

Análise de sentimentos com Naive Bayes e armazenamento em MongoDB

Diego Felipe Berg Mauricio Pardin
Jeronimo Alencar Barros
Rodrigo Carlos de Jesus Teodoro
Vilmar de Paula Nunes

Introdução

Objetivo deste trabalho é conseguir coletar e analisar os tweets criando um sistema de classificação baseado em sentimentos positivos e negativos demonstrados visualmente por meio de gráficos e relatórios.



Tecnologias

- MongoDB <https://www.mongodb.com/>  mongoDB
- Python 3 <https://www.python.org/>  python
- Tweepy (acesso a API do Twitter) <http://www.tweepy.org/> 
- Flask (Framework para Web) <http://flask.pocoo.org/> 
- Algoritmo de classificação Naive Bayes https://en.wikipedia.org/wiki/Naive_Bayes_classifier

Arquitetura

Armazenamento - MongoDB

Pelo MongoDB ser um banco de dados de alta performance orientado a documentos e sem necessidade de ter esquemas pré-definidos ele se torna adequado a armazenar os tweets recuperados.



Arquitetura

Algoritmo (Python)

Foi utilizada a Linguagem Python por ter uma boa biblioteca que faz a integração com a API do Twitter e Naive Bayes, além trabalhar muito bem com documentos em formato JSON e tratamento de textos.

Fontes estão disponíveis em

<https://github.com/rodrigoteodoro/analisesentimento>



Arquitetura

Visualização (Flask)

Flask é um framework simples e ágil para desenvolvimento WEB em Python. Se integra muito bem com diversas tecnologias utilizadas na internet (JavaScript, HTML, JQuery, ChartJS entre outros).

```
from flask import Flask  
app = Flask(__name__)
```

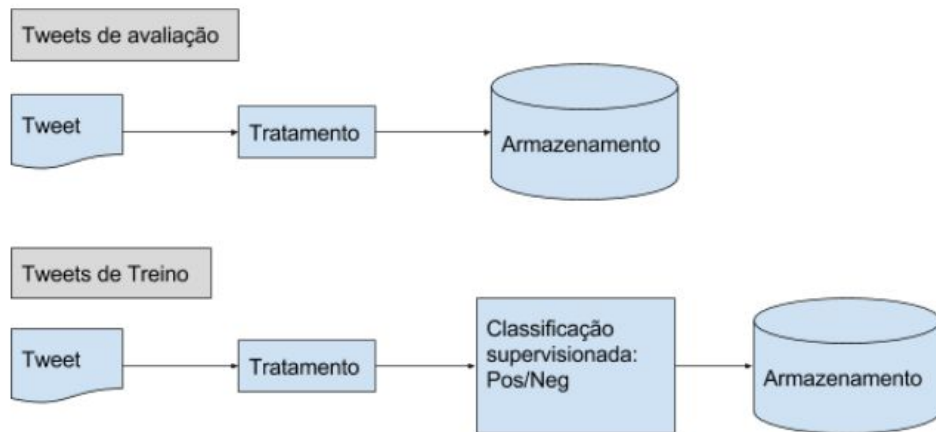
```
@app.route("/")  
def hello():  
    return u'Olá mundo!'
```

```
if __name__ == "__main__":  
    app.run()
```



Coleta

Fluxo de coleta e armazenamento (treino e avaliação)



Limpeza e tratamento dos textos

- Remover caracteres especiais exceto: acentuação, pontuação e espaços;
- Remover quebras de linhas
- Remover # e @
- Tweets de Treino - frases com mais de 5 palavras
- Tweets de avaliação - frases com mais de 2 palavras
- Ignorar re-tweets e iniciados com “rt”



Coleções e documentos no banco de dados

Tweets de treino (twitter_treino)

```
{  
  tag = "Nome da tag",  
  texto = "Texto do tweet",  
  track = "Informações de recuperação da API, exemplo: globo, temer"  
  sentimento = "pos = positivo e neg = negativo"  
}
```

Tweets de avaliação (twitter_collection)

```
{  
  tag = "Nome da tag",  
  texto = "Texto do tweet",  
  track = "Informações de recuperação da API, exemplo: globo, temer"  
}
```



