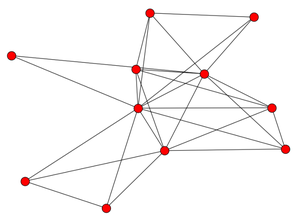

Redes de Relacionamentos Criminais na DeepWeb

Alexandre Albuquerque (IC) & Sebastián Gonçalves(Orientador)

Bacharelado em Física
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 classificar, de forma automática e consistente, mensagens entre usuários da DeepWeb.

3. Metodologia

O O algoritmo do Naives Bayes, é um dos mais simples algoritmos baseado no teorema de Bayes. Este algoritmo é caracterizado pela classe independente condicional, pois o valor de atributo de uma classe é independente dos valores dos outros atributos. Esse algoritmo tem uma alta acurácia e um menor tempo de execução, quando aplicado a grandes banco de dados. Existem dois modelos usados no aprendizado: modelo Naives Bayes Binário, onde considera a presença dos termos nos documentos; e o Naives Bayes Multinomial, que é o utilizado neste trabalho, onde são consideradas as frequências dos termos dentro de cada documento.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

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. Aplica-se o processo conhecido como ”stemming”. Consiste em reduzir palavras flexionadas (ou às vezes derivadas) ao seu tronco (stem), base ou raiz, geralmente uma forma da palavra escrita.

Com base na biblioteca NLTK (Natural Language Tool Kit) [1] do Python foi possível realizar o processo de stemming das palavras e assim reduzir o custo computacional.
Exemplo de Stemming:

Original	Radical
Presidente, Presidência, Presidencialismo	Presid
Maravilhosamente, Maravilhosa, Maravilha	Maravilh
Amoroso, Amor, Amado	Am

4. Resultados

PAra ter uma análise qualitativa do método, foi escolhido o texto conjunto de todos os discursos dos deputados votantes no impeachment da ex-presidenta Dilma Roussef, captado de [2]. Foi separado para cada deputado um arquivo específico de seu discurso. A base de treinamento é feita com escolhas aleatórias e também sequenciais dos discursos. A base de treinamento é sempre feita com o mesmo número de discursos favoráveis e contrários ao impeachment.

Os conjuntos de discursos foram analisados e depois confrontados com os resultados de predição, verificando se havia correspondência ou não com o votos dos deputados. Então, ao aumentar o tamanho da base de discursos.

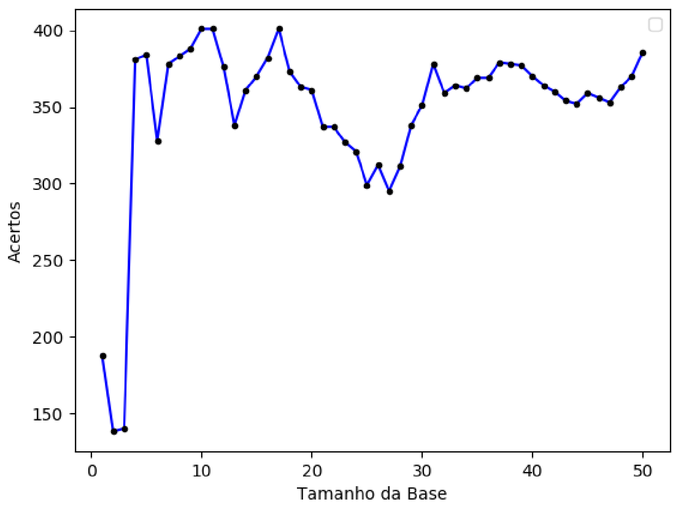

Fig. 1: *Número de acertos do algoritmo pelo número de discursos utilizados como base de treinamento.*

Foram verificadas 100 amostras aleatórias, utilizando como tamanho de base 5 discursos favoráveis e 5 contrários, conforme figura 2. E, igualmente feita para amostras com tamanho de base 10, conforme figura 3.

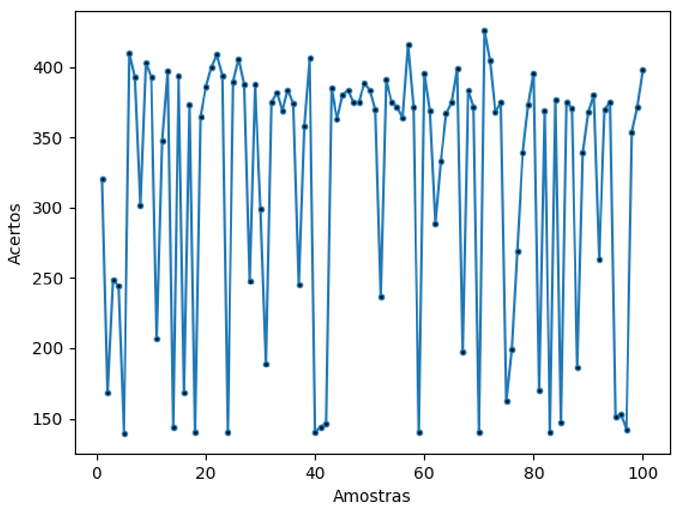

Fig. 2: *Cada amostra contém 5 discursos favoráveis ao impeachment e 5 contrários.*

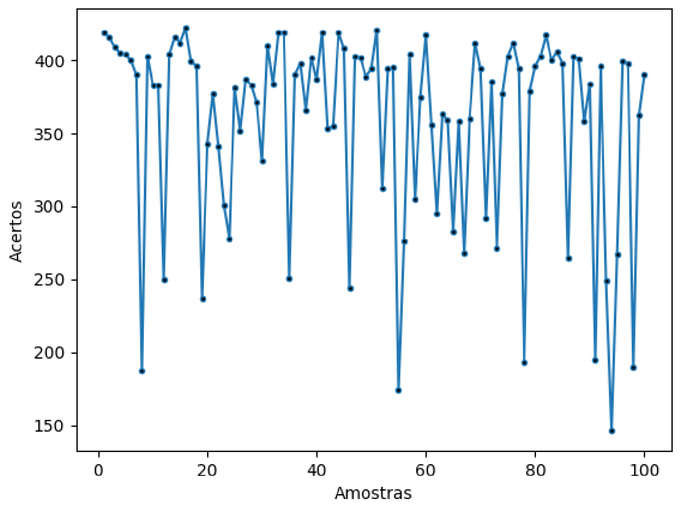

Fig. 3: *Cada amostra contém 10 discursos favoráveis ao impeachment e 10 contrários.*

Tamanho da Base	Desvio Padrão (σ)	Média de Acertos
5	95.18	317.23
8	82.27	342.15
10	66.33	358.42

5. Conclusão

A variabilidade dos termos dos discursos de base de treinamento influenciam de maneira significativa na predição. Porém quanto maior a base, menor é seu desvio e maior sua média de acertos. Alguns termos ganham maior importância que outros no discurso total, medido por suas frequencias, e nem sempre estes termos estão presentes nos discursos tomados como base.
Foi observado também que há um salto grande e rápido na acurácia do método. Com menos de 10% de toda a base se consegue um número de acertos satisfatório.

6. Discussão e Perspectivas

PAra uma análise mais profunda, deve ser replicada várias amostras aleatórias, aumentando o tamanho da base, afim de obter um leque do número óptimo (aquele em que há um rápido crescimento no número de acertos.
Para análise da rede criminal, é poupado tempo de classificação dos crimes envolvidos (dentre outras possíveis classificações) a partir de uma pequena base já classificada analiticamente. Isso traz maior agilidade aos investigadores que podem focar em usuários específicos da rede, por exemplo.

References

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

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

[3] http://cis.poly.edu/ mleung/FRE7851/f07/naiveBayesianClassifier.pdf.

[4] http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.65.9324rep