Inteligencia Artificial

Arthur, Danilo, Vinicius

Março 2020

Content

- Introdução
- Motivação
- Metodologia
- Pré-Processamento
- Banco de Dados
- Treinamento
- Conclusão
- Trabalhos Futuros

Introdução

- O presente trabalho tem como objetivo a análise de sentimentos de notícias que impactem no valor de ações.
- Ou seja, classificar as noticias para verificar se existe correlação entre a alta e baixa da bolsa e o sentimento daquela noticia.
- A linguagem de programação utilizada no trabalho foi Python com o auxilio de várias bibliotecas como a **Coogro4py**, **Selenium**, **nltk**, **pandas** entre outras.

Motivação e Metodologia

- A motivação do trabalho foi o interesse em analisar a bolsa de valores e trabalhar com processamento natural de linguagem que é uma área da IA que está crescendo muito no mercado de trabalho.
- Nosso trabalho foi dividido em 4 etapas.
- Inicialmente trabalhamos com um banco de dados inicial onde realizamos nosso pré-processamento.
- Feito isso, treinamos nossa rede com o banco de dados todo processado.
- Por fim, analisamos os resultados.

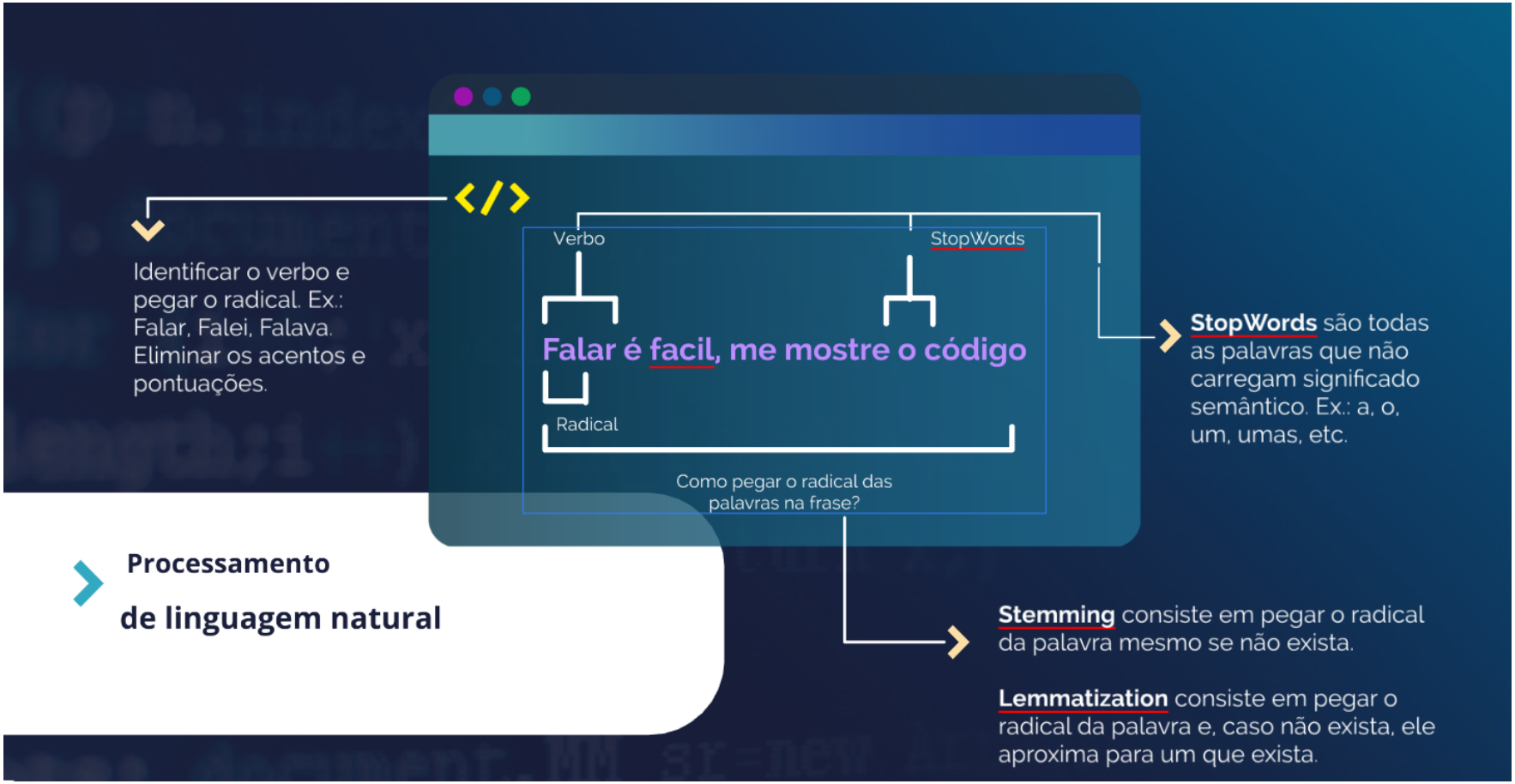
Pré-Processamento

Processamento Natural de Linguagem

- Para o nosso trabalho, como estamos trabalhando com texto, existe um tipo específico de pré-processamento.
- Este pré-processamento é chamado de Processamento Natural de Linguagem.
- Dado uma base de dados que seja formada por texto, o processamento natural de linguagem consiste em analisar apenas as palavras importantes nessa base de dados.
- Além disso, temos que converter a *string* do texto em uma representação numérica para podermos avaliar o sentimento daquela frase.

- **E quais foram as etapas de Pré-Processamento?**
- Primeiramente eliminamos as **StopWords** que são palavras que não contém valor semântico para determinada frase.
- Para isso transformamos todo o texto em letras minúsculas e com o auxilio da biblioteca **nltk** removemos a maioria das StopWords.
- Feito isso, também precisamos eliminar os prefixos e sufixos das palavras pois, uma frase pode conter as palavaras: infeliz, feliz, infelizmente.
- Daí, precisamos apenas dos radicais das palavras.

Pré-Processamento



Banco de Dados

- Nosso banco de dados para teste era composto por notícias com a maioria sendo sobre política.
- As notícias eram classificadas em Positivo (PO), Negativo (NG) ou Neutro(NE).
- A1, A2, A3 e A4 são avaliações de jornalistas e **M** a moda das avalia ções.

Banco de Dados

Out[5]:

	Text	M	A1	A2	A3	A4
0	Os candidatos à Presidência mais bem posiciona...	'NE',	'NE',	'NG',	'PO',	'',
1	Os dados foram disponibilizados neste sábado (...	'',	'',	'',	'',	'',
2	A lei permite que os presidenciáveis arrecadem...	'',	'NE',	'',	'NE',	'',
3	As maiores doadoras para a campanha de Dilma s...	'NE',	'NE',	'NE',	'NE',	'NG',
4	Os maiores colaboradores da campanha de Aécio ...	'NE',	'NE',	'NE',	'NE',	'NG',
5	A candidatura do PSB, inicialmente com Eduardo...	'NE',	'NE',	'NE',	'NE',	'NG',
6	A prestação de contas registrou também as desp...	'',	'',	'',	'',	'',
7	Em dois meses, a candidatura de Dilma foi que ...	'NE',	'NE',	'NG',	'NE',	'NG',
8	saiba mais Dilma, Aécio e Campos obtêm 94% das...	'',	'',	'',	'NE',	'',
9	Em MT, Ibope aponta: Dilma, 36%, Marina, 23%, ...	'',	'',	'',	'PO',	'PO',
10	Em MS, Ibope aponta: Marina, 35%, Dilma, 33%, ...	'',	'',	'',	'PO',	'PO',

Treinamento

- Para a criação da nossa base de dados extraímos notícias do site do G1 utilizando bibliotecas de automação e scraping.
- **Motivação para o uso das bibliotecas**
  Como precisávamos de uma base de dados grande para o treinamento ficaria muito cansativo e demorado extrair notícias uma a uma, então decidimos usar as bibliotecas *Selenium* (Automação) e *BeautifulSoup* (Scraping).
- **Selenium**
  O *Selenium* nos ajudou bastante na parte da extração das notícias pois o site no qual trabalhamos nesse processo precisa ser carregado por completo para podermos coletar todos os links das notícias. Com isso usamos a biblioteca citada acima para automatizar o acesso ao site. Em resumo o *Selenium* é usado basicamente para acessar o site e carregar todo o **JavaScript** e capturar links de próximas páginas.

Treinamento

- **BeautifulSoup**
  Após carregarmos todo o **JavaScript** do site e pegarmos os links para a próxima página o *BeautifulSoup* entra em ação acessando todo HTML para pegarmos o conteúdo do site. O *BeautifulSoup* aplicado em um site te retorna todo o HTML porém não queremos tudo, para isso tivemos que estudar toda a estrutura do site para poder pegar apenas o que nos interessa.
- **Sobre a Base**
  No final de todo processo obtivemos todas as notícias relacionadas a determinada empresa do site do G1. Aplicamos esse processo nas empresas: Petrobras e Vale. Com a Petrobras obtivemos um total de 1010 notícias. Já com a Vale obtivemos 324 notícias. Após isso transformamos todos os dados obtidos em um arquivo **CSV** para usarmos no treinamento.

Treinamento

BERT-Bidirectional Encoder Representations from Transformers

- Para esse treinamento utilizamos da arquitetura para processamento natural de linguagem do Google que tem o nome Bidirectional Encoder Representations from Transformers (BERT)
- Nessa arquitetura o *encouder* utilizado para transformar as palavras em vetores é um já previamente treinado.
- Para a utilização dessa arquitetura utilizamos a biblioteca K-train. Toda arquitetura é pré-definida, porém para esse treinamento utilizamos uma função de *callBack* de *Early Stop* com 5 épocas, o que significa que caso não haja uma alterção muito significativa na rede em 5 épocas, o treinamento é paralisado. Isto é preciso para evitar o super-ajustamento.
- Para o treinamento foi utilizado um split com 80% para treinamento e 20% para teste.

Conclusão

- Analisamos as notícias em um intervalo de tempo e caso não houvesse notícia naquele determinado dia definimos o sentimento da notícia como neutro.
- Um problema que obtivemos foi que em determinado dia poderia acontecer de haver mais de uma notícia, para isso definimos o sentimento das notícias naquele dia com a moda.
- Definimos uma porcentagem para avaliarmos se a bolsa subiu, desceu ou manteve em determinado dia.
- E por fim, analisamos a correlação entre o sentimento e a variação como pode ser observado nas figuras a seguir.

Conclusão

Matriz final de resultados e Matriz de Correlação

Out[232]:

Date	Sentimento_Geral	Vari
2019-01-02	PO	Subiu
2019-01-03	NG	Desceu
2019-01-04	NE	Subiu
2019-01-07	PO	Subiu
2019-01-08	NE	Desceu
...	...	...
2019-11-26	PO	Desceu
2019-11-27	NG	Manteve
2019-11-29	PO	Desceu
2019-12-02	NE	Desceu
2019-12-03	NE	Manteve

233 rows × 2 columns

	Sentimento_Geral	Vari
Sentimento_Geral	1.000000	-0.085707
Vari	-0.085707	1.000000

Trabalhos Futuros

- Um próximo passo para nosso projeto seria analisar os dados utilizando *shift* de um ou mais dias.
- Poderíamos também re-treinar a nossa rede com outro banco de dados com um viés abaixo do atual e não só com o titulo da notícia, mas também com o corpo da notícia.
- Mudar a porcentagem para saber se a bolsa subiu, desceu ou se manteve também poderia melhorar o resultado.
- Outra possível melhora seria extrair dados de sites específicos sobre economia para melhorar os nossos resultados.

The End