FACULDADE ANGLO-AMERICANO DE FOZ DO IGUAÇU

JORGE SILVA
SILVA MATHEUS

TCC SOBRE JAVA

FOZ DO IGUAÇU

2016

JORGE SILVA
SILVA MATHEUS

TCC SOBRE JAVA

Trabalho de conclusão de curso apresentado como requisito obrigatório para obtenção do título de Bacharel em Ciência da Computação da Faculdade Anglo-Americano de Foz do Iguaçu.

Orientador: Msc. Nome do Orientador

Coorientador: Prof. Esp. Nome do Coorientador

FOZ DO IGUAÇU

2016

SobreNome, Nome1 Nome2
TCC sobre Java / Jorge Silva
Silva Matheus – Foz do Iguaçu, 2016.
29 p. : il.

Orientador: Msc. Nome do Orientador

– Faculdade Anglo-Americano de Foz do Iguaçu. Curso de Ciência da Computação, 2016.

1. Palavra-chave1. 2. Palavra-chave2. I. Msc. Nome do Orientador. II. Faculdade Anglo-Americano de Foz do Iguaçu. III. Curso de Ciência da Computação. IV. TCC sobre Java

CDU

TERMO DE APROVAÇÃO

Jorge Silva
Silva Matheus

TCC sobre Java

Trabalho de conclusão de curso apresentado como requisito obrigatório para obtenção do título de Bacharel em Ciência da Computação da Faculdade Anglo-Americano de Foz do Iguaçu, pela seguinte banca examinadora:

Msc. Nome do Orientador
Faculdade Anglo-Americano de Foz do Iguaçu
(Orientador)

Prof. Banca 2
Faculdade Anglo-Americano de Foz do Iguaçu

Prof. Banca 3
Faculdade Anglo-Americano de Foz do Iguaçu

Foz do Iguaçu, 2016

Dedico este trabalho a meus pais,
Sr... e Sra...
que, com muito amor, me ensinaram os valores da vida.

AGRADECIMENTOS

Primeiramente agradeço a Deus por sua graça e salvação.

À minha família, por terem me proporcionado ...

À ... por me mostrar o caminho da ...

Aos meus grandes amigos....

A todos os professores que fizeram parte desta importante etapa da minha vida.

Aos meus orientadores.....

"Once you have eliminated the impossible, whatever remains,
however improbable, must be the truth."
- Mr. Spock, Star Trek (2009)

RESUMO

A rede social ....

Palavras-chaves: Dados. Data Mining. Twitter. Python.

ABSTRACT

The social network ....

Keywords: Data. Data Mining. Twitter. Python.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE CÓDIGOS

LISTA DE GRÁFICOS

LISTA DE ABREVIATURAS

API	*Application Programming Interface* - Interface de Programação de Aplicação
BMP	*Windows Bitmap*
CGI	*Common Gateway Interface* - Interface Comum de Entrada[1]
CSV	*Comma-Separated Values* - Valores Separados Por Vírgula[1]
FTP	*File Transfer Protocol* - Protocolo de Transferência de Arquivos
GIF	*Graphics Interchange Format* - Formato Para Intercâmbio de Gráficos[1]
HTTP	*Hypertext Transfer Protocol* - Protocolo de Transferência de Hipertexto
HTTPS	*Hyper Text Transfer Protocol Secure* - Protocolo de Transferência de Hipertexto Seguro
JPG	*Joint Photographic Experts Group*
PDF	*Portable Document Format* - Formato de Documento Portátil[1]
PNG	*Portable Network Graphics* - Rede Portável de Gráficos[1]
URL	*Uniform Resource Locator* - Localizador Padrão de Recursos
XHTML	*eXtensible Hypertext Markup Language* - Linguagem de Marcação de Hipertexto Extensiva
XML	*eXtensible Markup Language* - Linguagem de Marcação Extensiva
YML	*Yet Another Markup Language* - Uma Outra Linguagem de Marcação[2]

[1] Tradução do autor

Lista de símbolos

Γ	Letra grega Gama
Λ	Lambda
ζ	Letra grega minúscula zeta
$\in$	Pertence

SUMÁRIO

1 INTRODUÇÃO

Redes sociais se tornaram um termo comum e uma chave fundamental para o estilo de vida moderno. Hoje em dia, ...

