UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA

DEPARTAMENTO DE ESTATÍSTICA

# APROXIMANDO CIÊNCIA POLÍTICA E ESTATÍSTICA NA ERA DO *BIG DATA*

**EDUARDO SCHINDLER**

Porto Alegre
2016

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

INSTITUTO DE MATEMÁTICA E ESTATÍSTICA

DEPARTAMENTO DE ESTATÌSTICA

# APROXIMANDO CIÊNCIA POLÍTICA E ESTATÍSTICA NA ERA DO *BIG DATA*

**EDUARDO SCHINDLER**

Porto Alegre

2016

**EDUARDO SCHINDLER**

# APROXIMANDO CIÊNCIA POLÍTICA E ESTATÍSTICA NA ERA DO *BIG DATA*

Artigo submetido como requisito parcial para a obtenção do grau de Bacharel em Estatística

Orientador Metodológico

Prof[a]. Dr[a]. Luciana Neves Nunes

Porto Alegre

2016

Instituto de Matemática e Estatística

Departamento de Estatística

**Aproximando Ciência Política e Estatística na era do *Big Data***

Eduardo Schindler

Banca examinadora:

Prof[a]. Dr[a]. Márcia Helena Barbian

IME/UFRGS

Prof[a]. Dr[a]. Luciana Neves Nunes

IME/UFRGS

*Não deveríamos avaliar as ações humanas com base nos resultados.*

*Bernoulli, Jakob*

**Agradecimentos**

Este trabalho representa o ápice de 5 anos de estudos em que pude expandir os alicerces da minha formação acadêmica. Considero de grande sorte poder combinar formação em Ciências Humanas e Exatas em um período onde tanto a Estatística quanto a Ciência Política estão em rápido desenvolvimento.

Primeiramente gostaria de agradecer à Universidade Federal do Rio Grande do Sul, especialmente ao Instituto de Matemática e Estatística por fornecerem educação pública, gratuita e de qualidade; e à minha orientadora, professora Luciana Neves Nunes, abraçar a ideia, pelo voto de confiança que permitiu a realização desse trabalho.

Aos amigos e colegas Gabriel da Cunha, Lucas Cé, e Luciana Ghiggi e Ana Júlia Possamai pelas revisões e discussões durante a elaboração deste trabalho. Aos demais colegas e amigos estatísticos e cientistas políticos que contribuíram para o trabalho, mesmo que indiretamente.

E ainda, minha gratidão a todos aqueles que de algum modo fizeram parte deste período. Todos têm sua contribuição no desenvolvimento de um profissional de estatística. Especialmente aos familiares que mesmo distantes e sem compreender exatamente o que desenvolvi ao longo desse período, sempre depositaram fé e confiança.

**RESUMO**

Este trabalho propõe avançar o diálogo e a troca de conhecimento entre a Estatística e as Ciências Sociais apresentando, de forma introdutória, técnicas de coleta, análise e apresentação de dados utilizando ferramentas computacionais. Seu objetivo é apresentar o processo de criação de uma aplicação web que combina conhecimentos das duas disciplinas - técnicas de análise e processamento de dados de forma acessível aos cientistas sociais. Para tanto, inicialmente se discute o contexto do *Big Data* e, após, apresenta-se *o* desenvolvimento da aplicação, ocorrido em três etapas principais: o planejamento da base de dados, o planejamento da coleta de dados usando R, e o desenvolvimento da interface com o usuário utilizando Shiny Dashboard. Essa ferramenta contribui para o desenvolvimento da cidadania ao aumentar a transparência da atividade legislativa, facilitando o acesso de cidadãos às estatísticas de desempenho e realização de atividades no poder legislativo de uma determinada unidade da federação.

**Palavras-chave:** *Big Data*, Estatística. Ciência Política, *web scrap*, Estatística descritiva, Modelo Relacional, Base de dados, Shiny, R

**ABSTRACT**

This paper intents to go forward on the dialog and the exchange of knowledge between Statistics and Social Sciences demonstrating and introducing, briefly, web scrap, data base and data presentation techniques using computational resources. It intends to present the process of creating a web application that merges both subjects – data analysis and processing in approachable form for social scientists. First, the Big data context is presented, following by the application development, in three main steps: data base design within the relational data base approach, data harvest using R, and user application interface development using Shiny Dashboard. When applied on political institutions this framework contributes enhancing citizenship as it allows public accountability, providing data on public policies actors' performances.

**Keywords:** Big Data, Statistics. Political Science, web scrap, Descriptive Statistics, relational data base, data base, Shiny, R

**Sumário**

## 1. Introdução

As novas tecnologias têm trazido diversas oportunidades para a pesquisa científica: novas formas de coleta e análise de dados tornam-se mais acessíveis devido à redução do custo de equipamentos e ao desenvolvimento de softwares relativamente simples de serem utilizados capazes de processar grandes volumes de dados. Essa disseminação de técnicas de análises que contribuem para tornar os processos de tomada de decisão mais eficientes insere-se no contexto do fenômeno do *Big Data*. Tem sido atribuído (Kitchin, 2014; Yiu, 2012) ao *Big Data* uma capacidade praticamente ilimitada no que diz respeito à compreensão do comportamento humano em razão da sua produção de dados em massa, porém, para que esta capacidade seja desenvolvida, é preciso que pesquisadores de diferentes áreas consigam dialogar e trabalhar com bancos de dados em termos compreensíveis a ambas as partes.

Essa realidade é uma das faces do que vem sendo chamado de revolução do *Big Data* (Kitchin, 2014). O termo *Big Data* já foi utilizado em diversos congressos e eventos especializados de diferentes áreas, porém sem muita precisão quanto a sua definição. Ele tem sido utilizado para 1) descrever os fenômenos sociais, como a mudança de comportamento gerada pela informação em tempo real; 2) as características específicas dos dados, como grandes volumes gerados rapidamente, e de fontes e formas variadas; 3) as técnicas de análises e de armazenamento, que precisam ser pensadas e planejadas de modo a superar limites das técnicas utilizadas hoje em dia; e, ainda, 4) as necessidades tecnológicas, como dispositivos de captura e de processamento (De Mauro, 2014). Ou seja, a propagação rápida desse conceito tem levado a sua ampliação sem um núcleo claro de significado de o que é *Big Data* (Grimes, De Mauro, *et al* 2016).

Diebold (2012) realizou uma busca pela origem etimológica do termo sem chegar a conclusões definitivas. Contudo, em meados dos anos 1990, apontou possíveis origens do uso do termo com o sentido que se conhece hoje, que é descrever soluções para a armazenagem e processamento de grandes volumes de dados. Segundo o autor (Diebold, 2012), Cox e Ellsworth (1997) foram os responsáveis por uma das primeiras aparições do termo com o sentido atual ao se referir ao *big*

*Data problem* para descrever uma situação em que um conjunto de dados era maior do que a capacidade da memória principal[1], exigindo novas soluções.

A consolidação do fenômeno do *Big Data* enquanto atributo dos dados veio no início dos anos 2000 com o relatório de Douglas Laney (2001), em que estuda os desafios que o incremento do *e-commerce* trouxe para o mercado de gestão de dados, e afirma que "as condições e meios de negócios atuais estão empurrando os princípios tradicionais da gestão de dados para os seus limites, fazendo surgir novas abordagens" (Laney, 2001, p. 1). Os desafios da gestão de dados mencionados pelo autor são apresentados em três dimensões: Volume, Velocidade e Variedade. Laney não cunhou nem utilizou o termo *Big Data*, mas estabeleceu seus fundamentos: os "3 vês" são referidos constantemente como núcleo do conceito de *Big Data*. Sua caracterização tornou-se, assim, referência (De Mauro, 2014; Diebold, 2012).

Uma das primeiras utilizações estatísticas conscientes do termo *Big Data* foi implementada por Diebold (2003 *apud* Diebold, 2012). Por "conscientes" entende-se que o termo não continha apenas o sentido genérico de grande base de dados, mas também as outras implicações próprias do fenômeno, como, por exemplo, implicações metodológicas. O próprio Diebold (2012) reconhece que outros autores também utilizaram o termo na mesma época com a finalidade de discutir aspectos metodológicos da econometria em que eram necessárias novas técnicas e abordagens para lidar adequadamente com quantidades de variáveis, preditores e observações que levavam os métodos e recursos da época ao limite. Entretanto, mesmo antes da aplicação do termo, o fenômeno do *Big Data* já era conhecido: o autor cita como exemplo uma publicação de 1996, que trata de *massive data sets* (Anderson, 1996), em que é ressaltado o papel das técnicas de inteligência artificial que podem contribuir para lidar com grandes bases de dados. Nessa mesma publicação, Dumais (1996) discute problemas típicos do que hoje é conhecido por *text mining*, fazendo observações sobre as grandes matrizes que eram geradas nas análises de texto e seus limites.

No caso da estatística, tanto a necessidade de desenvolver novas técnicas para tratar da grande massa de dados disponíveis, quanto a oportunidade gerada por novas tecnologias de

[1] A memória principal de um computador é aquela que pode ser acessada diretamente pelo processador. É onde os programas e os dados necessários a sua execução são temporariamente armazenados.

processamento, permitiram a disseminação de softwares e o desenvolvimento de métodos computacionais intensivos, além do avanço e da aplicação em larga escala dessas técnicas. Diferentes técnicas estatísticas encontraram campos prósperos para seu desenvolvimento. Algumas delas, como a análise fatorial, a análise de componentes principais, a análise de cluster e a regressão logística tornaram se parte de técnicas conhecidas como *Data Minning*, *Machine Learning*, e foram incorporadas ao *Big Data* como funções de um cientista de dados[2].

Finalmente, a convergência entre os dados com novas características, as novas tecnologias capazes de armazenar e processar grandes massas de dados, e as novas técnicas de análise têm produzido algum grau de impacto social que pode ser observado em diversas áreas. O relatório do Mckinsey Global Institute (Manyka *et al*, 2011), por exemplo, chamou o *Big Data* de a próxima fronteira para a inovação, competitividade e produtividade. Segundo o Instituto, este fenômeno é capaz de aumentar a produtividade e a geração de valor em diversos setores econômicos, gerar um novo mercado e um novo tipo de profissional.

