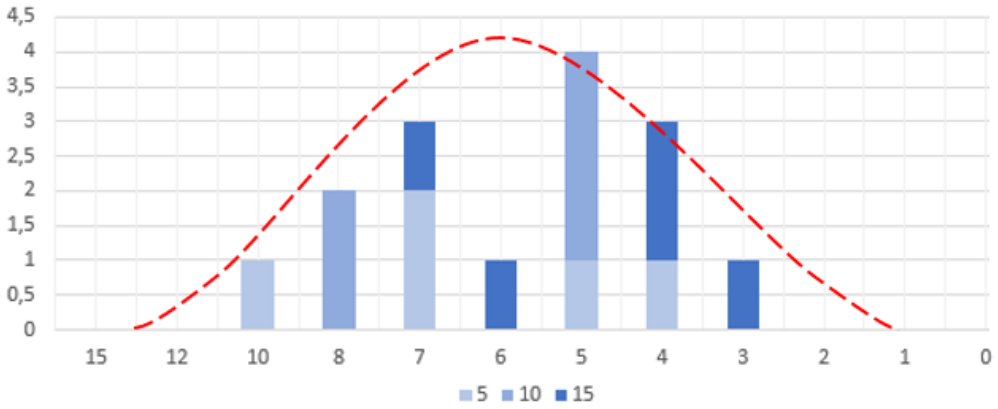
Tabela 6.72: Relação do ranque de consulta com as ocorrências do levantamento

Ranque		**Levantamento**	
Termo	ER	Ordem	Ocorrências
ardencia	1	3	5
pruriginosa	2	5	2
descamativa	3	5	2
mancha	4	N/A	0
mao	5	5	2
sangramento	6	5	2
eritematosa	7	5	2
placa	8	N/A	0
piora	9	6	1
descamacao	10	6	1
dorso	11	6	1
membro	12	N/A	0
hipercromica	13	N/A	0
sensibilidade	14	6	1
corpo	15	4	3

Fonte: Os Autores

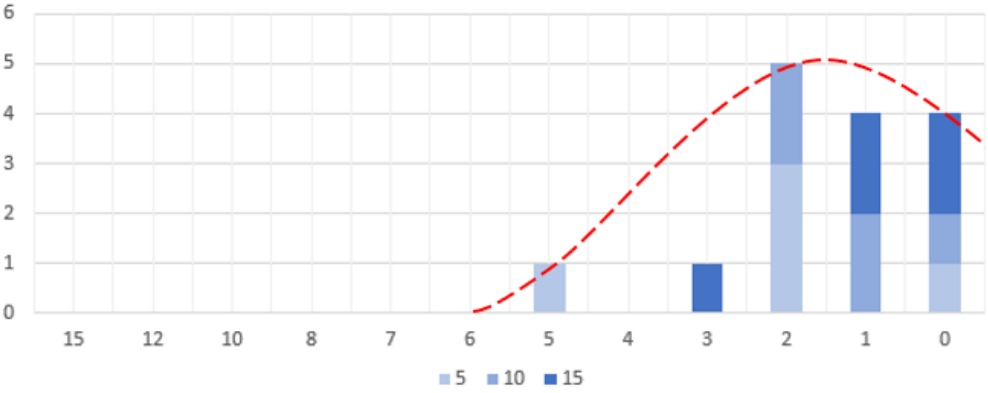
O fato que desloca a curva do gráfico da Figura 6.7 dos termos de **conduta** é por possuir apenas 84 associações e os termos utilizados na busca serem mais vulgares em relação aos da **conduta**.

Figura 6.6: Distribuição dos termos de conduta do ranque nos laudos



Fonte: Dos Autores.

Figura 6.7: Distribuição dos termos de consulta do ranque nos laudos



Fonte: Dos Autores.

6.3.2.3 Considerações das análises

Conforme apresentado, o comportamento do método WAR permitiu com que uma nova abordagem coerente seja considerada. O equilíbrio entre número de associações, métrica EF e total de associações TF para cálculo da probabilidade de associação permitiu que o ranque distribua os pesos evitando os extremos como termos vulgares ou *outliers* de uso raro.

Ao analisar o comportamento de busca com a própria base de recuperação foi possível identificar a precisão e atuação do método para trazer os resultados de maneira equilibrada em cenários de número de associações variadas.

7 CONCLUSÃO

Em cenários de mineração de texto o uso de algoritmos de associação como o Apriori faz com que as pesquisas percam precisão e se tornem inutilizáveis devido a abordagem de contabilizar somente as associações pela lógica de estruturação de pesos baseada na árvore de associações. Algoritmos de IR como o TD/IDF apresentam uma prioridade diferente aos pesos, pois realizam um ranque que equilibra o número de registros despriorizando as ocorrências de termos vulgares e raras.

Como proposta nesta dissertação, o método WAR apresentou duas contribuições relevantes, a capacidade de aplicar a busca de associações em registros multissegmentados por meio das matrizes multidimensionais MMRT e MMRI. Outro fator, foi aplicar a lógica que despriorizando as ocorrências de associações vulgares e raras. Isto significa que, a lógica que o TF/IDF aplica sobre os termos, o método WAR aplica sobre todo o conjunto de associações.

Nos experimentos foi possível identificar que o método WAR priorizou maior número de associações exclusivas entre os termos da busca e do ranque, despriorizando associações vulgares. Outro fator que interferiu diretamente no desempenho do método WAR é o número de termos e associações geradas no segmento de texto e a distribuição da quantidade de associações.

Desta forma, foi possível concluir que o método proposto possui um comportamento de regras de associação, não apresentando as limitações do Apriori pois apresenta pesos ponderados como das técnicas de recuperação de informações. Assim, o método WAR fornece uma forma diferente aos termos de associação clássicos.

7.1 Limitações

O método WAR apresentou bons resultados na recuperação das associações dos termos, entretanto a elaboração das matrizes multidimensionais exige um tempo considerável de processamento.

Sobre os recursos computacionais necessários, foi observado uma alta demanda por memória pelo fato de a implementação armazenar as métricas em memória. Como resultado o código necessitou de toda a capacidade de 16GB do computador utilizado para o processamento.

7.2 Aplicabilidade

A capacidade de permitir que a busca possa ser realizada em bases de múltiplos segmentos de texto, onde eles equalizem o peso do resultado para todos os segmentos do ranque, faz com que o método seja muito válido para análise de registros processuais. Em cenários onde cada etapa de um processo gera um registro, é possível analisar o impacto dos termos do começo de um processo em seu final e em via inversa também.

Com a pandemia do Covid-19 os sistemas de atendimento a distância foram mais requisitados e, consequentemente, geraram mais dados. Nesta dissertação, foi utilizada a plataforma de teleatendimento do TelessaúdeRS e, além de cumprir os requisitos da pesquisa, o método pode também auxiliar nos seguintes cenários:

- Permitir análises exploratórias nas bases de dados a fim de validar hipóteses sobre o comportamento do processo;
- Realizar o levantamento de qual medicamento está sendo mais associado a atendimentos de determinada doença; e
- Por ser um processo de múltiplas etapas que tem por fim a definição de encaminhar para uma unidade de saúde especializada quando há risco ao paciente, com base em um *threshold* o método WAR pode avaliar as etapas iniciais de forma correlacionada e verificar se há a probabilidade de encaminhamento na atividade final, solicitando, assim, uma revisão priorizada do atendimento. Esta abordagem pode evitar, por exemplo, que pacientes com tendência a ter câncer de pelo tenham que esperar o mesmo tempo que uma pessoa com alergia.

Outras aplicabilidades podem ser estendidas a todos os cenários de processos com múltiplas etapas, no qual cada uma gera uma saída direta ou indiretamente no formato de texto, tais como:

- A fim de entender os motivos de *feedbacks* negativos, é possível correlacionar os termos da descrição do produto, dúvidas dos clientes e *feedbacks* de compra de um *market place*;
- Ferramentas de *service desk* para RH e TI tramitam toda a comunicação via mensagens textuais e este método pode permitir melhor precisão no entendimento na relação entre causa e solução, identificando informações de termos no processo de atendimento de chamados de empresas;

- Buscar se há termos de uma petição inicial e informações dos trâmites processuais que possuem relação com o parecer do juiz. Caso haja, quais são os termos que acarretam mais em decisões positivas e negativas.

7.3 Trabalhos Futuros

Este trabalho contribuiu para a possibilidade de aplicar os mesmos experimentos em bases diferentes, preferencialmente que precisem de menos pré-processamento.

O desempenho do método está diretamente ligado com a qualidade e variabilidade dos termos da base. Dessa forma, é levantando como um ponto de pesquisa o impacto da lógica de ontologia no pré-processamento para o desempenho do método WAR. Por fim, este método pode contribuir na entrega de um *baseline* para demais técnicas.

REFERÊNCIAS

2007, S. B. P. n. 35 de 04 de janeiro de. *Institui, no âmbito do Ministério da Saúde, o Programa Nacional de Telessaúde. Diário Oficial da União 2007;Seçao 1:85*. 2007.

AGGARWAL, C. C. **Data Mining: The Textbook**. [S.l.]: Springer Publishing Company, Incorporated, 2015. ISBN 3319141414.

AI, Q. et al. Analysis of the paragraph vector model for information retrieval. In: **Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval**. New York, NY, USA: Association for Computing Machinery, 2016. (ICTIR '16), p. 133–142. ISBN 9781450344975. Available from Internet: <https://doi.org/10.1145/2970398.2970409>.

BAEZA-YATES, R.; RIBEIRO-NETO, B. **Recuperação de Informação - 2ed: Conceitos e Tecnologia das Máquinas de Busca**. [S.l.]: Bookman Editora, 2013. ISBN 9788582600498.

BERRY, M.; CASTELLANOS, M. **Survey of Text Mining II: Clustering, Classification, and Retrieval**. [S.l.]: Springer, 2008.

