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 foi a análise de um conjunto de textos resultantes de uma operação da Polícia Federal sobre pedofilia na DeepWeb. O conjunto dos textos tem como base 26.000 postagens em um fórum. Ferramentas de mineração de dados são necessárias para a análise semântica das palavras e para a medida das frequências das palavras.

3. Metodologia

ONa primeira fase do trabalho, usamos Python e a biblioteca NLTK para realizar, de fato, as estatísticas das palavras. Há que executar 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”.

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.
Na segunda fase é utilizado a base de texto já tratada para a representação de palavras que liga a compreensão humana da linguagem à de uma máquina. Essa representação é conhecida como embeddings. Word Embedding é muito usado em diferentes tarefas de processamento de linguagem natural tais como reconhecimento da linguagem natural, similaridade entre palavras, classificação de documentos, parsing, análise de sentimentos.

4. Resultados

PAlavras com maiores frequências, depois da retirada do conjuto de stopwords, podem nos mostrar algumas características do conjunto total de textos.
É observado um número bem maior para o termo ”http” em relação aos outros termos.

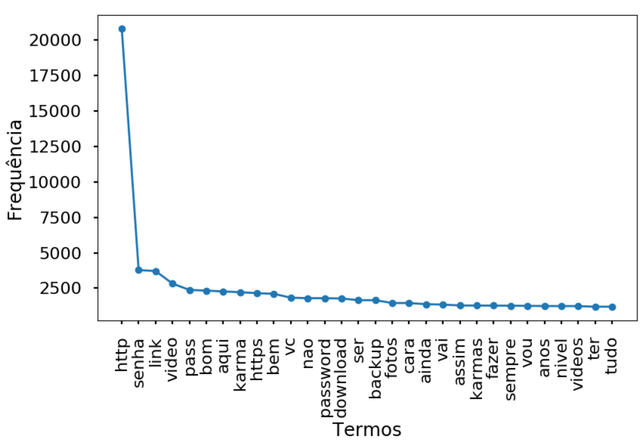

Fig. 1: *Palavras com maiores frequências*

Word2Vec [2] é um algoritmo que cria uma representação vetorial para palavras de um conjunto de textos. 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 à termos já conhecidos.

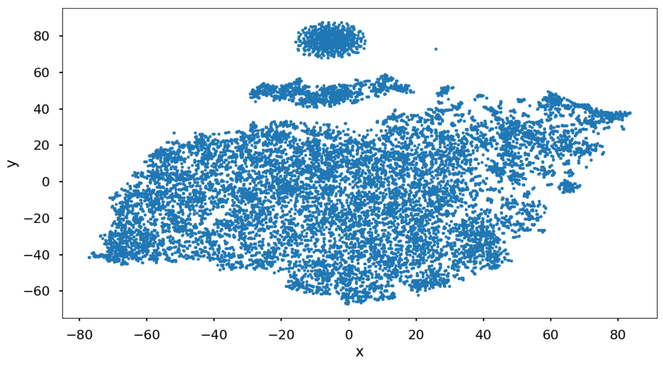

Fig. 2: *Word Embedding da base de texto criado utilizando Word2Vec.*

Quando a representação é visualizada com coordenadas específicas, podemos identificar regiões que nos mostram contextos similares. É o caso da figura 3 e 4. Na figura 3 nos é mostrado palavras com referência a mídias digitais.

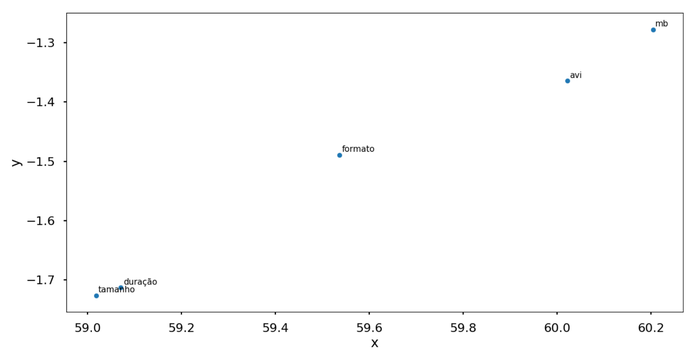

Fig. 3: *Um zoom no mostra termos com contextos similares.*

Na figura 4 é observada algumas exclamações positivas.