1.1 JUSTIFICATIVA

A rede social *Twitter* é um excelente ponto de partida para a mineração de dados em redes sociais, ...

A rede social possui um total de 289 milhões de usuários ativos no mundo inteiro, totalizando 58 milhões de *tweets* por dia (BRAIN, 2016).

1.2 OBJETIVOS

1.2.1 Objetivo Geral

Este trabalho tem como objetivo principal utilizar técnicas e algoritmos de *data mining*, para a análise e mineração de dados provenientes da rede social *Twitter*, utilizando os recursos e bibliotecas que a linguagem de programação Python possui.

1.2.2 Objetivos Específicos

- Identificar os conceitos sobre KDD e *data mining*;
- Descrever as técnicas de *data mining*;
- Explorar as funcionalidades das bibliotecas de mineração e visualização da linguagem Python;
- Examinar e utilizar a API da rede social *Twitter* para a coleta de dados;
- Encontrar padrões em dados provenientes do *Twitter*;
- Compreender e aplicar técnicas para apresentação e visualização de informações geográficas encontradas nos dados coletados;
- Apresentar testes e resultados obtidos da análise e mineração dos dados.

1.3 CRONOGRAMA DE ATIVIDADES

As atividades a serem executadas no decorrer do projeto visando o êxito do mesmo, estão listados a seguir e especificados em meses na Tabela 2:

TABELA 1 – Cronograma

Mês - Ano	**08/15**	**09/15**	**10/15**	**11/15**	**12/15**	**02/16**	**03/16**	**04/16**	**05/16**
Estudo e Pesquisa	X	X	X	X	X	X	X	X	
Análise de Requisitos	X	X	X	X	X	X	X	X	
Geração do Documento	X	X	X	X	X	X	X	X	X
Implementação				X	X	X	X	X	X
Testes				X	X	X	X	X	X
Elaboração de Artigos			X	X	X			X	X
Apresentação de Resultados					X				X

FONTE: Autor

- Estudo e Pesquisa: aquisição dos conhecimentos pertinentes e necessários para o desenvolvimento do projeto;
- Análise de Requisitos: levantamento dos requisitos do projeto;
- Geração do Documento: desenvolvimento das documentações para especificação do projeto;
- Implementação: desenvolvimento dos códigos para a análise de dados;
- Testes: execução dos testes que irão garantir a qualidade das informações a serem geradas;
- Elaboração de Artigos: parte do tempo destinado ao projeto será para desenvolver artigos visando a publicação em eventos da área;
- Apresentação de Resultados: etapas destinadas à apresentação dos resultados parciais e finais.

TABELA 2 – Cronograma de execução

Mês - Ano	**08/15**	**09/15**	**10/15**	**11/15**	**12/15**	**02/16**	**03/16**	**04/16**	**05/16**
Estudo e Pesquisa	X	X	X	X	X	X	X	X	
Análise de Requisitos	X	X	X	X	X	X	X	X	
Geração do Documento	X	X	X	X	X	X	X	X	X
Implementação				X	X	X	X	X	X
Testes				X	X	X	X	X	X
Elaboração de Artigos			X	X	X			X	X
Apresentação de Resultados					X				X

FONTE: Autor

1.4 ORGANIZAÇÃO DO TRABALHO

Além deste capítulo Tabela 1, este trabalho é composto de mais seis capítulos.

O Capítulo 2 apresenta os trabalhos que são referências para este estudo.

Os fundamentos teóricos, como os conceitos de *data mining* e base para o entendimento do tema proposto, estão descritos no Capítulo 3.

No Capítulo 5 são apresentadas as fases do desenvolvimento ....

Os resultados obtidos e a apresentação de planilhas e gráficos das soluções desenvolvidas são apresentados no Capítulo 6.