BRAMER, M. **Principles of Data Mining**. 3. ed. London: Springer, 2013. (Undergraduate Topics in Computer Science). ISSN 1863-7310. ISBN 978-1-4471-7306-9.

BRASIL. Portaria nº 35, de 4 de janeiro de 2007: Institui, no âmbito do ministério da saúde, o programa nacional de telessaúde. **Diário Oficial da República Federativa do Brasil**, Brasília, DF, 2007. Available from Internet: <http://portalarquivos2.saude.gov.br/images/pdf/2014/fevereiro/13/portaria35-04012007.pdf>.

BRASIL. Redefine e amplia o programa telessaude brasil,que passa a ser denominado programa nacional telessaude brasil red. **Diário Oficial da República Federativa do Brasil**, Brasília, DF, 2011. Available from Internet: <https://bvsms.saude.gov.br/bvs/saudelegis/gm/2011/prt2546_27_10_2011.html>.

BRASIL. Manual de telessaude para atencao basica/atencao primaria em saude - protocolo de resposta. **Diário Oficial da República Federativa do Brasil**, Brasília, DF, 2012. Available from Internet: <http://bvsms.saude.gov.br/bvs/publicacoes/manual_telessaude_protocolo_respostas_teleconsultorias.pdf>.

BRASIL. Manual de telessaude para atencao basica/atencao primaria em saude - protocolo de solicitacao. **Diário Oficial da República Federativa do Brasil**, Brasília, DF, 2012. Available from Internet: <http://bvsms.saude.gov.br/bvs/publicacoes/pataforma_telessaude_tutorial_solicitante.pdf>.

BRASIL. Manual de telessaude para atencao basica/atencao primaria em saude - protocolo de telerregulacao. **Diário Oficial da República Federativa do Brasil**, Brasília, DF, 2012. Available from Internet: <http://bvsms.saude.gov.br/bvs/publicacoes/manual_telessaude_atencao_basica_telerregulacao.pdf>.

COSTA, M. M.; GONçALVES, M. R.; BAKOS, R. M. **Mineração de Texto Para Descoberta do Conhecimento em Laudos de Teledermatologia**. Dissertation (Master) — Universidade Federal do Rio Grande do Sul, Porto Alegre, RS, 2021.

GARCIA-MOLINA, H.; ULLMAN, J. D.; WIDOM, J. **Database Systems: The Complete Book**. 2. ed. USA: Prentice Hall Press, 2008. ISBN 9780131873254.

GOSWAMI, S.; KUNDU, C. Xml based advanced distributed database: implemented on library system. **International journal of information management**, Elsevier, v. 33, n. 1, p. 28–31, 2013.

HADDAD, A. E. Experiência brasileira do programa nacional telessaude brasil. In: ____. [S.l.: s.n.], 2012. p. 12–42. ISBN 978-85-7511-238-0.

Heaton, J. Comparing dataset characteristics that favor the apriori, eclat or fp-growth frequent itemset mining algorithms. In: **SoutheastCon 2016**. [S.l.: s.n.], 2016. p. 1–7.

HERRERA, A. G. S. de et al. (Ed.). **33rd IEEE International Symposium on Computer-Based Medical Systems, CBMS 2020, Rochester, MN, USA, July 28-30, 2020**. [S.l.]: IEEE, 2020. ISBN 978-1-7281-9429-5.

HS, P. Teledermatology and teledermatopathology. In: **Semin Cutan Med Surg**. [S.l.: s.n.], 2002. (21(3)), p. 89–179.

JENSEN, P. B.; JENSEN, L. J.; BRUNAK, S. Mining electronic health records: towards better research applications and clinical care. **Nature Reviews Genetics**, v. 13, n. 6, p. 395–405, Jun 2012. ISSN 1471-0064. Available from Internet: <https://doi.org/10.1038/nrg3208>.

KULKARNI, A.; TOKEKAR, V.; KULKARNI, P. Identifying context of text documents using naïve bayes classification and apriori association rule mining. In: . [S.l.: s.n.], 2012. p. 1–4. ISBN 978-1-4673-2174-7.

LIPCZAK, M. et al. Selective retrieval for categorization of semi-structured web resources. In: SPRINGER. **Canadian Conference on Artificial Intelligence**. [S.l.], 2013. p. 126–137.

MADANI, A.; BOUSSAID, O.; ZEGOUR, D. E. Semi-structured documents mining: a review and comparison. **Procedia Computer Science**, Elsevier, v. 22, p. 330–339, 2013.

MANNING, C. D.; RAGHAVAN, P.; SCHUTZE, H. **Introduction to Information Retrieval**. [S.l.]: Cambridge University Press, 2008.

MEADOW, C.; BOYCE, B.; KRAFT, D. **Text Information Retrieval Systems**. [S.l.]: Academic Press, 2000. (Library and information science). ISBN 9780124874053.

MELLO, R. d. S. Dados semi-estruturados. 2000.

MELLO, R. d. S. Uma abordagem bottom-up para a integração semântica de esquemas xml. 2002.

NADIG, S.; BRASCHLER, M.; STOCKINGER, K. Database search vs. information retrieval: a novel method for studying natural language querying of semi-structured data. In: EUROPEAN LANGUAGE RESOURCES ASSOCIATION. **12th Language Resources and Evaluation Conference (LREC) 2020**. [S.l.], 2020.

PAKHS. Teledermatology and teledermatopathology. **Semin Cutan Med Surg**, v. 21, n. 3, p. 179–189, 2002.

SINGH, T.; SETHI, M. Sandwich-apriori: A combine approach of apriori and reverse-apriori. In: **2015 Annual IEEE India Conference (INDICON)**. [S.l.: s.n.], 2015. p. 1–4.

WEISS, S. M. et al. **Information Retrieval and Text Mining**. New York, NY: Springer New York, 2005. ISBN 978-0-387-34555-0.

APÊNDICE A — TABELA DE TERMOS INDEVIDAMENTE CONCATENADOS

VALUE
tempo
realizou
fatores de risco
historia
sinais e sintomas
ja foi
se sim
faz uso
descricao
comorbidades
realizou
acrescente
tempo
imunossupressao
pdf

APÊNDICE B — TABELA DE *TOKENS* PARA O SEGMENTO DE CONSULTA

VALUE	TOKEN
TEMPO DE EVOLUCAO: MENOS DE 1 SEMANA	tempo_sem_1_menos
TEMPO DE EVOLUCAO: ENTRE 1 E 4 SEMANAS	tempo_sem_1_4_entre
TEMPO DE EVOLUCAO: ENTRE 1 E 3 MESES	tempo_mes_1_3_entre
TEMPO DE EVOLUCAO: ENTRE 3 E 6 MESES	tempo_mes_3_6_entre
TEMPO DE EVOLUCAO: MAIS DE 6 MESES	tempo_mes_6_mais
IMUNOSSUPRESSAO?: [U'N\XE3O']	imunossup_sim
IMUNOSSUPRESSAO?: NAO	imunossup_nao
IMUNOSSUPRESSAO?:	
HISTORIA PREVIA DE CANCER DE PELE?: NAO	hist_cancer_nao
HISTORIA PREVIA DE CANCER DE PELE?: CARCINOMA BASOCELULAR	hist_cancer_sim
HISTORIA PREVIA DE CANCER DE PELE?: CARCINOMA ESPINOCELULAR	hist_cancer_sim
HISTORIA PREVIA DE CANCER DE PELE?: MELANOMA	hist_cancer_sim
HISTORIA PREVIA DE CANCER DE PELE?:	
HISTORIA FAMILIAR DE CANCER DE PELE?: NAO	hist_fam_cancer_nao
HISTORIA FAMILIAR DE CANCER DE PELE?: CARCINOMA BASOCELULAR	hist_fam_cancer_sim
HISTORIA FAMILIAR DE CANCER DE PELE?: CARCINOMA ESPINOCELULAR	hist_fam_cancer_sim
HISTORIA FAMILIAR DE CANCER DE PELE?: MELANOMA	hist_fam_cancer_sim
REALIZOU BIOPSIA?: NAO	biopsia_nao
REALIZOU BIOPSIA?: SIM	biopsia_sim
REALIZOU BIOPSIA?:	
ACRESCENTE OUTRAS INFORMACOES IMPORTANTES:	
DESCRICAO DA LESAO (LOCALIZACAO, TAMANHO, CARACTERISTICAS):	
SINAIS E SINTOMAS:	
SE SIM, INFORME QUAL(IS) TRATAMENTO(S) REALIZADO(S)TOPICO(S) E SISTEMICO(S):	
SE SIM, DESCREVA O RESULTADO E/OU ANEXE A FOTO E INFORME A DATA:	
JA FOI REALIZADO ALGUM TRATAMENTO PARA O QUADRO?: SIM	tratamento_para_quadro_sim
JA FOI REALIZADO ALGUM TRATAMENTO PARA O QUADRO?: NAO	tratamento_para_quadro_nao
JA FOI REALIZADO ALGUM TRATAMENTO PARA O QUADRO?:	
COMORBIDADES: NAO	comorbidades_nao
COMORBIDADES: SIM	comorbidades_sim
COMORBIDADES:	
FAZ USO DE MEDICACOES CONTINUAS OU EVENTUAIS: NAO	
FAZ USO DE MEDICACOES CONTINUAS OU EVENTUAIS:	
HIPOTESE DIAGNOSTICA:	
DESCRICAO DO QUADRO CLINICO (LOCALIZACAO, CARACTERISTICAS, EVOLUCAO):	
SE SIM, QUAL(IS) A(S) MEDICACAO(OES) UTILIZADA(S):	
SE SIM, DESCREVA AS COMORBIDADES DO PACIENTE:	
REALIZOU EXAMES COMPLEMENTARES?: SIM	example_comp_sim
REALIZOU EXAMES COMPLEMENTARES?: NAO	example_comp_nao
REALIZOU EXAMES COMPLEMENTARES?:	
CARACTERISTICAS:	
PRIMEIRA CONSULTA :	
FATORES DE RISCO PARA CANCER DE PELE:	
TEMPO DE EVOLUCAO: NAO SABE	
TEMPO DE EVOLUCAO:	

