Pontifícia Universidade Católica de Minas Gerais
Pós-Graduacao Lato Sensu – Ciencia de Dados e Big Data

Autor: Hernani Prates Costa Dias
Banco de Dados não Relationais – Gabriel Campos

Coleta de Informações da Rede Social Twitter Utilizando Python e MongoDB.

Objetivo

Este trabalho tem como objetivo extrair os *posts* do Twitter da região de Belo Horizonte, em um raio de 30km, para obter as informações dos termos mais frequentes, volume de dados por dia e volume de dados por hora.

Metodologia

Para extração das informações e inserção na base de dados do MongoDB, foi utilizada a linguagem de programação Python. Para conexão e extração dos *posts* do Twitter, foi utilizada a biblioteca “tweepy”. As bibliotecas “pymongo”, “pandas” e “json”, foram usadas para o tratamento e inserção dos dados no MongoDB.

Resultados

Após a coleta dos dados realizada pelo programa “extracao_dados_twitter.py”, foram utilizados os comandos do MongoDB “aggregate” e “mapReduce” para responder as perguntas:

Quais os termos mais frequentes?
Qual o volume de dados diário?
Qual o volume de dados por hora do dia?

Temos mais frequentes

O comando abaixo realiza um *MapReduce* da *collection* “collectionTwitter” onde foram armazenados todos os *posts* extraídos. Foi realizado o filtro para buscar as palavras com mais de 4 letras e o resultado foi limitado para os 100 termos mais frequentes:

```
var map = function() {
            this.Texto.split(' ').forEach(function(TextoAux) {
               emit(TextoAux,1);
                              });
 }

var reduce = function(key,value){
return Array.sum(value);
}

db.collectionTwitter.mapReduce(map, reduce, {query:{}, out: "resultado"})

db.resultado.find({$where: "this._id.length > 4"}).sort({value:-1}).limit(100)


{ "_id" : "minha", "value" : 6875 }
{ "_id" : "pessoas", "value" : 5795 }
{ "_id" : "@cabeyousada_:", "value" : 4782 }
{ "_id" : "quando", "value" : 4750 }
{ "_id" : "gente", "value" : 4301 }
{ "_id" : "muito", "value" : 4249 }
{ "_id" : "vídeo", "value" : 4246 }
{ "_id" : "@guilhrsme:", "value" : 4192 }
{ "_id" : "Lagoas,", "value" : 3848 }
{ "_id" : "fazer", "value" : 3708 }
{ "_id" : "perfil", "value" : 3349 }
{ "_id" : "@intedioso:", "value" : 3243 }
{ "_id" : "@SamCriscolo:", "value" : 3094 }
{ "_id" : "ainda", "value" : 2984 }
{ "_id" : "agora", "value" : 2983 }
{ "_id" : "@LucasRanngel:", "value" : 2915 }
{ "_id" : "coisa", "value" : 2893 }
{ "_id" : "tenho", "value" : 2874 }
{ "_id" : "pessoa", "value" : 2867 }
{ "_id" : "Panico", "value" : 2856 }
{ "_id" : "mesmo", "value" : 2839 }
{ "_id" : "mundo", "value" : 2822 }
{ "_id" : "@YouTube", "value" : 2770 }
{ "_id" : "melhor", "value" : 2677 }
{ "_id" : "ficar", "value" : 2598 }
{ "_id" : "quero", "value" : 2545 }
{ "_id" : "nunca", "value" : 2309 }
{ "_id" : "coisas", "value" : 2228 }
{ "_id" : "depressao,", "value" : 2212 }
{ "_id" : "Gostei", "value" : 2109 }
{ "_id" : "dormir", "value" : 2080 }
{ "_id" : "vezes", "value" : 2078 }
{ "_id" : "#umRei", "value" : 2068 }
```

