Interactive Twitter Data Visualization

Word Cloud

Mehmet Duman

Goals?

1. Get data from twitter data and generate most used word visualization

2. Find Historical twitter data for 2016 Presidential Election day

3. Create Word Cloud to analyze twitter social media for selected filter

What is Twitter ?

- Online news and social networking service
- 6,000 /sec - 500 Million /day - 200 Billion /year (2016)
- 40 million tweets sent by 10pm **about** 2016 U.S. election day
- Twitter platform offers access to that corpus of data, via Twitter APIs
- Query result would be in JSON format (Advantage of Python, php, Ruby)

Some of the Twitter restrictions;

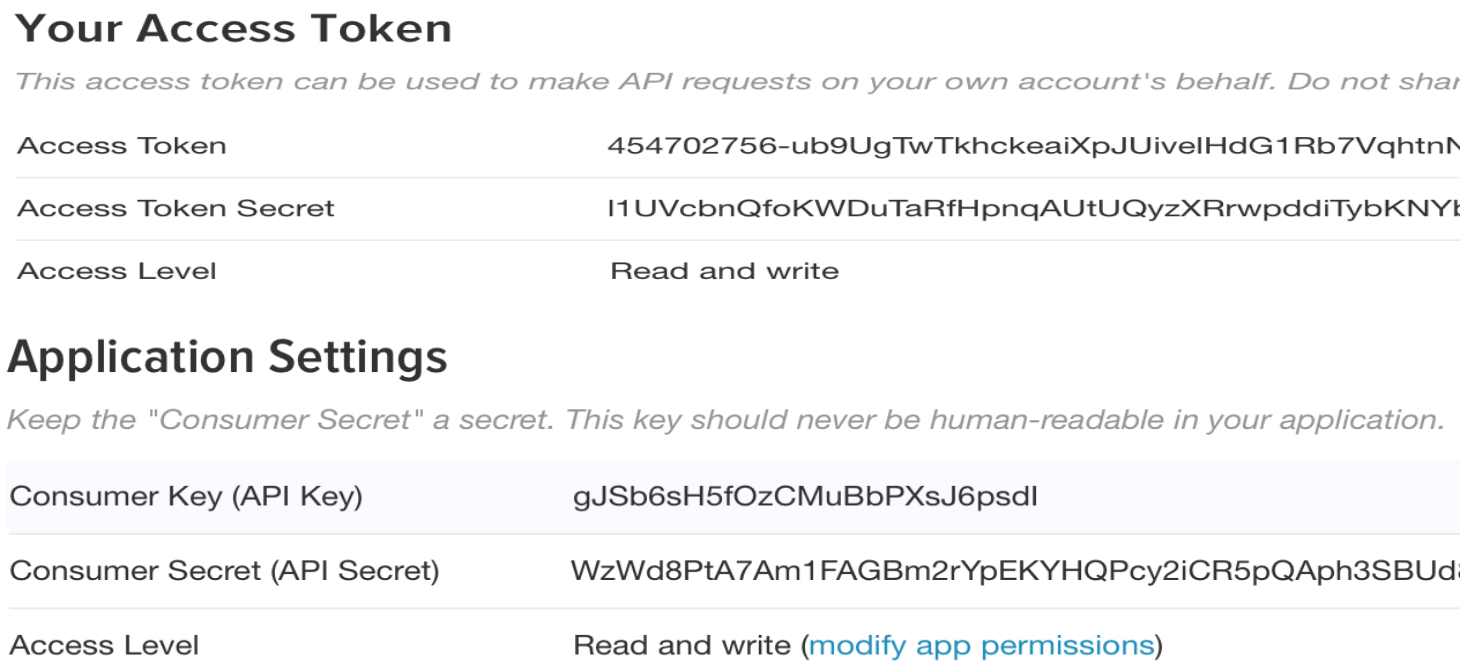
- 3,200 of a User's most recent Tweets (by screen_name or user_id)
- The Twitter Search API searches Tweets published in the past 7 days.
- User timelines belonging to protected users may only be requested

What Technologies Used for Project?

- Python
  - Tweepy library for accessing Twitter API
  - JSON
  - Wordcloud library
- JavaScript
- D3.js Visualization

How to get tweets? - Twitter Account

- Create a twitter account
- After logging create new "Applications“ to get communication data with Twitter API



Your Access Token

This access token can be used to make API requests on your own account's behalf. Do not shar

Access Token	454702756-ub9UgTwTkhckeaiXpJUivelHdG1Rb7VqhtnN
Access Token Secret	l1UVcbnQfoKWDuTaRfHpnqAUtUQyzXRrwpddiTybKNYb
Access Level	Read and write

Application Settings

Keep the "Consumer Secret" a secret. This key should never be human-readable in your application.

Consumer Key (API Key)	gJSb6sH5fOzCMuBbPXsJ6psdl
Consumer Secret (API Secret)	WzWd8PtA7Am1FAGBm2rYpEKYHQPcy2iCR5pQAph3SBUd
Access Level	Read and write (modify app permissions)

How to get tweets? - Tweepy – connection to Twitter API

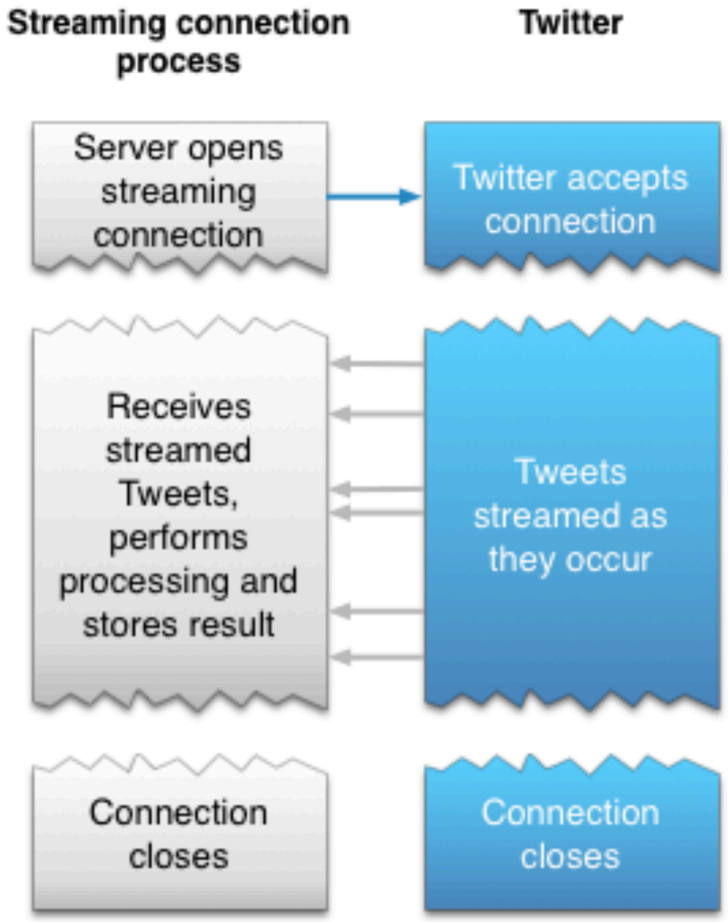
- Open-sourced Python library
- Enables to communicate with Twitter platform and uses Twitter API

A sample of how to access the Twitter API

```
import tweepy

#Variables that contains the user credentials to access Twitter API
access_token = "454702756-ub9UgTwTkhckeaiXpJUivelHdG1Rb7VqhtnNkMLv"
access_token_secret = "l1UVcbnQfoKWDuTaRfHpnqAUtUQyzXRrwpddi"
consumer_key = "gJSb6sH5fOzCMuBbPXsJ6psdI"
consumer_secret = "WzWd8PtA7Am1FAGBm2rYpEKYHQPcy2iCR5pQAph3SBUd"

#OAuth process, using the keys and tokens
auth = OAuthHandler(consumer_key, consumer_secret)
auth.set_access_token(access_token, access_token_secret)

#Creating of the actual interface, using authentication
api = tweepy.API(auth)
```

How to get tweets? - Tweepy StreamingAPI

- Monitoring for tweets and doing actions in real time and catches them when some event happens




```
class StdOutListener(StreamListener):
    ''' Handles data received from the stream. '''

    def on_status(self, status):
        # Prints the text of the tweet
        print('Tweet text: ' + status.text)

        # There are many options in the status object,
        # hashtags can be very easily accessed.
        for hashtag in status.entries['hashtags']:
            print(hashtag['text'])

        return true

    def on_error(self, status_code):
        print('Got an error with status code: ' + str(status_code))
        return True # To continue listening

    def on_timeout(self):
        print('Timeout...')
        return True # To continue listening
```

```
    try:
        auth = OAuthHandler(consumer_key, consumer_secret)
        auth.set_access_token(access_token, access_token_secret)
        if tfscr_name != '' :
            api = tweepy.API(auth)                    ## Get all tweeets for give tfscr_name on the setup file
            tw_tot = getTweetsByScreenName(api)
            createWordFile(tw_tot, 'U_'+tfscr_name )
        else:
            StdOutListener.ft_save = open(tf_text,'w')       ## Get new tweets
            l = StdOutListener()
            stream = Stream(auth, l)
            if tf_filter == ['']:
                stream.sample()
            else:
                stream.filter(track = tf_filter )
```

How to get tweets? - tweet sample

{"created_at":"Fri Nov 25 19:31:34 +0000 2016","id":802233284218368001,"id_str":"802233284218368001","text":"OH MY GOD: LOOK WHAT HAPPENED IMMEDIATELY AFTER TRUMP LANDED IN FLORIDA FOR THANKSGIVING! https:\/\/t.co\/WzB6edU0Rl https:\/\/t.co\/kTWKZxPCdF","display_text_range":[0,113],"source":"\u003ca href=\"http:\/\/dlvr.it\" rel=\"nofollow\"\u003edlvr.it\u003c\/a\u003e","truncated":false,"in_reply_to_status_id":null,"in_reply_to_status_id_str":null,"in_reply_to_user_id":null,"in_reply_to_user_id_str":null,"in_reply_to_screen_name":null,"user":{"id":772922637953667072,"id_str":"772922637953667072","name":"Students4Trump8","screen_name":"AnnaAlvarez1992","location":"South Carolina, USA","url":null,"description":"We believe students who think in honest objective terms make better citizens.\u201c.@students4trump8","protected":false,"verified":false,"followers_count":2254,"friends_count":2307,"listed_count":16,"favourites_count":7,"statuses_count":47840,"created_at":"Mon Sep 05 22:21:31 +0000 2016","utc_offset":-25200,"time_zone":"Arizona","geo_enabled":false,"lang":"es","contributors_enabled":false,"is_translator":false,"profile_background_color":"000000","profile_background_image_url":"http:\/\/abs.twimg.com\/images\/themes\/theme1\/bg.png","profile_background_image_url_https":"https:\/\/abs.twimg.com\/images\/themes\/theme1\/bg.png","profile_background_tile":false,"profile_link_color":"F58EA8","profile_sidebar_border_color":"000000","profile_sidebar_fill_color":"000000","profile_text_color":"000000","profile_use_background_image":false,"profile_image_url":"http:\/\/pbs.twimg.com\/profile_images\/772925759614750720\/D2dkyku0_normal.jpg","profile_image_url_https":"https:\/\/pbs.twimg.com\/profile_images\/772925759614750720\/D2dkyku0_normal.jpg","profile_banner_url":"https:\/\/pbs.twimg.com\/profile_banners\/772922637953667072\/1473123472","default_profile":false,"default_profile_image":false,"following":null,"follow_request_sent":null,"notifications":null},"geo":null,"coordinates":null,"place":null,"contributors":null,"is_quote_status":false,"retweet_count":0,"favorite_count":0,"entities":{"hashtags":[],"urls":[{"url":"https:\/\/t.co\/WzB6edU0Rl","expanded_url":"http:\/\/viid.me\/qq2WHi","display_url":"viid.me\/qq2WHi","indices":[90,113]}],"user_mentions":[],"symbols":[],"media":[{"id":802233280405721088,"id_str":"802233280405721088","indices":[114,137],"media_url":"http:\/\/pbs.twimg.com\/media\/CyIa6wmUUAApLWH.jpg","media_url_https":"https:\/\/pbs.twimg.com\/media\/CyIa6wmUUAApLWH.jpg","url":"https:\/\/t.co\/kTWKZxPCdF","display_url":"pic.twitter.com\/kTWKZxPCdF","expanded_url":"https:\/\/twitter.com\/AnnaAlvarez1992\/status\/802233284218368001\/photo\/1","type":"photo","sizes":{"medium":{"w":660,"h":330,"resize":"fit"},"thumb":{"w":150,"h":150,"resize":"crop"},"small":{"w":660,"h":330,"resize":"fit"},"large":{"w":660,"h":330,"resize":"fit"}}}]},"extended_entities":{"media":[{"id":802233280405721088,"id_str":"802233280405721088","indices":[114,137],"media_url":"http:\/\/pbs.twimg.com\/media\/CyIa6wmUUAApLWH.jpg","media_url_https":"https:\/\/pbs.twimg.com\/media\/CyIa6wmUUAApLWH.jpg","url":"https:\/\/t.co\/kTWKZxPCdF","display_url":"pic.twitter.com\/kTWKZxPCdF","expanded_url":"https:\/\/twitter.com\/AnnaAlvarez1992\/status\/802233284218368001\/photo\/1","type":"photo","sizes":{"medium":{"w":660,"h":330,"resize":"fit"},"thumb":{"w":150,"h":150,"resize":"crop"},"small":{"w":660,"h":330,"resize":"fit"},"large":{"w":660,"h":330,"resize":"fit"}}}]},"favorited":false,"retweeted":false,"possibly_sensitive":false,"filter_level":"low","lang":"en","timestamp_ms":"1480102294388"}

How to get tweets? - tweet sample

{"created_at":"Fri Nov 25 19:31:34 +0000 2016",

"id":802233284218368001,

"id_str":"802233284218368001",

"text":"**OH MY GOD: LOOK WHAT HAPPENED IMMEDIATELY AFTER TRUMP LANDED IN FLORIDA FOR THANKSGIVING!**
https:\/\/t.co\/WzB6edU0Rl
https:\/\/t.co\/kTWKZxPCdF",

"display_text_range":[0,113],

"source":"\u003ca href=\"http:\/\/dlvr.it\" rel=\"nofollow\"\u003edlvr.it\u003c\/a\u003e",

"truncated":false,

"in_reply_to_status_id":null,

"in_reply_to_status_id_str":null,

"in_reply_to_user_id":null,

"in_reply_to_user_id_str":null,

"in_reply_to_screen_name":null,

"user":{"id":772922637953667072,

"id_str":"772922637953667072",

"name":"Students4Trump8",

"screen_name":"AnnaAlvarez1992",

"location":"South Carolina, USA",

"url":null,

"description":"We believe students who think in honest objective terms make better citizens.\u201c.@students4trump8",

"protected":false,

"verified":false,

"followers_count":2254,

"

friends_count":2307,

"listed_count":16,

"favourites_count":7,

"statuses_count":47840,

"created_at":"Mon Sep 05 22:21:31 +0000 2016",

"utc_offset":-25200,

"time_zone":"Arizona",

"geo_enabled":false,

"lang":"es",

"contributors_enabled":false,

"is_translator":false,

"profile_background_color":"000000",

"profile_background_image_url":"http:\/\/abs.twimg.com\/images\/themes\/theme1\/bg.png",

"profile_background_image_url_https":"https:\/\/abs.twimg.com\/images\/themes\/theme1\/bg.png",

"profile_background_tile":false,

"profile_link_color":"F58EA8",

"profile_sidebar_border_color":"000000",

"profile_sidebar_fill_color":"000000",

"profile_text_color":"000000",

"profile_use_background_image":false,

"profile_image_url":"http:\/\/pbs.twimg.com\/profile_images\/772925759614750720\/D2dkyku0_normal.jpg",

"profile_image_url_https":"https:\/\/pbs.twimg.com\/profile_images\/772925759614750720\/D2dkyku0_normal.jpg",

"profile_banner_url":"https:\/\/pbs.twimg.com\/profile_banners\/772922637953667072\/1473123472",

"default_profile":false,

"default_profile_image":false,

"following":null,

"follow_request_sent":null,

"notifications":null},

"geo":null,

"coordinates":null,

"place":null,

"contributors":null,

"is_quote_status":false,

"retweet_count":0,

"favorite_count":0,

"favorite_count":0,

"entities":{"hashtags":[],

"urls":[{"url":"https:\/\/t.co\/WzB6edU0Rl",

"expanded_url":"http:\/\/viid.me\/qq2WHi",

"display_url":"viid.me\/qq2WHi",

"indices":[90,113]}],

"user_mentions":[],

"symbols":[],

"media":[{"id":802233280405721088,

"id_str":"802233280405721088",

"indices":[114,137],

"media_url":"http:\/\/pbs.twimg.com\/media\/CyIa6wmUUAApLWH.jpg",
"media_url_https":"https:\/\/pbs.twimg.com\/media\/CyIa6wmUUAApLWH.jpg","url":"https:\/\/t.co\/kTWKZxPCdF",
"display_url":"pic.twitter.com\/kTWKZxPCdF",
"expanded_url":"https:\/\/twitter.com\/AnnaAlvarez1992\/status\/802233284218368001\/photo\/1",
"type":"photo",
"sizes":{"medium":{"w":660,"h":330,"resize":"fit"},
"thumb":{"w":150,"h":150,"resize":"crop"},
"small":{"w":660,"h":330,"resize":"fit"},
"large":{"w":660,"h":330,"resize":"fit"}}}]},
"extended_entities":{"media":[{"id":802233280405721088,
"id_str":"802233280405721088",
"indices":[114,137],
"media_url":"http:\/\/pbs.twimg.com\/media\/CyIa6wmUUAApLWH.jpg",
"media_url_https":"https:\/\/pbs.twimg.com\/media\/CyIa6wmUUAApLWH.jpg",
"url":"https:\/\/t.co\/kTWKZxPCdF","display_url":"pic.twitter.com\/kTWKZxPCdF",
"expanded_url":"https:\/\/twitter.com\/AnnaAlvarez1992\/status\/802233284218368001\/photo\/1",
"type":"photo",
"sizes":{"medium":{"w":660,"h":330,"resize":"fit"},
"thumb":{"w":150,"h":150,"resize":"crop"},
"small":{"w":660,"h":330,"resize":"fit"},
"large":{"w":660,"h":330,"resize":"fit"}}}]},
"favorited":false,
"retweeted":false,
"possibly_sensitive":false,
"filter_level":"low",
"lang":"en",
"timestamp_ms":"1480102294388"}

How to get tweets? Python - Multiprocessing

Multiprocessing to read and process tweets by worker

```
import multiprocessing
from multiprocessing import Pool
from functools import partial

stopWords = getStopWordList(tf_stopword)

pool = multiprocessing.Pool(4)
part_func = partial(processTweet, stopWords)
with open(tf_text) as source_file:
    result = pool.map(part_func, source_file,4)

pool.close()
pool.join()
```

multiprocessing.Pool(*processes*)

A process pool object which controls a pool of worker processes to which jobs can be submitted

```
def processTweet(stopWords,line ):
    try:
        tweet = line.strip()
    except:
        return
    WordList=''

    tweet = tweet.lower()                                          #Convert to lower case
    tweet = re.sub('((www\.[^\s]+)|(https?://[^\s]+))','URL',tweet) #Convert www.* or https?://* to URL
    tweet = re.sub('@[^\s]+','AT_USER',tweet)                      #Convert @username to AT_USER
    tweet = re.sub('[\s]+', ' ', tweet)                            #Remove additional white spaces
    tweet = re.sub(r'#([^\s]+)', r'\1', tweet)                     #Replace #word with word
    tweet = tweet.strip('\'"')                                     #trim
    words = tweet.split()                                          #split tweet into words

    for w in words:
        w = replaceTwoOrMore(w)                                    #replace two or more with two occurrences
        w = w.strip('\'"?,.')                                      #strip punctuation
        val = re.search(r"^[a-zA-Z][a-zA-Z0-9]*$", w)              #check if the word stats with an alphabet
        if(w in stopWords or val is None):                         #ignore if it is a stop word
            continue
        else:
            WordList = WordList +  w.lower() + ','

        if WordList == None:
            WordList =''

        if WordList[:-1].endswith(','):
            WordList = WordList[:-1]

    return WordList
```

map(func, iterable[, chunksize])

A parallel equivalent of the map() built-in function (it supports only one iterable argument though). The Pool.map will lock the main program until all a process is finished, which is quite useful if we want to obtain resuls in a particular order for certain applications.

How to get tweets? Python - WordCloud library

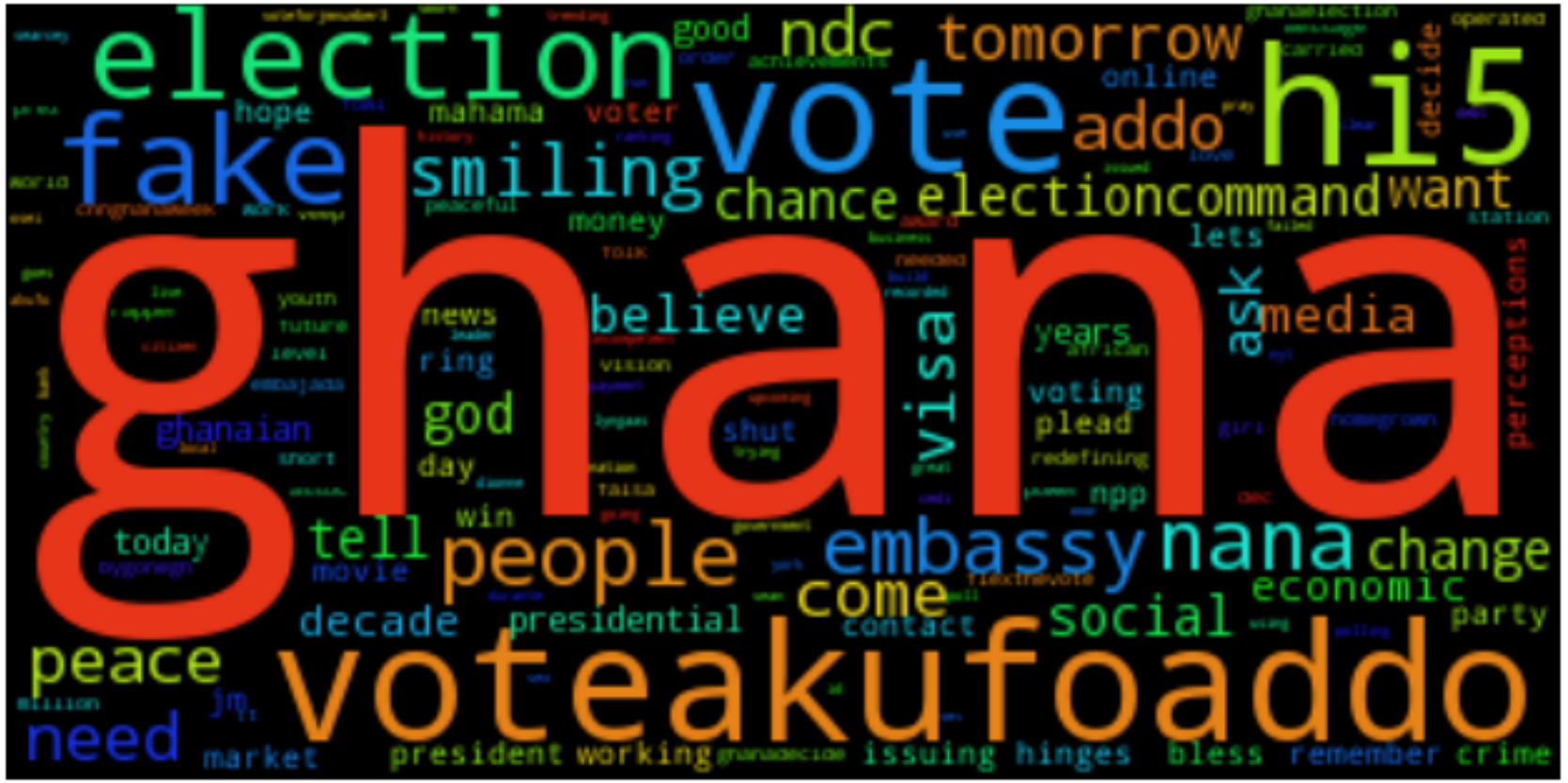
Twitter API authentication, retrieve real-time tweets and record it.

- Listen the Twitter API for new tweets
- Record tweets text properties by time-period or after certain number of tweets retrieved
- Run second python program by time-period or after certain number of tweets retrieved

```
# Generate a word cloud image  the matplotlib way:
WordListAll_text = " ".join(str(x) for x in result)

wordcloud = WordCloud().generate( WordListAll_text )

plt.imshow(wordcloud)
plt.axis("off")
plt.title('Number of tweets : ' + str(tw_tot) ,  fontsize=20)
```

How to get tweets? Python - WordCloud library

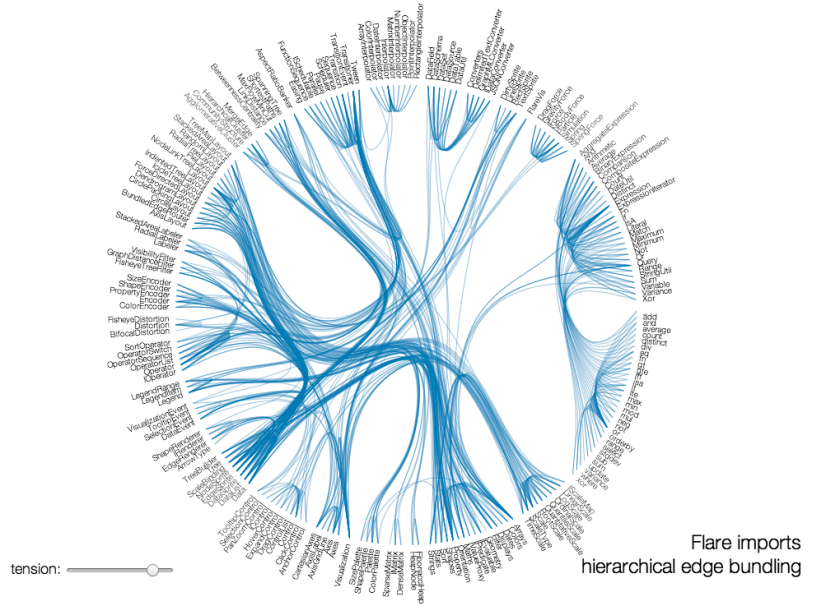
Number of tweets : 18816

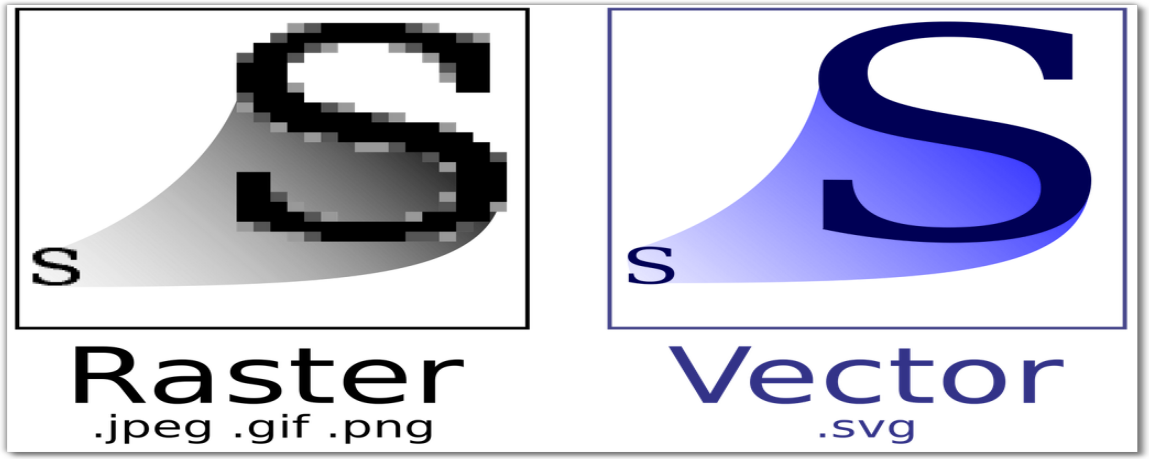


D3.js - D3 for Data-Driven Documents

- Producing dynamic,interactive data visualization

- Solves the fundamental problem of data visualization
  - Supports differential data update
  - Creates SVG (Scalable Vector Graphics) elements
  - helps to facilitate zoom-in, zoom-out functions





How to get tweets? Python – Word frequencies for D3.js WordCloud

Find word frequencies and calculate the weight based on frequencies

$$\text{LinearSize} = f_{min} + (\frac{(\text{WordFrequency} - w_{min})}{(w_{max} - w_{min})} * (f_{max} - f_{min})$$

$$\text{Weight} = (\frac{(\text{LinearSize})}{8})^3$$

```
WordListAll ={}
for text in result:
    if text is not None:
        for word in text.split(","):
            if word != "":
                if word in WordListAll:
                    WordListAll[word] +=1
                else:
                    WordListAll[word] = 1

df = pd.DataFrame(list(WordListAll.items()), columns=['word', 'size'])
df.sort_values(['size'], ascending=[False], inplace = True)
df_top = df.head(30)

f_min = 25
f_max = 35
w_min = df_top['size'].min().astype(int)
w_max = df_top['size'].max().astype(int)

df_top['weight'] = ( ( ( f_min +
    ( (df_top['size'] - w_min)  / (w_max - w_min) ) * (f_max -f_min) ) / 8 ) ** 3 ).apply(math.ceil)

dlist = df_top.to_json(orient='records')
open(tf_word,"w").writelines(dlist)
```

For weight calculation :http://www.bennybottema.com/2016/03/02/showcase-your-skillset-with-an-interactive-colorful-d3-js-tag-cloud/

JavaScript D3.js

- Interactive Visualization for text files (most popular top 30 words)
- Interactive updates for new top words
- Transition animation brings focus on the important changes

Twitter Screen Name.....:
Active Filter.....:
Recieved tweet.....: 200

Conclusion - 1

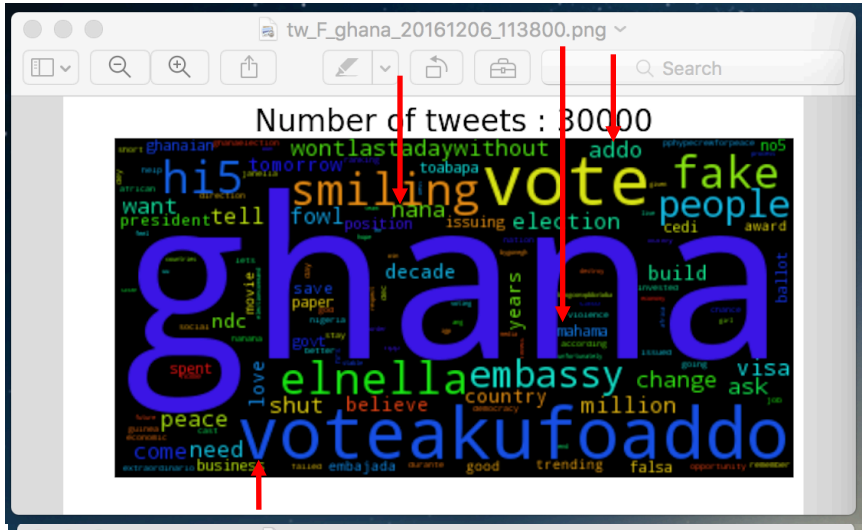
- Get data from twitter and generate most used word visualization
  - Yes, I was able to get new tweets and most recent 3200 tweets for selected user
- Find Historical twitter data for 2016 Presidential Election day
  - Twitter has one week restriction for old tweets, but if you have tweet ID's you can get it . 2016 Election day there was more than 40 million tweets. I found tweet ID's for 20 Million tweets, but It was to slow to retrieve it one by one.
- Create Word Cloud to analyze twitter social media for selected filter
  - Yes
- It is great to find out what social media talks about a subject by just using a simple filter

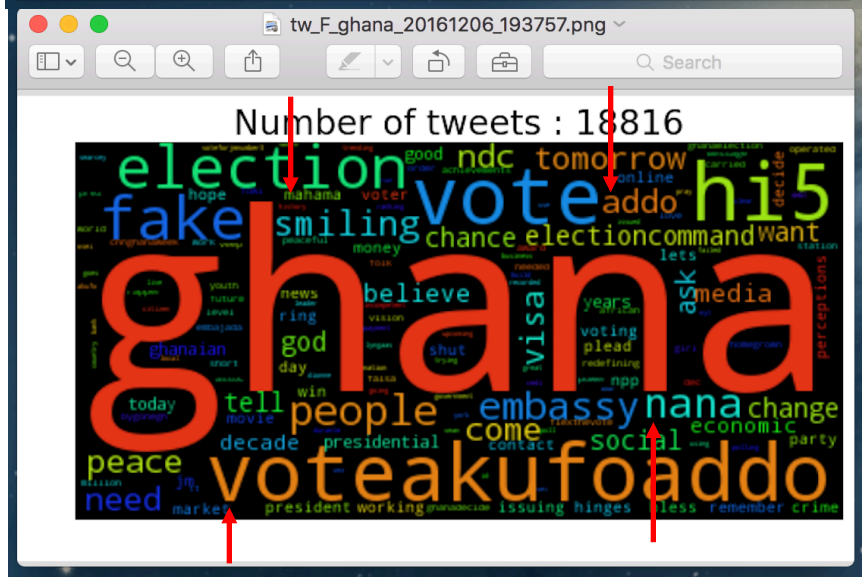
Conclusion - 2

- Python Multithreading needs to improve, so complex, I could’t use it
- Python Multiprocessing increased the performance significantly
- **C** OpenMP and **C** MPI are more powerful
- Tweepy is a great open-source library to access to the Twitter API, heavily relies on it and has great Streaming API support

Filter: ‘Ghana’ Tweets for Dec 6 2016 (8:25am-4:38pm)

Ghana Election
Dec 6,2016





Presidential Candidate
*John Draman **Mahama**
Ivor Greenstreet
Nana **Akufo-Addo (addo , voteakufoaddo)**
Pae Kwesi Nduon
Edward Mahama
Nana Konadu A. R
Jacob Osei Yeboah

Dec 7,2016 (today)

Python-WordCloud @realDonaldTrump Tweets:34.1K

Python-WordCloud @HillaryClinton Tweets:9.8K

The End

Thank You

Feel free to ask Questions