APÊNDICE C — TABELA DE *TOKENS* PARA O SEGMENTO DE CONDUTA

VALUE	TOKEN
DEDOS EM SALSICHA	dedos_em_salsicha

APÊNDICE D — *WORDLIST*: LISTA DE *TOKENS* DE CONSULTA VÁLIDOS

Tokens válidos: prurido, ardencia, pruriginosas, sangramento, descamacao, dorso, descamativas, manchas, aumento, sensibilidade, maos, surgimento, piora, cabeludo, crescimento, drenagem, eritematosas, membros, pruriginosa, torax, pus, corpo, secrecao, mancha, descamativa, agricultor, placas, hipercromicas, papulas, laboral, pedreiro, mao, tronco, motoboy, contato, progressivo, aparecimento, bordas, hipercromica, perna, hipocromicas, eritematosa, febre, coceira, asma, dermatite, pernas, aumentando, cor, bracos, nariz, depressao, irregulares, placa, cervical, hiperemia, carcinoma, acne, antebraco, bordos, coloracao, rosto, sol, abdome, hipotireoidismo, unhas, crostas, psoriase, hiv, costas, hiperemiadas, basocelular, obesidade, nasal, braco, coxa, mmii, tabagista, vesiculas, diabetes, alergia, edema, mmss, calor, nevos, cabelo, surgiram, elevada, eritema, dorsal, irregular, cronica, ansiedade, bilateral, maculas, crosta, hiperemiada, nevo, verruga, dm2, notou, dolorosa, papula, dislipidemia, nascimento, alopecia, localizada, tabagismo, sifilis, antebracos, surgiu, hipocromica, multiplas, elevadas, mama, alergias, abdomen, cotovelos, plantar, coxas, ulcerada, unha, avermelhadas, escura, verrugas, dedo, relevo, acastanhada, axilas, bolhas, membro, novas, rapido, artrite, rinite, alergica, inguinal, macula, temporal, verao, labio, progressiva, regulares, sinal, agua, lombar, dolorosas, pruridose, pustulas, regular, apareceu, borda, purulenta, aumentou, cronico, nadegas, crostosa, recidiva, toracica, verrucosa, avermelhada, axilar, cotovelo, disseminadas, ombro, bilateralmente, nodular, palpebra, hipertensa, crostosas, onicomicose, nodulo, olho, artralgia, cabeca, fissuras, ulcera, apareceram, queimacao, infeccao, marrom, rediscussao, halo, rostro, gestacao, seborreica, abdominal, aspera, aumentado, flogisticos, intermitente, cicatriz, delimitadas, joelho, maxilar, sangra, colo, esbranquicadas, halux, nova, palma, papulares, sangrante, seca, tosse, cocar, planta, queimadura, crescendo, endurecida, vitiligo, cicatrizacao, clara, luvas, escurecida, insuficiencia, micose, picada, surgem, eritematoso, novo, pontos, tornozelo, assintomatica, definidos, dpoc, escuras, gestante, mamas, pruriginoso, recorrentes, verrucosas, actinica, aparicao, brancas, cafe, inverno, nodulos, rural, escabiose, neoplasia, coriza, eritemato, espinocelular, peito, ardenciase, auricular, avc, bordes, fossa, hiperpigmentada, linfonodomegalias, odinofagia, puntiformes, ungueal, crises, hcv, hipertenso, planas, recorrente, ressecamento, escoriacoes, hiperceratose, mellitus, mid, nadega, venosa, vermelhas, cicatriza, cisto, diabetica, distal, retroauricular, cronicas, esbranquicada, feridas, hiperpigmentadas, inseto, praia, acastanhadas, assintomaticas, bolhosas, ferida, crise, definidas, mal, molusco, penis, ulceracao, vermelha, arredondada, facial, progressao, progressivamente, quente, cicatriciais, doloroso, retirar, circular, cocadura, nodulares, pior, preocupada, regressao, reumatoide, rosacea, textura, voltam, claro, consistencia, eczema, encaminhar, escurecidas, fronte, indolores, pavilhao, plantas, quiro-

dactilo, reacao, remocao, testa, varizes, vascular, cardiopatia, elevados, emagrecimento, higido, palmas, ulceras, anorexia, artrose, barba, elevacao, incomodo, nuca, ombros, percebe, surgidas, amarelada, circulares, etilismo, extremidades, gluteos, herpes, intermitentes, cardiaca, cauterizacao, depressivo, desaparece, hialina, hiperceratoticas, inframamaria, permanece, pigmentada, procedimento, subito, alimentos, assimetrica, aumentaram, cabelos, doi, escapular, genital, limites, maiores, melanocitico, multiplos, orais, plano, rachaduras, perioral, persiste, poplitea, surgida, tumoracao, vegetante, vesicula, amarronzada, bolha, cocam, diabetico, enegrecida, escuro, estetico, extabagista, fungo, gastrite, labial, papular, papulosas, piorou, pododactilo, prostata, verrugosa, arredondadas, bolinhas, descamativo, dobras, hiperemiado, incomoda, irritacao, palpebras, pustulosas, ressecada, ulceradas, vermelhidao, vermelho, anemia, asperas, assintomatico, boca, capilar, cicatricial, elevado, epilepsia, eritematodescamativas, escamosas, exacerbacao, gradual, hepatica, hipotiroidismo, maculares, melasma, mudancas, orelhas, palmar, persistencia, sangram, sudorese, vesiculares, alergico, apresentam, atrito, aumentam, cresceu, hemiface, inflamacao, isquemica, msis, recidivantes, atopia, bordo, contornos, dores, impetigo, msss, odor, parietal, prurito, quadril, tirar, contorno, cubital, delimitados, digitopressao, espinhas, excisao, hiperceratotica, hipercolesterolemia, hiperpigmentacao, mamaria, manipulacao, pigmentacao, pitiriase, diabete, escurecimento, espalhadas, iniciadas, iniciando, periorbital, polegar, ponto, preta, tinea, ardor, arritmia, articulacoes, campo, cirurgica, confluentes, cores, espessamento, estetica, ficaram, fissura, fumante, glutea, piorando, recidivas, ulcerado, violacea, acamada, alimentacao, aparencia, areata, bochecha, calcaneo, cbc, comedoes, corporal, erisipela, espalhando, evoluiram, hipocromico, liquen, lupus, macular, mento, panturrilha, periferica, recidivante, secrecoes, tornozelos, verrucoso, versicolor, cimento, desaparecimento, desconhece, despigmentacao, estrias, exantema, flanco, foliculite, hipercromia, icc, labios, leito, narina, percebido, puntiforme, sobrancelha, vaginal, vulva, agrupadas, calcanhar, descama, desejo, drenam, espessura, exofitica, fratura, glaucoma, gravidez, interdigital, lisa, melanocitica, movel, nevus, nodulacao, observa, parcialmente, parto, seio, traumatismo, vulgar, zoster, acidente, agricultora, alergenos, amamentando, amareladas, aparecer, bilaterais, branco, calosidades, ceratotica, cicatrizes, cirurgiao, cirurgico, escamacao, escrotal, hipertrofica, hpv, machas, melanociticas, nascenca, novos, observou, palpaveis, panico, pediculada, perineo, quirodactilos, rompem, saiu, sangramentos, sangramentose, sanguinolenta, sobrancelhas, sulco, suor, temperatura, urticaria, violaceas, congenita, distribuidas, doem, escamas, escoriacao, estase, ferimento, formato, imunossupressor, mantendo, margens, massa, obesa, ocular, petequias, pododactilos, prurigo, pulsos, punhos, queixo, rarefacao, recentes, seios, verrugosas, virilhas, zigomatica, alterada, articular, asmatica, aspero, aumentar, ceratoticas, claras, crescer, crioterapia, desidrose, dolorida, eritematosos, fetido, fungica, generalizada, hiperplasia, malignidade, man-