{ "_id" : "manha", "value" : 2047 }
{ "_id" : "assim", "value" : 2010 }
{ "_id" : "horas", "value" : 2000 }
{ "_id" : "Ansiedade,", "value" : 1929 }
{ "_id" : "menos", "value" : 1863 }
{ "_id" : "queria", "value" : 1785 }
{ "_id" : "alguém", "value" : 1765 }
{ "_id" : "sobre", "value" : 1751 }
{ "_id" : "sempre", "value" : 1724 }
{ "_id" : "natal", "value" : 1664 }
{ "_id" : "@boudelaires:", "value" : 1613 }
{ "_id" : "\n#HarmosHatesSyco", "value" : 1589 }
{ "_id" : "todos", "value" : 1532 }
{ "_id" : "semana", "value" : 1525 }
{ "_id" : "tempo", "value" : 1525 }
{ "_id" : "terapia", "value" : 1476 }
{ "_id" : "@tardedesetembro:", "value" : 1450 }
{ "_id" : "Psicologo", "value" : 1428 }
{ "_id" : "visto", "value" : 1389 }
{ "_id" : "depois", "value" : 1367 }
{ "_id" : "acabar", "value" : 1361 }
{ "_id" : "Atlético", "value" : 1351 }
{ "_id" : "kkkkk", "value" : 1344 }
{ "_id" : "Minha", "value" : 1333 }
{ "_id" : "estao", "value" : 1316 }
{ "_id" : "falar", "value" : 1316 }
{ "_id" : "demais", "value" : 1294 }
{ "_id" : "voltar", "value" : 1292 }
{ "_id" : "tanto", "value" : 1286 }
{ "_id" : "durmo", "value" : 1241 }
{ "_id" : "falando", "value" : 1238 }
{ "_id" : "estou", "value" : 1237 }
{ "_id" : "nesse", "value" : 1197 }
{ "_id" : "\"hoje", "value" : 1155 }
{ "_id" : "porque", "value" : 1141 }
{ "_id" : "#H4RMONYParty", "value" : 1127 }
{ "_id" : "últimas", "value" : 1121 }
{ "_id" : "Quando", "value" : 1116 }
{ "_id" : "PRECISA", "value" : 1109 }
{ "_id" : "tinha", "value" : 1106 }
{ "_id" : "NISSO", "value" : 1105 }
{ "_id" : "SENTIDA", "value" : 1096 }
{ "_id" : "https://t.co/0vOvjKdkax", "value" : 1096 }
{ "_id" : "Camila", "value" : 1095 }
{ "_id" : "saber", "value" : 1090 }
{ "_id" : "@RodP13:", "value" : 1072 }
{ "_id" : "contra", "value" : 1067 }
{ "_id" : "Minas", "value" : 1064 }

{ "_id" : "música", "value" : 1064 }
{ "_id" : "Gente", "value" : 1049 }
{ "_id" : "antes", "value" : 1043 }
{ "_id" : "também", "value" : 1042 }
{ "_id" : "mulher", "value" : 1034 }
{ "_id" : "kkkkkk", "value" : 1025 }
{ "_id" : "desse", "value" : 1012 }
{ "_id" : "entao", "value" : 1008 }
{ "_id" : "Cruzeiro", "value" : 1000 }
{ "_id" : "nossa", "value" : 1000 }
{ "_id" : "feliz", "value" : 996 }
{ "_id" : "férias", "value" : 996 }
{ "_id" : "essas", "value" : 983 }
{ "_id" : "deixem", "value" : 979 }
{ "_id" : "deixa", "value" : 978 }
{ "_id" : "passar", "value" : 972 }
{ "_id" : "parece", "value" : 963 }
{ "_id" : "vontade", "value" : 958 }
{ "_id" : "fazendo", "value" : 946 }

Volume dos dados diário

Para mostrar o volume diário, foi utilizado o comando “aggregate” onde a coluna “data”, derivada da coluna “Created”, é a chave para o agrupamento. Para que a hora fosse desconsiderada na chave, a coluna “data”foi adicionada na *collection* “collectionTwitter” com o tratamento para retirar a hora:

```
var cursor = db.collectionTwitter.find({"Created":/2016/,});

while (cursor.hasNext()) {
  var currentDocument = cursor.next();

  var data_aux = currentDocument['Created'].substr(0,10);
  currentDocument['data'] = data_aux;

   db.collectionTwitter.update({_id: currentDocument._id} ,{$set: {"data": currentDocument['data'] } } )

  }

 db.collectionTwitter.aggregate([{$match:{"Created":/2016/}} , {$group: { "_id":"$data", "total":{$sum:1} } }  ])
```

{ "_id" : "2016-12-15", "total" : 10472 }
{ "_id" : "2016-12-14", "total" : 12756 }

Volume dos dados por hora do dia

Para mostrar o volume dos dados por hora do dia, foi utilizado o comando “aggregate” onde a coluna “hora”, derivada da coluna “Created”, é a chave para o agrupamento. A coluna “hora” foi adicionada na *collection* “collectionTwitter” com o formato “YYYY-MM-DD HH”:

```
var cursor = db.collectionTwitter.find({"Created":/2016/,});

while (cursor.hasNext()) {
  var currentDocument = cursor.next();

  var data_aux = currentDocument['Created'].substr(0,13);
  currentDocument['hora'] = data_aux;

    db.collectionTwitter.update({_id: currentDocument._id} ,{$set: {"hora": currentDocument['hora'] } } )

  }


db.collectionTwitter.aggregate([{$match:{"Created":/2016/}} , {$group: { "_id":"$hora", "total":{$sum:1} } }  ])


{ "_id" : "2016-12-15 20", "total" : 2510 }
{ "_id" : "2016-12-15 21", "total" : 5162 }
{ "_id" : "2016-12-15 00", "total" : 2651 }
{ "_id" : "2016-12-15 01", "total" : 149 }
{ "_id" : "2016-12-14 21", "total" : 1928 }
{ "_id" : "2016-12-14 22", "total" : 5511 }
{ "_id" : "2016-12-14 23", "total" : 5317 }
```