Trabalho Prático - Banco de dados não relacionais

Aluno: Eduardo Augusto Nascimento Lima

Matrícula: 71727

Enunciado do trabalho

Coletar informações de rede sociais ou importar dados externos e armazenar 1M de dados em um banco NoSQL.

Extrair informações do tipo:

- Termos mais frequentes;

- Volume x dia;

- Volume x hora do dia

Sofwtares aplicados

- Ambiente de desenvolvimento integrado (IDE) ==> Jupyter Notebook 3.5 (python);

- Banco de dados não relacionais (NoSQL) ==> Mongodb 3.2.10;

- Pacotes requeridos ==> pip;

- Comandos requeridos ==> pip install tweepy==3.3.0 / pip install pymongo

- módulos python ==> tweepy, Datetime e Json;

Palavras chaves para pesquisa

Palavras chaves para pesquisa

Foram escolhidos as seguintes palavras chaves, a partir do assunto mais recente sobre indicação do Mr. Scott Pruitt para administrar Agência de Proteção Ambiental (EPA) no EUA.

Fonte: https://www.theguardian.com/us-news/2016/dec/07/trump-scott-pruitt-environmental-protection-agency (https://www.theguardian.com/us-news/2016/dec/07/trump-scott-pruitt-environmental-protection-agency)

Keywords:

'***Donald Trump***' ***(Novo presidente eleito nos Estados Unidos)***

'***Scott Pruitt***' ***(Indicação para ministerio do meio ambiente)***

'***administrator of the Environmental***' ***(ministerio do meio ambiente)***

'***climate***' ***(termo climatico)***

'***fuel industry***' ***(Indústria de combustíveis)***

'***air pollution***' ***(poluição do ar)***

'***pollution***' ***(poluição)***

'***EPA***' ***(Agência de Proteção Ambiental)***

'***Barack Obama***' ***(Ultimo presidente eleito nos Estados Unidos)***

Código Fonte

Coletando Dados do Twitter

In [11]:
```
# Importando os módulos Tweepy, Datetime e Json

from tweepy.streaming import StreamListener
from tweepy import OAuthHandler
from tweepy import Stream
from datetime import datetime
from pymongo import MongoClient
import json


# Definições de acesso ao API Twitter
```

```
# Consumer Key
consumer_key = "mUZEhiqD4n2OwtWmEPDgpENwr"

# Consumer Secret
consumer_secret = "j9NcQ8bM5yPQC9tX9ALaTRHfufqDcjrvDVW28IKuJQ09OHIjci"

# Access Token
access_token = "809188036667965441-S9M5AzTINYLpJtZ3HTEOqSND4C4v0UX"

# Access Token Secret
access_token_secret = "bbY22BM8Lcd6ld0X84f2sGzMOh87Iq792xxtJP5zho83q"

# Criando as chaves de autenticação
auth = OAuthHandler(consumer_key, consumer_secret)

auth.set_access_token(access_token, access_token_secret)

# Criando uma classe para capturar os stream de dados do Twitter e arm
azenar no MongoDB
class MyListener(StreamListener):

    def on_data(self, dados):
        tweet = json.loads(dados)
        created_at = tweet["created_at"]
        id_str = tweet["id_str"]
        text = tweet["text"]
        obj = {"created_at":created_at,"id_str":id_str,"text":text,}
        tweetind = col.insert_one(obj).inserted_id
        print (obj)
        return True

# Criando o objeto pythonlistener
mylistener = MyListener()


# Criando o objeto stream
mystream = Stream(auth, listener = mylistener)

# Criando a conexão ao MongoDB
client = MongoClient('localhost', 27017)

# Criando o banco de dados Trabalho_Pratico
db = client.twitterdb

# Criando a collection "twitter"
col = db.tweets

# Criando uma lista de palavras chave para filtrar os Tweets
keywords = ['Donald Trump', 'Scott Pruitt', 'administrator of the Envi
ronmental', 'climate', 'fuel industry','air pollution','pollution', 'E
PA', 'Barack Obama']

mystream.filter(track=keywords)
```

```
# Encerrando o stream de dados do Twitter
mystream.disconnect()
```

```
{'id_str': '812447975632936960', 'created_at': 'Sat Dec 24 00:01:06 +00
00 2016', '_id': ObjectId('585dbac3efd71702f79d9ecb'), 'text': 'RT @dga
rdner: Too young to remember the Cold War? For a sense of what a nuclea
r arms race feels like, watch The Day After. https://t.co/b3d…'}
{'id_str': '812447976131989504', 'created_at': 'Sat Dec 24 00:01:06 +00
00 2016', '_id': ObjectId('585dbac3efd71702f79d9ecc'), 'text': 'RT @vet
eranjames1: @bessbell Look bess! i just ordered the greatest Mug of Don
ald Trump from&gt;&gt; https://t.co/z5scukf5Pn'}
{'id_str': '812447975435804673', 'created_at': 'Sat Dec 24 00:01:06 +00
00 2016', '_id': ObjectId('585dbac3efd71702f79d9ecd'), 'text': 'GO #sco
ttpruitt https://t.co/1YJ5BQ6wMR'}
{'id_str': '812447975540748288', 'created_at': 'Sat Dec 24 00:01:06 +00
00 2016', '_id': ObjectId('585dbac3efd71702f79d9ece'), 'text': 'RT @nig
gi: Hillary Clinton hat die Wahl mit knapp drei Millionen Stimmen Vorsp
rung vor Donald Trump verloren. https://t.co/vOHfySAWPP'}
{'id_str': '812447976291377152', 'created_at': 'Sat Dec 24 00:01:06 +00
00 2016', '_id': ObjectId('585dbac3efd71702f79d9ecf'), 'text': 'RT @jho
ugh80: Great read. The true story of how Teen Vogue got mad, got woke,
and began terrifying men like Donald Trump https://t.co/H2wi7…'}
{'id_str': '812447977625092096', 'created_at': 'Sat Dec 24 00:01:07 +00
00 2016', '_id': ObjectId('585dbac3efd71702f79d9ed0'), 'text': "RT @NPR
: The Rockettes To Perform At Donald Trump's Inauguration, Whether They
Like It Or Not https://t.co/Zj6rHLW2cd"}
{'id_str': '812447977646260224', 'created_at': 'Sat Dec 24 00:01:07 +00
00 2016', '_id': ObjectId('585dbac3efd71702f79d9ed1'), 'text': 'Climate
Link to Glacier Retreat Now Irrefutable https://t.co/VRvv09np1z'}
{'id_str': '812447978669473795', 'created_at': 'Sat Dec 24 00:01:07 +00
00 2016', '_id': ObjectId('585dbac3efd71702f79d9ed2'), 'text': 'RT @Cro
wdFundGurus: Check out "Donald Trump Your President" #Trump2016 #TrumpT
rain by Rick Poppe - https://t.co/j52uUugGVO'}
{'id_str': '812447979772620801', 'created_at': 'Sat Dec 24 00:01:07 +00
00 2016', '_id': ObjectId('585dbac3efd71702f79d9ed3'), 'text': "Alec Ba
ldwin Offers to Sing AC/DC's 'Highway to Hell' at Donald Trump's Inaugu
ration https://t.co/h7VXTTaDaS"}
{'id_str': '812447979852431360', 'created_at': 'Sat Dec 24 00:01:07 +00
00 2016', '_id': ObjectId('585dbac3efd71702f79d9ed4'), 'text': 'RT @CEP
Calgary: #Climatechange could have devastating impact on #globalfisheri
es | You could see fish in #YVR you never have before https:/…'}
{'id_str': '812447979831492608', 'created_at': 'Sat Dec 24 00:01:07 +00
00 2016', '_id': ObjectId('585dbac3efd71702f79d9ed5'), 'text': 'RT @jbo
uie: .@SandyDarity\'s response to "My President Was Black" is worth you
r full attention. https://t.co/PoYr6dijdO'}
{'id_str': '812447980318035968', 'created_at': 'Sat Dec 24 00:01:07 +00
00 2016', '_id': ObjectId('585dbac4efd71702f79d9ed6'), 'text': 'Astrona
ut and climate scientist Piers Sellers dies at 61 https://t.co/HRBMtUPH
o7 via @houstonchron'}
{'id_str': '812447982843031552', 'created_at': 'Sat Dec 24 00:01:08 +00
00 2016', '_id': ObjectId('585dbac4efd71702f79d9ed7'), 'text': 'Donald
Trump recently called for the criminalization of burning the American F
lag. Freedom of speech or illegal? https://t.co/U1dRtXTWXf'}
```