dibular, maquiagem, nitrogenio, parestesia, purulento, pustula, rugosa, simetricas, temporaria, tumor, ungueais, voltaram, abandonou, ansiosa, antigas, atopias, calosidade, crostoso, dermatofitose, diarreia, efurix, erupcao, esternal, etilista, extensas, falange, glande, gluteo, impressiona, imunobiologicos, incomodam, inflamatoria, intertrigo, liquenificacao, maligna, melhorando, micropapulas, minima, minimo, mucosa, mudou, notado, ocasionalmente, ovalada, papulo, picadas, raiz, rosea, sebaceo, sensivel, simetrica, surge, xerose, alcoolismo, amarronzadas, arroxeada, ata, atrofia, avermelhado, axilares, barriga, borde, borracha, cheiro, cistos, clavicular, diminutas, disidrose, endurecidas, endurecido, epidermoide, escoriadas, esquizofrenia, extensora, fibromialgia, financeiras, fossas, fumo, heterogenea, insonia, marron, medo, metotrexato, nasceu, palpavel, parkinson, plurido, preocupacao, quentes, sangrantes, sequelas, subcutaneo, telangiectasias, varicela, abscesso, actinicas, agrotoxicos, ardem, assimetria, bactrim, brancos, caspa, cec, cefaleia, cessar, cicatrizou, ciclopirox, cigarros, coalescentes, corrimento, depilacao, descamam, desodorante, doloridas, dolorosos, drge, enxaqueca, esteatose, falha, hipercromico, hipocromia, inflamatorios, infraorbitaria, itu, lavoura, limao, mamilo, obeso, ocre, osteopenia, ouvido, palpebral, papulosa, pedunculada, pigmentadas, preauricular, preto, rash, reapareceram, retornaram, seborreia, sida, sumiram, surgido, tireoide, umbigo, varicosa, vascularizacao, vasos, visiveis, xerodermia, abscessos, aguda, aleitamento, alzheimer, arde, areola, arroxeadas, atividade, ativo, bolhosa, calcanhares, casquinha, comecam, cravo, definida, delimitado, dermica, desapareceram, descamacoes, detergente, discretas, enegrecidas, erosao, espalharam, espera, exereses, fetida, firme, flancos, flogose, hiperemicas, hipopigmentadas, infiltracao, insetos, isquemico, mandibula, manipulado, manipular, necrose, numerosas, orbitaria, peitoral, peniana, perfume, pioraram, pulmao, reumato, rins, rubor, sangrou, sensibilidadese, sono, superficiais, visual, abuso, acantose, acromica, agricultura, alimentar, amarronadas, anelar, angulo, antecubital, asmatico, assimetricas, bochechas, bolinha, calo, caroco, ceftriaxona, cicatrizam, cintura, cirrose, comissura, cosmeticos, costras, crescem, dermatites, descartar, desenvolveu, domestica, enxerto, escapula, esclerose, espalhou, esporadico, estagios, fisgadas, garganta, gestacional, grossa, infectada, inflamatorio, interdigitais, interglutea, lipoma, lombalgia, luva, luz, maleolar, maleolo, manipulada, mudando, oncologia, orificio, osteoartrose, periocular, picadura, pintas, plantares, pustulosa, reaparecem, recorrencia, resistente, ressecadas, sintomatica, surgir, tibial, ulceracoes, ulcerosa, urgencia, vegetantes, vulvar, alergicas, algia, amarronada, amolecida, amoxi, ansioso, ativas, atuais, autoestima, bexiga, castanho, ceratoses, cresce, crosticulas, deformidade, eczematosas, emocional, equimoses, erupcoes, esparsas, espessa, expandindo, fisgada, foliculos, friccao, genitais, hematoma, hiperlipidemia, hiperqueratose, hipertireoidismo, hipertroficas, inchaco, infeccioso, inumeros, ipsilateral, isoladas, liquenificada, lisas, machuca, maleolos, marcas, margem, medicamentoso, melanose, micoses, mosquito,

negra, observar, oleosa, parestesias, periorbitaria, perolada, perolado, plastica, prepucio, progredindo, progressivas, prostatica, protese, psiquiatra, quebradicas, queratose, residual, retornam, retornando, retracao, rosa, roxa, sacra, secam, seguem, seroso, subcutanea, supraclavicular, supraprubica, supurativa, surgindo, suspeito, telangectasias, termica, tinta, vagina, vesiculosas, vomitos, vulnerabilidade, acrocordons, acromicas, alcoolica, alergeno, amarelado, amarelo, angina, aparentes, aranha, articulares, atipica, azul, cachorro, cardiopata, cardiovascular, cascas, catarata, cavidade, celulite, chagas, cilios, circunscrita, cistica, cocando, cocera, coco, condiloma, congenito, constitucionais, continuam, crecimento, difuso, diminui, disseminacao, endometriose, endurecimento, eritematoescamosas, esbranquicado, extremidade, fios, flexora, formigamento, furunculos, gatos, hematica, hematomas, hemitorax, hialino, hipersensibilidade, hipertensivos, hipertrigliceridemia, hipocondrio, hipotireodismo, incomodando, infecao, infeccoes, infraclavicular, irregularidade, lentigo, linfoma, negou, nodulacoes, normocromica, ocipital, panturrilhas, paroniquia, pelvica, periungueal, permanecem, plaquetopenia, pruridos, pruriginosos, reaparecimento, regridem, submandibular, torso, abaulamento, abcessos, acrocordon, adenopatias, aderida, agudizacao, alergista, alimentares, alimento, amarela, anus, articulacao, artralgias, assimetrico, aumentados, bariatrica, brilhante, brilhosa, cantoplastia, castanhas, cauterizou, cloasma, comencou, cravos, crio, cubitais, descamada, descamativos, descasca, despigmentadas, diarista, diminuir, discoide, dolorido, dormir, drena, drenado, ecodoppler, eritematopapulosas, esfoliacao, femur, fotoexposicao, fototerapia, frente, fronto, furunculose, generalizado, hipercromicos, hipopigmentacao, homogenea, incomodada, infiltrado, infraorbital, irregularidades, irritantes, linfonodomegalia, maculopapulares, malares, manicure, marcapasso, marrons, meias, narinas, peroladas, policromatica, polpa, pontilhado, popliteas, profunda, progrediram, psiquiatrico, puerpera, purpura, rachadura, resseccao, rosada, sacral, sangrando, satelites, subcutaneos, sulfametoxazol, supercilio, tabaco, tacrolimo, talco, tempora, tintas, tintura, transitoria, trombose, tvp, umidas, verrucoide, abrupto, adenopatia, agudo, amamentacao, amarelas, amigdalite, amox, autoimunes, avci, azulada, bacteriana, benigno, caes, calos, cama, candida, catapora, cavalo, ceratosica, cicatrizar, circinadas, cirurgicamente, clavicula, cosera, costal, costrosas, cozinheira, cranio, cresceram, crescido, cuidadora, dermatoscopia, descamcao, descolamento, disseminado, drenando, edemaciado, enegrecido, erosoes, escalpo, escamativa, escamativas, escroto, espessadas, espremer, exantematicas, exsudacao, exsudativa, extenso, extracao, falanges, faxineira, flexuras, fotodano, fotoexposta, fungicas, furunculo, generalizadas, granuloma, hbv, hemorragias, hidradenite, hipercronica, hiperemica, hiperidrose, hipertrofia, hipocoradas, hipoestesia, hipotireidismo, imprecisos, indolora, infectadas, inflamatorias, irradia, latex, mamilar, mamilos, melicericas, mentoniana, muscular, obscuras, occiptal, onicolise, orvalho, papuloeritematosas, paralisia, pediosos, pelve, perianal, periumbilical,

permanencia, pescador, pintura, piorar, pioras, pioria, pioro, poeira, populosas, procedimentos, proeminencia, profundas, pustulares, queloides, radio, reacoes, rebordo, recidivando, redondeada, refrataria, repetido, romperam, rompimento, safena, sarna, simetricos, sobreelevadas, tireoidectomia, trocanter, tuberculose, umbilical, unguel, vascularizada, veia, veneno, venoso, verrugoso, vesicular, acamado, acastanhado, acneicas, acneiformes, afastar, agravamento, alastrando, amamenta, amoxa, amputacao, antiga, anulares, apetite, arranhadura, auriculares, bolhoso, carcinomas, cardiomegalia, cauterizada, cesarea, cicatrizada, cigarro, circundante, circunscritas, clav, constrangimento, costrosa, crescente, debridamento, degenerativas, deltoide, dentes, depressivos, desaparecendo, desapareceu, descamtivas, desenvolvimento, despigmentada, difusamente, digitais, digital, diseminadas, disseminada, disseminados, disseminaram, eczematosa, endurada, enegrecidos, eritematocrostosa, eritematopapulares, esbranquecidas, escama, escamosa, escarificadas, esclerodermia, esclerodermiforme, escurecidos, escuros, espalhandose, espalharamse, espinho, espontaneos, esporao, espremeu, estende, estendeu, estoura, esverdeada, exofiticas, exsudato, faxina, febril, febris, fedorenta, flexoras, flogistico, flutuacao, foliculo, fragilidade, genitalia, granulacao, hematicas, hemiparesia, hemorragia, hemorragico, hemorroidas, hiperemias, hiperpigmentado, infecciosos, infectado, infiltrada, inflamacoes, inflamada, inguinais, insiste, interescapular, irregulais, irregularmente, limitada, linfonodos, litiase, lombociatalgia, maconha, maculopapular, maligno, mamario, manipula, manuseia, mastectomia, matriz, mebros, mecanica, melanocedticos, metatarso, micofenolato, migrans, musculares, neoplasias, normocromicas, operado, operou, orbita, osteomielite, ovaladas, palmares, papulopustulas, paranasal, pedunculadas, percepcao, perfurante, periauricular, peribucal, perinasal, pigmentado, pilosos, polegares, psoriasica, pubiana, pubica, pulso, purido, purpuras, queimaduras, redonda, refratario, ressurgimento, retinopatia, reto, reumatica, reumatologia, rigidez, sintomaticas, sintomaticos, sublingual, supraciliar, tonalidades, toracico, tornamse, tornando, tornase, tornou, toxoplasmose, ulcerosas, umbilicacao, umbilicada, uretral, varicosas, vasculares, vegetacao, vela, vesiculosa, violaceo, virais, vitropressao, vomito, vulgares, xantelasma, zigomatico, zileri, abcesso, abdomem, acastanhas, acneica, agricola, agricolas, agrotoxico, alcoolatra, alicate, amorolfina, amoxiclav, amoxiclavulanato, androgenica, antibiotica, ardentia, arredondado, arrendondadas, assadura, atopico, avermelhamento, basocelulares, bichos, bordoes, brancoamareladas, broncoespasmo, bronquica, calcificacoes, calvice, capitis, cardiaco, caspas, casquinhas, ceborreica, ceftriaxone, celiaca, cervicais, circinada, circundada, cirugia, cirurgicas, cirurgicos, clamidia, clavulim, crecendo, crostras, crostrosas, crural, depressiva, descamadas, descamando, descamar, descamarias, drenou, dst, dsts, efelides, endurecidos, enfisema, engrossada, enjoo, eritematocrostosas, escoriar, escureceu, escurecido, esfacelo, espesso, espontaneas, esporadica, esporotricose, esteticas, estomatite, evolutiva, exacerba-