Por fim, a conclusão deste trabalho se dá no Capítulo 7, onde são abordadas e analisadas as dificuldades, além de determinar as possibilidades para trabalhos futuros.

2 REVISÃO BIBLIOGRÁFICA

Alguns trabalhos serviram como ajuda e inspiração para este estudo. Porém durante o período de busca por bibliografias .... Capítulo 1.

De acordo com Lemos (2003), um dado se transforma em informação ....

Em seu estudo, Lemos (2003) aborda duas técnicas ...

...

O reconhecimento de padrões permite .... (SILVA; BOSCARIOLI; PERES, 2003). Para o desenvolvimento ....

3 FUNDAMENTAÇÃO TEÓRICA

A mineração de dados é um assunto totalmente interdisciplinar, ...

3.1 DESCOBERTA DE CONHECIMENTO EM BASE DE DADOS E *DATA MINING*

Muitas pessoas tratam a mineração de dados ... O processo de KDD é demonstrado através da Figura 1 e, posteriormente, listada como uma sequência interativa e iterativa dos seguintes passos:

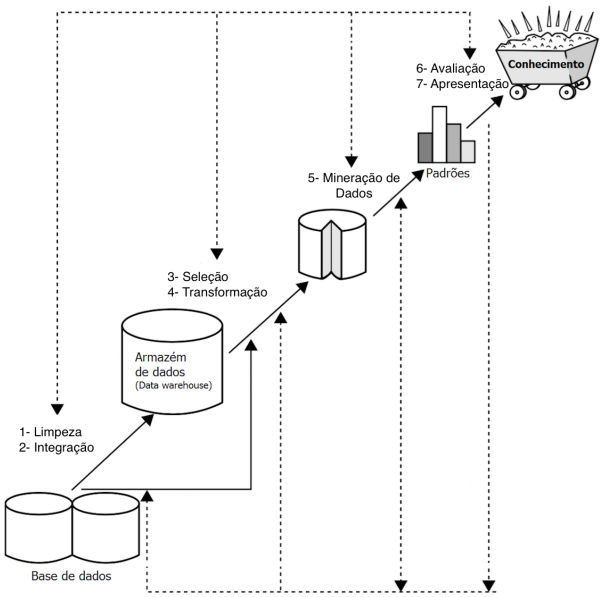

FIGURA 1 – Etapas do processo de KDD
FONTE: Adaptado de Han et al. (2012)

1. *Data cleaning* (Limpeza de dados);
2. *Data integration* (Integração de dados);
3. *Data selection* (Seleção de dados);
4. *Data transformation* (Transformação de dados);
5. *Data mining* (Mineração de dados);
6. *Pattern evaluation* (Avaliação de padrões);
7. *Knowledge presentation* (Apresentação de conhecimento).

É importante notar que algum dos processos acontecem na mesma etapa: Limpeza e integração; Seleção e transformação; Avaliação e apresentação.

De acordo com Brachman et al. (1996 apud FAYYAD et al., 1996-b), as etapas são interativas ....

4 MATERIAIS E MÉTODOS

Após a revisão bibliográfica de outros estudos e os fundamentos teóricos necessários ....

Este capítulo apresenta os materiais e métodos utilizados ...

4.1 TECNOLOGIAS E FERRAMENTAS

Tecnologias e ferramentas para a implementação de *scripts* e utilização dos algoritmos.

4.1.1 Bibliotecas da Linguagem Python

Um dos grandes diferenciais da linguagem Python é o seu enorme conjunto de bibliotecas para soluções de diversos problemas.

4.1.1.1 *Biblioteca NumPy*

NumPy é o pacote fundamental para computação científica em Python. É o acrônico para *Numerical Python*. Esta biblioteca provê:

4.1.1.2 Biblioteca *pandas*

A biblioteca *pandas* ... (MCKINNEY, 2013):

Uma simples *Series* é formado por uma única matriz de dados, conforme a Figura 2.

```
[In [5]: obj = Series([4, 7, -5, 3])

[In [6]: obj
 Out[6]:
 0    4
 1    7
 2   -5
 3    3
 dtype: int64
```

FIGURA 2 – Exemplo de uma *Series*
FONTE: McKinney (2013)

DataFrame representa uma tabela...

4.1.2 Rede Social *Twitter*

Para definir o que seria ...

4.1.2.1 API do *Twitter*

Twitter é caracterizado como um serviço ...

4.1.2.2 Bibliotecas Para o Consumo de Dados da API do *Twitter*

O acesso a API acontece através da criação ....

O CÓDIGO 1 exemplifica o consumo da API segundo Tweepy (2009).

CÓDIGO 1 – Acesso à API do *Twitter*

```
1  import tweepy
2
3  auth = tweepy.OAuthHandler(consumer_key, consumer_secret)
4  auth.set_access_token(access_token, access_token_secret)
5
6  api = tweepy.API(auth)
7
8  public_tweets = api.home_timeline()
9  for tweet in public_tweets:
10     print tweet.text
```

5 IMPLEMENTAÇÃO DAS TÉCNICAS

Este capítulo tem como finalidade apresentar, com um maior nível de detalhamento as técnicas utilizadas neste trabalho, com o objetivo de se atingir as metas propostas já descritas na Seção 1.2.

5.1 COLETA DE DADOS

Uma característica comentada anteriormente ....

As primeiras linhas mostradas no CÓDIGO 2 servem para ...

CÓDIGO 2 – *Script* coletar-hashtags.py

```
1  from tweepy.streaming import StreamListener
2  from tweepy import OAuthHandler
3  from tweepy import Stream
4
5  access_token = "131556934-LrYRiXzAL3QcRyFN0fdN53EDWhNGfZFnVX59NCnT"
6  access_token_secret = "JraMtps5lB98d8XoelAF71KHn8ZQ4nshdoSKiFlTz6OHd"
7  consumer_key = "P4XZ2GUkeqdhIlQMOredBuW05"
8  consumer_secret = "r5TPb2UcM8bzxq7t5zflRPMHUrCfwNG4GRuVPXypowrpHhTmue"
9
10
11 class StdOutListener(StreamListener):
12
13     def on_data(self, data):
14         print data
15         return True
16
17     def on_error(self, status):
18         print status
19
20
21 if __name__ == '__main__':
22
23     l = StdOutListener()
24     auth = OAuthHandler(consumer_key, consumer_secret)
25     auth.set_access_token(access_token, access_token_secret)
26     stream = Stream(auth, l)
27
28     stream.filter(track=['ImpeachmentDay', 'NaoVaiTerGolpe', 'ForaDilma'↩
       ])
```

...

O comando *stdout* permite redirecionar a saída do código anterior, no caso a execução do *script* coletar-hashtags.py, para um novo arquivo ou um arquivo já existente, conforme ilustrado pela Figura 3.

```
scripts git:(master) ✗
› python coletar-hashtags.py > ../data/coleta-impeachment.json
```

FIGURA 3 – Execução do *script* para coleta de dados
FONTE: Autor

...

5.2 ANÁLISE DE DADOS

Após a coleta dos dados foi gerado, então, um arquivo ....

6 ANÁLISE DOS RESULTADOS

Este capítulo tem como finalidade apresentar os resultados obtidos através das implementações demonstrados no Capítulo 5.

6.1 APRESENTAÇÃO DOS RESULTADOS

Após o mapeamento das informações do *DataFrame*, ... conjunto de dados coletados, Gráfico 1.

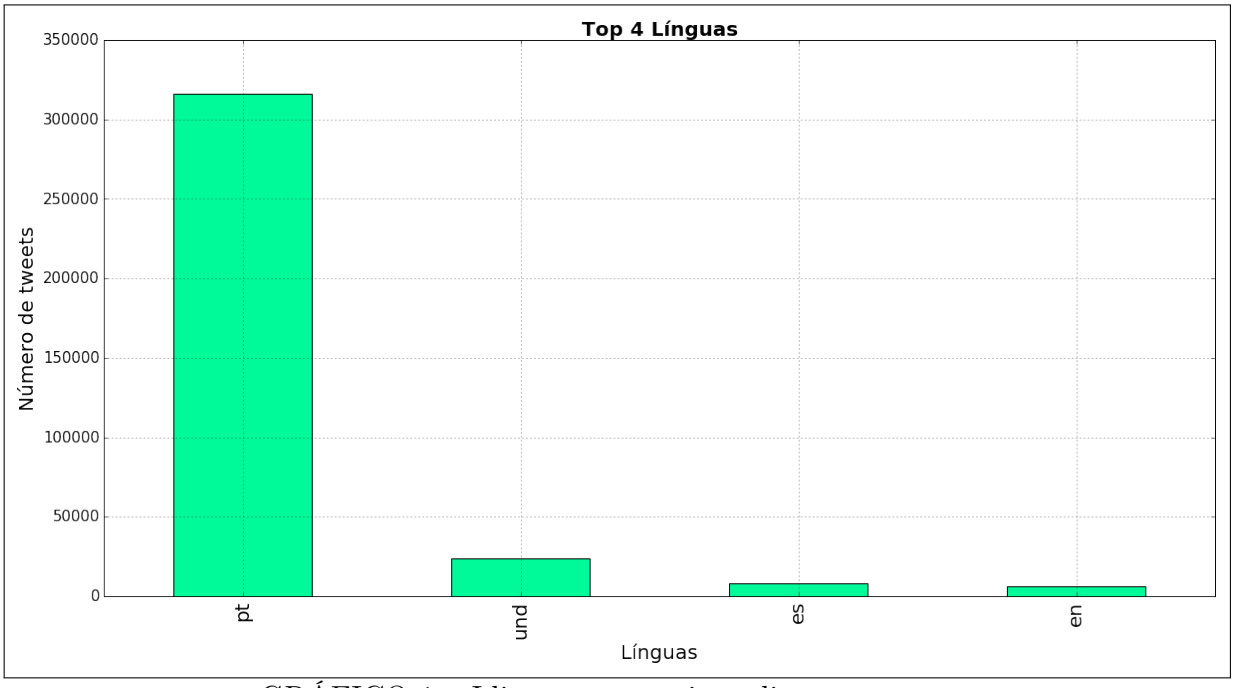

GRÁFICO 1 – Idiomas que mais realizaram *tweets*
FONTE: Elaborado pelo autor

7 CONCLUSÕES E SUGESTÕES PARA FUTUROS TRABALHOS

7.1 CONCLUSÕES

O uso das bibliotecas que Python oferece para a mineração de dados...

- Resgatar o objetivo
- Comentar as ferramentas estudadas
- Comentar as ferramentas utilizadas
- Breve resumo dos resultados
- Pontos positivos e negativos (O fato de não ter o perfil real)

7.2 SUGESTÕES PARA FUTUROS TRABALHOS

REFERÊNCIAS

BRACHMAN, R. J. et al. The process of knowledge discovery in databases. 1996. Acesso em 23 de outubro de 2015. Disponível em: <https://www.aaai.org/Papers/Workshops/1994/WS-94-03/WS94-03-001.pdf>.

BRAIN, S. **Twitter Statistics**. 2016. Acesso em 20 de abril de 2016. Disponível em: <http://www.statisticbrain.com/twitter-statistics/>.

FAYYAD, U. et al. Advances in knowledge discovery in data mining. 1996–b.

HAN, J. et al. **Data Mining: Concepts and Techniques**. [S.l.]: Elsevier, 2012.

LEMOS, E. P. **Análise de crédito bancário com o uso de data mining: redes neurais e árvores de decisão**. Tese (Doutorado) — Universidade Federal do Paraná, 2003.

MCKINNEY, W. **Python for Data Analysis**. [S.l.]: O'Reilly, 2013.

SILVA, M. P. da; BOSCARIOLI, C.; PERES, S. M. Análise de logs da web por meio de técnicas de data mining. 2003.

TWEEPY, D. **Biblioteca Tweepy - 3.5.0**. 2009. Acesso em 03 de abril de 2016. Disponível em: <http://tweepy.readthedocs.org/en/v3.5.0/index.html>.