{'id_str': '812447983228882944', 'created_at': 'Sat Dec 24 00:01:08 +00
00 2016', '_id': ObjectId('585dbac4efd71702f79d9ed8'), 'text': 'RT @Cro
wdFundGurus: Check out "Donald Trump Your President" #Trump2016 #TrumpT
rain by Rick Poppe - https://t.co/j52uUugGVO'}
{'id_str': '812447984122142720', 'created_at': 'Sat Dec 24 00:01:08 +00
00 2016', '_id': ObjectId('585dbac4efd71702f79d9ed9'), 'text': 'RT @Cro
wdFundGurus: Check out "Donald Trump Your President" #Trump2016 #TrumpT
rain by Rick Poppe - https://t.co/j52uUugGVO'}
{'id_str': '812447984499757056', 'created_at': 'Sat Dec 24 00:01:08 +00
00 2016', '_id': ObjectId('585dbac4efd71702f79d9eda'), 'text': 'RT @El_
Universal_Mx: Benjamín Netnayahu, acusó a Barack Obama de conspirar con
los palestinos contra #Israel https://t.co/orWZ2UZEMi'}
{'id_str': '812447984499773440', 'created_at': 'Sat Dec 24 00:01:08 +00
00 2016', '_id': ObjectId('585dbac5efd71702f79d9edb'), 'text': 'RT @Cro
wdFundGurus: Check out "Donald Trump Your President" #Trump2016 #TrumpT
rain by Rick Poppe - https://t.co/j52uUugGVO'}
{'id_str': '812447985250410496', 'created_at': 'Sat Dec 24 00:01:08 +00
00 2016', '_id': ObjectId('585dbac5efd71702f79d9edc'), 'text': 'RT @Cro
wdFundGurus: Check out "Donald Trump Your President" #Trump2016 #TrumpT
rain by Rick Poppe - https://t.co/j52uUugGVO'}
{'id_str': '812447986655498240', 'created_at': 'Sat Dec 24 00:01:09 +00
00 2016', '_id': ObjectId('585dbac5efd71702f79d9edd'), 'text': 'RT @Cro
wdFundGurus: Check out "Donald Trump Your President" #Trump2016 #TrumpT
rain by Rick Poppe - https://t.co/j52uUugGVO'}
{'id_str': '812447986445758464', 'created_at': 'Sat Dec 24 00:01:09 +00
00 2016', '_id': ObjectId('585dbac5efd71702f79d9ede'), 'text': 'RT @Cro
wdFundGurus: Check out "Donald Trump Your President" #Trump2016 #TrumpT
rain by Rick Poppe - https://t.co/j52uUugGVO'}
{'id_str': '812447986806423552', 'created_at': 'Sat Dec 24 00:01:09 +00
00 2016', '_id': ObjectId('585dbac5efd71702f79d9edf'), 'text': 'RT @Cro
wdFundGurus: Check out "Donald Trump Your President" #Trump2016 #TrumpT
rain by Rick Poppe - https://t.co/j52uUugGVO'}
{'id_str': '812447986781392896', 'created_at': 'Sat Dec 24 00:01:09 +00
00 2016', '_id': ObjectId('585dbac5efd71702f79d9ee0'), 'text': 'America
ns Still Split On Climate Change https://t.co/E1lpI8xJFv'}
{'id_str': '812447986697588736', 'created_at': 'Sat Dec 24 00:01:09 +00
00 2016', '_id': ObjectId('585dbac5efd71702f79d9ee1'), 'text': 'RT @Cro
wdFundGurus: Check out "Donald Trump Your President" #Trump2016 #TrumpT
rain by Rick Poppe - https://t.co/j52uUugGVO'}
{'id_str': '812447987100172288', 'created_at': 'Sat Dec 24 00:01:09 +00
00 2016', '_id': ObjectId('585dbac5efd71702f79d9ee2'), 'text': 'RT @Cro
wdFundGurus: Check out "Donald Trump Your President" #Trump2016 #TrumpT
rain by Rick Poppe - https://t.co/j52uUugGVO'}
{'id_str': '812447987683053568', 'created_at': 'Sat Dec 24 00:01:09 +00
00 2016', '_id': ObjectId('585dbac5efd71702f79d9ee3'), 'text': 'RT @Cro
wdFundGurus: Check out "Donald Trump Your President" #Trump2016 #TrumpT
rain by Rick Poppe - https://t.co/j52uUugGVO'}
{'id_str': '812447987662077952', 'created_at': 'Sat Dec 24 00:01:09 +00
00 2016', '_id': ObjectId('585dbac5efd71702f79d9ee4'), 'text': 'RT @Cro
wdFundGurus: Check out "Donald Trump Your President" #Trump2016 #TrumpT
rain by Rick Poppe - https://t.co/j52uUugGVO'}
{'id_str': '812447987548884992', 'created_at': 'Sat Dec 24 00:01:09 +00
00 2016', '_id': ObjectId('585dbac5efd71702f79d9ee5'), 'text': 'RT @Cro
wdFundGurus: Check out "Donald Trump Your President" #Trump2016 #TrumpT

```
rain by Rick Poppe - https://t.co/j52uUugGVO'}
{'id_str': '812447988601626624', 'created_at': 'Sat Dec 24 00:01:09 +00
00 2016', '_id': ObjectId('585dbac6efd71702f79d9ee6'), 'text': 'RT @lau
renduca: Eat shit, @realDonaldTrump. https://t.co/ZAVif26SYI'}
{'id_str': '812447989868285955', 'created_at': 'Sat Dec 24 00:01:09 +00
00 2016', '_id': ObjectId('585dbac6efd71702f79d9ee7'), 'text': 'RT @Cro
wdFundGurus: Check out "Donald Trump Your President" #Trump2016 #TrumpT
rain by Rick Poppe - https://t.co/j52uUugGVO'}
```