coes, expandiu, extende, exuberante, ferimentos, fibroelastico, fibrose, fincadas, fisuras, flexao, flogisticas, fogachos, fraqueza, grossas, grosso, insidiosa, insidioso, intensificacao, intercostal, interfalangeanas, intergluteo, intermamilar, intermitentemente, irradiando, irritabilidade, irritada, irritavel, labiais, laringe, larva, latejamento, leucodermicas, leuconiquia, limitadas, lineares, linfadenomegalias, linfonodo, linhas, liquenificadas, lombossacra, machucar, maculopapulosas, melanocitos, melanomas, melasmas, melito, melitus, micoticas, micoticos, microvesiculas, miiase, miomas, miomatose, modificou, moles, molhado, momentanea, morena, morfologia, muda, mudar, mudaram, multipla, multiplicacao, nasais, nascer, nasogeniano, nasolabial, nauseas, nodoso, nucal, numulares, operacao, orbital, oval, papulopustular, papulopustulosa, paraceratose, perifericas, perilabial, perioculares, perivascular, pigmentados, pilar, piloso, piodermite, popular, populosa, porfiria, precaria, preocupados, pretibial, proliferacao, proximais, psoriases, psoriasiforme, psoriatica, queilite, queima, queimada, queimor, raspados, raspar, rastreamento, reaparece, reaparecer, reativa, reavaliado, recidivam, recorrencias, reduzir, regrediram, regrediu, regride, repentino, repetitivo, repetiu, reposta, repouso, resgate, residuais, resolve, respondeu, restrita, retroauriculares, retroesternal, reumatologista, revascularizacao, rinites, roseas, roseo, saliente, sanguineos, sanguinolento, sardas, sebacea, seborreico, serosanguinolenta, soropositiva, subjetivo, teleangectasias, toracoabdominal, translucido, trapezio, tumoral, tumores, ulcerara, urticariformes, vegetativas, vergonha, vermelhos, vertigem, vesicobolhosas, volumosas, 5fluoruracila, abrupta, acentuacao, acentuado, acuminado, adenocarcinoma, adenoma, adenomegalia, adere, aderencia, aderidos, adiantando, adiantou, agravando, albinismo, alcoolicas, alergicos, alergologista, alisamento, aliviam, aliviaram, alodinia, alopecicas, alopesia, anabolizantes, anel, anestesia, anestesico, angioedema, angioma, angustiada, anogenital, antialergica, antigamente, aredondadas, areolar, areolas, arnica, aroeira, arranhar, arrendondada, arteria, assintomaticos, aureola, bicho, blastoconideos, bocelada, brilhosas, brinco, cafalexina, calcaneos, calcar, cansaco, ceratoacantoma, ceratosicas, ceratotico, cerume, cicatrizadas, cicatrizado, ciclicas, cocado, cocaduras, coccar, coccigea, colageno, colarete, confluencia, contraindicacao, costelas, cozeira, cozera, cronicamente, cronificado, crostosos, crostrosa, deltoidea, dermatitie, dermatofitos, dermatomo, descamatica, descamtiva, dispareunia, eczemas, eczematoso, edemaciada, edematosa, episodica, equimoticas, erecao, eritematodescamativaspruriginosas, eritematodescamativo, eritematoescamosa, eritematopruriginosa, eritematopruriginosas, eritematovesiculares, eritemstosas, eritrmatosa, eritrmatosas, eritrodescamativas, erosada, eruptivas, escabicida, escamadas, escamoso, escapulares, escara, escleroatrofico, escurecendo, esferica, espessas, espiculada, espinhos, espino, estora, estoram, estouradas, estranha, estressada, estressado, estressores, esverdeadas, evolutivas, evolutivo, evoucao, exacerbada, exacerbados, exacerbar, exalcoolista, excisional, excluir, exofitico, expaliando, expande,

expandem, expandiram, expansiva, extendendo, extendendose, extravasamento, exudacao, exudado, faces, faciais, farmaco, farmacodermia, farmacologicos, farmacos, fasciite, fedor, femoral, fistulas, flexura, folicular, formiga, fotoexpostas, fotossensibilidade, fragrancias, fumando, gangrenosa, ginecomastia, granulomas, granulomatoso, granulos, granuloso, grossos, hallux, hemangiomas, hemiabdome, hemiplegia, hemorroidaria, henna, hernias, herpeszoster, herpetica, heterogeneas, heterogeneidade, heterogeneo, hipercomicas, hipercromias, hiperemidas, hiperglicemia, hipermiada, hiperplasica, hiperqueratinizacao, hiperqueratinizada, hiperqueratosica, hiperqueratosicas, hiperqueratosis, hipertensiva, hipertermia, hipertoficas, hipertrofico, hipocorado, hipocromaticas, hipocromicaa, hipocromicos, hipocronicas, hipogastrica, hipogastrio, hipogastro, hipoglicemia, hipoplasicas, hipotireoidea, hipovitaminose, inchada, inchado, infecciosa, infiltradas, inflama, inflamadas, inflamado, infraaxilar, infraescapular, inframamario, infrapalpebral, infraumbilical, ingerido, inguinocrural, insectos, interdigitos, interfalangiana, intermamario, introito, invertida, joanete, lacrimal, lactacao, leishmaniose, leucopenia, leveduras, linfadenopatias, linfondomegalia, linfonodomegaliasse, liso, maceracao, macerado, machucando, macia, maculo, maculosa, malignizar, mamilares, manchar, manchinhas, manipuladas, manuseio, maquiagens, marrao, marromescuro, mastite, maxilares, maxilas, metacarpofalangeana, metastase, micologica, micotico, micropapular, microscopica, migracao, migranea, migratoria, migratorias, milium, minimos, mioma, mitotico, mmiis, mmis, modificado, modificando, modulacao, modulares, mucosas, multicolor, multiformes, musculo, nacarada, nacimento, naopruriginosa, naris, neoformacao, nervo, nervos, nervosa, nervoso, neuralgia, neurologica, nevicas, nevralgia, nitidamente, nodularidade, oncomicose, onicorrexe, palido, palmoplantar, pantorrilha, papilomatose, papulocostrosas, papuloeritematosa, papuloescamosas, papulonodulares, papulopustulares, papulovesiculares, parasitas, pardas, pardo, parkinsonismo, parotida, pediculado, pedunculado, pendulada, penfigo, peniano, perfusao, periareolar, perineal, periorbicular, peritonite, periungueais, pigmentacoes, pilonidal, poliartralgia, polineuropatia, polipoide, postero, posterosuperior, pretas, pruginosas, prurginosas, pruriginoas, psicologo, psicotico, psicoticos, psiquiatricas, psoriaticas, pulgas, purpuricas, purulentas, pustuloso, queimado, queimante, queimar, queimou, queratosica, racha, racham, recidivar, recidivou, refratarias, relevos, remove, removeu, removida, ressecam, ressecao, retalho, reticular, reticulocitos, retomaram, retornara, reumaticas, reumatismo, reumatologicas, rompemse, rompida, rosadas, rubi, saborreica, sagramento, saliencia, sanguinea, sanguineo, sanguinolentas, sebaceos, sebo, seborreicas, secando, secura, sedentario, senbilidade, sepse, serohematica, seropurulenta, serpiginosas, siflis, sobreelevados, sobrelevada, sobrelevadas, sobrepostas, subcostal, subcutaneas, subita, subitamente, subjetiva, subjetivopaciente, submamarias, submentoniana, suicida, suicidio, sujo, supralabial, surgimentos, tardiamente, telangiectasia, teleangiectasias, tendinobursite, tendi-

nopatia, tomografia, translucidas, transmissao, tristeza, tumefacao, tumoracoes, tunel, tunelizadas, ulceram, ulcerativas, ulcerou, ulna, umbilicadas, unilateral, uretra, urticarias, varicosidades, versiculares, veruga, vesiculo, volumosa, volumoso, hipotrofica, hiprocomica e hispertensao.

APÊNDICE E — *WORDLIST*: LISTA DE *TOKENS* DE CONDUTA VÁLIDOS

Tokens válidos: perfume, prescrever, gercon, fotoprotecao, acido, camada, limitacao, fina, esponjas, mometasona, benignidade, girassol, micologico, hidroxizine, protocolos, banhos, cetoconazol, lavagem, acompanhamento, maos, aferir, reforcar, cuidados, discussao, irritacao, antihistaminicos, algodao, dermatoscopia, adversos, excesso, questionar, pequena, clobetasol, brevidade, miconazol, reavaliacao, couro, xampu, observar, conforme, afastar, rediscutir, cor, risco, cremes, salicilico, excessiva, dexclorfeniramina, locao, corticoide, modificacao, hidratante, cefalexina, continuo, desencadeantes, exames, luvas, perfumes, cronicidade, cronico, unhas, suspender, furoato, plantas, sabonete, procedimento, lavar, coleta, somente, prioridade, urgencia, confirmacao, priorizar, recidivas, ardencia, neoplasica, ureia, terbinafina, atrofia, metais, corticoides, ligar, buchas, recidivante, recorrente, fungos, higiene, borracha, exerese, propionato, remocao, manutencao, repetir, sinteticas, fragrancias, cronica, alimentos, informar, cremosa, chapeu, excessivo, rotina, cuidado, dexametasona, neutros, sistemico, acidos, dimeticone, suspeitas, suspensao, abrasivos, retirar, teste, unha, biopsia, raspado, perfumados, peso, amaciantes, bone, cancer, fotoprotetor, hemograma, familiar, protetor, benigna, perilesional, rigorosa, especialidade, toalhas, cronicas, explicar, curativo, cosmeticos, manipular, secundaria, doxiciclina, ceramidas, cultural, antibiotico, exacerbacao, diario, atopica, neutro, regulasus, cultura, uva, estetico, patologia, cobrir, extraseca, crescimento, vdrl, itraconazol, calcados, plastico, benignas, compressas, desconforto, elevacao, esparadrapo, chapeus, meias, antihistaminico, micropore, efurix, fluconazol, protecao, resolucao, suspeita, aumento, secretaria, vinilicas, fisico, lixar, temperatura, retirada, alcool, atividade, posinflamatoria, sifilis, lixa, protegida, queimadura, quente, significativa, atopia, estresse, benzoila, peroxido, sol, filme, capilar, delicadamente, free, oil, tetraciclina, aumentar, examinar, investigar, aparecimento, edema, hiperpigmentacao, obesidade, venosa, investigacao, leite, gradual, oclusao, sabao, semente, diminuicao, exercicios, seguimento, espontaneamente, excluir, meg, modificado, anamnese, antifungico, nivea, antitireoglobulina, antitpo, mantendo, residual, sorologias, diferencial, potencia, trauma, dermoscopy, verificar, mupirocina, enxague, vaselina, limpeza, imerso, transaminases, ansiedade, telefonico, exclusao, telecondutas, irritantes, loratadina, taquifilaxia, repigmentacao, tabagismo, adapaleno, emolientes, sobrepeso, atrito, elasticas, autoimunes, vasenol, enxaguar, gestacao, benigno, machucar, recorrencia, espontanea, fungica, lcd, acompanhar, arterial, seborreica, 5fluorouracil, contraindicado, alvejantes, familiares, manchar, elucidacao, hidroquinona, lenta, reducao, gestantes, ressecamento, hepaticos, revisar, hepatite, pulsos, azelaico, coletar, hepatica, empirico, tretinoina, ungueal, dobras, infeccoes, persistir, remover, confortaveis, contraindicada, detergentes, fronhas, estimular, regredir, amaciante, asma,