Outra faceta do impacto social do fenômeno inclui a comunicação instantânea e as novas formas de interação social, já que estas alteram o funcionamento de uma sociedade e o comportamento dos indivíduos, uma vez que "passam a ser produtores incessantes de dados digitais" (Nascimento, 2016, p. 223).

A dimensão do impacto social do *Big data* seria, por sua natureza, objeto de estudo das Ciências do comportamento. Entretanto, nas Ciências Sociais brasileiras, a disponibilidade de dados sobre temas específicos e aplicação de técnicas estatísticas para além das medidas descritivas é escassa. Em 2005, quando o fenômeno do *Big Data* já estava caracterizado e batizado, a Ciência Política brasileira, por exemplo, estava "na contramão da história" (Soares, 2005), pois a estatística, como técnica de análise, não era bem recebida nos círculos acadêmicos. Segundo o autor, havia uma falsa dicotomia entre estudos quantitativos e qualitativos que era, na verdade, reflexo de uma formação deficiente na capacidade dos cientistas políticos trabalharem com métodos rigorosos. Como consequência, Gláucio Soares (2005) argumenta que a Ciência Política tende ao isolamento, e está sujeita à invasão de outras áreas.

[2] Cientista de dados é como tem sido referido o profissional que trabalha com Estatística, processamento de dados e apresentação de relatórios no contexto de Big Data.

Nesse mesmo sentido, Barboza e Godoy (2014) realizaram um estudo a fim de verificar o oferecimento de disciplinas metodológicas nos programas de pós-graduação em Ciência Política e concluíram que o problema apontado 10 anos antes por Soares (2005) ainda persistia. Cabe destacar que ambos os autores utilizam apenas estatísticas descritivas nos seus trabalhos.

Baltar e Baltar (2013) também apontam para o isolamento das Ciências Sociais brasileiras em termos de interdisciplinaridade e, assim como Ramos (2013), destacam os limites dos pesquisadores em relação ao uso de ferramentas computacionais em pesquisa social. O desafio que as Ciências Sociais têm pela frente é, portanto, desenvolver um *know-how* de coleta, armazenamento e análise de dados numa escala com a qual ainda não é capaz de lidar. Considerando que todo esse fenômeno descrito gera impactos que, por sua natureza, são objetos de estudo das Ciências Sociais (Nascimento, 2016), a solução é a aproximação com outras áreas do conhecimento que detenham as técnicas de análise e processamento de dados, pondo fim ao isolamento e se tornando interdisciplinar de fato.

Há, contudo, aqueles que acreditam que essas técnicas tornam o método científico tradicional obsoleto e que apontam inclusive o “fim da teoria” e da ciência como ela vem sendo feita (Anderson, Prensky, Dyche *apud* Kitchin, 2014), uma vez que apenas a identificação de relações estatisticamente significativas seriam suficientes para desvendar padrões, sem a necessidade de um marco teórico. Kitchin (2014) destaca que os próprios dados são gerados em contextos sociais e a partir de construções teóricas. Sendo assim, o fim da teoria seria uma falácia. Análises teóricas sociais ainda são fundamentais, dada a complexidade das relações sociais e do comportamento humano.

Dessa forma, cabe aos cientistas sociais conhecer os métodos para compreender os tipos de técnicas empregadas na literatura e ler com propriedade os trabalhos de seus pares; além de conhecer as ferramentas que têm à disposição para suas investigações científicas. O estatístico, por sua vez, tem nas técnicas o seu objeto de estudo, surgindo daí a responsabilidade de tornar estas ferramentas acessíveis aos pesquisadores de outras áreas. Por diversas vezes em congressos e seminários de Estatística, a falta de comunicação da Estatística com outras áreas tem sido

apontada como responsável por parte desta barreira[3]. Podemos notar assim, que o isolamento é uma característica da relação não só entre a Estatística e as Ciências Sociais, mas também destas com outras áreas.

Embora a oportunidade de fusão entre as áreas seja frequentemente destacada, a escassez da literatura em português que aborda métodos quantitativos de maneira acessível aos cientistas sociais atua como uma barreira de entrada na área, dificultando sua aplicação e uso.

Este trabalho propõe avançar o diálogo e a troca de conhecimento entre a Estatística e as Ciências Sociais apresentando de forma introdutória técnicas de coleta, análise e apresentação de dados utilizando ferramentas computacionais utilizadas no contexto do *Big Data*. Seu objetivo é apresentar o processo de criação de uma aplicação *web* que associa fatores das duas disciplinas – técnicas de análise e processamento de dados de forma acessível aos cientistas sociais. Essa ferramenta também contribui para o desenvolvimento da cidadania ao aumentar a transparência da atividade legislativa, pois facilita o acesso de cidadãos às estatísticas de desempenho e realização de atividades no poder legislativo de uma determinada localidade. O único requisito para a aplicação da técnica é que os dados de interesse estejam dispostos em páginas da *web*. De forma a demonstrar que é possível implementar este tipo de aplicação longe dos grandes centros populacionais e econômicos, o município selecionado para esta pesquisa foi Ijuí, cidade localizada no noroeste do Rio Grande do Sul e com aproximadamente 80 mil habitantes.

Este texto está estruturado em três partes principais além desta introdução. Inicialmente, apresentam-se, de forma introdutória, conceitos básicos de bancos de dados relacionais e regras de normalização, que se aplicam de forma a garantir o armazenamento de dados de maneira eficiente. Esta etapa tem a finalidade de apresentar o marco teórico por trás da construção do banco de dados. Na segunda parte, introduz-se uma ferramenta capaz de coletar dados (semi) estruturados de sítios *web*, o *Web Scraping*, que consiste na leitura do código fonte para retirar informações específicas para a construção de bancos de dados. A utilização da ferramenta será exemplificada utilizando um pacote desenvolvido na linguagem R feito para esta finalidade, o

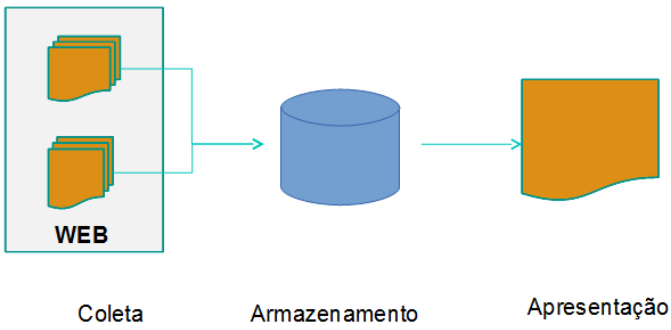
[3] Esta afirmação foi feita em diversas situações em palestras do XXII SINAPE, de 2016. Por exemplo, David Spiegelhalter na conferência de abertura; na mesa redonda *ISI Session: Expanding the Frontiers of Statistical Practice*; e no evento satélite “Perspectivas e Desafios na Produção e Divulgação de Estatísticas Públicas no Brasil”.

rvest (Wickham, 2016). Por fim, discute-se a criação de uma interface *web* para a apresentação dos dados obtidos utilizando Shiny Dashboard (Chang, 2016).

## 2. Metodologia

O desenvolvimento da aplicação ocorreu em três etapas principais: o planejamento da base de dados, o planejamento da coleta de dados, e o desenvolvimento da interface. A interação entre as etapas da aplicação pode ser visualizada no esquema a seguir.

Figura 1- Fluxograma da aplicação




O planejamento da base de dados corresponde à identificação das variáveis de interesse e a estruturação do esquema de armazenamento de dados. Optou-se pela criação de um banco de dados dentro do marco teórico das bases de dados relacionais. A escolha por esse modelo justifica-se pela versatilidade e otimização do espaço que a estratégia provê (Heuser, 1998).

O planejamento da coleta significa localizar e definir a estratégia de extração dos dados. Para tanto é necessário identificar a localização dos dados e estabelecer uma rotina capaz realizar a atividade. Esta tarefa foi desenvolvida em R versão 3.3.2 (RCORE, 2015), com auxílio do pacote

rvest (Hadley, 2016). Inicialmente considerou-se o uso do RSelenium (Harrison, 2016), porém, dificuldades técnicas[4] tornam seu uso mais restrito em relação ao rvest.

Por fim, o desenvolvimento da interface foi feito utilizando o Shiny Dashboard (Chang, 2016), que permite a criação de páginas e painéis de controle *web* no ambiente R. A aplicação foca na apresentação de estatísticas descritivas da atividade legislativa. Uma vez que a técnica utilizada na coleta dos dados permite a análise de todo o universo a utilização de estatísticas descritivas fornece uma visão completa da situação. Dessa forma, são apresentadas estatísticas descritivas, como médias, proporções e frequências – inclusive na forma de nuvem de palavras. Os gráficos apresentados foram elaborados utilizando os pacotes highcharter (Kunst, 2016), bubbles (Cheng, 2015), e wordcloud (Fellows, 2014).

Todo o trabalho foi desenvolvido com a linguagem R versão 3.3.2, no ambiente de desenvolvimento RStudio (2015) por ser uma ferramenta de código aberto que garante a reprodutibilidade da aplicação.

## 2.1 Estruturação do banco de dados

Dados coletados são, na verdade, representações de atributos ou características pertencentes a alguma unidade. Por exemplo, cada vereador possui um nome e está filiado a um partido. O nome e o partido são, portanto, atributos de um vereador. Coletando os atributos de diversos vereadores do município, pode-se construir um banco de dados. Um banco de dados é, portanto, o conjunto de unidades reunidas segundo uma mesma lógica de relação entre variáveis (Heuser, 1998).

