



Evento	Salão UFRGS 2013: SIC - XXV SALÃO DE INICIAÇÃO CIENTÍFICA DA UFRGS
Ano	2013
Local	Porto Alegre - RS
Título	Desenvolvimento de ferramentas para auxiliar análise semântica no minerador de texto Sobek
Autor	RAFAEL FINKELSTEIN
Orientador	ELISEO BERNI REATEGUI

O objetivo da ferramenta de mineração de textos Sobek é servir de modo a facilitar na compreensão dos principais temas de um texto ou de um conjunto de textos. A ferramenta faz uma análise estatística sobre o conteúdo dos textos e gera uma base de conhecimento que pode ser treinada e aprimorada. Essa base é composta pelos principais conceitos contidos nos textos. Após a análise, a base resultante é utilizada na criação de grafos que demonstram visualmente as informações extraídas.

Para aprimorar e melhor selecionar os nodos do grafo, foram adicionados ao conjunto do Sobek duas novas ferramentas, o Tree Tagger e o Thesaurus, essa primeira é um classificador de palavras quanto à classe, sendo possível trabalhar com os conceitos e adicioná-los ou excluí-los do grafo considerando a sua classe. A segunda ferramenta, Thesaurus, citada anteriormente, analisa todos os conceitos e busca sinônimos, fundindo os nodos em apenas um nodo com a palavra de maior frequência de aparição no texto fornecido pelo usuário.

A metodologia usada no desenvolvimento das duas ferramentas foi muito parecida, foi dividida em 3 etapas básicas. A mesma metodologia foi usada nas duas ferramentas desenvolvidas, entretanto a primeira a ser concluída foi o Tree Tagger só então foi iniciado o projeto da segunda ferramenta.

Para a primeira etapa do que resultou no Tree Tagger e o Thesaurus, foi buscado o melhor método de análise gramatical e comparação de sinônimos disponíveis e a compatibilidade com a aplicação, a segunda etapa foi implementar o método selecionado, separadamente do software principal, verificando a sua completude e se supria as necessidades inicialmente pensadas e, por fim, testá-lo até que estivesse estável o suficiente para a realização da terceira etapa, a qual era basicamente anexação ao software principal.

Após as 2 ferramentas estarem funcionais no programa principal, visando deixar o executável da aplicação o menor possível implementamos uma versão que buscava as informações das ferramentas na internet, mais precisamente no servidor na UFRGS.

Até o momento estas duas ferramentas estão completas, agora estamos pensando em uma continuação da melhoria do Sobek na parte dos grafos, contudo estamos apenas na parte de projeto, sem nada concluído.