ceratose, desonida, inflamatorio, viral, quentes, renal, afinamento, cama, delicada, prednisona, regulacao, seco, autolimitado, bones, hiv, morno, prolongar, psoriatica, troca, dexpantenol, diametros, melanoma, preventiva, reumatologista, tintas, dactilite, duofilm, elastica, ferro, sensibilidade, tenossinovite, tolerancia, discutir, emocional, amendoas, metronidazol, tranquilizar, articulacoes, bordas, certificarse, cotonete, mornas, sedativo, entesite, regressao, agravantes, arteriais, doces, oligoarticular, palpaveis, plaquetas, salsicha, autorizacao, mineral, neoplasia, remissao, suaves, 5fluoruracila, descartar, esmalte, frio, ivermectina, prometazina, minimizar, olhos, repelentes, subungueal, tireoide, comidas, corticoterapia, domiciliar, gentamicina, hidroxido, seca, tgo, antibioticos, autolimitada, compressiva, diproprionato, extrasseca, regressivo, rinite, sintomatico, tgp, apimentadas, barreira, glicemia, oncologia, verrux, desencadeante, lactico, sedativos, alergica, consumo, encaminhada, especialista, hanseniase, substancias, traumas, absorcao, hipercromia, hipocromia, insetos, sulfato, ciclopirox, equ, erosao, friccao, gaze, clindamicina, fluocinolona, perguntas, queda, roupa, tacrolimo, triagem, antihcv, antihiv, cirurgia, curativos, deseje, material, solventes, antibioticoterapia, emocionais, ferritina, infiltracao, intenso, progressao, tecidos, adesao, ambientais, completamente, resultados, secundarias, ata, certeza, detalhada, expostas, inverno, sensiveis, zinco, aciclovir, posterior, afastarse, aquosa, hidroxizina, age, anticoncepcional, esfoliantes, gradualmente, hidroalcoolica, ingestao, laboratoriais, lixamento, metal, mosquiteiros, repilacao, seguir, virais, calor, escabiose, fuorato, janelas, lavadas, ocular, pesquisar, tecido, termica, vsg, compressao, consiga, imediatamente, irritativa, otimizado, definicao, diametro, neuropatia, pedis, sapatos, sindrome, telas, tireoideopatia, tricloracetico, hidrocortisona, minoxidil, tacrolimus, antitireoperoxidase, arejamento, dieta, drogas, esmaltes, hormonais, malignidade, odor, prolongada, verao, adstringentes, antifungicos, assintomaticas, aumenta, cheiro, esfoliacao, fisiologico, soro, cimento, cirurgiao, creatinina, graxos, interacoes, lamina, molhar, neutrogena, depressao, ensaboarse, medicamentosas, regularmente, ambiente, cessacao, cirurgica, controlada, crioterapia, frutas, hipertricose, previo, recorrencias, tcm, ulceracao, alergia, bebidas, contactantes, distancia, engravidar, intervalo, laminas, restringir, testes, anatomopatologico, antiinflamatorios, clorexidine, comorbidades, desaparecer, descamativas, emoliente, espuma, falha, hidratada, imunossupressao, pressao, prognostico, renais, amplo, anemia, anteriores, continuamente, depilacao, escoriar, estase, fan, intralesional, laboratorial, oleosidade, palpacao, satisfatoria, solares, colaterais, comparacao, fotoexpostas, gelado, oxido, umidade, vigorosa, esclarecimento, hipertensao, aderencia, alternativa, camomila, cardiaca, chinelos, confirmar, contaminar, corretamente, distribuicao, domestico, esteticos, exato, fototipos, fungo, gastrointestinais, olamina, paracetamol, previamente, procedimentos, queimacao, secrecao, silicone, sinteticos, tracao, ungueais, camiseta, cha, cicatriz, desodorantes, despigmentacao, doi, eliminar, fusidico, gentil,

interacao, localmente, nitrilicas, orelhas, pequenas, previos, revisao, saboes, sintoma, umidas, vascular, agravamento, artralgia, causador, cicatricial, dersani, epf, idosos, leve, leves, motivos, otimizar, proxima, questionamentos, acompanhados, despigmentante, glicose, herpes, imunocompetentes, inseticidas, negativos, prevencao, previas, surgir, tetraciclinas, transmissao, venha, alba, aparelhos, assintomaticos, cauterizacao, clotrimazol, colirio, controlar, cuticulas, descolonizacao, despigmentantes, diferenciacao, ftaabs, hepatopatia, regula, sanitaria, secos, apresentem, benzatina, cavidade, ciproterona, cobertores, cortar, disseminacao, domesticas, efetivo, escamas, lavados, localizadas, mangas, meia, ocorrem, palmoplantar, plastica, progressivo, proximo, rotulo, significativo, tesoura, thiersch, vinil, cobertura, domesticos, dura, eletrocoagulacao, fotoexposicao, fotossensibilidade, funciona, ggt, gravidade, inflamatorias, maligna, maligno, maquina, neoplasias, quimica, reduz, xarope, alcalina, apertados, cilios, cirurgico, definir, dentes, dislipidemia, ecografia, fosfatase, grave, hirsutismo, isotretinoina, longa, maquiagem, medir, permanganato, quais, realizadas, sandalias, teleangiectasias, actinicas, apertadas, indeterminado, molusco, opioides, primaria, resinas, sonolencia, sudorese, acetonida, amorolfina, ataduras, blusas, consecutivos, elevados, estetica, historico, ingesta, mellitus, ocluido, orificio, pruriginosas, radiacao, refeicao, sintomatologia, teledermatologia, b12, bilirrubinas, borda, camisetas, comedogenicos, contagioso, detalhes, dexclorferinamina, dificuldade, discreto, fotodano, graves, hbsag, hepatites, hospital, maquiagens, profilatico, proteger, raiox, rapida, recorrentes, recorrer, rediscutido, reinfeccao, temporariamente, vitamina, 5fluoracil, cessar, cicatriciais, eletroforese, emagrecimento, episodio, hepatopatas, semelhante, subsequente, tintura, xerose, actinica, aderidas, aderidos, alcoolica, anteriormente, autorizado, bepantol, blefarite, calazio, capitis, capsaicina, complementares, conjuntival, cosmetica, extenso, gradativamente, hordeolo, irritativos, limpar, litio, longas, margem, nascimento, oftalmica, oftalmologista, palpar, papel, passados, picadas, protetores, psiquiatrica, retinoides, sangramento, sodio, suave, surgindo, titulos, zileri, 2012protocolo, acantose, ampla, anticoagulantes, antinflamatorios, atualizado, bacterianas, candida, chamada, cirurgias, compressivas, contraindicacao, crise, efetivos, emergencia, exsudacao, griseofulvina, ibuprofeno, limao, localizada, lubrificante, medicamentosa, persistente, pesquisando, sintomaticos, anestesico, bicarbonato, coabitantes, farmaco, fungicas, oleosos, rebote, recidivar, retinoico, reversivel, ambientes, animais, centro, citricos, deficiencia, drospirenona, equimoses, fragrancia, friccionar, granuloma, linfonodais, lipidico, lupus, mantenha, melhorarem, molhadas, mucosa, neomicina, niquel, psicologico, rapidamente, sexuais, treponemico, treponemicos, assepsia, comportamento, condilomas, corporis, desenvolver, ftabs, habitos, lidocaina, prevenir, pruriginosa, resistencia, retornar, seguindo, suadas, calamina, cancelado, caneta, escola, codeina, desencadeando, detalhadas, dipirona, disidrose, domiciliares, efelides, efluvio, fitofotodermatite, fraldas, funcional, horizontal,