O modelo Relacional foi desenvolvido por Edgar Frank Codd em 1970, e publicado no artigo *Relational Model of Data for Large Shared Data Banks*. Codd concebeu um modelo de organização dos dados com base em uma estrutura fundamental chamada de relação ou entidade, que é comumente conhecida por tabela. Uma relação é constituída por um ou mais atributos (colunas da tabela) que traduzem os tipos de dados a armazenar. As listas ordenadas destes

[4] O pacote apresentou incompatibilidades com a versão 49 do Firefox. Foram consideradas alternativas como o uso do phantomJS. O uso deste navegador viabilizou a atividade, entretanto, exigiu uma série de conhecimentos específicos para configurar adequadamente o software. Por isso, optou-se pela troca do pacote. Para uma aplicação utilizando o RSelenium ver (Meirelles, 2016)

atributos são chamadas de tuplas. Cada tupla preenchida é um registro (uma linha na tabela) (Heuser, 1998). Considere por exemplo a relação abaixo que descreve atributos das matérias no modelo lógico e em forma de tabela.

```
Matéria(id, Tipo, Autor)
```

Tabela 1 – Exemplo de uma relação

| id | Tipo | Autor |
|---|---|---|
| 5135 | Projeto de Lei | Poder Executivo |
| 5136 | Requerimento | Vereador A |
| 5137 | Indicação | Vereador B |

Cada elemento de uma tupla deve ser atômico. Isto é, não é permitido que o componente seja uma estrutura de dados em que os valores possam ser decompostos em unidades menores. No exemplo acima pode ser atribuído apenas um Tipo e um Autor a cada matéria, representada por um id. É preciso, ainda, que os valores dos componentes pertençam ao domínio do respectivo atributo. Isto é, há um conjunto especificado de autores e de tipos que podem ocupar os respectivos atributos. Existe ainda nesta tabela a variável id, que assume um valor único em cada matéria. Essa característica permite a adoção dessa variável como chave primária, um elemento que pode ser usado como índice para criar relacionamentos com outras tabelas e atributos. A relação entre atributos é chamada de dependência funcional. O atributo B possui dependência funcional do atributo A se para cada valor do atributo A existir um único valor do atributo B. No exemplo anterior, como cada matéria possui exatamente um tipo, pode-se dizer que Tipo depende funcionalmente do id. Ou seja, conhecendo o id, pode-se obter o respectivo tipo associado.

As associações entre tabelas são definidas pelas regras de compartilhamento de informações. Essas relações se dão por meio de chaves que indicam que as tupla estão conectadas. Considere, por exemplo,

```
Vereador(Nome, Partido)
Matéria(id, Tipo, Autor = Vereador.Nome)
```

Neste exemplo, tem-se uma relação entre tabelas. Um vereador é caracterizado por um nome e um partido, enquanto uma matéria tem uma identificação e é caracterizada por um tipo e um autor. Nome e id são as respectivas chaves primárias, e a relação entre as duas tabelas está no fato de que os autores das matérias são os vereadores.

A relação entre tabelas e atributos deve ser definida de modo a evitar aspectos indesejáveis, como a repetição de informação, a incapacidade de representar parte da informação ou a perda de informação, garantindo, assim, a organização e a eficiência no banco de dados. Para tanto, existe um conjunto de regras com a finalidade de evitar estes problemas. A aplicação dessas regras é conhecida como processo de normalização.

No processo de normalização, cada regra é chamada de uma "forma normal". O banco é considerado na primeira forma normal se a primeira regra é observada, o banco de dados é considerado na "segunda forma normal", se a segunda regra for observada, e assim por diante. Existem mais níveis de normalização.

A primeira forma normal (FN1) consiste em eliminar grupos repetidos em tabelas individuais, criando tabelas separadas para cada conjunto de dados e identificando cada uma por meio de chaves primárias. Esta forma é obtida se todos os atributos forem atômicos e contiverem apenas um valor. A relação abaixo, por exemplo, no caso em que o requerimento foi de autoria do Vereador A e do Vereador C, deve ser dividida em duas. Assim, as novas tabelas possuem apenas um valor em cada atributo, cumprindo os critérios da primeira forma normal.

Figura 2 – Primeira forma normal: exemplo da decomposição de uma tabela em 2 tabelas na primeira forma normal.

| id | Tipo | Autor |
|---|---|---|
| 5135 | Projeto de Lei | Poder Executivo |
| 5136 | Requerimento | Vereador A, Vereador C |
| 5137 | Indicação | Vereador B |

| id | Tipo |
|---|---|
| 5135 | Projeto de Lei |
| 5136 | Requerimento |
| 5137 | Indicação |

| id | Autor |
|---|---|
| 5135 | Poder Executivo |
| 5136 | Vereador A |
| 5136 | Vereador C |
| 5137 | Vereador B |

A segunda forma normal (FN2) argumenta que a única dependência funcional deve ser da chave primária, além de estarem na FN1. No banco Vereador, onde Nome é chave primária

`Vereador(`**`Nome`**`, Partido, Espectro político do partido)`

Observa-se que o espectro político do partido está associado diretamente ao partido e não ao vereador. Isso significa que ele depende funcionalmente de Partido e não da chave primária id. Portanto, a relação pode ser dividida em

`Vereador(`**`Nome`**`, Partido)`

`Partido(`**`Nome`**`, Sigla, Espectro político)`

A terceira forma normal (FN3), por sua vez, argumenta que os atributos não chave devem ser independentes entre si, e dependentes única e exclusivamente da chave primária. Considere o banco abaixo que relaciona Vereador, Nome, Partido e Frequência.

`Vereador(`**`Nome`**`, Partido, Frequência)`

`Sessões (`**`Data`**`, Tipo)`

Considerando a existência da relação que diz respeito às Sessões, Frequência é uma função do número de sessões que o vereador compareceu. Portanto, apesar de depender da chave primária, ela não depende única e exclusivamente dela. Sendo assim, este atributo deve ser removido da relação e uma relação das sessões e dos vereadores presentes deve ser criada.

`Vereador(`**`Nome`**`, Partido)`

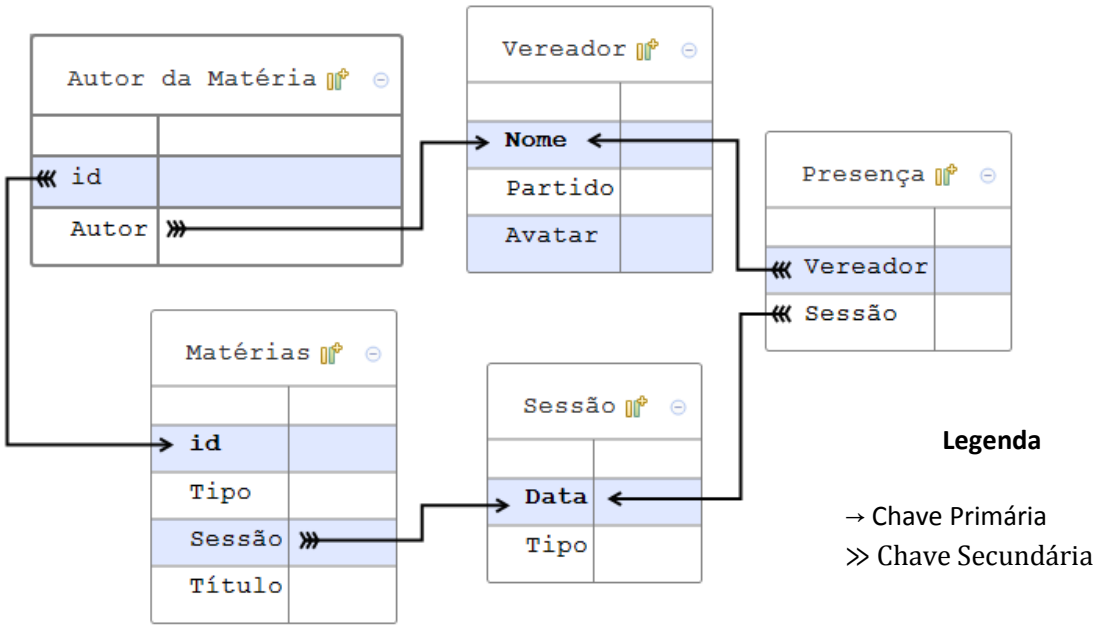
Sessões (**Data**, Tipo)

Presença(Sessão, Vereador)

Dessa forma, é possível explicitar a origem da estatística “Frequência”, que é a soma da ocorrência do nome do vereador em cada Sessão dividido pelo total de sessões ocorridas. Ao acrescentar as novas Sessões e os vereadores presentes na tabela Presença, o valor da estatística Frequência pode ser atualizado. Note que no caso anterior, seria preciso realizar a alteração dos valores para cada vereador.

Considerando os princípios dos bancos de dados relacionais e as regras de normalização apresentadas, a seguinte estrutura relacional foi adotada para o banco de dados:

Figura 3 – Modelo relacional: Proposta de armazenamento de dados de acordo com as três primeiras formas normais.




A estrutura proposta evita repetições e permite a realização de consultas, por meio das ferramentas adequadas, de modo a gerar as tabelas e as estatísticas de interesse. Além disso, permite maior eficiência na atividade de *Web Scrap*, uma vez que somente os dados necessários serão coletados.

## 2.2 *Web scrap*

Na maior parte do tempo, navegar pela internet significa visitar sítios *web* utilizando um software adequado à tarefa. Estes softwares são chamados de navegadores (browsers), e são diversos os navegadores disponíveis nos computadores e nos dispositivos móveis que permitem acesso à internet. A tarefa essencial de um navegador é comunicar-se com um servidor e exibir adequadamente a página indicada pelo usuário.

A indicação é feita através de uma linha de texto chamada de Uniform Resource Locator[5] (URL). Esta linha deve conter, no mínimo, a indicação de um protocolo de transmissão (ftp, http, https, mailto...) e de um host[6], que deve ser um nome registrado ou um endereço de IP[7]. Por exemplo o endereço "http://www.camaraijui.rs.gov.br/" indica o esquema HTTP[8] e o host camaraijui.rs.gov.br. Dessa forma, ao fornecer essa URL para o navegador, espera-se o retorno do conteúdo referente à câmara. Adicionalmente, o endereço poderia conter ainda um caminho indicando conteúdos dentro do host, como por exemplo, adicionando "vereadores/" ao final da URL. Nesse caso, teríamos acesso à página de vereadores do sítio da câmara de vereadores de Ijuí.

