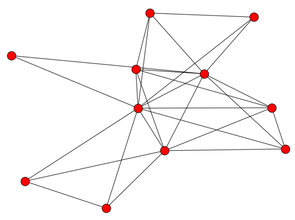

Redes de Relacionamentos Criminais na DeepWeb

Alexandre Albuquerque (IC) & Sebastián Gonçalves(Orientador)

Bacharelado em Física Computacional
UFRGS/Instituto de Física
husky.hannuky@gmail.com

1. Introducão

MINERAÇÃO de textos é um processo que utiliza algoritmos computacionais para que seja analisada uma enorme coleção de documentos texto, da qual sejam extraídas informações valiosas para os mineradores. Muitos algoritmos utilizam da abordagem estatística para que palavras importantes sejam capturadas destas coleções. Atualmente, a mineração de texto ou de dados (forma mais geral) é importante pois com o avanço da computação há um alto volume de conteúdo possível para analisar.

2. Objetivos

O Objetivo final do trabalho é analisar um conjunto de textos resultantes de uma operação da Polícia Federal sobre pedofilia na DeepWeb. O conjunto dos textos ocupa aproximadamente 1 Terabyte de dados, de tal forma que, manualmente, a análise não poderia ser executada. Portanto, ferramentas de mineração de dados são necessárias.
O objetivo desta fase da pesquisa foi desenvolver ferramentas e algoritmos que nos permitissem analisar grandes quantidades de textos.

3. Metodologia

O Programa SOBEK [5], desenvolvido por pesquisadores do Instituto de Informática da UFRGS, é uma das ferramentas que analisa textos e relaciona as palavras identificadas como palavras importantes, construindo uma rede semântica entre elas como resultado final do programa. Esta ferramenta mostrou-se interessante, porém seu uso é limitado a pequenos textos inseridos em um formulário Java on-line. Por isso, na primeira fase do trabalho, usamos Python para replicar os resultados do programa SOBEK em diversos textos testes, obtendo uma rede semelhante final. Para nosso algoritimo realizar, de fato, as estatísticas das palavras ele também executa o tratamento de todas as palavras para que sejam analisadas de forma precisa.

1. Todos os caracteres devem ser colocados no padrão minúsculo.
2. Artigos, pronomes, preposições e outras palavras e/ou caracteres sem valor semântico são retirados por meio de uma lista conhecida como “stopwords”.
3. As palavras selecionadas são aquelas cuja ocorrencia no texto (f), é tal que $\sqrt{f_{max}} <= f <= f_{max}$, onde f_{max} é a maior ocorrencia de uma palavra no texto. Caso contrário, todas as palavras apareceriam na rede.
4. Duas palavras selecionadas são conectadas se a distância entre elas forem de até 3 palavras quaisquer do texto.

Com base na biblioteca NLTK (Natural Language Tool Kit) [1] do Python foi possível verificar as palavras mais frequentes dos textos, além de ser possível verificar outros dados estatísticos, como suas probabilidades. E, com o auxílio da biblioteca Networkx, foi possível gerar arquivos de extensão ‘XML’, para que fossem lidas pelo software Cytoscape. Este arquivo é o responsável por criar graficamente as conexões entre as palavras e assim a rede ser apresentada de forma mais elegante.

4. Resultados

PAra verificar as redes de palavras com maiores quantidade de palavras e ter uma análise qualitativa, foi escolhido o texto conjunto de todos os deputados votantes favoravelmente ao impeachment da ex-presidenta Dilma Roussef, captado de [7], que coincide com a análise feita por cientistas políticos. [4]

Além da rede, foi dimensionada a fonte de cada palavra com a frequência de aparição neste mesmo texto conjunto.

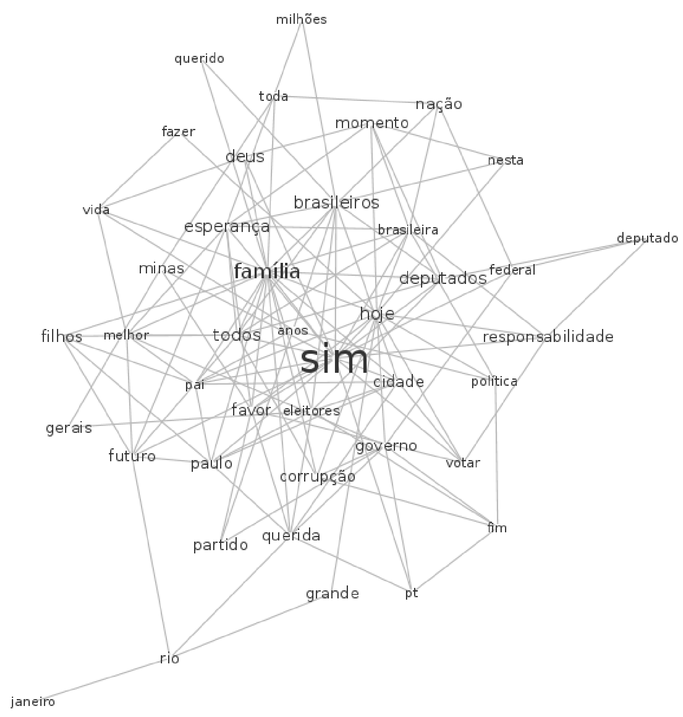

Fig. 1: *Rede de palavras gerada a partir do discurso de todos deputados de posição favorável ao Impeachment de Dilma Rousseff.*

Outro exemplo verificado foi uma rede feita a partir de reportagem [6] sobre a morte do cantor Belchior. É possível identificar o assunto da reportagem mesmo sem ter lido a notícia completamente.