lateral, linear, palmares, papaina, pcr, pediculose, pendente, recontaminacao, refratario, revendo, solicite, sorologia, telogeno, tireoidopatia, volateis, abrasivo, antiga, antiviral, aureus, bergamotas, betabloqueadores, bronzeamento, cetirizina, ciclos, cisto, condicionador, contraindicacoes, crescer, curetagem, dependendo, desencadeadores, escoriacoes, espessura, estender, esteril, estressores, exposicoes, falencia, fototoxica, frias, gastrointestinal, geladas, hemostasia, hepaticas, limoes, linfadenomegalias, matriz, minimizamse, mornos, pitting, tabagista, acrocordons, androgenetica, anticoncepcionais, antimalaricos, bizarras, cadeiras, calvicie, capilares, captopril, celulite, cimentos, cloridrato, clormadinona, cocadura, colodio, complementar, complicacoes, comportamentais, confirmada, furocumarinas, glicada, agendamento, alergeno, alho, antiandrogenico, anticonvulsivantes, artificial, azitromicina, caminhar, cervicais, clareadores, contaminacao, cores, dermatofitos, desloratadina, detergente, hemoglobina, hidroclorotiazida, hipotireoidismo, insulinica, interglutea, involucao, latex, mencionar, mostrando, mudar, nauseas, nodulos, pigmento, predisponentes, questionado, serotonina, sintetico, sintomatica, suficiente, surge, tiverem, transformacao, trocando, ultrapassando, umedecer, vasculite, vernizes, vestir, zoster, acuracia, afazeres, alergenos, aluminio, cafecomleite, cebola, cenoura, clareamento, cloreto, disseminamse, ecodoppler, extratos, fotoprotetores, fotossensibilizante, fps50, hidratar, hiperidrose, imunodeprimidos, incidencia, ingerir, lactato, plasticos, refratarios, residuais, sulfametoxazol, tireoidopatias, tolere, ultravioleta, urinaria, uteis, varrer, virilhas, amonio, amortecimento, antibacterianos, assimetria, atingida, atingido, claudicacao, coceira, erradicacao, escoriacao, esteroides, filtros, flushing, hemangioma, levocetirizina, luva, oculares, olho, passado, pergunta, photodermatoses, puberdade, reacoes, reavaliando, removedores, repouso, resolve, sanitario, superficies, tireoidiana, triglicerideos, trocas, ubs, vir, visualizar, afastamento, alimentacao, anafilaxia, anca, antidepressivos, aquecidas, arvores, bloqueadores, calcinhas, colagenase, congenitos, contagiosum, continue, corticosteroides, dermatites, dermatofitoses, desaparecendo, desencadeiam, epidermoide, esfregar, especializada, estimulo, evita, excessivamente, favorecam, fototerapia, fps30, genetica, glicemico, infecciosos, informada, irritante, limpo, manipule, mantemse, moldura, naoandrogenico, notado, optativo, ovarios, paroniquia, pefume, permanecerem, pigmentar, poiquilodermia, policisticos, ponderal, pseudomonas, reavaliarmos, reservatorios, resolverse, rigorsa, sexualmente, simetria, similar, suplementacao, suporte, surgiram, tornem, totalmente, transpiracao, travesseiros, vasculites, 5fluoruracil, aas, adenomegalias, albendazol, amoxicilina, analisar, arvore, assento, autoimune, balanceada, barbear, calorica, carreadores, cellulitis, colesterol, comprida, corticoesteroide, crescendo, crioglobulinas, cuticula, descontaminacao, detalhado, dexclofeniramina, diluicao, domicilio, eletrocauterizacao, envelhecimento, esteroidais, etilismo, exposta, fragilidade, fralda, fungico, hidrantes, hiperceratotica, hipertireoidismo, histopatologica, hormonal, hospe-

deiro, idiopatica, idiopatico, inibidor, inseto, irregulares, koeber, laboratorio, largas, lavagens, necessitamos, neurofibromatose, pinca, ponteira, posparto, possua, posteriormente, procurando, procure, prodromo, progestagenos, propiltiuracil, protegidas, proteica, provas, prurigo, purpuricas, reavaliacoes, refratariedade, removida, restabelecer, restritiva, reumatoide, sacral, sangue, temperaturas, tracionamse, urina, utilizou, vaso, acontecer, afastado, alternativos, amamentando, amareladas, amendoim, andar, angioedema, antivirais, aprovados, articulares, aspereza, atadura, atingir, bilateral, bisturi, brincos, bronzeada, bsa, cerato, contraindicados, contrario, correcao, creche, decubito, definitivo, deformidades, dermatoscopica, desbridamento, descolamento, desidrose, deslocamento, difusa, diminuam, diminuindo, enalapril, encontra, encontrado, envolvimento, eosinofilia, epidemiologia, erosadas, espacadamente, especializado, especificamente, especificas, esperada, esporadico, esteticamente, exacerbacoes, excisada, expectante, exsudativas, fenitoina, fezes, grao, gravidez, haluces, has, hidrogel, higienico, higienizar, hipoalergenico, histologica, imunomediado, incerto, isoniazida, lesao2, leucodermia, maldefinidas, maopeboca, multifatorial, nefropatia, nodular, oftalmologia, ortopedicas, parasitarias, tinea, precedido, proeminente, progestageno, promover, provocar, punhos, regridem, repeticao, resistente, responsivos, ruptura, saliva, sedacao, sinvastatina, staphylococcus, suscetibilidade, suspenda, tireoidianas, tpo, tranquilizacao, traumatizar, triciclicos, trigo, umarediscussao, agulha, alcalis, alergicas, alimento, anestesia, anotar, antihbs, aparadas, aslo, barbituricos, cadeias, centrifugo, demencia, dermoscopic, desapareceram, descamativa, descartaveis, descontinuar, desejem, desencadeada, diluir, discoide, eritematoso, esclarecer, estagio, exuberantes, fadiga, fluoracil, formando, genetico, genitalia, glutea, gluteos, gordurosos, gratuitamente, gutata, habitual, hexahidratado, hidratando, hidrocistoma, hiperandrogenismo, histopatologico, icaridina, impermeavel, impetiginacao, inesteticas, interrupcao, irregularidade, irritativas, labial, laticinios, limpas, linfonodos, malassezia, manicure, mantoux, mastocitos, metformina, migratorias, milk, nodulo, numular, nutricional, objetos, ocupacionais, oftalmologica, organica, parceria, parcerias, parto, paucibacilar, penicilinas, penis, pentes, periferico, perifericos, perioral, pioram, piscina, primarias, profilaxia, progressivamente, propriedades, psoriasedevido, psoriasiformes, purpura, queilite, quirodactilo, recipientes, recreacionais, scabies, seguranca, seguro, sessoes, surgirem, surjam, suspensa, tireoglobulina, toalha, toxicidade, traumatizadas, trimetoprim, umidos, uvb, varicela, veterinario, violaceas, vitais, absorvente, abuso, aerossois, afirmativo, agendada, alopecias, anabolizantes, anestesica, animal, anotando, antibiotica, antihipertensivo, antissepticos, aparar, assintomatico, betabloqueador, biotina, boricada, branco, brasil, brasilia, calcipotriol, cautela, clavulanato, comida, condiloma, confirme, consistencia, contenha, continuado, corticoesteroides, cosmetico, cronicamente, cultivo, depilatorio, dermatofitose, dermatoneurologico, dieteticas, dimensoes, domestica, drosperinona, eczemas, eletrocauterio, en-

caminhadas, enxofre, eruptivos, escama, espatula, especificar, eter, farmacodermia, flexivel, forro, fotografar, fotossensibilizantes, fungicos, geleia, glandulas, granulacao, hcv, hdl, hidradenite, hidroxicloroquina, higienizacao, hipocromica, hipomelanose, imediata, imediato, imperceptiveis, impetiginizacao, incerteza, incomodada, infecciosa, insulina, interromper, intertriginosas, intestinal, intramuscular, irregular, kalloplast, laser, lavado, lavalas, levemente, limitada, limpa, lingua, lojas, malformacao, malignidades, maquinas, mariscos, melanociticas, metoprolol, movimentos, nadega, naproxeno, narcoticos, neurite, neurologicas, neurologicos, nigricans, nodulares, notalgia, notarmos, obras, oftalmologico, oleaginosas, oleosas, palpebra, papulosas, periorbital, permanece, persista, piogenico, piorem, popliteas, preservativos, psiquiatricas, psiquiatrico, psiquiatricos, reavaliado, recomendado, recontaminar, repetitivo, residuos, respiratorios, rubis, sangram, shampoo, sintomaticamente, social, telediagnositco, tireoperoxidase, tolerados, trabalhar, transitorio, tricloroacetico, trimetropim, tumor, umedecidos, umida, umido, varicoses, vasoconstritor, veias, 5fluoracila, acompanharmos, acuminado, adenocarcinoma, adjacentes, alantoina, albumina, alergico, alicates, alimentares, aliviar, ambiental, amitriptilina, angiomas, anorexia, antigas, antigenos, antihbc, antitermicos, anualmente, atipicos, auxilia, axilares, bacitracina, borrachas, botoes, calosidade, calosidades, carbamazepina, ceratoliticos, ceratotica, cigarro, clara, clarear, claritromicina, colar, coletada, composicaolimpeza, condimentados, congenito, conjuntivo, consecutivo, contralateral, contribuindo, controverso, criocirurgia, dapsona, deficiencias, delimitadas, demorada, desagradaveis, desaparecam, desaparecerem, desaparecido, descolorir, descontinuacao, desodorante, despigmentoso, detalhados, diane, diprosalic, discrepancias, diureticos, dobra, dolorosas, dominante, eczematosa, elastico, elementar, elevar, encaminha, endurecida, envolvido, enzima, epidermico, escamosos, escura, esfregaco, espinocelular, espondilite, estancar, estimulos, estrongiloides, evoluiram, exacerbar, examinador, excluida, extremas, faces, farmacologicas, ferroso, fibrose, fibroticos, flexoras, folicular, generosa, gestacional, glicolico, global, gougerot, hepatico, hipersensibilidade, hiponiquio, hobbies, idoso, imunocompetente, imunofixacao, inespecificas, infeccioso, infectada, infectadas, insuficienciavenosa, investigacoes, involvement, irritativo, larva, laterais, lavando, lentigos, leucopenia, limitacoes, loceryl, macrolideo, maduras, mama, mamaria, maneiras, metotrexato, mialgia, minociclina, necrose, necrotico, nervoso, neurologico, neuropatica, nistatina, nitido, nitidus, omeprazol, oncologico, onicolise, onicomadese, ota, otorrinolaringologista, penteado, penteados, perceber, perguntar, periodicamente, permanecem, permanecendo, perpetuacao, persiste, pilosos, pinturas, piolhos, plaquetopenia, precipitantes, profilatica, profundos, puericultura, recidivantes, rediscutindo, repetidamente, repetitiva, replicacoes, requerem, requerer, rigidez, roer, sacroiletite, sensibilizacao, travesseiro, triclosan, tuberculose, ulcerada, ulceradas, unilateral, urgente, varicosas, vaselinada, vasos, vermelha, vinagre, vomito, xantomas, yasmin,