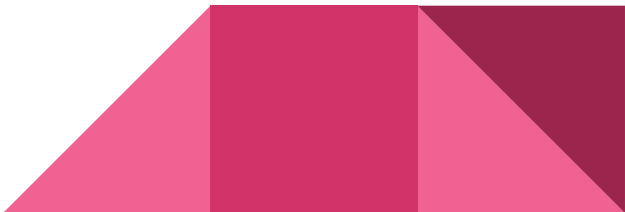
Coleções e documentos no banco de dados

Pesquisa análise (pesquisa_analise)

```
{  
  tag = "Nome da tag",  
  qtd_pos = "Quantidade de tweets positivos",  
  qtd_neg = "Quantidade de tweets negativos"  
  total = "Quantidade total de tweets analisados"  
}
```

Tweets analisados (twitter_analise)

```
{  
  tag = "Nome da tag",  
  texto = "Texto do tweet",  
  classe = "pos = positivo e neg = negativo"  
  prob = "Probabilidade"  
}
```



Análise (Naive Bayes)

Naive Bayes é um dos mais simples e bem difundidos classificadores baseados no teorema de Bayes. Pode ser usado em modelagem de previsão e exploratória. Por não considerar dependências, suas suposições são consideradas ingênuas.

Uma característica atraente deste classificador é a sua capacidade de produzir estimativas de probabilidade, isto significa que, para cada rótulo de classe o classificador irá gerar uma estimativa.

Para poder classificar os textos, o algoritmo deverá ser treinado com uma base pré-analisada, seja por meios computacionais (no caso deste trabalho) ou por meios humanos (pessoa avaliar cada texto).



Análise (Naive Bayes)

Tweet	Sentimento
Estou feliz hoje	pos
Estou muito triste	neg

Análise (Naive Bayes)

No caso, iremos ter as classes de sentimentos: Positivos (acima de 50%) e Negativos (abaixo de 50%)

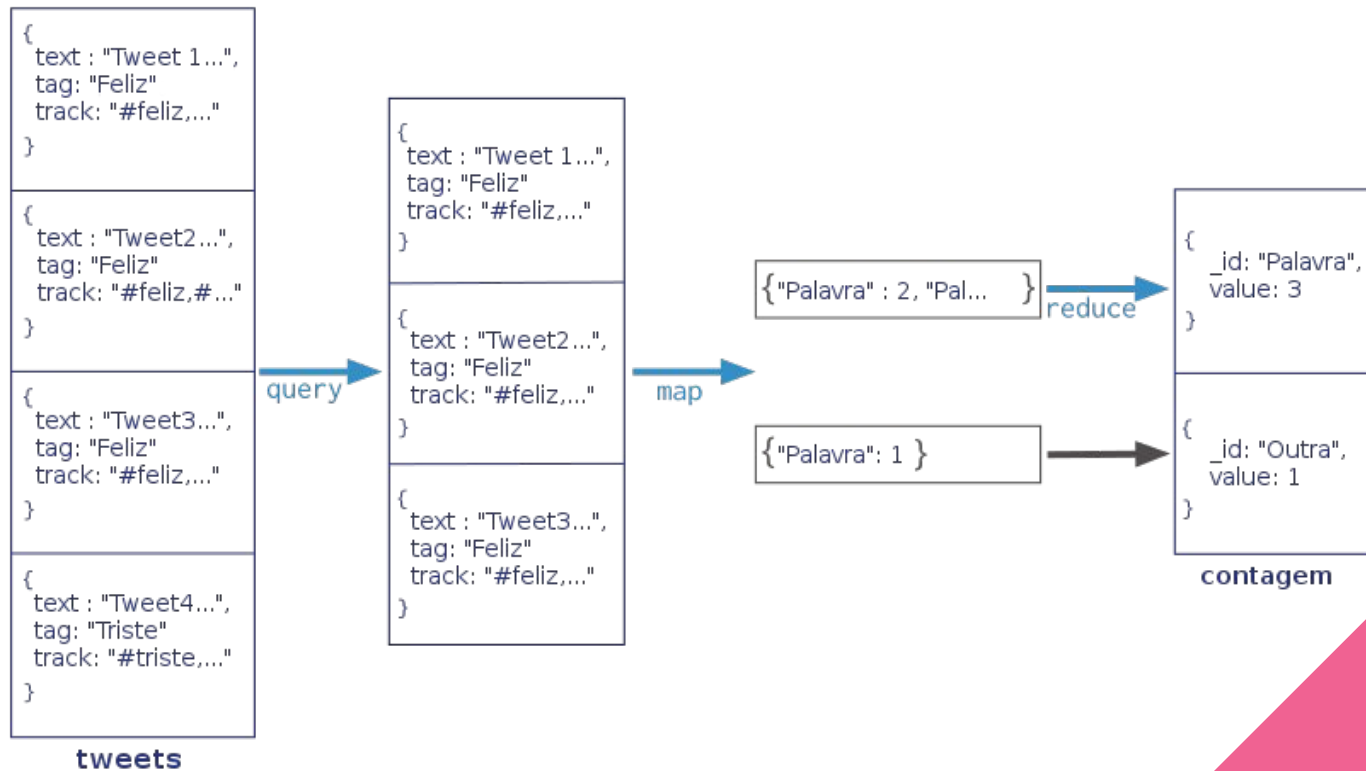
Tweet	Sentimento	Probabilidade
Que alegria, arrumei novo emprego	pos	80%
Meu time perdeu hoje de 7x1, que infelicidade	neg	10%

