
Extração e Classificação de Dados Semânticos do Twitter

Clarissa Castellã Xavier
Marlo Souza

Agenda

- Twitter
- Extraíndo dados do Twitter
- Análise de Polaridade
- Extração de Entidades

Repositório curso:

<https://github.com/clarissacastella/twittercourse>

Twitter

- Twitter
 - Serviço de microblog e rede social lançado no final de 2006
 - Mensagens de até 280 caracteres
 - Média de 336 milhões usuários mensais (primeiro trimestre de 2018)
 - Mídia informativa para o público brasileiro
 - Coletar e transmitir informações do que conversar
 - 62% dos Tweets têm conteúdo informativo
 - 48% são de natureza conversacional
 - 10% com ambas as características.

Twitter

Timeline

<https://twitter.com>

The screenshot shows the Twitter web interface in Portuguese. At the top, there's a navigation bar with links for 'Página Inicial', 'Moments', 'Notificações' (67), and 'Mensagens'. A search bar and a 'Tweetar' button are on the right. The main content area is divided into three columns. The left column features 'Assuntos do momento: Brasil' with a list of trending topics and their tweet counts. The middle column displays a timeline of tweets, including a tweet from Eliot Higgins about a profile on Radio 4, a tweet from Hacker News about a CSS Layout cookbook, and a tweet from Tierney Siren about the Node.js Collaborators' Summit. The right column shows a 'Quem seguir' section with profiles like HailTo, stormideas, and NSA/CSS, followed by a section for 'Encontre pessoas que você conhece'.

Assuntos do momento: Brasil

- Uribe: 58,6 mil Tweets
- #SomosTodosReginaDuarte: 35,3 mil Tweets
- #fcseguindofo: 6.070 Tweets
- Norton: 311 mil Tweets
- Leo Duarte: 8.278 Tweets
- #WomanLikeMeSelfie: 3.118 Tweets
- Everton Ribeiro: 2.636 Tweets
- Cuellar: 10,1 mil Tweets
- Marcos Jr: 2.049 Tweets
- Holanda: 14,3 mil Tweets

Timeline:

Eliot Higgins @EliotHiggins · 1 min
Here's Radio 4's profile on my, on their appropriately named "Profile".

BBC Radio 4 - Profile, Eliot Higgins
Mark Coles looks at the man behind the investigative website Bellingcat.
bbc.co.uk

Hacker News @newscombinator · 32 min
CSS Layout cookbook
The CSS layout cookbook aims to bring together recipes for common layout patterns, things you might need to implement in your own sites. In addition to providing ...
developer.mozilla.org

Tierney Siren @bitandbang · 2 h
Working on what our values, programs, and questions for a merged @nodejs foundation and @the_jsf would look like:

Twitter

Página Usuário

<https://twitter.com/TwitterBrasil>

<https://twitter.com/TwitterBrasil>

aduz G Whats Linguee NLPNews note media-actions travis Filmes webmail.inf Foto - Google

Página Inicial Momentos Notificações Mensagens

Buscar no Twitter

Tweetar

Twitter Brasil @TwitterBrasil

Bem-vindos à conta oficial do Twitter Brasil! Precisa de ajuda? Acesse support.twitter.com/forms

Brasil

blog.twitter.com/pt/brasil

Participa desde março de 2011

Nasceu em 21 de março

Tweetar para

5 Seguidores que você conhece

685 Fotos e vídeos

Tweets Tweets e respostas Midia

Tweet Fixado

Twitter Brasil @TwitterBrasil · 11 de out

Não deixe que sua preocupação com consumo de dados te mantenha longe do Twitter. Com o Twitter Lite, você fica informado sobre tudo o que está acontecendo, economiza dados e espaço em seus dispositivos e salva Tweets para ver depois. Baixe agora na Google Play Store

Use o Twitter Lite e saiba o que está acontecendo.

Ele ocupa menos espaço, e você pode usar seu pacote de dado à vontade, até em redes 2G/3G.

Twitter Lite

12 4 20

Mostrar esta sequência

Twitter Brasil rebroadcastou

Brasil Game Show #BGS2018 @BrasilGameShow · 12 de out

Shota Nakama, produtor musical, e Daniel Pesina, intérprete de diversos personagens de Mortal Kombat, vão responder perguntas de vocês aqui no @TwitterBrasil. Basta usar a tag #BGSResponde e enviar a sua pergunta. Valendo! #BGS2018

Quem seguir · Atualizar · Ver todos

Twitter @Twitter

Twitter Support @Twit...

Twitter Live @TwitterLive

Encontre pessoas que você conhece

Importe seus contatos de Gmail

Conectar outras listas de contatos

Assuntos do momento: Brasil

Alterar

#TWD9naFOX

O povo não está aguentando ver 'The Walking Dead' sem chorar

#NBSTourSaoPaulo

Que momento! A Anitta cantou no palco com a Camila Cabello

#Fantástico

6.202 Tweets

#NFLnaESPN

7.316 Tweets

#ARMYIndependenceDay

546 mil Tweets

Fausto

6.189 Tweets

Guilherme de Pádua

3.500 Tweets

Twitter

- Twitter
 - Textos com perfil informativo:
 - +- 25% opinativos
 - Opiniões ou sentimentos.
 - 72% dos brasileiros residentes em áreas urbanas usam mídias sociais como fonte de notícias
 - 13% utilizam o Twitter como principal rede social para esse fim

Twitter

Estratégias para lidar com as particularidades de textos do Twitter:

- Normalização de texto:
 - Técnicas de análise automática para converter o texto ruidoso do Twitter em uma variante mais formal da língua, como corrigir erros de grafia
- Agrupamento de Tweets:
 - Agrupamento automático de Tweets tratando de um mesmo assunto, ou que estejam contextualmente ligados
 - Criar textos capazes de fornecer informação contextual mais relevante às ferramentas de Análise de Texto

Twitter

Estratégias para lidar com as particularidades de textos do Twitter:

- Séries temporais:
 - Como Tweets possuem uma estrutura de fluxo, i.e. estão deslocados no tempo, alguns métodos para processamento utilizam modelagens baseadas em séries temporais tentando capturar organicamente o contexto pela sua informação temporal
- Ferramentas específicas:
 - Alguns métodos e ferramentas parecem precisar de estratégias específicas para textos provenientes de Tweets, que levam em consideração a pobreza contextual e as variações lexicais próprias desses textos

Twitter

Twitter é excelente corpus para análise de sentimentos e mineração de opinião {pak2010twitter}:

- Utilizado para expressar seu ponto de vista sobre diferentes tópicos, sendo assim, uma fonte valiosa de opiniões
- Contém uma quantidade imensa de postagens de texto que cresce a cada dia - tamanho
- Audiência variada - é possível coletar mensagens de texto de usuários de diferentes grupos sociais e de interesses.
- Audiência representada por usuários de vários países e idiomas

Twitter

- Que tipo de informações vamos lidar neste curso?
 - Classificação de Polaridade
 - Obter a polaridade do sentimento transmitido pela informação
 - Se o texto expressa um sentimento positivo, negativo ou neutro.
 - Dadas as limitações de tamanho dos Tweets é normalmente executada no nível da sentença
 - Linguagem informal e especializada faz com que esta seja uma tarefa singular

Twitter

- Que tipo de informações vamos lidar neste curso?
 - Reconhecimento de Entidades
 - Localizar e classificar entidades nomeadas no texto.
 - O estilo breve, informal e ruidoso do Twitter apresenta desafios.

Extraindo dados do Twitter

Extraindo dados do Twitter

É possível coletar informações do Twitter utilizando

- API pública
- Aplicativos e bibliotecas alternativas

Extraindo dados do Twitter

- API pública
 - REST APIs:
 - Estratégia pull
 - O usuário deve explicitamente fazer uma solicitação
 - Streaming APIs:
 - Estratégia push
 - Depois que uma solicitação de informações é feita, a API fornece um fluxo contínuo de atualizações sem necessitar nenhuma outra solicitação

Extraindo dados do Twitter

- Cinco grupos principais:
 - Contas e usuários
 - Mensagens diretas
 - Anúncios
 - Ferramentas de publisher e SDKs
 - **Tweets e respostas**
 - Torna os Tweets e as respostas públicas disponíveis para os desenvolvedores e permite que estes também postem Tweets
 - Os Tweets podem ser acessados pesquisando por palavras-chave específicas ou solicitando conteúdo de contas específicas

Extraindo dados do Twitter

Acessando API

- As aplicações (consumidores) precisam se registrar no Twitter
 - Neste processo a aplicação recebe uma chave e um *token* que o aplicativo deve usar para se autenticar.

The screenshot shows the Twitter Developer Portal for an application named 'getTweetsMaslab'. The URL in the browser is <https://apps.twitter.com/app/14718959/keys>. The page has tabs for 'Details', 'Settings', 'Keys and Access Tokens', and 'Permissions', with 'Settings' currently selected. Under 'Application Settings', there is a note: 'Keep the "Consumer Secret" a secret. This key should never be human-readable in your application.' The form fields show: Consumer Key (API Key) as 'Y8o7', Consumer Secret (API Secret) as 'Z6KI', Access Level as 'Read and write (modify app permissions)', Owner as 'cla_cx', and Owner ID as '3525322276'. Below this is the 'Application Actions' section with buttons for 'Regenerate Consumer Key and Secret' and 'Change App Permissions'. The 'Your Access Token' section includes a note: 'This access token can be used to make API requests on your own account's behalf. Do not share your access token secret with anyone.' The form fields show: Access Token as '35', Access Token Secret as 'gYrF', Access Level as 'Read and write', Owner as 'cla_cx', and Owner ID as '3525322276'. The browser's taskbar at the bottom shows several open applications, including 'twittercourse' and 'chaves.txt'.

Extraindo dados do Twitter

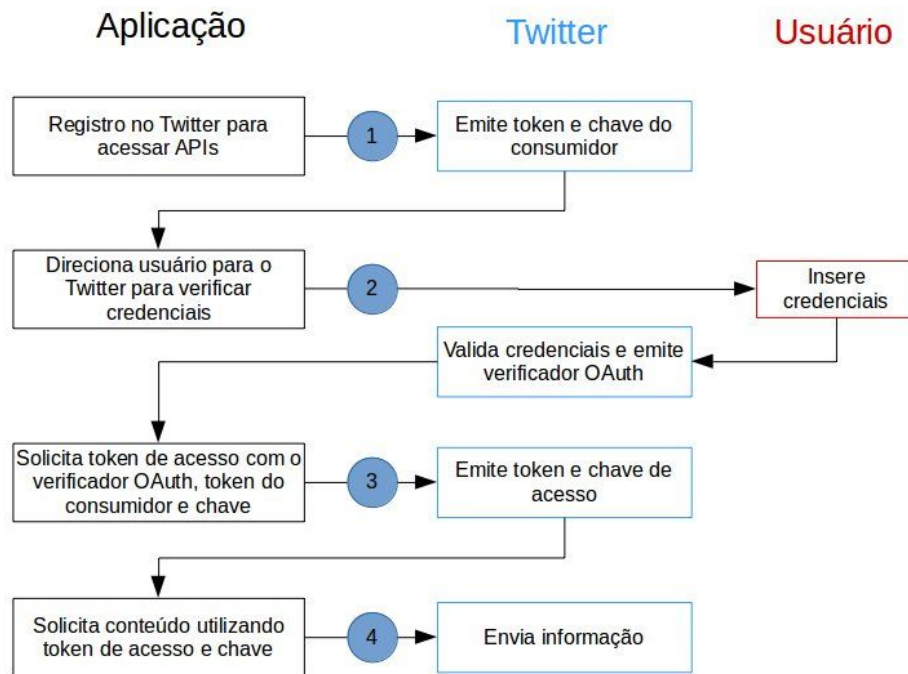
Acessando API

- O usuário informa o PIN para o aplicativo. O aplicativo usa o PIN para solicitar um token e uma chave de acesso exclusivos para o usuário.
- Utilizando o token e a chave de acesso o aplicativo autentica o usuário no Twitter e chama a API em nome do usuário.

```
1 import oauth2
2
3 CHAVE_CONSUMO = '4xxxxxxxxxxxB'
4 TOKEN_CONSUMO = 'h2xxxxxxxxxxxp'
5
6 TOKEN_ACESSO = '35xxxxxxxxxxxLYIe'
7 CHAVE_ACESSO = '67d9xxxxxxxxxxxko0'
8
9 def oauth_req(url, CHAVE_ACESSO,
10               TOKEN_ACESSO, http_method="GET",
11               post_body="", http_headers=None):
12     token = oauth2.Token(key=CHAVE_ACESSO,
13                          secret=TOKEN_ACESSO)
14     consumo = oauth2.Consumer(key=
15                               CHAVE_CONSUMO, secret=TOKEN_CONSUMO)
16     cliente = oauth2.Client(consumo, token)
17     resp, conteudo = cliente.request( url,
18                                     method=http_method, body=post_body,
19                                     headers=http_headers )
20     return conteudo
```

Extraindo dados do Twitter

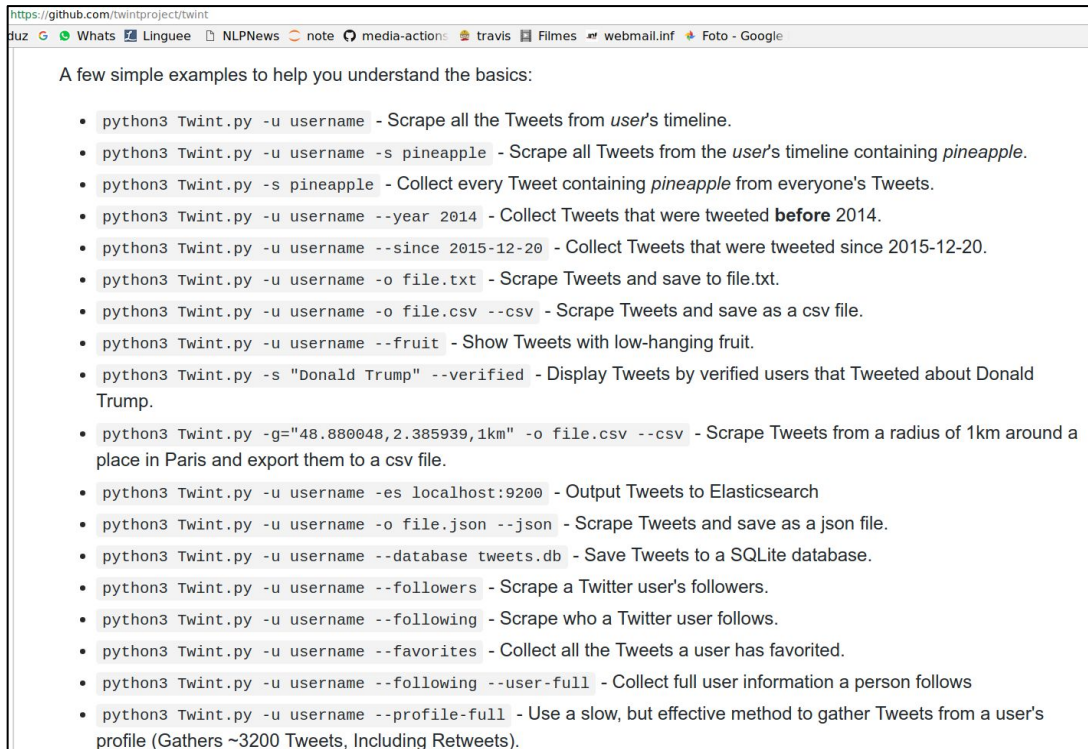
Revendo Processo de acesso à API



Extraindo dados do Twitter

Twint

- Ferramenta avançada de captura de texto do Twitter
- Permite extrair tweets de perfis do Twitter sem usar a API do Twitter
- Escrita em Python
- Usa os operadores de pesquisa do Twitter
- <https://github.com/twintproject/twint>



The screenshot shows the GitHub repository for Twint. The page title is "twintproject/twint". Below the title, there is a section titled "A few simple examples to help you understand the basics:". This section contains a list of 15 command-line examples for using Twint. The examples are as follows:

- `python3 Twint.py -u username` - Scrape all the Tweets from *user's* timeline.
- `python3 Twint.py -u username -s pineapple` - Scrape all Tweets from the *user's* timeline containing *pineapple*.
- `python3 Twint.py -s pineapple` - Collect every Tweet containing *pineapple* from everyone's Tweets.
- `python3 Twint.py -u username --year 2014` - Collect Tweets that were tweeted **before** 2014.
- `python3 Twint.py -u username --since 2015-12-20` - Collect Tweets that were tweeted since 2015-12-20.
- `python3 Twint.py -u username -o file.txt` - Scrape Tweets and save to file.txt.
- `python3 Twint.py -u username -o file.csv --csv` - Scrape Tweets and save as a csv file.
- `python3 Twint.py -u username --fruit` - Show Tweets with low-hanging fruit.
- `python3 Twint.py -s "Donald Trump" --verified` - Display Tweets by verified users that Tweeted about Donald Trump.
- `python3 Twint.py -g="48.880048,2.385939,1km" -o file.csv --csv` - Scrape Tweets from a radius of 1km around a place in Paris and export them to a csv file.
- `python3 Twint.py -u username -es localhost:9200` - Output Tweets to Elasticsearch
- `python3 Twint.py -u username -o file.json --json` - Scrape Tweets and save as a json file.
- `python3 Twint.py -u username --database tweets.db` - Save Tweets to a SQLite database.
- `python3 Twint.py -u username --followers` - Scrape a Twitter user's followers.
- `python3 Twint.py -u username --following` - Scrape who a Twitter user follows.
- `python3 Twint.py -u username --favorites` - Collect all the Tweets a user has favorited.
- `python3 Twint.py -u username --following --user-full` - Collect full user information a person follows
- `python3 Twint.py -u username --profile-full` - Use a slow, but effective method to gather Tweets from a user's profile (Gathers ~3200 Tweets, Including Retweets).

Extraindo dados do Twitter

Exemplos:

- Informações sobre um usuário
- Seguidores de um usuário
- Quem o usuário segue
- Tweets publicados
- Resultados de uma pesquisa

Extraindo dados do Twitter

Informações sobre um usuário - API

- A API principal do Twitter é responsável pela manipulação e consulta dos dados.
- Qualquer método dessa API é precedido da URI <http://api.twitter.com/version/>
 - Version é a versão da API (atualmente 1)

Extraindo dados do Twitter

Informações sobre um usuário - API

- Cada usuário do Twitter está associado a um identificador *screen_name* e um *user_id*
- O método users/show retorna as informações do perfil do usuário
 - Aceita um nome de usuário válido como parâmetro e retorna o perfil deste usuário no Twitter.

```
22 def info_usr(usuario):
23     GET_USR_URL = "https://
        api.twitter.com/1.1/users/
        show.json?screen_name="+usuario
24     req = oauth_req(GET_USR_URL,TOKEN_ACESSO
        ,CHAVE_ACESSO)
25     return req
26
27 d = json.loads( info_usr('twitterbrasil'))
28 print json.dumps(d, indent=4, sort_keys=True
    )
```

Extraindo dados do Twitter

Informações sobre um usuário - API

```
tmp@ccx-Inspiron-7559: /media/tmp/faa8fb7c-c759-4e71-8525-d0d7a7a6dbba/ufrgs/twittercourse$ python info_usr.py
{
  "contributors_enabled": false,
  "created_at": "Thu Mar 10 22:54:23 +0000 2011",
  "default_profile": false,
  "default_profile_image": false,
  "description": "Bem-vindos \u00e0 conta oficial do Twitter Brasil! Precisa de ajuda? Acesse https://t.co/Nu5ZS0w4UD",
  "entities": {
    "description": {
      "urls": [
        {
          "display_url": "support.twitter.com/forms",
          "expanded_url": "https://support.twitter.com/forms",
          "indices": [
            71,
            94
          ],
          "url": "https://t.co/Nu5ZS0w4UD"
        }
      ]
    },
    "url": {
      "urls": [
        {
          "display_url": "blog.twitter.com/pt/brasil",
          "expanded_url": "https://blog.twitter.com/pt/brasil",
          "indices": [
            0,
            22
          ],
          "url": "http://t.co/GuzH0naY84"
        }
      ]
    }
  },
  "favourites_count": 1553,
  "follow_request_sent": false,
  "followers_count": 2199386,
  "following": false,
  "friends_count": 269,
  "geo_enabled": true,
  "has_extended_profile": true,
  "id": 263884490,
```

Extraindo dados do Twitter

Seguidores de um usuário - API

- Método followers/list
- Retorna uma coleção de objetos de usuário contendo os usuários que seguem o perfil
- Resultados são fornecidos em grupos de 20 usuários
 - Páginas de resultados navegadas usando o valor next_cursor
 - https://api.twitter.com/1.1/followers/list.json?screen_name=theSeanCooks&cursor=137400477753100783

3

```
def seguidores_usr(usuario):
    GET_USR_URL = "https://api.twitter.com/1.1/
        followers/list.json?cursor=-1&skip_status=true&include_user_entities=false&screen_name="+usuario
    req = oauth_req(GET_USR_URL, TOKEN_ACESSO,
        CHAVE_ACESSO)
    return req

d = json.loads(seguidores_usr('twitterbrasil'))
print json.dumps(d, indent=4, sort_keys=True)
```


Extraindo dados do Twitter

Seguidores de um usuário - API

```
tmp@ccx-Inspiron-7559:/media/tmp/faa8fb7c-c759-4e71-8525-d0d7a7a6dbba/ufrrgs/twittercourse$ python seguidores_usr.py
{
  "next_cursor": 1614243531635697877,
  "next_cursor_str": "1614243531635697877",
  "previous_cursor": 0,
  "previous_cursor_str": "0",
  "total_count": null,
  "users": [
    {
      "blocked_by": false,
      "blocking": false,
      "contributors_enabled": false,
      "created_at": "Sat Oct 13 19:08:19 +0000 2018",
      "default_profile": true,
      "default_profile_image": false,
      "description": "",
      "favourites_count": 0,
      "follow_request_sent": false,
      "followers_count": 2,
      "following": false,
      "friends_count": 209,
      "geo_enabled": false,
      "has_extended_profile": false,
      "id": 1051187898496835585,
      "id_str": "1051187898496835585",
      "is_translation_enabled": false,
      "is_translator": false,
      "lang": "pt",
      "listed_count": 0,
      "live_following": false,
      "location": "",
      "muting": false,
      "name": "Cintiabezerra13",
      "notifications": false,
      "profile_background_color": "F5F8FA",
      "profile_background_image_url": null,
      "profile_background_image_url_https": null,
      "profile_background_tile": false,
      "profile_image_url": "http://pbs.twimg.com/profile_images/1051239033500774404/wpJzHMMWz_normal.jpg",
      "profile_image_url_https": "https://pbs.twimg.com/profile_images/1051239033500774404/wpJzHMMWz_normal.jpg",
      "profile_link_color": "1DA1F2",
      "profile_sidebar_border_color": "C0DEED",
      "profile_sidebar_fill_color": "DDEEFF",
```

Extraindo dados do Twitter

Seguidores de um usuário - Twint

- `python3 Twint.py -u username --followers -`

```
tercourse/twint$ python3 Twint.py -u twitterbrasil --followers
Cintiabezerra14
JeffersonDoni94
thelabtocafe
Junior85115130
SteelRockTech
AgnadjaM
AhRuduit
TheusSousa12
imcarlosalves
Sid11708608
nathansandy2020
erica_pabline
Fudidoaoquadra1
GreiceKlahr
CassioRoberto07
Aadarsh560065
Clevert07758608
pamy_vitoriaofc
GilStyle17
rajraj3
gabrielhpgama
Edinalv85200691
mark_pkasi
```

Extraindo dados do Twitter

Quem o usuário segue - API

- Método friends/list
- Retorna uma coleção de objetos de usuário contendo os usuários seguidos pelo perfil
- Resultados são fornecidos em grupos de 20 usuários
 - Páginas de resultados navegadas usando cursor

```
def seguidos_usr(usuario):  
    GET_USR_URL = "https://api.twitter.com/1.1/  
        friends/list.json?cursor=-1&skip_status=  
true&include_user_entities=false&screen_name  
="+usuario  
    req = oauth_req(GET_USR_URL, TOKEN_ACESSO,  
        CHAVE_ACESSO)  
    return req  
  
d = json.loads(seguidos_usr('twitterbrasil'))  
print json.dumps(d, indent=4, sort_keys=True)
```

Extraindo dados do Twitter

Quem o usuário segue

```
course$ python seguidos_usr.py
{
  "next_cursor": 1542758452149154338,
  "next_cursor_str": "1542758452149154338",
  "previous_cursor": 0,
  "previous_cursor_str": "0",
  "total_count": null,
  "users": [
    {
      "blocked_by": false,
      "blocking": false,
      "contributors_enabled": false,
      "created_at": "Wed Jun 06 14:21:58 +0000 2018",
      "default_profile": false,
      "default_profile_image": false,
      "description": "#MiCasaEsTuCasa \ud83e\udd51 \u00a1Bienvenidos a la cuenta oficial de Twitter M\u00e9xico! Si necesitas ayuda o tienes dudas de tu cuenta accede al Centro de Ayuda de Twitter.",
      "favourites_count": 55,
      "follow_request_sent": false,
      "followers_count": 41941,
      "following": false,
      "friends_count": 8,
      "geo_enabled": false,
```

Extraindo dados do Twitter

Quem o usuário segue - Twint

- `python3 Twint.py -u twitterbrasil --following`

```
course/twint$ python3 Twint.py -u twitterbrasil --following
TwitterMexico
tvaparecida
RSF_pt
abraji
jack
edupanzi
ONUMulheresBR
NBB
clayton_melo
TwitterVideo
Amoramamora
anthonymoto
ResenhaESPN
CopadoBrasil
NFLBrasil
Thaynara0G
danielamercury
zanetti_arthur
hypolitoginasta
arthurnory
NicheBrasil
balancogeral
Pele
```

Extraindo dados do Twitter

Tweets publicados - API

- Método statuses/timeline
- REST API
- Retorna coleção dos Tweets mais recentes postados por um usuário
- Retorna até 3.200 mensagens
- Resultados são fornecidos em grupo (até 200 mensagens)
 - Páginas de resultados navegadas usando max_id

```
def tweets_usr(usuario):  
    GET_USR_URL = "https://api.twitter.com/1.1/  
        statuses/  
        user_timeline.json?count=1&screen_name="  
        +usuario  
    req = oauth_req(GET_USR_URL, TOKEN_ACESSO,  
        CHAVE_ACESSO)  
    return req  
  
d = json.loads(tweets_usr('twitterbrasil'))  
  
print json.dumps(d, indent=4, sort_keys=True)
```

Extraindo dados do Twitter

Tweets publicados - API

```
course$ python tweets_usr.py
```

```
[
  {
    "favorite_count": 33,
    "favorited": false,
    "geo": null,
    "id": 1051125468945113088,
    "id_str": "1051125468945113088",
    "in_reply_to_screen_name": "NBB",
    "in_reply_to_status_id": 105111500392174387,
    "in_reply_to_status_id_str": "105111500392174387",
    "in_reply_to_user_id": 18079515,
    "in_reply_to_user_id_str": "18079515",
    "is_quote_status": false,
    "lang": "pt",
    "place": null,
    "retweet_count": 4,
    "retweeted": false,
    "source": "<a href='\"http://twitter.com/dov
twitter for iPhone\"'/>
    \"text\": \"@NBB ... vai ter muita cesta e mu
a #D\\u00e9lJogo #NBBnoTwitter\",
    \"truncated\": false,
    \"user\": {
      \"contributors_enabled\": false,
      \"created_at\": \"Thu Mar 10 22:54:23 +000
      \"profile_background_image_url_https\": \"https://abs.twimg.com/images/them
      \"profile_background_tile\": true,
      \"profile_banner_url\": \"https://pbs.twimg.com/profile_banners/263884490/1
      \"profile_image_url\": \"http://pbs.twimg.com/profile_images/10084609441781
      \"profile_image_url_https\": \"https://pbs.twimg.com/profile_images/1008460
      \"profile_link_color\": \"1DA1F2\",
      \"profile_sidebar_border_color\": \"C0DEED\",
      \"profile_sidebar_fill_color\": \"DDEEF6\",
      \"profile_text_color\": \"333333\",
      \"profile_use_background_image\": true,
      \"protected\": false,
      \"screen_name\": \"TwitterBrasil\",
      \"statuses_count\": 7570,
      \"time_zone\": null,
      \"translator_type\": \"regular\",
      \"url\": \"http://t.co/GuzH0naY84\",
      \"utc_offset\": null,
      \"verified\": true
    }
  }
]
```

Extraindo dados do Twitter

Tweets publicados - API

- Método statuses/filter
- Retorna coleção dos Tweets que correspondem a um ou mais parâmetros de filtro
- Streaming
 - O cliente usa uma única conexão, persistindo a conexão com a API
- O parâmetro é o termo que será sendo seguido
 - Se o parâmetro for twitter, o método irá imprimir os Tweets contendo este termo sendo criados publicamente na plataforma.

Extraindo dados do Twitter

Tweets publicados - API

```
def segue_tweets(termo):
    url = "https://stream.twitter.com/1.1/statuses/filter.json?track="+termo
    http_method="POST"
    post_body=""
    http_headers=None
    token = oauth2.Token(key=CHAVE_ACESSO, secret=TOKEN_ACESSO)
    consumo = oauth2.Consumer(key=CHAVE_CONSUMO, secret=TOKEN_CONSUMO)
    cliente = oauth2.Client(consumo, token)
    headers = {}
    req = oauth2.Request.from_consumer_and_token(
        cliente.consumer, token=cliente.token,
        http_method="POST", http_url=url)
    req.sign_request(cliente.method, cliente.consumer, cliente.token)
    headers.update(req.to_header())
    body = req.to_postdata()
    headers['Content-Type'] = 'application/x-www-form-urlencoded'
    req = urllib2.Request(url, body, headers=headers)
    try:
        f = urllib2.urlopen(req)
    except urllib2.HTTPError, e:
        data = e.fp.read(1024)
        raise Exception(e, data)

    for line in f:
        d = json.loads(line)
        try:
            print d["user"]["name"], d["text"]
        except:
            print d.get("id")

segue_tweets('twitter')
```

Extraindo dados do Twitter

Tweets publicados - API

```
ourse$ python segue_tweets.py
elle♡지민 @99JMN THEY POSTED FOR YOU
Luiz E. Bendotti RT @paulacamara_: "Pegaram minha foto no meu blog e postaram num po
rtal de notícias da Bahia. Porém, a Foto é de 1992, e o conteúdo não tem...
PrincessJessica @iWantClips is easy to use, come feed your addiction with Princess J
essika https://t.co/pSLLPagfjZ https://t.co/rxGYHYUnKV
แบบชัก RT @janjc84: มีคำถามค่ะพี่ร่อน
แนนไหมคะ 😊

อยากจีเป็นเพื่อนจริงๆ https://t.co/xZP55vppD7
[반홀~15일]☆재현포카☆는 일단 나에게로 RT @_020205: 꼭 봐줘 https://t.co/GnsgvCor2s
ことみ🍡 RT @Sharara0427: アニメの香澄!!! ☺主人公だからぜひ見てね☺ https://t.co/5
HuwTjsyM1
taekook🐼🐼💖 RT @becauseofV95: 181013 LOVE YOURSELF TOUR
AMSTERDAM

#방탄소년단 #뷔 #V #BTS @BTS_twt https://t.co/0MVlR3gfjZ
💎raihannaㅇㅈㅇ🍷 RT @billboard: BTS is hitting the big screen #BillboardNews https
://t.co/8TDTWCWNYE5 https://t.co/BtQnhCEoVE
saki RT @kawaii_d2: ☆☆猫背も治っちゃう https://t.co/3akHr3tiEb
siska😺 | jimin's day🐼 RT @seokjinspout: seokjin: i think i'm going to fall in love
Yoongi:Did i say i was going to date you?I don't want to date you
```

Extraindo dados do Twitter

Tweets publicados - Twint

- O comando que retorna todos os Tweets da timeline de um usuário informado como parâmetro:
- `python3 Twint.py -u username`

```
course/twint$ python3 Twint.py -u twitterbrasil
1051125468945113088 2018-10-13 12:00:15 -03 <TwitterBrasil> @NBB ... vai ter muita c
esta e muito Tweet! 🙌 #DáJogo #NBBnoTwitter
1050485642864414721 2018-10-11 17:37:48 -03 <TwitterBrasil> Habilite o modo noturno
para ler melhor enquanto estiver se atualizando sobre os Tweets da noite 🌙 pic.twitt
ter.com/eCLTmewqPJ
1050485636409503746 2018-10-11 17:37:47 -03 <TwitterBrasil> Quer assistir a um vídeo
mas não pretende consumir seus dados?A funcionalidade de Salvar Tweets permite que
você deixe para assistir a vídeos ou ler artigos quanto estiver conectado ao WiFi, o
u no momento que for mais conveniente para você. pic.twitter.com/PDLakhjDTE
1050485632831803397 2018-10-11 17:37:46 -03 <TwitterBrasil> Reduza seu consumo de da
dos com o Twitter Lite. Depois de abrir o aplicativo, clique em sua foto de perfil p
ara acessar o menu e escolha a opção de economia de dados. pic.twitter.com/uSv3dv0mg
W
1050485628822056963 2018-10-11 17:37:45 -03 <TwitterBrasil> Não deixe que sua preocu
pação com consumo de dados te mantenha longe do Twitter. Com o Twitter Lite, você fi
ca informado sobre tudo o que está acontecendo, economiza dados e espaço em seus dis
positivos e salva Tweets para ver depois. Baixe agora na Google Play Store pic.twitt
er.com/tr3kb5Rftu
1050409930891689984 2018-10-11 12:36:57 -03 <TwitterBrasil> Quer saber tudo o que es
tá acontecendo na Brasil Game Show e entrar nas conversas sobre essa super feira? Us
e as hashtags #BGS2018 e #BrasilGameShow e ative um emoji exclusivo!
1050388174969692160 2018-10-11 11:10:30 -03 <TwitterBrasil> Acompanhe aqui o primeir
o debate entre os candidatos ao Governo do Rio neste segundo turno #Eleições2018 htt
```

Extraindo dados do Twitter

Resultados de uma pesquisa - API

- Método search/tweets
- Retorna os Tweets que correspondem aos parâmetros da consulta.
 - Parâmetros podem incluir palavras-chave, hashtags, frases, regiões, nomes de usuários ou ids

```
course$ python busca_tweets.py
{
  "search_metadata": {
    "completed_in": 0.061,
    "count": 15,
    "max_id": 1051028261193601024,
    "max_id_str": "1051028261193601024",
    "query": "webmedia",
    "refresh_url": "?since_id=1051028261193601024&q=webmedia&include_entities=1"
  },
  "statuses": [
    {
      "contributors": null,
      "iso_language_code": "fr",
      "result_type": "recent"
    },
    {
      "place": null,
      "possibly_sensitive": false,
      "retweet_count": 6,
      "retweeted": false,
      "source": "<a href='\"http://twitter.com/download/android\"' rel='\"nofollow\">Twitter for Android</a>",
      "text": "Les \u00e9l\u00e8ves du lyc\u00e9e Amiral de Grasse s'engagent pour un journalisme citoyen avec la plate-forme webmedia\u2013",
      "truncated": true,
      "user": {
        "contributors_enabled": false,
        "created_at": "Wed Jan 25 09:55:47 +0000 2012",
        "default_profile": false,
        "default_profile_image": false,
        "description": "Compte officiel du Clemi de l'Acad\u00e9mie Nice / Cap'Radio, la #webradio et Cap'TV, la #webtv de l'acad\u00e9mie de Nice \n#EMI #webmedia #InfoIntox",
        "entities": {
          "description": {
            "urls": []
          }
        }
      }
    }
  ]
}
```


Extraindo dados do Twitter

Resultados de uma pesquisa - Twint

- python3 Twint.py -s [palavra-chave]

```
tmp@ccx-Inspiron-7559:/media/tmp/faa8fb7c-c759-4e71-8525-d0d7a7a6dbba/ufrgs/twitterc
ource/twint$ python3 Twint.py -s 'dia das crianças'
1051286863099637761 2018-10-13 22:41:34 -03 <fwatrin_> Foto do dia das crianças pra
lembra a epoca mais feliz da minha vida, 12 anin pic.twitter.com/YzYbJ4QCEm
1051286834221916160 2018-10-13 22:41:27 -03 <EvelenAbreu> Vota no Bolsonaro, mas foi
para festa do dia das Crianças que o Tráfico financiou!!! Num é que foi mexmo 🤔
1051286779519795200 2018-10-13 22:41:14 -03 <Cabelinlt_> Ação Social do dia das cria
nças tudo lindo ♥️📷 pic.twitter.com/TosmnYQHTm
1051286672548290566 2018-10-13 22:40:49 -03 <vinistone_> Gostei de um vídeo @YouTube
http://youtu.be/MjZJWZetYV4?ap ós Dia das Crianças - DESCONFINADOS
1051286568135249920 2018-10-13 22:40:24 -03 <gaby_gcsm> Minha mãe me deu 100 reais d
e presente de dia das crianças♥
1051286515190571008 2018-10-13 22:40:11 -03 <josejrsanfona> Feliz dia das crianças.
Manuel Carvalho https://www.facebook.com/100005781758499/posts/871688969700487/ ...
1051286509763141632 2018-10-13 22:40:10 -03 <juhikki> Dia das crianças foi ontem...
🤔🤔🤔🤔 mas tá valendo hahaha... 1985 #diadascrianças #latepost #80s #criançaafeliz
#bebe #bebezinho #tbt https://www.instagram.com/p/Bo5Tm-0hu8H/?utm_source=ig_twitte
r_share&igshid=aqfl46cu3mp6 ...
1051286455979569152 2018-10-13 22:39:57 -03 <jrgoiis> Gostei de um vídeo @YouTube h
ttp://youtu.be/KaSSURrPYVM?ase MEU PAI FOSSE CRIANÇA NO DIA DAS CRIANÇAS - PRETEND
TO BE A CHILD BY EARNING
1051286264291446784 2018-10-13 22:39:12 -03 <cordeirobru_> @LizaSan16 @emycristinar_
@ViGbriele feliz dia das crianças rapaziada
```

Análise de Polaridade

Análise de Polaridade

- Obter a polaridade do sentimento transmitido pela informação
 - Se o texto expressa um sentimento positivo, negativo ou neutro.
- AP de Tweets:
 - Dadas as limitações de tamanho é normalmente executada no nível da sentença
 - A linguagem informal e especializada faz com que esta seja uma tarefa singular

Análise de Polaridade

Usuário	@ivetesangalo
<i>Tweet</i>	Estou muito feliz e muito agradecida por todo esse amor ♥♥♥
Polaridade	Positiva

Usuário	@EPTC_POA
<i>Tweet</i>	Neste momento, bem complicado o acesso a Rodoviária no Largo Vespasiano Julio Veppo, pelo Túnel da Conceição.
Polaridade	Negativa

Análise de Polaridade

Abordagens:

- Baseada em Regras
- Baseada em Léxico
- ➔ **Aprendizado de Máquina**

Análise de Polaridade

Abordagem Baseada em Regras:

- São definidas regras ou padrões para compreender as opiniões sobre o texto
- Ex:
 - Faz a tokenização
 - Inicia com pontuação neutra (0)
 - Para cada padrão encontrado, aplica uma classificação
 - A sentença é considerada positiva se a pontuação de polaridade final for maior que zero ou negativa se a pontuação geral for menor que zero.

Análise de Polaridade

Abordagem Baseada em Regras:

“Estou muito feliz e agradecida por todo este amor”

(Estou muito feliz) = +1

(agradecida) = +1

(Todo este amor) = +1

3 = positiva

“Neste momento, bem complicado o acesso a Rodoviária no Largo Vespasiano Julio Veppo, pelo Túnel da Conceição.”

(bem complicado) = -1

-1 = negativa

Análise de Polaridade

Abordagem Baseada em Léxico

- Parte do pressuposto de que a polaridade ou o sentimento expresso por uma frase ou documento pode ser identificada pelas polaridades das unidades lexicais que a compõem.
- Unidade lexical é a unidade de significado no léxico mental e pode ser composta de uma ou mais palavras
 - cabeça, guarda-chuva, dinheiro sujo, etc.

Análise de Polaridade

Abordagem Baseada em Léxico

- Primeiro efetuar o pré-processamento
 - Reduzir o volume dos dados
- O método precisa representar cada texto e suas palavras
 - Cada palavra é representada por meio de um peso
 - Pode ser simplesmente sua frequência, ou, por exemplo, o valor TF-IDF
 - O peso da palavra positiva ou negativa, aumenta proporcionalmente à medida que aumenta o número de ocorrências
 - É possível distinguir o fato de alguns termos serem geralmente mais comuns que outros no âmbito positivo ou negativo

Análise de Polaridade

Abordagem Baseada em Léxico

“Estou muito feliz e agradecida por todo este amor”

```
estou = 0
muito = 0
feliz = 3
e = 0
agradecida = 1
por = 0
todo = 0
este = 0
amor = 2
```

Positiva

“Neste momento, bem complicado o acesso a Rodoviária no Largo Vespasiano Julio Veppo, pelo Túnel da Conceição.”

```
neste = 0
momento = 0
bem = 0
complicado = -2
acesso = 0
rodoviária = 0
largo = 0
vespasiano = 0
julio = 0
veppo = 0
pelo = 0
túnel = 0
conceição = 0
```

Negativa

Análise de Polaridade

Aprendizado de Máquina

- 2 etapas:
 - Primeiro treina a si mesmo utilizando um conjunto de dados de treinamento
 - Depois realiza o aprendizado em si

Análise de Polaridade

- Problema de Classificação
 - Aplicar o rótulo correto para uma determinada entrada
 - Iremos considerar a análise de polaridade como uma tarefa de classificação onde os Tweets podem ser classificados como Positivos, Negativos ou Neutros.

Análise de Polaridade

Classificação supervisionada

- Um classificador é denominado supervisionado quando utiliza um corpora de treinamento.
- Se *framework* é dividido em duas etapas:
 - Treino
 - Aprendizado

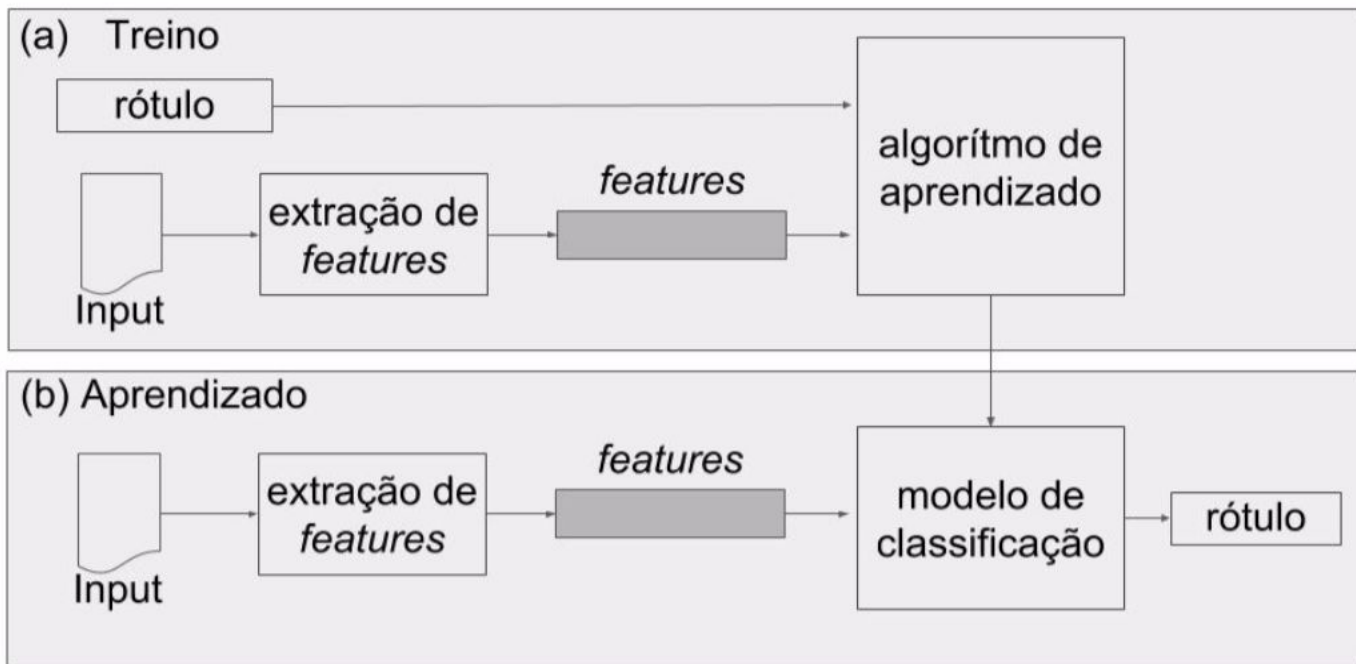
Análise de Polaridade

Classificação supervisionada

- Treinamento: um extrator de *features* converte cada valor de entrada em um *feature set*. Pares de feature sets e rótulos alimentam o algoritmo de aprendizado de máquina para gerar um modelo
- Aprendizado: o mesmo extrator de *features* é usado para converter novas entradas . Os *feature sets* são alimentados no modelo, o que gera os novos rótulos (classificação)

Análise de Polaridade

Framework de Classificação Supervisionada



Análise de Polaridade

Aprendizado de Máquina

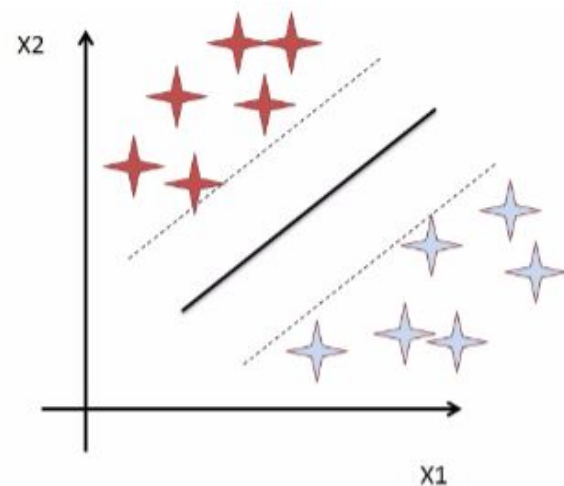
- Três Modelos:
 - Máxima Entropia
 - Naive Bayes
 - SVM - Support Vector Machine

Análise de Polaridade

Aprendizado de Máquina

Support Vector Machine - SVM

- Dado um conjunto de treino, cada um marcado como pertencente a uma categoria, o algoritmo de treinamento cria um modelo que atribui novos exemplos a uma categoria
- Representação dos exemplos como pontos em um hiperplano.
 - Encontrar uma linha de separação (hiperplano) entre dados de duas classes, buscando maximizar a distância entre os pontos mais próximos em relação a cada uma das classes

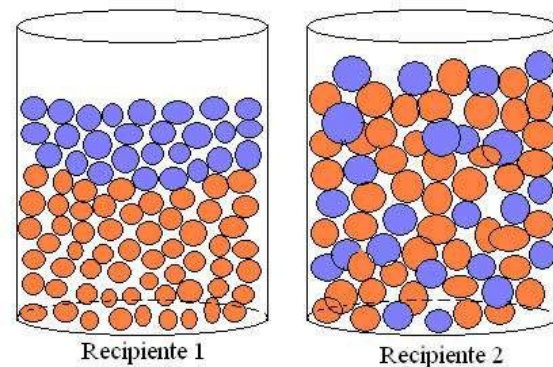


Análise de Polaridade

Aprendizado de Máquina Máxima Entropia

- Segundo o princípio da máxima entropia, a distribuição correta é aquela que maximiza a entropia ou sujeito incerto às restrições que representam a "evidência", isto é, os fatos conhecidos pelo experimentador

A grandeza Entropia (S) é a medida da desordem de um sistema.



Recipiente 1: mais organizado, menor entropia
Recipiente 2: menos organizado, maior entropia

Análise de Polaridade

Aprendizado de Máquina

Naive Bayes

- Modelo de probabilidade que assume a independência entre os recursos de entrada, baseado na aplicação do teorema de Bayes.
- Assume que a presença de uma feature específica não se relaciona com a existência de qualquer outra feature

Análise de Polaridade

Exemplo Prático

- Implementado em Python
- Biblioteca NLTK
 - Plataforma para trabalhar com dados em linguagem humana
 - Fornece mais de 50 recursos corpora e lexicais como o WordNet, juntamente com um conjunto de bibliotecas de processamento de texto para:
 - **Classificação**
 - Tokenização
 - Stemming
 - Tagging
 - Análise semântica

Análise de Polaridade

Exemplo Prático

Extração de Features

- Criação do vetor de entrada (*bag-of-words*) que será utilizado tanto pelo algoritmo de aprendizado, como pelo modelo de classificação
- O texto é representado como um conjunto de suas palavras, desconsiderando a gramática e até a ordem das palavras, mas mantendo a multiplicidade.

```
"Ana gosta de Zeca, Zeca gosta de Lia e Lia não gosta de  
ninguém"
```

```
{ ["Ana":1, "gosta":3, "de":3, "Zeca":2, "Lia":2, "e":1,  
  "não":1, "ninguém":1] }
```

Análise de Polaridade

Exemplo Prático

Extração de Features

- **divide()** e **bag_of_words** recebe como entrada um vetor de frases e retorna um vetor de palavras para cada frase no vetor de entrada
- Entrada: ["Trânsito acentuado nos dois sentidos da Av. Carlos Gomes x Campos Sales."]
- Retorno: ['trânsito', 'acentuado', 'nos', 'dois', 'sentidos', 'da', 'av.', 'carlos', 'gomes', 'x', 'campos', 'sales.']

```
def divide(dados):  
    dados_new = []  
    for palavra in dados:  
        palavra_filter = [i.lower() for i in  
                           palavra.split()]  
        dados_new.append(palavra_filter)  
    return dados_new[0]  
  
def bag_of_words(palavras):  
    return dict([(palavra, palavras.count(  
        palavra)) for palavra in palavras])
```

Análise de Polaridade

Exemplo Prático

Treino

- **treina_classificadores()** utiliza as *features bag-of-words* geradas pelas funções **divide()** e **bag_of_words** e retorna um modelo treinado para cada um dos três algoritmos de classificação: SVM, Naive Bayes e Maxent
- Os dados de treino são *Tweets* da conta @EPTC_POA rotulados como positivo, negativo e neutro

Análise de Polaridade

Exemplo Prático

```
def treina_classificadores():
    posdados = []
    with open('./dadostreino/train_EPTC_POA_v3nbal_1.data', 'rb') as myfile:
    negdados = []
    with open('./dadostreino/train_EPTC_POA_v3nbal_0.data', 'rb') as myfile:
    neudados = []
    with open('./dadostreino/train_EPTC_POA_v3nbal_2.data', 'rb') as myfile:
    negfeats = [(bag_of_words(f), 'neg') for f in divide(negdados)]
    posfeats = [(bag_of_words(f), 'pos') for f in divide(posdados)]
    neufeats = [(bag_of_words(f), 'neu') for f in divide(neudados)]
    treino = negfeats + posfeats + neufeats
    #Maximum Entropy
    classificadorME = MaxentClassifier.train(treino, 'GIS', trace=0, encoding=None
        , labels=None, gaussian_prior_sigma=0, max_iter = 1)
    #SVM
    classificadorSVM = SklearnClassifier(LinearSVC(), sparse=False)
    classificadorSVM.train(treino)
    # Naive Bayes
    classificadorNB = NaiveBayesClassifier.train(treino)
    return ([classificadorME, classificadorSVM, classificadorNB])
```

Análise de Polaridade

Exemplo Prático

Extração de Features

- **classifica()** recebe como entrada um vetor de frases e os três modelos gerados no treino e retorna um vetor com as polaridades identificadas por cada um dos classificadores para cada frase
- **Entrada:** ['Fluxo muito congestionado na Osvaldo Aranha no acesso para o Túnel. Agora, tá chovendo também. Então, atenção!', 'Não use o celular ao volante, 80% da sua atenção é desviada']
- **Retorno:** [['neg', 'pos', 'neg'], ['neu', 'neu', 'neu']]
(SVM, NB, Maxent)

```
def classifica(sentencas, classificadores):  
    ret = []  
    for s in sentencas:  
        c = divide([s])  
        feats = bag_of_words(c[0])  
        classificacao = []  
        classificacao.append(classificadores[1]  
                               ].classify(feats))  
        classificacao.append(classificadores[2]  
                               ].classify(feats))  
        classificacao.append(classificadores[0]  
                               ].classify(feats))  
        ret.append(classificacao)  
    return ret
```

Extração de Entidades

Extração de Entidades

- EE é uma sub-tarefa da área de extração de informações
- Seu objetivo é localizar e classificar entidades nomeadas no texto
- Também conhecida como Reconhecimento de Entidades Nomeadas (REN)

Extração de Entidades

- Essa tarefa é frequentemente uma das primeiras etapas na análise semântica de um texto
 - As entidades mencionadas transmitem bastante informação sobre o conteúdo do texto em si

`"Neste momento, bem complicado o acesso a Rodoviária no Largo Vespasiano Julio Veppo, pelo Túnel da Conceição"`

`[Rodoviária (LOC)] [Largo Vespasiano Julio Veppo (LOC)] [Túnel da Conceição (LOC)]`

Extração de Entidades

- Os sistemas de EE mais recentes têm utilizado técnicas de aprendizado de máquina
 - Em contraste com os sistemas mais antigos que utilizavam regras manualmente codificadas e heurísticas para efetuar esta tarefa
- EE vem sendo tratada como uma tarefa de rotulagem sequencial

Extração de Entidades

- O formato dos Tweets impõe algumas dificuldades
 - Por esse motivo, os métodos de EE que são aplicados em outros tipos de texto podem não funcionar bem
 - Um anotador de última geração como o Stanford NER tem seu desempenho bastante reduzido quando aplicado a textos do Twitter
 - Textos curtos oferecem pouca informação contextual para a identificação de entidades
 - Recursos importantes para identificar nomes próprios, como capitalização, não são utilizados com rigor nos textos do Twitter
 - É comum subcapitalização e supercapitalização
- Os métodos de EE devem ser adaptados a esse contexto.

Extração de Entidades

Exemplo Prático

- Implementado em Python
- Biblioteca Spacy
 - Biblioteca de código aberto para o processamento avançado de linguagem natural
 - Solução baseada em *deep learning* que interage com as bibliotecas do ecossistema de AI do Python
- Sistema de EE estatístico
 - Atribui rótulos a extensões contíguas de tokens.

Extração de Entidades

Exemplo Prático

- Cada decisão, como por exemplo se uma palavra é uma entidade, é uma previsão
- Essa previsão é baseada nos exemplos que o modelo viu durante o treinamento
- Para treinar um modelo é necessário um conjunto de dados de treinamento que consistem de exemplos de texto e de rótulos que o modelo deve prever
- Como treinar um modelo é gerar um padrão que possa ser generalizado em outros exemplos, os dados de treinamento devem ser representativos dos dados que serão processados

Extração de Entidades

Exemplo Prático

- Extração de Features
- `cria_dados_treino()` transforma os dados de treino contidos em um arquivo texto para o formato de entrada do modelo

```
def cria_dados_treino(arq='./data/
dadosTreinoLoc.txt'):
    dados_treino = []
    fin = open(arq, 'rb')
    n=0
    post = u''
    for val in fin:
        d = {}
        n = n + 1
        if (n % 2 == 1) :
            post = val.replace('\n', '')
        else :
            d['entities']=ast.literal_eval(val
                .replace('\n', ''))
            dados_treino.append((post, d))

    fin.close()
    return dados_treino
```

Entrada	... 17h53 - ATENÇÃO! Acidente entre carro e moto na R. Dom Pedro II, sentido sul/norte, próximo a R. Barão do Cotegipe. [(48,63,'LOC'), (94,114,'LOC')]
Saída	{ [...('17h53 - ATENÇÃO! Acidente entre carro e moto na R. Dom Pedro II, sentido sul/norte, próximo a R. Barão do Cotegipe.', ({ 'entities': { [(48,63,'LOC'), (94,114,'LOC')] } })] }

Extração de Entidades

Exemplo Prático

Criação do Modelo

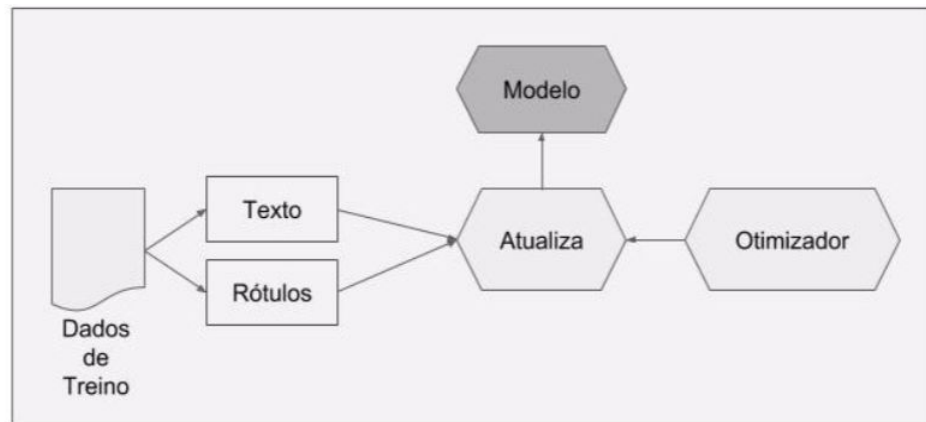
- Treina com os dados importados pela função **cria_dados_treino()** seguindo os seguintes passos:
 - Cria modelo em branco
 - Adiciona ao modelo o pipeline NER, responsável por executar a EE
 - Lê os dados de treino
 - Treina o modelo

Extração de Entidades

Exemplo Prático

Criação do Modelo

- Treino é feito em várias iterações
- Em cada iteração os dados de treinamento são embaralhados para que o modelo não faça generalizações com base na ordem dos exemplos
- Também é definida uma taxa de desistência para "descartar" recursos e representações individuais aleatoriamente
 - Exemplo, um abandono de 0,25 significa que cada recurso tem $\frac{1}{4}$ de probabilidade de ser descartado



Extração de Entidades

Exemplo Prático

```
dir = "./modeloEE" # cria modelo em branco
dir = Path(dir)
nlp = spacy.blank('pt')
dir.mkdir()
ner = nlp.create_pipe('ner')
nlp.add_pipe(ner, last=True)
```

```
import dados_treino as tr
dados_treino = tr.cria_dados_treino('./data/testP.txt') # dados de treino
for _, annotations in dados_treino:
    for ent in annotations.get('entities'):
        ner.add_label(ent[2])
# treina EE
optimizer = nlp.begin_training()
n_iter = 2 #nro iteracoes
for itn in range(n_iter):
    c = 0
    losses = {}
    random.shuffle(dados_treino) #embaralha para nao viciar
    for text, annotations in dados_treino:
        c = c + 1
        nlp.update(
            [unicode(text)], # batch de textos
            [annotations], # batch de anotacoes correspondentes aos textos
            sgd=optimizer,
            losses=losses)
    print(losses)
```


Extração de Entidades

Exemplo Prático

Aprendizado

- Passos:

- Carrega modelo
- Lê arquivo com posts
- Para cada post: aplica o modelo de EE

```
modelo_dir = Path("./model10iter")
modelo = spacy.load(modelo_dir) # roda modelo

fin = open('./data/postsP.txt', 'rb')
for text in fin:
    doc_model = modelo(unicode(text))
    locs = []
    for ent in doc_model.ents:
        aux = (ent.start_char, ent.end_char, str(ent.label_))
        locs.append(text[ent.start_char:ent.end_char+1])
    print text, locs
```

Extração de Entidades

Exemplo Prático

Aprendizado

```
ccxavier@laplace:~/work/EE/spacy$ python2.7 aprendeEE.py
Acesso ao aeroporto, pela 3ª Perimetral, tem fluxo acentuado, mas ainda sem pontos de lentidão. https://t.c
['aeroporto,', '3ª Perimetral']
Acidente c/ danos materiais entre dois carros na Av. Assis Brasil próx. a Av. Bernardino Silveira Amorim se
['Av. Assis Brasil ', ' Av. Bernardino Silveira Amorim']
Acidente c/ danos materiais entre dois carros na Av. Ipiranga próx. a R. Silva Só, sentido bairro/centro. E
['Av. Ipiranga ', ' R. Silva S']
Acidente com danos materiais entre dois carros na Av. Benjamin Constant próx. cruz. Av. Cristóvão Colombo,...
['Av. Benjamin Constant ', ' Av. Cristóvão Colom']
Acidente entre carro e moto na Av. Otto Niemeyer entre as ruas Silvio Silveira Soares x Tv. Escobar. EPTC e
['Av. Otto Niemeyer ', 'Silvio Silveira Soares ', 'Tv. Escobar.']
Acidente entre carro e táxi c/ danos materiais na Av. Ipiranga cruz. a R. Guilherme Alves, sentido centro/b
[' Av. Ipiranga', ' R. Guilherme Alves']
Acidente entre dois carros com danos materiais no cruz. das R. Quintino Bocaiúva e R. Casemiro de Abreu, ba
['R. Quintino Bocaiúva', ' R. Casemiro de Abreu']
Acidente entre dois carros na pista da direita da Av. Cel. Marcos, bairro Pedra Redonda, próximo a R. Evar
['Av. Cel. Marcos,', 'bairro Pedra Redonda,']
Acidente entre dois carros na R. Dom Pedro II, próx. a Marquês do Pombal. EPTC e SAMU no local.
['R. Dom Pedro II,', ' Marquês do Pomba']
Acidente entre moto e bicicleta no cruzamento da Av. Praia de Belas com Aureliano de Figueiredo Pinto. EPTC
['Av. Praia de Belas ', 'Aureliano de Figueiredo ']
```

FIM

<https://github.com/clarissacastella/twittercourse>
clarissacastella@gmail.com

Autores



Clarissa Castellã Xavier é pesquisadora de pós-doutorado na Universidade Federal do Rio Grande do Sul (UFRGS), mestre e doutora em Ciências da Computação pela Pontifícia Universidade Católica do Rio Grande do Sul (PUCRS). Começou a pesquisar em Processamento da Linguagem Natural (PLN) em 1999 no Grupo de Pesquisa em PLN da PUCRS. Desde então, trabalhou para empresas multi-culturais em todo o mundo, desenvolvendo ferramentas de processamento de linguagem com foco em mídias sociais e redes. Seu trabalho atual tem como foco a extração semântica de dados na área do transporte urbano a partir de redes sociais. Também é pesquisadora convidada do grupo FORMAS da Universidade Federal da Bahia (UFBA).

clarissacastella@gmail.com



marlovss@gmail.com



Marlo Souza é professor adjunto na Universidade Federal da Bahia (UFBA), mestre em Ciências da Computação pela Pontifícia Universidade Católica do Rio Grande do Sul (PUCRS) e doutor em Ciências da Computação pela Universidade Federal do Rio Grande do Sul (UFRGS). Suas atividades de ensino e pesquisa concentram-se nas áreas de computação teórica, representação do conhecimento, lógica aplicada e PLN. Iniciou suas pesquisas em PLN no ano de 2007 no contexto do projeto CoGROO - Corretor Gramatical livre para o OpenOffice, passando posteriormente pelo Grupo de Pesquisa em PLN da Pontifícia Universidade Católica do Rio Grande do Sul, em que estudou métodos de identificação de entidades e mineração de opiniões em textos do Twitter para monitoramento de marcas. Na UFBA, integra o grupo FORMAS trabalhando com métodos de extração de informações semânticas em texto.