Criando um Indice

In [6]:
```
%%bash

mongo

use twitterdb

db.tweets.createIndex({"text":1})
```

```
MongoDB shell version: 3.2.10
connecting to: test
switched to db twitterdb
{
        "createdCollectionAutomatically" : false,
        "numIndexesBefore" : 1,
        "numIndexesAfter" : 2,
        "ok" : 1
}
bye
```

Obter termos mais frequentes (top 500+)

In [11]:
```
# Trabalhando com módulo Pandas para datasets e Importando o módulo Sc
ikit-Learn

from pymongo import MongoClient
import pandas as pan
from sklearn.feature_extraction.text import CountVectorizer


# Conexão ao MongoDB
client = MongoClient('localhost', 27017)

# o banco de dados Trabalho_Pratico
db = client.twitterdb

# O collection "twitter"
col = db.tweets

# criando um dataset com dados retornados no collection Tweets
```

```
# Criando um dataset com dados retornados no collection Tweets

dataset = [{"created_at": item["created_at"],
            "text": item["text"],
           } for item in col.find()]

data = pan.DataFrame(dataset)

# o método CountVectorizer para criar uma matriz de documentos

cv = CountVectorizer()
count_matrix = cv.fit_transform(data.text)



# Contando o número de ocorrências das principais palavras no dataset

word_count = pan.DataFrame(cv.get_feature_names(), columns=["word"])
word_count["count"] = count_matrix.sum(axis=0).tolist()[0]
word_count = word_count.sort_values("count", ascending=False).reset_in
dex(drop=True)
word_count[:500]
```

Out[11]:

	word	count
0	https	790889
1	co	737585
2	rt	613886
3	trump	610153
4	donald	488682
5	the	408061
6	to	326033
7	of	220737
8	is	186201
9	and	161767
10	in	160467
11	for	126838
12	climate	120169
13	on	95782
14	president	90319
15	this	88647
16	it	84963
17	you	73333

18	he	73028
19	that	72218
20	obama	66504
21	be	64191
22	change	62767
23	with	57574
24	we	55910
25	by	54890
26	his	54336
27	not	54194
28	has	51984
29	from	51954
...	...	...
470	nice	3797
471	voters	3793
472	also	3783
473	interest	3780
474	done	3778
475	york	3773
476	power	3770
477	soundcloud	3768
478	100	3756
479	rogerjstonejr	3750
480	anyone	3746
481	agree	3745
482	kids	3744
483	claims	3744
484	oil	3734
485	remember	3723
486	swamp	3713
487	free	3711