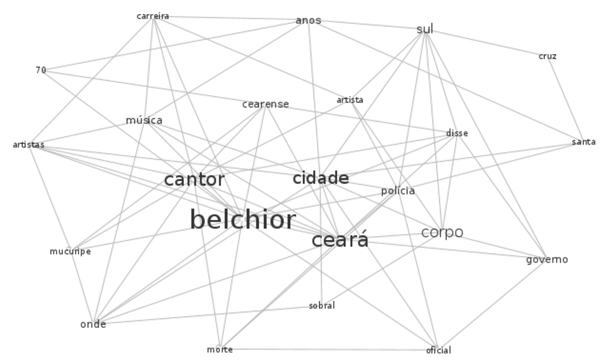

Fig. 2: *Rede de palavras gerada a partir de uma notícia.*

5. Conclusão

A Representação gráfica das redes semânticas nos trazem uma idéia abrangente e rápida de qual conteúdo iremos encontrar. Além de qual importância de cada termo pelo tamanho de sua fonte na imagem. Para 1 terabyte de dados, ela será útil para verificar características essenciais destas informações como assuntos, sujeitos e suas possíveis ligações e importância. Com grandes quantidades de dados, em forma de texto, a rede gerada tem um potencial de apresentar um panorama geral de forma satisfatória.

6. Discussão e Perspectivas

PAra uma análise mais profunda, outras ferramentas serão utilizadas. Word2Vec [2] é um algoritmo que cria uma representação vetorial para palavras de um texto, como a que se pode ver na Fig. 3. A idéia é que palavras parecidas, que estão sobre um mesmo contexto, se posicionem próximas. Uma funcionalidade para este algoritmo, principalmente em uma rede de relacionamentos criminais, é que termos desconhecidos possam ser vinculados contextualmente a termos já conhecidos.

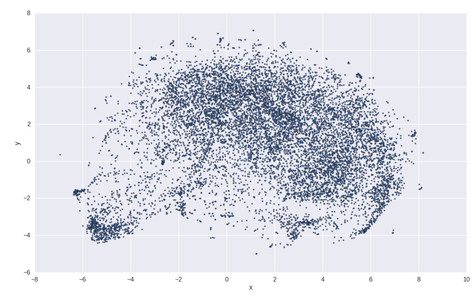

Fig. 3: *Representação gráfica de todas as palavras do livro Game of Thrones realizada pelo algoritmo de Word2Vec.*

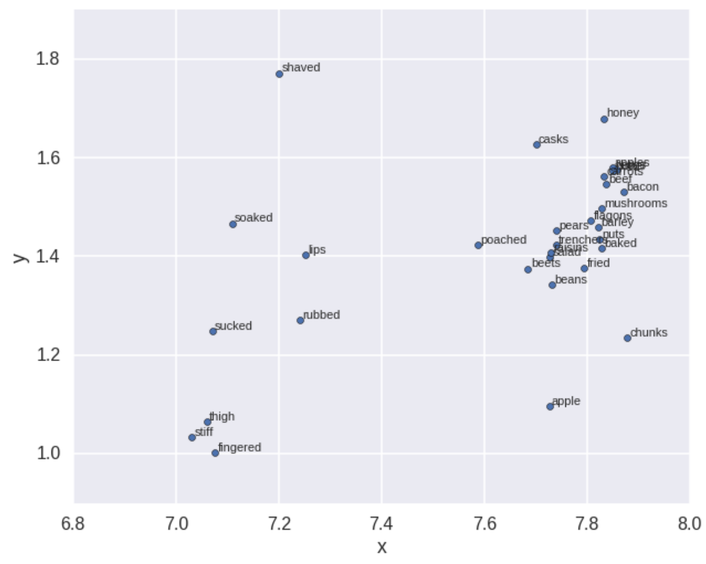

Fig. 4: *Aproximação da Fig. 3 em uma região específica. Palavras com contexto relacionado à alimentação.*

Uma outra forma de abordagem sobre a informação das palavras vem da Teoria da Informação de Claude Shannon [3], cuja idéia chave seria o cálculo da Entropia da Informação. Com essa abordagem, o valor de informação das palavras não seria somente pela frequência, mas também pelo posicionamento delas no texto geral comparado ao mesmo texto ordenando aleatoriamente. A perspectiva é que com uma medida de informação para cada palavra seja possível identificar quando dois textos provavelmente estejam falando sobre o mesmo assunto.

References

[1] S.Bird, E.Klein, E. Loper. 2009. Natural Language Processing With Python. O’Reilly Media: Sebastopol, CA.

[2] Tomas Mikolov , Ilya Sutskever , Kai Chen , Greg Corrado , Jeffrey Dean, Distributed representations of words and phrases and their compositionality, Proceedings of the 26th International Conference on Neural Information Processing Systems, p.3111-3119, December 05-10, 2013, Lake Tahoe, Nevada

[3] Shannon, C. E., Prediction and entropy of printed english, Bell System Technical Journal 30 (1951) 50–64.

[4] http://www.gazetaonline.com.br/noticias/politica2016/04cientistas-politicos-criticam-argumentos-de-deputados-em-votacao-do-impeachment-1013939231.html

[5] http://sobek.ufrgs.br/index.html

[6] https://g1.globo.com/ceara/noticia/cantor-cearense-belchior-morre-aos-70-anos-no-rio-grande-do-sul.ghtml

[7] http://www2.camara.leg.br/