Caso a URL não seja localizada ou esteja incorreta, o navegador exibe uma mensagem de erro. Caso o navegador tenha sucesso em localizar, solicitar e receber o conteúdo da página, este será exibido. O conteúdo exibido pelo navegador ao acessar uma página é gerado a partir de uma interpretação que o navegador faz de um arquivo fornecido pelo servidor. Este arquivo é chamado de código fonte. Ele contém uma série de orientações para que o navegador disponha corretamente as figuras, textos, tabelas e outros elementos que compõe a página. Este arquivo é

---

[5] Uniform Resource Locator é uma referência a um recurso na web que especifica sua localização e a forma de acesso a esse recurso.

[6] Hospedeiro. Qualquer computador conectado a rede que ofereça recursos, serviços e aplicações.

[7] Endereço no formato xxx.xxx.xxx.xxx que identifica dispositivos na rede.

[8] Hiper Text Transfer Protocol. Protocolo de transferência de páginas web utilizado na internet.

normalmente escrito em uma classe de linguagens conhecidas como linguagens de marcação[9] e, por isso, pode ser lido em editores de texto. Na comunicação por páginas da internet convencionou-se utilizar estruturas e marcações chamadas de *tags*, dispostas no texto do código fonte de acordo com algum padrão e com marcas descritivas, que definem o início e o fim do texto marcado como unidade ou elemento de informação. Por exemplo:

```
<div>
 <p> Olá Mundo! </p>
</div>
```

No código acima, as *tags* são marcadas pelo "<" e ">", indicando o início e o fim de uma estrutura. Além disso, também é possível inserir estruturas dentro de estruturas, e definir características especiais, chamadas de atributos. As estruturas têm a função de definir como uma informação será apresentada, qualificando blocos de conteúdo. Desse modo, as linguagens utilizadas nos códigos fontes podem ser tratadas como objetos estruturados, o que permite não só a compreensão pelo navegador, mas também a realização de buscas automatizadas e a estruturação dos objetos em bancos de dados.

O World-Wide Consortium[10] (W3C) é o órgão internacional responsável por desenvolver e definir especificações técnicas e orientações mínimas comuns para garantir a interoperabilidade da internet. O W3C trabalha com princípios de padrão aberto que garantem a abertura e transparência dessas especificações. No caso das linguagens como HTML, XML, etc., a adoção destes princípios significa que elas não são propriedades sujeitas aos interesses de corporações, "permitindo assim a criação de documentos portáveis, […] que não são dependentes de um determinado software, hardware, ou sistema operacional" (Bax, 2000, 34).

Portanto, sabendo que as páginas exibidas na internet são documentos estruturados, pode-se definir estratégias para extrair estas informações, ou seja, executar o *web scrap*. A ideia por trás do *web scrap* é de que se a página exibe dados de forma sistemática, ela provavelmente foi

[9] Markup Languages. São um conjunto de (meta) linguagens que definem marcas para a representação de textos. O objetivo destas linguagens é separar conteúdo, estrutura e formatação. Nesse conjunto encontram-se SGML, HTML, XML, XHTML, por exemplo. (BAX, 2000)

[10] No sítio Www.w3c.org é possível encontrar mais informações sobre o trabalho da w3c e os princípios de padrão aberto.

gerada a partir de um banco de dados, e, portanto, pode-se tentar reconstruir este banco a partir do conteúdo exibido na página. Munzert (2015) sugere a execução dessa tarefa em 6 passos, a saber 1) identificar a informação que se deseja extrair; 2) escolher uma estratégia adequada à realização da tarefa; 3) obter os dados; 4) extrair a informação desejada; 5) preparar os dados; 6) validar os dados. Contudo, Munzert (2015) alerta que um *web scrap* eficiente é dado pela combinação adequada destas etapas em cada caso específico, nem sempre sendo necessária a execução de todos os passos, podendo, até mesmo, serem executados simultaneamente.

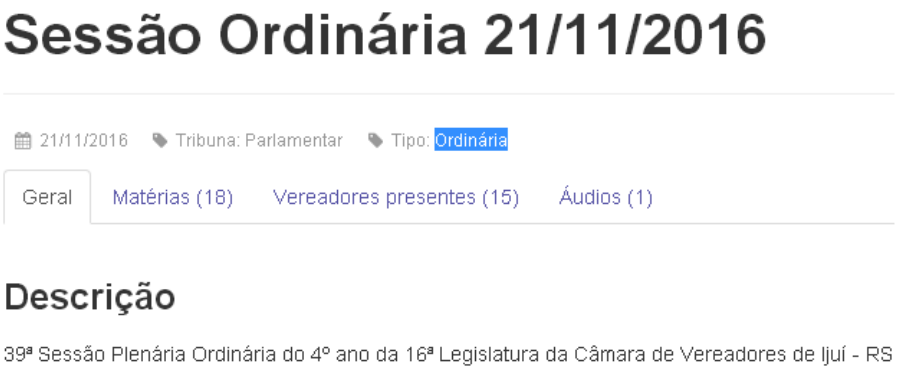
Identificar a existência da informação que se deseja extrair em um *website* é o passo inicial para se optar pela utilização do *web scrap*. Além da informação em si, é preciso localizar a partir da estrutura exibida no código fonte quais são as *tags* e atributos que apontam para a informação que se deseja coletar. A informação pode estar disposta em uma única página, ou distribuída através de diversas páginas. No primeiro caso, é necessário identificar apenas a estrutura da página na qual os dados de interesse estão dispostos. No segundo caso, é necessário, além da etapa anterior, identificar uma estrutura por trás das URLs das páginas.

No primeiro caso, quando a informação está disposta em uma única página, o acesso às estruturas de *tags* do documento é feito pelo código fonte. Em navegadores como o Google Chrome ou o Microsoft Edge, o código fonte pode ser visualizado clicando com o botão direito e selecionando "inspecionar" ou "exibir código-fonte", e a identificação do dado de interesse pode ser feita através de uma busca simples.  A partir daí, o item pode ser localizado pelo X-Path ou pelo Selector[11].

Por exemplo, a página da sessão ordinária de 21/11/2016 (Câmara municipal de Ijuí, 2016) exibe a informação de diversas características da sessão, como o tipo da sessão, a data, o número de matérias votadas e o número de vereadores presentes. A figura 4.a mostra localização da informação de interesse na página, enquanto a figura 4.b mostra a correspondência no código fonte.

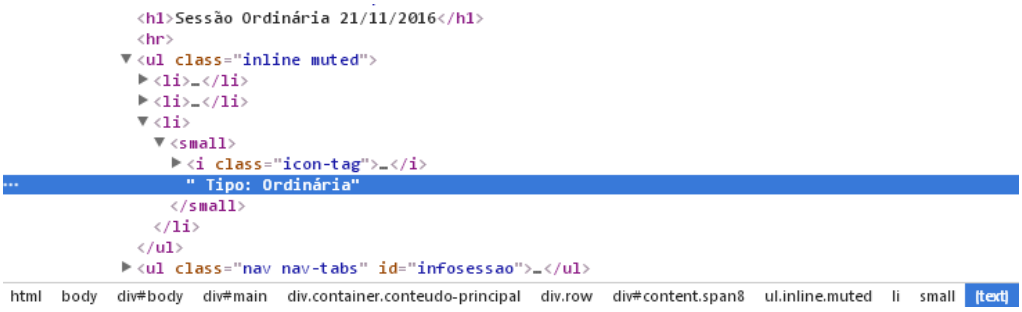
Figura 4.a – Exemplo de página de sessão.

[11] *Selector* são expressões que selecionam atributos em *tags* HTML definidos a partir de folhas de estilo. Já o *X-Path* é uma sintaxe para definir e selecionar as partes (nós) de um documento XML, inclusive seu conteúdo.




Fonte: Câmara municipal de Ijuí (2016)

Figura 4.b – Código fonte




Fonte: Google Chrome

No rodapé exibido na figura 4.b é possível observar o caminho percorrido em toda a estrutura HTML até a *tag* onde está localizado o tipo da sessão. Este caminho é a forma como se indica ao *scrapper* onde a *tag* de interesse está localizada e é representado pelo Seletor CSS, que pode ser obtido facilmente no Google Chrome clicando em cima da *tag* com o botão direito e selecionando "copy CSS path". Assim, será copiado para o *clipboard* "#content > ul.inline.muted > li:nth-child(3) > small". Este Seletor indica a seguinte estrutura mínima:

```
<id = "content">
     <ul>
          <li> … </li>
          <li> … </li>
          <li>
               <small>...</small>
          </li>
```

```
        </ul>
    </div>
```

O Seletor pode ser lido da direita para a esquerda, e o ">" pode ser entendido como "está contido". Assim, a *tag* "small" está contida no terceiro nó "li" da tag "ul.inline.muted", contido na *tag* com âncora "content". Por fim, para localizar adequadamente a informação dentro da *tag*, é necessário, ainda, a distinção entre conteúdos, que são os pedaços de texto exibidos na tela, como neste caso, ou atributos, como no caso dos links, que são armazenados nos atributos href[12].

Para o segundo caso, quando a informação está distribuída ao longo de diversas páginas, é necessário identificar a localização das páginas no servidor pela estrutura da URL. No caso de coleta de informação dos vereadores, a URL "http://www.camaraijui.rs.gov.br/vereadores" lista todas as páginas de vereadores da câmara. Assim, poder-se-ia coletar as URLs como um atributo, para, em seguida, solicitar ao *scrapper* que visite todas as páginas percorrendo a lista e colete a informação desejada em cada uma delas. Uma alternativa seria inspecionar o link para cada página a fim de identificar o mecanismo que estrutura a sintaxe das URLs. Neste caso, a estrutura das URLs é "http://www.camaraijui.rs.gov.br/vereadores/nome_do_vereador". Dessa forma, a URL das páginas dos vereadores poderia ser construída a partir de uma lista de nomes.