abscessos, acromicas, alergias, alopurinol, amostras, ampliar, amtriptilina, anticorpos, antihiv2, antisepticos, anulares, bacteriologico, banheiro, calafrios, citomegalovirus, coagulacao, coalescer, colchao, compressa, cronificacao, demodex, depilatorios, descoloracao, difenidramina, distribuidas, doce, doxicilina, educacao, efeitotaquifilaxia, effaclar, elementares, elevado, endocrino, endogeno, endurecimento, enfermagem, enzimas, escoriadas, escova, estereis, estreptococcica, esverdeada, etinilestradiol, gravidas, hepatoesplenomegalia, hiperhidrose, htlv, ictiose, imiquimod, imiquimode, impetinizacao, imunomoduladores, malformacoes, margens, mefenamico, melanoniquia, melanose, mensalmente, micologicos, moradores, morfina, niacina, normaderm, onicotilomania, ortopedicos, ortopedista, palmilhas, pelagra, penfigoide, perianal, permanencia, pulsada, quinolonas, removendo, repigmentar, rubor, rubra, sangramentos, sindromes, sintomaticas, sulfametoxazoltrimetopril, suppurativa, suspenso, sutia, tacrolimos, tarv, tatuagem, teleoftalmo, tiabendazol, traumatica, trok, vasculares, alcoolgel, alcoolismo, alergicos, alongamentos, amiantacea, amiloidose, amolecida, antiandrogenica, anticoncepcao, antidna, antifosfolipideo, antihipertensivas, antimicrobiano, antimicrobianos, antinuclear, antiro, aroeira, arranhar, atipia, atroficas, autossomica, cetaphil, classification, clorexidina, cloridroxido, cloroquina, contanto, continuacao, continuam, contraceptivos, convulsoes, culturas, cushing, doppler, dorsiflexao, elixir, elucidar, eritrasma, erosada, eruptivo, espironolactona, linfoma, linfopenia, lipodermatoesclerose, lubrificantes, menopausa, migratorio, modificar, monitoramento, monitorando, morango, morder, mosaicismo, mosquitos, mude, multibacilar, multiforme, mutacoes, necrotica, neonatos, nervos, neurofibromas, neuropsicomotor, niacinamida, nicotinico, nigra, nitidez, osteomielite, ouro, paget, palidez, parentesco, parestesica e prenatal.

APÊNDICE F — OCORRÊNCIA DE TERMOS DE CONDUTA NO LEVANTAMENTO DE LAUDOS

Ordem	Ocorrências	Termo
1	15	sabonete
1	15	roupa
2	12	sintoma
3	10	antihistaminico
4	8	hidroxizine
4	8	diario
4	8	algodao
5	7	dexclorfeniramina
5	7	neutro
5	7	perfume
5	7	girassol
6	6	conforme
6	6	excessivo
6	6	peso
6	6	locao
7	5	hidratante
7	5	cuidado
7	5	mometasona
7	5	abrasivo
7	5	emoliente
7	5	fragrancia
7	5	unha
8	4	atopica
8	4	dexclofeniramina
8	4	diminuicao
8	4	exacerbacao
8	4	amaciante
8	4	questionar
8	4	secundaria
8	4	hepatopatia
8	4	sistemico
8	4	limpeza
8	4	estresse
8	4	emocional
8	4	sintomatico
8	4	cremosa

Ordem	Ocorrências	Termo
9	3	reforcar
9	3	furoato
9	3	quente
9	3	tireoide
9	3	nefropatia
9	3	hiv
9	3	cancer
9	3	psiquiatrica
9	3	sudorese
9	3	hemograma
9	3	fosfatase
9	3	alcalina
9	3	ggt
9	3	creatinina
9	3	ureia
9	3	raiox
9	3	antihiv
9	3	antihcv
9	3	hbsag
9	3	morno
9	3	irritante
9	3	sedativo
9	3	prometazina
9	3	hidroxizina
9	3	risco
9	3	pequena
10	2	dexametasona
10	2	cronicamente
10	2	somente
10	2	consecutivo
10	2	extraseca
10	2	nivea
10	2	cronica
10	2	ansiedade
10	2	lavagem
10	2	acompanhamento
10	2	revisar
10	2	cefalexina
10	2	corticoide
10	2	investigacao
10	2	confirmacao

Ordem	Ocorrências	Termo
10	2	material
10	2	cronicidade
10	2	escabiose
10	2	ivermectina
10	2	cama
10	2	ferro
10	2	venosa
10	2	ressecamento
10	2	ulceracao
10	2	prolongar
10	2	repetir
10	2	atrofia
10	2	aumento
10	2	edema
10	2	cirurgia
10	2	vascular
10	2	hipertensao
10	2	elevacao
10	2	reducao
10	2	suave
10	2	vaselina
10	2	umida
10	2	candida
10	2	antifungico
10	2	cetoconazol
10	2	miconazol
10	2	posterior
10	2	posinflamatoria
10	2	regredir
10	2	espontaneamente
10	2	plastico
10	2	cocadura
10	2	primaria
10	2	xerose
10	2	seca
10	2	sedacao
10	2	efetivo
10	2	luva
10	2	coleta
10	2	micologico
10	2	toalha

Ordem	Ocorrências	Termo
11	1	neutrogena
11	1	adesao
11	1	manipular
11	1	explicar
11	1	hipoalergenico
11	1	lavar
11	1	sabao
11	1	priorizar
11	1	prescrever
11	1	psiquiatrico
11	1	funcional
11	1	grave
11	1	previamente
11	1	domiciliar
11	1	reumatologista
11	1	potencia
11	1	continuo
11	1	assintomatico
11	1	retirar
11	1	higienizar
11	1	assento
11	1	vaso
11	1	sanitario
11	1	alcool
11	1	modificado
11	1	diferencial
11	1	regula
11	1	vdrl
11	1	eletroforese
11	1	glicose
11	1	camomila
11	1	consiga
11	1	anatomopatologico
11	1	clobetasol
11	1	manutencao
11	1	delicada
11	1	sintetico
11	1	patologia
11	1	recorrente
11	1	recidivante
11	1	espinocelular

Ordem	Ocorrências	Termo
11	1	biopsia
11	1	doce
11	1	proximo
11	1	recontaminacao
11	1	estase
11	1	prevencao
11	1	mupirocina
11	1	intenso
11	1	loratadina
11	1	umidade
11	1	atrito
11	1	frio
11	1	absorvente
11	1	dobra
11	1	sobrepeso
11	1	clotrimazol
11	1	nistatina
11	1	falha
11	1	fluconazol
11	1	recorrencia
11	1	vasenol
11	1	filme
11	1	surge
11	1	detergente
11	1	cultura
11	1	itraconazol
11	1	interacao
11	1	medicamentosa
11	1	hepatica
11	1	colirio
11	1	vinagre
11	1	branco
11	1	acido
11	1	descolonizacao
11	1	dependendo
11	1	odor
11	1	rapidamente
11	1	borracha
11	1	alimento
11	1	excessiva
11	1	dimeticone

Ordem	Ocorrências	Termo
11	1	retirada
11	1	cuticula
11	1	informar
11	1	lenta
11	1	silicone
11	1	episodio
11	1	ocorrem
11	1	aciclovir
11	1	renal
11	1	sexualmente
11	1	brasilia
11	1	contrario
11	1	idoso
11	1	lavado
11	1	passado
11	1	maquina
11	1	suspensao
11	1	cultural
11	1	fungo
11	1	raspado
11	1	resolucao
11	1	refratario
11	1	terbinafina
11	1	higiene
11	1	hiperpigmentacao
11	1	meg
11	1	dexclorferinamina

APÊNDICE G — OCORRÊNCIA DE TERMOS DE CONSULTA NO LEVANTAMENTO DE LAUDOS

Ordem	Ocorrências	Termo
1	15	mal
1	15	sinal
2	12	prurido
3	5	ardencia
4	3	braco
4	3	corpo
4	3	perna
5	2	eritematosa
5	2	pruriginosa
5	2	descamativa
5	2	sangramento
5	2	dermatite
5	2	asma
5	2	nadega
5	2	coceira
5	2	crosta
5	2	perineo
5	2	papula
5	2	escabiose
5	2	dedo
5	2	mao
5	2	abdome
6	1	rediscussao
6	1	antebraco
6	1	glutea
6	1	persistencia
6	1	endurecida
6	1	liquenificacao
6	1	crostosa
6	1	puntiforme
6	1	impetigo
6	1	alergeno
6	1	sensibilidade
6	1	tabagista
6	1	reumatologia
6	1	inflamatorio
6	1	coxa
6	1	cronico
6	1	hipertensa
6	1	vagina

Ordem	Ocorrências	Termo
6	1	queimacao
6	1	hiperpigmentada
6	1	vaginal
6	1	corrimento
6	1	depressao
6	1	dislipidemia
6	1	rompida
6	1	cocadura
6	1	cabeludo
6	1	pior
6	1	rinite
6	1	alergica
6	1	itu
6	1	surgimento
6	1	fossa
6	1	cubital
6	1	poplitea
6	1	prurigo
6	1	cronica
6	1	piora
6	1	hipocromica
6	1	esquizofrenia
6	1	torax
6	1	unha
6	1	fungo
6	1	seca
6	1	recidiva
6	1	endurecido
6	1	insuficiencia
6	1	cardiaca
6	1	eritematoso
6	1	descamacao
6	1	vesicula
6	1	popular
6	1	apareceu
6	1	vermelha
6	1	cotovelo
6	1	barriga
6	1	dorso
6	1	escoriacao
6	1	assadura
6	1	disseminado
6	1	cachorro
6	1	sarna