Pontifícia Universidade Católica de Minas Gerais
Unidade Praça da Liberdade
Curso de Pós-Graduação em Ciências de Dados e Big Data

Trabalho Prático

Fernando Francis Costa Maria

Professor: Gabriel Campos

BANCO DE DADOS NÃO RELACIONAL
Belo Horizonte
2016

Contexto e aplicação

Foram coletados tweets, da rede social Twitter, por ser hoje uma das principais redes sociais, são aproximadamente 310 milhões de usuários ativos.

Metodologia utilizada

Foi utilizada a linguagem de programação Python com o módulo tweepy armazenando os tweets diretamente no banco de dados Mongodb.

Código Python:

```
import tweepy
import pymongo
import json
from codecs import EncodedFile
from tweepy import OAuthHandler

consumer_key = 'oQKCzrq4PMemskj207****'
consumer_secret = 'C3fxgWnpwrdCapJa6QBQLTCqMPTVkCcz1KEi******'
access_token = '180141585-dk9OTGCDokRN6cWA6C2IOn91fCOV*****'
access_secret = 'cnnaq2u8nfoE2iQJoQt3mZAixaJLv6YOeHH4****'

auth = OAuthHandler(consumer_key, consumer_secret)
auth.set_access_token(access_token, access_secret)

api = tweepy.API(auth)

class CustomStreamListener(tweepy.StreamListener):
    def __init__(self, api):
        self.api = api
        super(tweepy.StreamListener, self).__init__()

        self.db = pymongo.MongoClient('10.0.0.15', 27017).nosql

    def on_data(self, tweet):
        self.db.tweets_collection.insert_one(json.loads(tweet))

    def on_error(self, status_code):
        return True

    def on_timeout(self):
        return True


sapi = tweepy.streaming.Stream(auth, CustomStreamListener(api))
sapi.filter(track=["a"],languages=['en','pt'])
```

Consultas Mongodb:

```
db.tweets_collection.count()

db.tweets_collection.aggregate( [ { $group: { _id: {},
minPrice: { $max: "$created_at" } } } ] );
```

Resultados

- Foram coletados aproximadamente um milhão de tweets.
- As informações extraídas foram à data do tweet e o tweet.
- Temos mais frequentes:
    - #HAPPYBYULSDAY
    - #5HDeservesBeUnited
    - #LuanSantanaDesaparecido
    - #VerãoComDetremuraSDV
    - #2017PrecisaDe
- A coleta foi realizada em cinco dias.
- Para a quantidade de tweets coletada não foi preciso a criação de índices na base de dados.
- As consultas foram executadas instantaneamente.

Conclusões

Fazendo este trabalho pude perceber que mineração de dados e bancos Nosql não é nenhum bicho de sete cabeças, é possível coletar dados de inúmeras fontes na web, sendo de redes sociais ou não, a informação esta disponível para todos, mas a grande sacada é saber oque fazer com ela, oque de valor ela pode gerar.