A obtenção dos dados consiste no núcleo do *web scrap*. Uma vez que a estrutura das URLs tenha sido desvendada e uma lista das páginas a serem percorridas tenha sido obtida, a atividade de coleta pode ser desdobrada em uma sequência de ações. 1) Mapeia-se e recupera-se o endereço de todas as páginas de interesse; uma vez de posse da lista de endereços, 2) visitam-se as páginas uma a uma para a obtenção do código-fonte. Obtido o código fonte, 3) localiza-se as estruturas que contêm os conteúdos ou atributos de interesse e extraem-se essas informações. Esse passo a passo a ser realizado automaticamente pelo software. É equivalente ao passo número 4 (Munzert, 2015).

[12] o principal atributo do elemento <a>, chamado de âncora, é o link de referência, designado pela sigla de href. Ele guarda a URL para a qual o elemento aponta. Toda a página é referida por um endereço URL.

De posse das informações desejadas, elas podem ser armazenadas conforme o esquema disposto anteriormente. Uma vez que o banco de dados for consolidado, é possível realizar consultas a partir da interface do usuário.

### 2.3 A interface do usuário

Neste trabalho, foi desenvolvida uma interface do usuário que consiste em uma aplicação *web* desenvolvida utilizando o Shiny Dashboard (Chang, 2016). Este pacote é uma expansão do Shiny, desenvolvido pela equipe Rstudio, que permite a criação de painéis de controle de forma simples e direta no ambiente de desenvolvimento em R.

A interface com o usuário consiste em uma página *web* com um menu onde podem ser acessadas as informações coletadas, e um painel principal, onde as informações são exibidas. Optou-se por um menu que contém links para páginas de suporte, que são uma página de apresentação, a página da Câmara de vereadores de Ijuí, e uma página sobre o autor; e links para as páginas de conteúdo, que são uma página com informações sobre a composição da câmara, uma página com informações sobre as matérias e uma página com informações sobre as seções. Os links para as páginas individuais de cada vereador foram gerados de forma dinâmica. Isso quer dizer que se for adicionado ou alterado algum vereador, tanto o número de páginas quanto a lista de nomes são automaticamente gerados a partir dos novos dados inseridos[13].

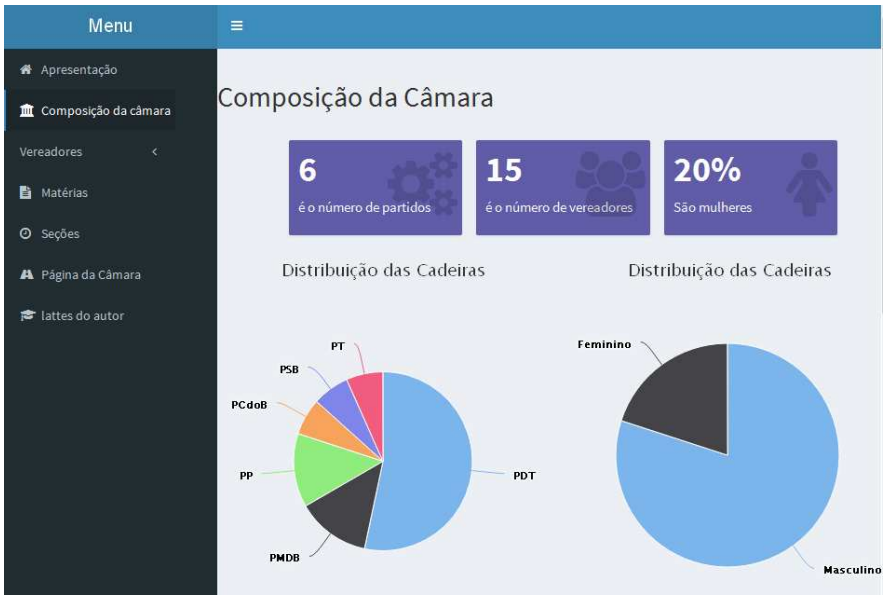
As estatísticas descritivas apresentadas para os usuários são geradas a partir de consultas ao banco de dados criado e armazenado na *web*. A coleta e o armazenamento descritos nas seções 2.1 e 2.2 são realizados em uma etapa separada. Os dados coletados são armazenados junto da aplicação pois a execução da atividade de *web* scrap gera necessidade de espera por parte do usuário.

Como informações para o usuário, são apresentadas estatísticas descritivas específicas em cada página. Na página da composição da câmara (figura 5), são apresentadas duas medidas de frequência (o número de partidos diferentes e o número de vereadores), uma medida resumo (a

[13] uma versão da interface pode ser acessada em https://eduschindler.shinyapps.io/Veredex

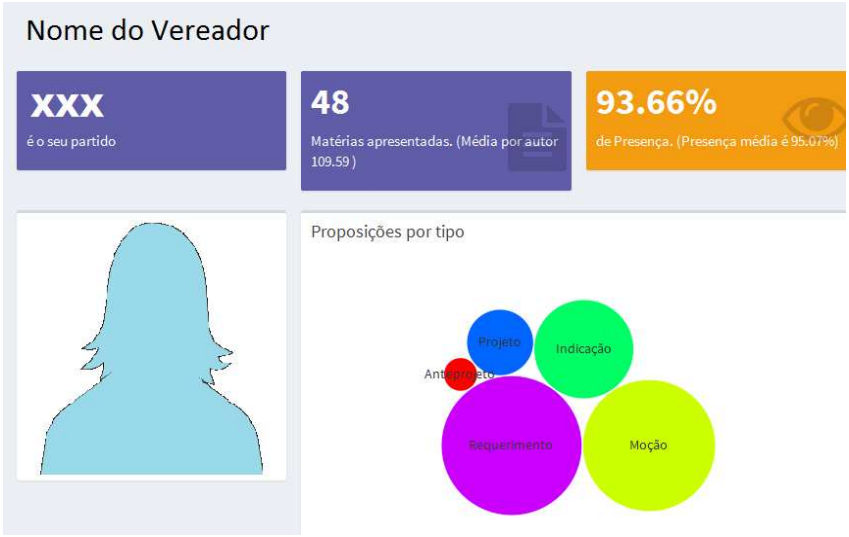
proporção de mulheres no total de vereadores), e dois gráficos de setores para as variáveis qualitativas (partido ao qual o vereador pertence e sexo do vereador).

Figura 5 – Menu e aba composição da câmara




Nas páginas dos vereadores, há a descrição do nome e do partido, o total de matérias apresentadas, ao lado do número médio de matérias apresentadas por autor, e a frequência relativa (em percentual) da presença do vereador nas sessões. Também há uma foto de cada vereador que é carregada diretamente do site da câmara e um gráfico de bolhas que mostra a frequência dos tipos de matérias que o vereador propõe.

Figura 6 – Aba vereadores



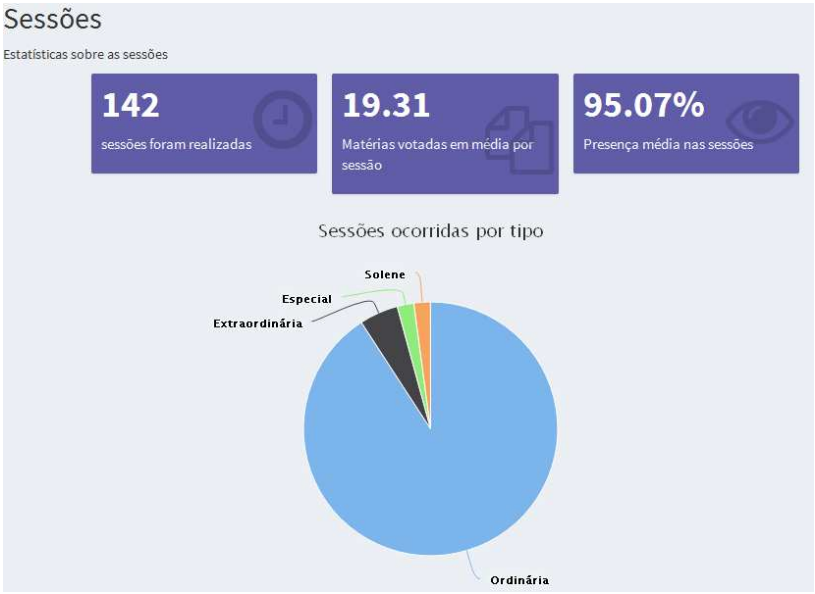

Na página das matérias (figura 7) são apresentadas duas medidas de frequência, o número de matérias votadas e o número de autores diferentes; e uma medida de tendência central, a média de autores por matéria; além de dois gráficos: um de barras mostrando a frequência de cada tipo de matéria, e uma nuvem de palavras elaborada com os títulos das matérias[14].

Figura 7 – Aba matérias


[14] A elaboração da nuvem de palavras requer algum tratamento do texto – retirada de pontuação, letras maiúsculas, de *stop words*, e retirada de palavras que são repetidas e causariam distorções nas frequências. Para maiores detalhes, ver a documentação do pacote wordcloud (Fellows, 2016), e o código em anexo.

Na página das Sessões, são apresentadas três medidas resumo e um gráfico. As medidas apresentadas são o número total de sessões ocorridas, o número médio de matérias votadas por sessão, e a proporção média de vereadores presentes por sessão. O gráfico apresentado é um gráfico de pizza que mostra a proporção de sessões por tipo.

Figura 8 – Aba sessões




## 3. Considerações finais

As técnicas aplicadas neste trabalho representam apenas parte do conhecimento que foi introduzido tanto na Estatística quanto na Ciência Política pelo cenário que a era do *Big Data* trouxe. O domínio destas ferramentas dá autonomia para ambas as disciplinas em todas as etapas do desenvolvimento da pesquisa e permite sinergia entre equipes multidisciplinares.

O *web scrap* representa a autonomia na obtenção de dados. Para os cientistas políticos e sociais em geral, significa que não é mais necessário que uma base de dados estruturada seja

disponibilizada. A existência dos dados dispostos em páginas da *web* é suficiente para que estes possam se tornar variáveis de análise nas questões da área e permitir o aumento da capacidade de análise e a integração de dados de diversas fontes. Para os estatísticos, a técnica simplifica esforços repetitivos e aumenta a chance de significância dos resultados, pois permite a obtenção de amostras maiores.

Conhecer as características da estruturação dos bancos de dados para cientistas políticos insere domínio sobre boas práticas na organização e construção da informação, evitando o surgimento de problemas que possam prejudicar as análises e, assim, comprometer o trabalho. Para estatísticos, por sua vez, aumenta a eficiência no processamento e armazenamento de dados.

Apresentar e disponibilizar os dados para ambos representa a forma de comunicação e acesso das informações pelo grande público, eliminando um grande obstáculo na disseminação de resultados de pesquisas científicas.

Uma etapa seguinte no desenvolvimento dessa pesquisa é a integração completa do *web scrap*, que incluiria a execução sistemática da rotina de *scrap* e o armazenamento do banco de dados em servidores sem a necessidade de interação com o usuário. Além disso, a consulta e descarga das tabelas com os dados brutos poderiam ser disponibilizadas.

Outra evolução possível é o tratamento mais apurado das informações fornecidas. A verificação da aceitação pública da aplicação e a adaptação a demandas do público específico, incluindo indicadores de interesse ou dados de outras fontes, corroborariam para sua consolidação. Bem como a apresentação de análises estatísticas mais complementares.

Por fim, incentiva-se a replicação deste trabalho em diferentes locais e unidades da federação.

**Referências Bibliográficas**


ANDERSON, A. F.. Statistics and Massive Data Sets: One View from the Social Sciences. In *Massive Data Sets*: Proceedings of a Workshop, Committee on Applied and Theoretical Statistics, National Research Council. National Academies Press, 1996 Disponível em: < http://www.nap.edu/catalog. php?record_id=5505 > acesso em: 24/11/2016

BALTAR, ronaldo; SIQUEIRA BALTAR, cláudia . As Ciências Sociais na Era do Zettabyte. *Mediações* - Revista de Ciências Sociais, v. 18, p. 11, 2013.

BARBOZA, D. P.; GODOY, S. R.. Superando o "calcanhar metodológico"? Mapeamento e evolução recente da formação em métodos de pesquisa na pós-graduação em Ciência Política no Brasil. In: *IV Seminário Discente da Pós-Graduação em Ciência Política da USP*, 2014, São Paulo. Anais do IV Seminário Discente da Pós-Graduação em Ciência Política da USP, 2014.

BAX, Marcello P.. Introdução às linguagens de marcas. *Ci. Inf*. vol.30 no.1 Brasília Jan./Apr. 2001. Disponível em: < http://dx.doi.org/10.1590/S0100-19652001000100005 > acesso em: 24/11/2016

CÂMARA MUNICIPAL DE IJUÍ. *Câmara municipal de Ijuí.* 2016. Disponível em: < http://www.camaraijui.rs.gov.br/ > acesso em: 24/11/2016

CERICOLA, Osvaldo Vicente. *Banco de Dados Relacional e Distribuído*. 4ª ed. Rio de Janeiro: LTC, 1991 p. 115

CHANG, Winston. *shinydashboard*: Create Dashboards with "Shiny". R package version 0.5.3. 2016. Disponível em: < https://CRAN.R-project.org/package=shinydashboard > acesso em: 24/11/2016

CHENG, Joe. *bubbles*: d3 Bubble Chart htmlwidget. R package version 0.2. 2015

COX, M., ELLSWORTH, D. *Application-Controlled Demand Paging for Out-of-Core Visualization*. Report NAS-97-010, july, 1997

De MAURO, A., GRECO, M., & GRIMALDI, M.. *What is big data?* A consensual definition and a review of key research topics. In International Conference on Integrated Information (IC-ININFO 2014) AIP Conf. Proc. 1644 (pp. 97104). Madrid, Spain: AIP Publishing LLC. 2015. http://doi.org/10.1063/1.4907823

De MAURO, Andrea; GRECO, marco; GRIMALDI, Michele. *A Formal definition of Big Data based on its essential Features*. Library Review. ISBN - 65: 122–135. 2016. doi:10.1108/LR-06-2015-0061

DIEBOLD, F.X., *A Personal Perspective on the Origin(s) and Development of 'Big Data'*: The Phenomenon, the Term, and and the Discipline, Manuscript, Department of Economics, University of Pennsylvania, 2012

DUMAIS, S. Information Retrieval: Finding Needles in Massive Haystacks. In *Massive Data Sets*: Proceedings of a Workshop, Committee on Applied and Theoretical Statistics, National Research Council. National Academies Press, 1996. Disponível em: < http://www.nap.edu/read/5505/chapter/6 > acesso em: 24/11/2016

FELLOWS, Ian. *wordcloud*: Word Clouds. R packages version 2.5. 2014. Disponível em: < https://CRAN.R-project.org/package=wordcloud > acesso em: 24/11/2016

HARRISON, John. *RSelenium*: R Bindings for Selenium WebDriver. R package version 1.4.2. 2016. Disponível em: < http://ropenschi.github.io/RSelenium > acesso em: 24/11/2016

HEUSER, Carlos A. *Projeto de Banco de Dados*. Série Livros Didáticos, 6ª edição. Editora Bookman. 1998. ISBN: 979-85-7780-382-8

KITCHIN, Rob,. Big Data, new epistemologies and paradigm shifts. *Big Data & Society* Jun 2014, 1 (1); DOI: 10.1177/2053951714528481

KUNST, Joshua. *highcharter*: A Wrapper for the "Highcharts" Library. R packages version 0.4.0. Disponível em: < https://CRAN-R-project.org/package=highcharter > acesso em: 24/11/2016

Laney,D., *3-D Data Management*: Controlling Data Volume, Velocity and Variety, META Group Research Note, February 6. 2001

MANYIKA, James; CHUI, Michael; BUGHIN, Jaques; BROWN, Brad; DOBBS, Richard; ROXBURGH, Charles; BYERS, Angela Hung. *Big Data*: The next frontier for innovation, competition, and productivity. McKinsey Global Institute. 2011.

MEIRELLES, F., Silva, D. Ciência Política na era do Big data: automação na coleta de dados digitais. *Revista Ciência Política Hoje*. 2016. 2 ed. Vol. 24. p. 87-101. Disponível em < www.revista.ufpe.br/politicahoje/index.php/politica/article/viewArticle/401 > acesso em: 24/11/2016

NASCIMENTO, L. F.. A Sociologia Digital: um desafio para o século XXI. *Sociologias* (UFRGS. Impresso), v. 18, p. 216-241, 2016. Disponível em: < http://www.seer.ufrgs.br/sociologias/article/viewFile/53754/37173 > acesso em: 24/11/2016

RAMOS, M. P.. Métodos quantitativos e pesquisa em Ciências Sociais: lógica e utilidade do uso da quantificação nas explicações dos fenômenos sociais. *Revista Mediações* (UEL), v. 12, p. 55-65, 2013.

RCORE TEAM. *R*: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2015. Disponível em: < https://www.R-project.org/ > acesso em 24/11/2016

RSTUDIO TEAM. *RStudio*: Integrated Development for R. RStudio, Inc., Boston, MA 2015. disponível em < http://www.rstudio.com/ > acesso em: 24/11/2016

SOARES, Gláucio. O calcanhar metodológico da ciência política no Brasil. *Sociologia, Problemas e Práticas*, Lisboa, n. 48, p. 27-52, 2005.

WICKHAM, Hadley. *rvest*: Easily Harvest (Scrap Web Pages). R package version 0.3.2. Disponível em: < htpps://cran.r-project.org/package=rvest > acesso em: 24/11/2016

YIU, Chris. *The Big Data opportunity:* making government faster, smarter and more personal.Policy Exchange: 2012 ISBN: 978-1-9076899-22-2

**Anexo I – Código fonte para realização de *Web Scrap***

```
library(xml2)
library(dplyr)
library(rvest)
require(XML)
setInternet2()
#Conecta na internet e define a url inicial
webscrap <- function (){

url.sessoes <- "http://www.camaraijui.rs.gov.br/sessoes/"
pagina <- read_html(url.sessoes)

#########
#Pega informação das sessões (tipo, data, presentes, n de
materias)
########

sessoes <-  pagina %>%
  html_nodes(css = "#main > div.container.conteudo-principal >
div > #content > table > tbody") %>%
  html_nodes("a>p") %>%
  html_text()

sessao.tipo <- pagina %>%
  html_nodes(css = "#main > div.container.conteudo-principal >
div > #content > table > tbody") %>%
  html_nodes(css = "tr>td:nth-child(2) > ul > li:nth-
child(3)")%>%
  html_text()

sessao.tipo <- unlist(lapply(strsplit(sessao.tipo," "),'[[',3))

sessao.data <- pagina %>%
  html_nodes(css = "#main > div.container.conteudo-principal >
div > #content > table > tbody") %>%
  html_nodes(css = "tr>td:nth-child(2) > ul > li:nth-
child(1)")%>%
  html_text(trim = TRUE)

sessao.tribuna <- pagina %>%
  html_nodes(css = "#main > div.container.conteudo-principal >
div > #content > table > tbody") %>%
  html_nodes(css = "tr>td:nth-child(2) > ul > li:nth-
child(2)")%>%
  html_text(trim = TRUE)

sessao.link <- pagina %>%
```

```
  html_nodes(css = "#main > div.container.conteudo-principal >
div > #content > table > tbody") %>%
  html_nodes(css = "a")%>%
  html_attr("href")


#########
#Pega informação dos Vereadores (nome, link.foto, link.pagina, n
de materias)
########

url.ver <- "http://www.camaraijui.rs.gov.br/vereadores/"
pagina.ver <- read_html(url.ver)

vereadores <- pagina.ver %>%
                  html_nodes("#content > ul") %>%
                  html_nodes("a > h5") %>%
                  html_text(trim=TRUE)

vereador.foto <- pagina.ver %>%
                  html_nodes("#content > ul") %>%
                  html_nodes("a > img") %>%
                  html_attr("src")

vereador.link <- pagina.ver %>%
                  html_nodes("#content > ul") %>%
                  html_nodes("a") %>%
                  html_attr("href")


vereador.nome    <- lapply(strsplit(vereadores," [(]"),'[[',1)
vereador.partido <- gsub("[)]","",lapply(strsplit(vereadores,"
[(]"),'[[',2))

materias.por.autor   <- list()
materias.nom.autor   <- list()

for(i in 1:length(vereador.link)){

  print(paste("Verificando vereador",vereador.nome[i],".."))

     vereador.pagina         <- read_html(vereador.link[i])
       materias.por.autor[[i]] <- vereador.pagina %>%
                                  html_nodes("#a_materias") %>%

html_nodes("div.accordion-heading > a") %>%
                                        html_attr("href")

       materias.nom.autor[[i]] <- rep(vereador.nome[[i]],
length(materias.por.autor[[i]]))

       print("..ok")
```

```
}


#########
#Pega informação das Materias (tipo, data, presentes, n de
materias)
########

#link <- sessao.link[1]
chamada_over  <- list()
Materia_over  <- list()
sessao_over   <- list()
outrosautores <- list()

for(link in sessao.link){
  chamada     <- list()
  Materias    <- list()
  sessao      <- list()

  print(paste("verificando",sessao.data[sessao.link ==
link],".."))

  sub.sessao  <- read_html(link)

  chamada[[1]] <- sub.sessao %>%
    html_nodes("#content > h1") %>%
    html_text()
  chamada[[2]] <- sub.sessao %>%
    html_nodes("#vereadores > ul >li ") %>%
    html_nodes("a > img") %>%
    html_attr("alt")
  sessao[[1]] <- rep(chamada[[1]],length(chamada[[2]]))

  #Materias_id
  Materias[[1]] <- sub.sessao %>%
    html_nodes("#a_materias") %>%
    html_nodes("div > div.accordion-heading > a") %>%
    html_attr("href")

  sessao[[2]] <- rep(chamada[[1]],length(Materias[[1]]))

  #Autor Individual
  Materias[[4]] <- sub.sessao %>%
    html_nodes("#a_materias") %>%
    html_nodes("div > div:nth-child(2) > div > div > ul > li:nth-
child(2)") %>%
    html_text(trim=T)

  aux <- cbind(Materias[[4]][!(Materias[[4]] %in%
vereador.nome)],Materias[[1]][!(Materias[[4]] %in%
vereador.nome)])
```

```
  #Titulo materia
  Materias[[2]] <- sub.sessao %>%
    html_nodes("#a_materias") %>%
    html_nodes("div > div.accordion-heading > a") %>%
    html_text(trim = TRUE)

  Materias[[2]] <- gsub("[^[:alnum:]]"," ",Materias[[2]])
  Materias[[2]] <- substr(Materias[[2]],7,nchar(Materias[[2]])-1)

  #tipo materia
  Materias[[3]] <- sub.sessao %>%
    html_nodes("#a_materias") %>%
    html_nodes("div > div:nth-child(2) > div > ul > li:nth-
child(2) > small") %>%
    html_text()
  materias.tipo.sub.str.aux <- Materias[[3]]
  Materias[[3]] <-
substr(materias.tipo.sub.str.aux,8,nchar(materias.tipo.sub.str.au
x))

  chamada_over  <- c(chamada_over, list(chamada))
  Materia_over  <- c(Materia_over, list(Materias))
  sessao_over   <- c(sessao_over,  list(sessao))
  outrosautores <- c(outrosautores,list(aux))

  print("ok")
}

materia.id     <- unlist(sapply(Materia_over,'[[',1))
materia.titulo <- unlist(sapply(Materia_over,'[[',2))
materia.tipo   <- unlist(sapply(Materia_over,'[[',3))
materia.autor  <-
cbind(unlist(sapply(outrosautores,function(X){X[,ncol(X)-
1]})),unlist(sapply(outrosautores,function(X){X[,ncol(X)]})))
materia.sessao <- unlist(sapply(sessao_over, '[[',2))

presenca.sessao    <-  unlist(sapply(sessao_over,'[[',1))
presenca.vereador  <-  unlist(sapply(chamada_over,'[[',2))


df.sessoes <- data.frame(tipo       = sessao.tipo,
                         data       =
as.Date(sessao.data,"%d/%m/%Y"),
                         tribuna    = as.numeric(sessao.tribuna),
                         link       = sessao.link)

df.materia <- data.frame(id       = materia.id,
                         tipo     = materia.tipo,
                         sessao   = materia.sessao,
                         titulo   = materia.titulo)
```

```
df.vereadores <- data.frame(nome     =
as.character(vereador.nome),
                            partido  = vereador.partido,
                            foto     = vereador.foto)

df.Autor.Materia   <-
rbind(cbind(unlist(materias.nom.autor),unlist(materias.por.autor)
),materia.autor)

colnames(df.Autor.Materia) <- c("Autor","Mataria")

df.vereador.sessao <- data.frame(nome     = presenca.vereador,
                                 sessao   = presenca.sessao)

write.csv(df.materia,         "materias.csv",         fileEncoding
= "UTF-8")
write.csv(df.sessoes,         "sessoes.csv",          fileEncoding
= "UTF-8")
write.csv(df.vereadores,      "vereadores.csv",       fileEncoding
= "UTF-8")
write.csv(df.Autor.Materia,   "materias_autor.csv",   fileEncoding
= "UTF-8")
write.csv(df.vereador.sessao,"vereador_sessao.csv", fileEncoding
= "UTF-8")

}


system.time( webscrap())
```

**Anexo II – Código fonte para interface Shiny**

```
require(shiny)
library(shinydashboard)
library(ggplot2)
library(dplyr)
#devtools::install_github("jcheng5/bubbles")
library(bubbles)
library(highcharter)
library(tm)
library(wordcloud)
library(rvest)

data.ver <- read.csv("vereadores.csv",          sep = ",",
stringsAsFactors = FALSE, fileEncoding = "UTF-8")
data.mat <- read.csv("materias.csv",            sep = ",",
stringsAsFactors = FALSE, fileEncoding = "UTF-8", header = T)
data.ses <- read.csv("sessoes.csv",             sep = ",",
stringsAsFactors = FALSE, fileEncoding = "UTF-8")
data.mat.aut <- read.csv("materias_autor.csv",  sep = ",",
stringsAsFactors = FALSE, fileEncoding = "UTF-8")
data.ver.sec <- read.csv("vereador_sessao.csv", sep = ",",
stringsAsFactors = FALSE, fileEncoding = "UTF-8")
data.ver.sex <- read.csv("cargo_sexo.csv",      sep = ";",
stringsAsFactors = FALSE, fileEncoding = "UTF-8")

server <- function(input, output) {
  ntabs <- length(data.ver$nome)
  tabnames <- paste0("tab", 1:(ntabs))
  a <- vector("list", ntabs)

  for(i in 1:ntabs){
    a[[i]] <- menuSubItem(data.ver$nome[i], tabName =
tabnames[i], icon = icon('user'))
  }

  output$menu <- renderMenu({
    sidebarMenu(
      menuItem("Apresentação", tabName = "Apresentacao", icon =
icon("home"), selected = TRUE),
      menuItem("Composição da câmara",     tabName = "Camara",
icon = icon("university")),
      menuItem("Vereadores",   tabName = "Vereador",
               a),
      menuItem("Matérias",     tabName = "Materias",     icon =
icon("file-text")),
```

```
      menuItem("Seções",       tabName = "Secoes",       icon =
icon("clock-o")),
      menuItem("Página da Câmara", icon = icon("road"),
               href = "http://www.camaraijui.com.br/"),
      menuItem("lattes do autor", icon = icon("graduation-cap"),
               href = "http://lattes.cnpq.br/0933237287129219")
    )
  })

  output$valueboxMaterias <- renderValueBox({
    valueBox(
      paste0(length(data.mat$id)), "Metérias votadas", icon =
icon("file-text"),
      color = "purple"
    )
  })

  output$nVereadores <- renderValueBox({
    valueBox(
      paste0(length(data.ver$nome)), "é o número de vereadores",
icon = icon("users"),
      color = "purple"
    )
  })

  output$nPartidos <- renderValueBox({
    valueBox(
      paste0(length(unique(data.ver$partido))), "é o número de
partidos", icon = icon("gears"),
      color = "purple"
    )
  })

  output$repFem <- renderValueBox({
    percentMulheres <-
round(100*data.ver.sex$Eleito[2]/data.ver.sex$Eleito[3],2)
    valueBox(
      paste0(percentMulheres,"%"), "São mulheres", icon =
icon("female"),
      color = "purple"
    )
  })

  output$valueboxAutorPorMateria <- renderValueBox({

    valueBox(
      paste0(length(unique(data.mat.aut$Autor))), "Autores
diferentes", icon = icon("users"),
      color = "purple"
    )
  })
```

```
  output$valueboxAutores <- renderValueBox({
    media.AutPorMat <-
round(length(data.mat.aut[,1])/length(data.mat[,1]),2)
    valueBox(
      paste0(media.AutPorMat), "Autores por matéria em média",
icon = icon("line-chart"),
      color = "purple"
    )
  })

  output$nSessoes <- renderValueBox({
    valueBox(
      paste0(length(data.ses$X)), "sessões foram realizadas",
icon = icon("clock-o"),
      color = "purple"
  )
  })
  output$nMateriasporSessao<- renderValueBox({
    media.MatPorSes <-
round(length(data.mat$X)/length(data.ses$X),2)
    valueBox(
      paste0(media.MatPorSes), "Matérias votadas em média por
sessão", icon = icon("files-o"),
      color = "purple"
  )
  })
  output$nPresenca<- renderValueBox({
    media.VerPorSes <-
round(100*length(data.ver.sec$X)/(length(data.ses$X)*length(data.
ver$X)),2)
    valueBox(
      paste0(media.VerPorSes,"%"), "Presença média nas sessões",
icon = icon("eye"),
      color = "purple"
  )
  })


  output$WordCloudTit <- renderPlot({

    myCorpus <- Corpus(VectorSource(data.mat$titulo))
    myCorpus <- tm_map(myCorpus, content_transformer(tolower))
    myCorpus <- tm_map(myCorpus, removePunctuation)
    myCorpus <- tm_map(myCorpus, removeNumbers)
    myCorpus <- tm_map(myCorpus, removeWords, c(stopwords(kind =
"pt"),"ijuí","outras","providencia","providência","providências",
"providencias"))
    myDTM <- TermDocumentMatrix(myCorpus, control =
list(minWordLength = 1))
    m     <- as.matrix(myDTM)
    m1    <- sort(rowSums(m),decreasing = TRUE)
```

```
    wordcloud(names(m1), m1, max.words = 60, colors=brewer.pal(8,
"Dark2"))
  })

  output$TiposAutorBarra <- renderHighchart({
    autor.tipo <- data.mat %>%
      count(tipo) %>%
      ungroup() %>%
      arrange(desc(n)) %>%
      mutate(x = row_number()) %>%
      rename(nome = tipo,
             y = n) %>%
      select(y,nome) %>% list_parse()

    hcbar <- highchart() %>%
      hc_xAxis(categories = unlist(pluck(autor.tipo,2))) %>%
      hc_title(text = "Matérias apresentada por tipo") %>%
      hc_yAxis(title = "Matérias apresentada por tipo") %>%
      hc_add_series(data = autor.tipo, type = "bar", showInLegend
= FALSE,
                    name = "Tipos de Matéria")
    hcbar
  })

  output$PartidosPizza <- renderHighchart({
    partido.tipo <- data.ver %>%
      count(partido) %>%
      ungroup() %>%
      arrange(desc(n)) %>%
      mutate(x = row_number()) %>%
      rename(partido = partido,
             y = n) %>%
      select(y,partido)

    hcbar <- highchart() %>%
      hc_chart(type = "pie") %>%
      hc_title(text = "Distribuição das Cadeiras") %>%
      hc_add_series_labels_values(partido.tipo$partido,
partido.tipo$y, name = "Número de vereadores",
                                  type = "pie")

    hcbar
  })

  output$SessoesTipo <- renderHighchart({
    sessoes.tipo <- data.ses %>%
      count(tipo) %>%
      ungroup() %>%
      arrange(desc(n)) %>%
      mutate(x = row_number()) %>%
      rename(tipo = tipo,
             y = n) %>%
```

```
      select(y,tipo)

    hcbar <- highchart() %>%
      hc_chart(type = "pie") %>%
      hc_title(text = "Sessões ocorridas por tipo") %>%
      hc_add_series_labels_values(sessoes.tipo$tipo,
sessoes.tipo$y, name = "Sessões ocorridas por tipo",
                                  type = "pie")

    hcbar
  })

  output$VereadorSexo <- renderHighchart({
    Sexo.tipo <- data.ver.sex[-3,c(3,5)] %>%
      arrange(desc(Eleito)) %>%
      mutate(x = row_number()) %>%
      rename(Genero = Sexo,
             y = Eleito) %>%
      select(y,Genero)

    hcbar <- highchart() %>%
      hc_chart(type = "pie") %>%
      hc_title(text = "Distribuição das Cadeiras") %>%
      hc_add_series_labels_values(Sexo.tipo$Genero, Sexo.tipo$y,
name = "Distribuição por gênero",
                                  type = "pie")

    hcbar
  })

  output$mytabs <- renderUI({
    Tabs <- vector("list", ntabs+3)
    for(i in 1:ntabs){

      materias <- as.character(data.mat.aut[ data.mat.aut[,2] ==
data.ver$nome[i],3])
      filtro1  <- subset(data.mat, data.mat$id %in% materias)
      tab.tipo.mat <- table(filtro1$tipo)
      tab.tipo.mat <- sort(tab.tipo.mat[tab.tipo.mat > 0],
decreasing = TRUE)

      vereador.partido <- data.ver$partido[i]
      vereador.materias.n <- sum(data.mat.aut[,2] ==
data.ver$nome[i])
      vereador.presenca   <-
round(100*sum(as.character(data.ver.sec[,2]) ==
as.character(data.ver$nome[i]))/length(data.ses[,1]),2)
      materias.por.autor  <-
round(length(data.mat.aut[,1])/length(unique(data.mat.aut[,2])),2
)
```

```
      media.VerPorSes <-
round(100*length(data.ver.sec$X)/(length(data.ses$X)*length(data.
ver$X)),2)

      Tabs[[i]] <- tabItem(tabName = tabnames[i],
                           h2(data.ver$nome[i]),br(),

                           fluidRow(
                             valueBox(
                               paste0(vereador.partido), "é o seu
partido", icon = icon("gear-large"),
                               color = "purple"
                             ),
                             valueBox(
                               paste0(vereador.materias.n),
paste("Matérias apresentadas. (Média por
autor",materias.por.autor,")"), icon = icon("file-text"),
                               color = "purple"
                             ),
                             valueBox(
                               paste0(vereador.presenca,"%"),
paste0("de Presença. (Presença média é ",media.VerPorSes,"%)"),
icon = icon("eye"),
                               color = "yellow"
                             )
                             ),

                           fluidRow(
                             box(width = 4,
                                 HTML(paste0("<img
src='",data.ver$foto[i],"' align = 'middle' width = '100%'>"))
                             ),
                             box(width = 8, height = 400,
                                 title = "Proposições por tipo",
                                 bubbles(value   =
as.numeric(tab.tipo.mat),
                                         label   = gsub("de
Lei","",rownames(tab.tipo.mat)),
                                         tooltip =
rownames(tab.tipo.mat),
                                         width   = "100%",
                                         height  = "300",
                                         color   =
rainbow(length(tab.tipo.mat),
alpha=NULL)[sample(length(tab.tipo.mat))])
                             ))
      )
    }

    Tabs[[i+1]] <- tabItem(tabName = "Materias",
                           fluidRow(
                             column(width = 1),
```

```
                              column(width = 10,
                              h2("Matérias"),
                              p("Estatísticas sobre as matérias
apresentadas na câmara.")
                            ),
                            fluidRow(
                              # Dynamic infoBoxes
                              valueBoxOutput("valueboxMaterias"),
                              valueBoxOutput("valueboxAutores"),

valueBoxOutput("valueboxAutorPorMateria")
                            ),
                            fluidRow(
                              column(width = 6,

highchartOutput("TiposAutorBarra")
                              ),
                              column(width = 6,
                                     span("Palavras frequentes nos
títulos das matérias", style="font-weight:bold"),br(),
                                     plotOutput("WordCloudTit")
                              )
                            ),
                            column(width = 1)
    ))

    Tabs[[i+2]] <- tabItem(tabName = "Camara",
                            fluidRow(
                              h2("Composição da Câmara"),
                              br(),
                              fluidRow(
                                column(width = 1),
                                column(width = 10,
                                valueBoxOutput("nPartidos"),
                                valueBoxOutput("nVereadores"),
                                valueBoxOutput("repFem")
                                ),
                                column(width = 1)
                              ),
                              fluidRow(
                                column(width = 6,
                                highchartOutput("PartidosPizza")),
                                column(width = 6,
                                highchartOutput("VereadorSexo"))
                              )
                            ))

    Tabs[[i+3]] <- tabItem(tabName = "Secoes",
                            fluidRow(
                              h2("Sessões"),
                              p("Estatísticas sobre as sessões"),
                              fluidRow(
```

```
                                column(width = 1),
                                column(width = 10,
                                       valueBoxOutput("nSessoes"),
valueBoxOutput("nMateriasporSessao"),
                                       valueBoxOutput("nPresenca")
                                ),
                                column(width = 1)
                              ),
                              fluidRow(
highchartOutput("SessoesTipo")
                                )
                            ))
    do.call(tabItems, Tabs)
  })
}


ui <- dashboardPage(
  dashboardHeader(title = "Menu"),
  dashboardSidebar(
    sidebarMenu(style = "position:fixed; overflow: visible",
      sidebarMenuOutput("menu")
    )
  ),
  dashboardBody(
    uiOutput('mytabs'),
    tabItem(tabName = "Apresentacao",
            fluidRow(
              column(width = 1),
              column(width = 10,
                     h2("Apresentação"),br(),
                     div(style = "text-align: justify",
                         p("Esta aplicação é parte do Trabalho de
Conclusão de curso do Bacharelado em Estatística da Universidade
Federal do Rio Grande do Sul, apresentado em 2016, intitulado",
                           span("Aproximando a estatística e as
Ciência Política na Era do Big Data.", style = "font-style:
italic"),
                           "Ela tem o objetivo de apresentar
estatísticas descritivas da atuação legislativa do município de
ijuí de forma simples e direta."),
                         p("A aplicação funciona em duas etapas
principais. A coleta de dados e a apresentação de estatísticas
descritivas de maneira acessível ao grande público. A coleta de
dados foi realizada utilizando técnicas de webscrap, e apenas a
composição da câmara por gênero foi obtida a partir do TSE."),
                         p("A aplicação foi desenvolvida
utilizando a linguagem R, com auxílio do pacote Shiny by
Rstudio."),
```

```
                        p("Todos os dados apresentados são
públicos e foram extraídos do sítio oficial da câmara de
vereadores de Ijuí e do repositório do TSE.")
                      ),
                      br(),
                      p(style= "text-align:center",
                        a(href =
"http://www.camaraijui.rs.gov.br/",
                          img(
src='http://static.camaraijui.rs.gov.br/imagens/zc1_w300_h250__c3
7beff059d0f2d20ded9a708f7555aef0c1f7618f51511fc6b2ab2ee4ad58bf_z6
o7p.png', height = '120', width = '140')
                        ),
                        a(href =
'http://www.tse.jus.br/eleicoes/estatisticas/repositorio-de-
dados-eleitorais',

img(src="http://metodologiapolitica.com/wp-
content/uploads/2012/10/tse.jpeg", height = '120', width = '120')
                        ),
                        a(href = 'http://shiny.rstudio.com/',

img(src="https://d21ii91i3y6o6h.cloudfront.net/gallery_images/fro
m_proof/9306/large/1447177198/rstudio-hex-shiny-dot-psd.png",
height = '120', width = '120')
                        )
                      ),
                      br(),br(),br()
              ),
              column(width = 1)
            )
    )

  )
)

shinyApp(ui = ui, server = server)
```