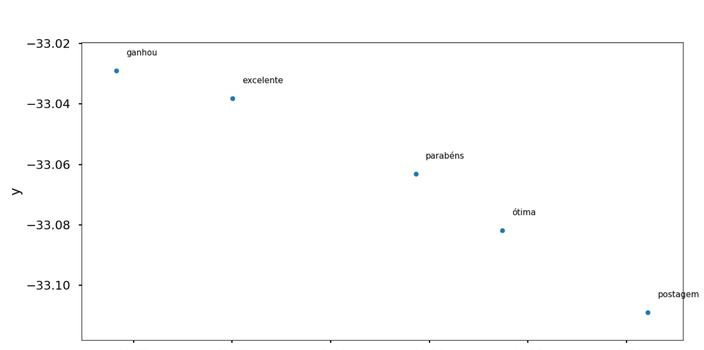

Fig. 4: *Nesta região palavras de congratulações são representadas próximas.*

Como cada palavra é representada em um vetor, é possível verificar e criar uma tabela de palavras similares com outra que queremos comparar pela similaridade de cossenos. Na tabela verificamos algumas palavras próximas a palavra ”pedofilia”.

Similares: Pedofilia	Valor (similaridade de cossenos)
organização	0.9089035987854004
regra	0.9040614366531372
brasileiros	0.9039642810821533
aceitação	0.9014383554458618
saibam	0.9000749588012695
analise	0.8987252712249756
interação	0.8953568935394287
solução	0.8895344138145447
participação	0.8862175941467285
necessidade	0.8844061493873596

5. Conclusão

AFrequência das palavras nos mostra que há uma extensa quantidade de material digital sendo trocado. Termos como ”http” e ”https” geralmente são para links de mídias digitais e que são protegidos por senhas.

As técnicas de Word Embeddings, e mais especificamente o algoritmo do Word2Vec, mapeiam termos de maneira que auxiliam na investigação de crimes cometidos através de sites e fóruns de discussões. A mineração de dados aliada a redes neurais ampliam largamente o horizonte investigativo.

6. Discussão e Perspectivas

COmo exemplo interessante de análise, pode-se verificar que Word Embeddings também é atualmente utilizado em tradutores. Na figura 5, é observado uma região específica em que palavras do idioma alemão são representadas em uma região próxima dentro da base.

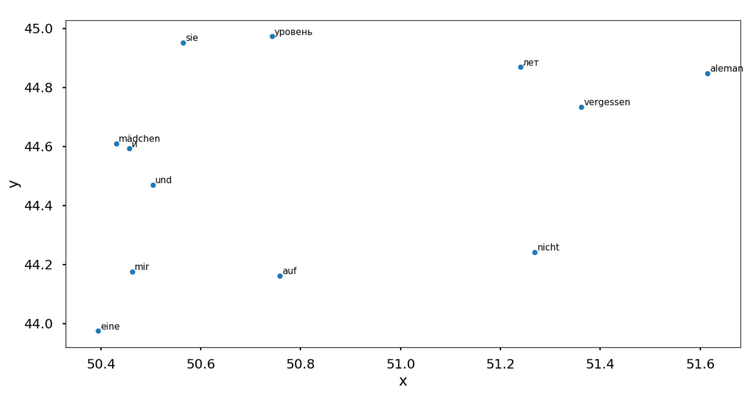

Fig. 5: *Termos agrupados pelo idioma.*

Uma outra forma de abordagem investigativa pode ser através de mensagens por áudio ou escutas. Já existem diversos mecanismos que transformam áudios em textos. Torna-se mais eficiente não somente a transcrição como também a análise desse áudio e por fim a investigação.

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.