Análise (contagem de palavras)

- Map-Reduce é um paradigma de processamento de dados para condensar grandes volumes de dados em resultados agregados.
- O MongoDB provê MapReduce como um comando do SGBD.
- O código é enviado via BSON (*Binary* JSON) do driver do python para o SGBD.

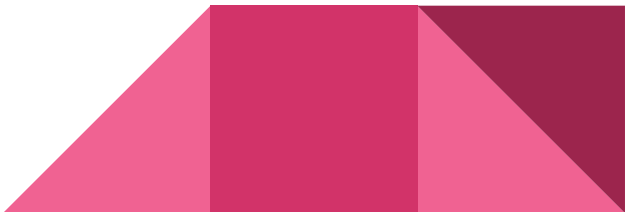
```
Collection
↓
db.tweets.mapReduce(
  map    → function() { emit(...); },
  reduce → function(key, values) { return...},
  query  → {
  output →   query: {tag: "Feliz" },
            out:  "contagem"
            }
)
```

Análise (contagem de palavras)



Análise (contagem de palavras) - MAP

```
function() {  
  var texto = this.texto;  
  if (texto) {  
    texto = texto.toLowerCase().split(" ");  
    for (var i = texto.length - 1; i >= 0; i--) {  
      if (texto[i]) {  
        emit(texto[i], 1);  
      }  
    }  
  }  
};
```



Análise (contagem de palavras) - REDUCE

```
function( key, values ) {  
    var count = 0;  
    values.forEach(function(v) {  
        count +=v;  
    });  
    return count;  
}
```



Demonstração

Análise Twitter

Usuário: rleodoro

Pesquisas

Tweets treino

Logout

Pesquisas

Cadastrar

Nome (tag)

É o nome da pesquisa (somente letras)

Lista (track)

Lista de valores separados por vígula

Cadastrar

Recarregar

Auto

Assuntos

Pesquisar

Nome (tag)	Lista (track)	Quantidade	Ativa	Tweets	Análise
globo	globo	120	Não	Visualizar	Visualizar
namorados	namorados,amor	1110	Não	Visualizar	Visualizar

Mostrando de 1 até 2 de 2 registros

Demonstração

Análise Twitter

Usuário: rteodoro

Pesquisas

Tweets treino

Logout

Lista dos Tweets relacionados à pesquisa: **globo**

Recuperar

Limite

Quantidade de tweets a recuperar

Recuperar

Recarregar

Tweets

Pesquisar

Texto

10h30 sp globo 7 7 record 6 2 sbt 4 4 cultura 1 5 gazeta 1 1 band 0 4 redetv 0 2

10h53 sp globo 7 2 record 5 5 sbt 4 3 cultura 2 1 gazeta 1 3 band 0 8 redetv 0 5

666 só se o flamengo usar a cota de globo de 2055 hahahaha

a mídia internacional anda sem assunto pq olha publiquem a situação dos servidores públicos

a nova versão da escolinha do professor raimundo é um sucesso nos canais viva e globo e já tem garantida uma

adolescente que teve testa tatuada é encontrado por amigos caminhando perto de casa, no abc

alguem me explica esse argumento time da globo

amanda de godoi é confirmada em nova novela da globo

amor nem existe é tudo invenção da globo

anta agoniza e globo parte para o desespero como é feito a imprensa se expor ao ridículo por uma notícia via

Mostrando de 1 até 10 de 120 registros

Anterior

1

2

3

4

5

...

12

Próximo

Demonstração

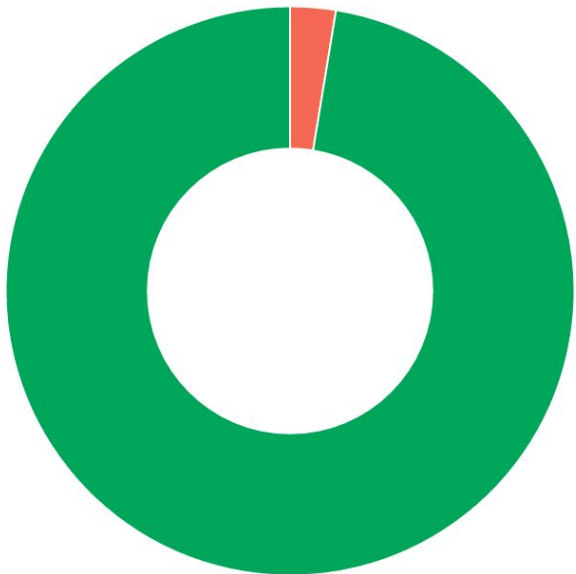
Análise dos Tweets relacionados à pesquisa: **globo**

Opções

Analisar

Visualizar

Total de Tweets analisados: 618



professor pessoas choror apaixonados adolescente pensa políticos deve decisiva oferta rico
toda curitiba paulinho câmara novo assim jogos coisa mundo ajuda
desde sempre rádio médica flamengo matéria cara aecolúcia fatima
juizes jornal cultura feliz aqui palmeiras lula jeito
conversa confira supremo record vídeo jogo retro
final nota post anos timbeta temer pmdb antes entra
dias anular força time juiz amor ficar bebê gente cesar
leva atender abre time rede filme novo pode caso
love ligadissoveja real tudo acheli dilma deixar caso beto
gripe nesta brasil acham alves paulo news
brasileiro atraído fala sobre um rei hoje humano apola
notícia futebol aaaaaaaa link após band apoio copa fiscal
carnaval doria contra psdb chapa pede costa filho ação
sabe julgamento quer todos taca youtube chapa anda morto
fica mostra bicicleta abin ambev condena gostei
golpe bal improbidade depoimento esporte namorados anal
omitiu diego improbidade dono escolhidos bate fofocalizandonosbt casal
consegui pedir viva alunos uerj bate

Demonstração

Tweets

Pesquisar

Texto	Sentimento	Probabilidade
é disso que povo gosta	neg	4.80
ze ricardo só escapará dentro de 1 mês então	pos	7.70
vídeo emprego com carteira assinada reverte queda e volta a crescer em abril globo play	pos	6.20
vou dar um berro toda vez que entrar na puc e vir o ambulante de matte e biscoito globo	neg	4.80
vai estamos com vcs	pos	8.00
vacinação contra gripe tem início nesta segunda feira para toda população em 33 cidades do ceará	pos	7.60
tresportes mas porque o narrador da globo falou vai consultar a gente quando o juiz foi falar com o árbitro de linha	neg	4.50
torço demais tbm pq essa menina merece muito	pos	6.20
todo ano é essa roubalheira, timezinho da globo	pos	6.50
temer desconfia de movimentos de rodrigo maia	pos	5.10

Mostrando de 1 até 10 de 120 registros

Anterior 1 2 3 4 5 ... 12 Próximo

Conclusão

O MongoDB se demonstrou útil para o armazenamento dos tweets, principalmente se levarmos em conta que os documentos armazenados podem ter campos distintos entre cada registro. Possui acesso simples e ágil em recuperar as informações por meio de filtros.

O algoritmo Naive Bayes se mostrou adequado na classificação, porém necessita ter uma base de treino bem avaliada para ter melhor resultado.