488	chinese	3697
489	red	3695
490	rickpoppe	3694
491	xsabdu9wse	3694
492	guy	3692
493	joyannreid	3683
494	selling	3682
495	para	3676
496	full	3671
497	filiilibertatis	3666
498	oh	3665
499	al	3660

500 rows × 2 columns

Volume x dia

In [19]:
```
%%bash

mongo

use twitterdb

var map = function() {
    var datetime = this._id.getTimestamp();

    var analise_por_dia = new Date(datetime.getFullYear(),
                                   datetime.getMonth(),
                                   datetime.getDate());
    emit(analise_por_dia, {count: 1});
}

var reduce = function(key, values) {
    var tot = 0;
    for(var x = 0; i < values.length; x++) { tot += values[x].count; }
    return {count: tot};
}

db.tweets_collection.mapReduce( map, reduce, { "out": "Volume_Dia" } )
;

db.Volume_Dia.find().limit(50).sort({"_id":-1})

# NOTA: TENTEI INUMERAS TENTATIVAS PARA SOLUCIONAR O ERRO, FAVOR CONSI
```

```
# NOTA: TENTEI INUMERAS TENTATIVAS PARA SOLUCIONAR O ERRO. FAVOR COSNI
DERAR A IMPLEMENTACAO DO CODIGO
```

```
MongoDB shell version: 3.2.10
connecting to: test
switched to db twitterdb
2016-12-23T23:26:59.842-0200 E QUERY    [thread1] Error: map reduce fai
led:{ "ok" : 0, "errmsg" : "ns doesn't exist" } :
_getErrorWithCode@src/mongo/shell/utils.js:25:13
DBCollection.prototype.mapReduce@src/mongo/shell/collection.js:1405:1
@(shell):1:1

2016-12-23T23:26:59.846-0200 E QUERY    [thread1] SyntaxError: illegal
character @(shell):1:0

bye
```

Volume x hora do dia

In [20]:
```
%%bash

mongo

use twitterdb


var map = function() {
    var datetime = this._id.getTimestamp();

    var analise_por_hora  = new Date(datetime.getFullYear(),
                                     datetime.getMonth(),
                                     datetime.getDate(),
                                     datetime.getHours());
    emit(analise_por_hora , {count: 1});
}

var reduce = function(key, values) {
    var tot = 0;
    for(var x = 0; i < values.length; x++) { tot += values[x].count; }
    return {count: total};
}

db.tweets_collection.mapReduce( map, reduce, { "out": "Volume_Hora" }
);

db.Volume_Hora.find().limit(10).sort({"_id":-1})


# NOTA: TENTEI INUMERAS TENTATIVAS PARA SOLUCIONAR O ERRO. FAVOR COSNI
DERAR A IMPLEMENTACAO DO CODIGO
```

```
MongoDB shell version: 3.2.10
connecting to: test
```

```
switched to db twitterdb
2016-12-23T23:29:06.967-0200 E QUERY    [thread1] Error: map reduce fai
led:{ "ok" : 0, "errmsg" : "ns doesn't exist" } :
_getErrorWithCode@src/mongo/shell/utils.js:25:13
DBCollection.prototype.mapReduce@src/mongo/shell/collection.js:1405:1
@(shell):1:1

bye
```

Referencias

Data Science Academy, Analisando stream de dados do Twitter com Mongdodb, URL: http://datascienceacademy.com.br/blog/2016/stream-de-dados-do-twitter-com-mongodb-pandas-e-scikit-learn/ (http://datascienceacademy.com.br/blog/2016/stream-de-dados-do-twitter-com-mongodb-pandas-e-scikit-learn/)

Eduardo Santos, Extraindo dados de redes sociais com Python, URL: http://www.eduardosan.com/2015/06/09/extraindo-dados-de-redes-sociais-com-python/ (http://www.eduardosan.com/2015/06/09/extraindo-dados-de-redes-sociais-com-python/)

GitHub Ana Paula Gomes, Repositorio codelab-analise-redes-sociais, URL: https://github.com/anapaulagomes/codelab-analise-redes-sociais/wiki/Coletando-Dados (https://github.com/anapaulagomes/codelab-analise-redes-sociais/wiki/Coletando-Dados)

In [ ]: