Twitter Data Mining & Sentiment Analytics
Keyword: Bitcoin

Code Description:

```
import json
credentials={"CONSUMER_KEY" : "iipQwmfCKvnXXKGgdiXspkkII",
 "CONSUMER_SECRET" : "EjFQNBrmYQEuaHhSQ0XZCN8wFRTIc2TdCaEsAXG87kwHa0x9pN",
 "ACCESS_TOKEN" :  "592652764-mF3Xd05rM7k7cJoafcxXWrtaqGpUmU7UjVXmI9Sv",
 "ACCESS_TOKEN_SECRET" : "1hzOqnym0JdEfS7rxBMPZsIzqIuAP3CCtYAy8dGhPQE5P"
}
```

The credentials were obtained after creating an application on https://apps.twitter.com/. Once, the credentials (which include Consumer Key, Consumer Secret, Access Token and Access Token Secret) were extracted, they were then fed into a json file, which was saved in the same directory, where the Python code was previous stored.

```
from twython import TwythonStreamer
import sys
# global variable to store tweets
tweets = []
```

In the next step, we installed a Python library 'Twython' to access the search and streaming twitter APIs. To install the 'Twython' library, we used pip into our terminal. After installing 'Twython', we imported the 'TwythonStreamer' module, since we want to access the Twitter streaming APIs. Next, we created a global variable called 'tweets' to store our tweets.

```
class MyStreamer(TwythonStreamer):
    '''our own subclass of TwythonStreamer'''

    # overriding
    def on_success(self, data):
        # check if the received tweet dictionary is in English
        if 'lang' in data and data['lang'] == 'en':
            tweets.append(data)
            print('received tweet #', len(tweets), data['text'][:100])

        # if we have enough tweets, store it into JSON file and disconnect API
        if len(tweets) >= 10000:
            self.store_json()
            self.disconnect()
```

Next, we create our own subclass of 'TwythonStreamer' i.e. 'MyStreamer' by inheriting from 'TwythonStreamer' class. Using the subclass, we then fetch 10,000 tweets and write these 10,000 tweets of data to a json file named 'tweet_stream_{*keyword*}_{*length of tweets*}.json' and save it in the same directory, where the python code was previously stored.

```
if len(sys.argv) > 1:
    keyword = sys.argv[1]
else:
    keyword = 'bitcoin'
```

Next, we define the keyword 'bitcoin' for our Twitter streaming API to extract tweets 10,000 with keyword 'bitcoin'.

```
import string
p = string.punctuation
d = string.digits
table_p = str.maketrans(p, len(p) * " ")
table_d = str.maketrans(d, len(d) * " ")
textstr=textstr.translate(table_p).translate(table_d)
```

For the string that contains all the wordlines for the 10,000 tweets, we first remove the punctuations and digits in those texts.

```
import nltk
stopwords=nltk.corpus.stopwords.words('english')
#add some meaningless words such as 'https' to the stopwords list
mystopwords=stopwords+['https','co','rt','the','will','btc','amp','get']
textlist_no_stop=[w for w in textstr.lower().split() if w not in mystopwords and len(w) > 1]
```

Then we further remove all the stoppers from our texts. For our analysis purpose, we add some high-frequency meaningless words to our stopper list.

```
dic['entities']['hashtags']
dic['entities']['user_mentions']
dic['user']['name']
```

For part b to part d in question B, the codes follow the similar logic. The only change is keys we use.

```
id_list=[]
for dic in rawlist:
    id_list.append(dic['id'])
#create dictionary if the id is in id_list
retweet_dic_insample={}
for dic in rawlist:
    if 'retweeted_status' in dic and dic['retweeted_status']['id'] in id_list:
        retweet_dic_insample[dic['retweeted_status']['retweet_count']+
                  dic['retweeted_status']['quote_count']+
                  dic['retweeted_status']['reply_count']]=dic['retweeted_status']['text']
#get the most influential tweet in my sample
max_count_insample=max(retweet_dic_insample.keys())
favor_tweet_insample=retweet_dic_insample[max_count_insample]
```

To get the most influential tweet, we first check whether a tweet a retweet; if yes, we then check whether it is retweeted from an earlier one in our sample. Then we create a dictionary with the sum of retweet_count, quote_count and reply_count as the key and the text as the values.

```
bull_mask=np.array(Image.open('bullmask.jpg')) #set the shape for mask
textstr_no_stop=''
for word in textlist_no_stop:
    textstr_no_stop +=' {}'.format(ss.stem(word))
wordcloud=WordCloud(background_color="white",max_font_size=60,mask=bull_mask).generate(textstr_no_stop)
plt.figure()
plt.imshow(wordcloud)
plt.axis("off")
plt.show()
```

Here we use the WordCloud module to draw the wordcloud figure. Here the shape we use is a bull.

```
from textblob import TextBlob
polarity=[]
subjectivity=[]
for dic in rawlist:
    t=dic['text'].translate(table_p).translate(table_d)
    tb_pos=TextBlob(t)
    polarity.append(tb_pos.sentiment.polarity)
    subjectivity.append(tb_pos.sentiment.subjectivity)
```

We first calculate the polarity and subjectivity score for each tweet and create two lists for polarity and subjectivity separately.

```
import matplotlib.pyplot as plt

plt.hist(polarity, bins=20) #, normed=1, alpha=0.75)
plt.xlabel('polarity score')
plt.ylabel('tweet count')
plt.grid(True)
plt.savefig('polarity.pdf')
plt.show()

plt.hist(polarity, bins=20) #, normed=1, alpha=0.75)
plt.xlabel('subjectivity score')
plt.ylabel('tweet count')
plt.grid(True)
plt.savefig('subjectivity.pdf')
plt.show()
```

```
mean_polarity=sum(polarity)/len(polarity)
mean_subjectivity=sum(subjectivity)/len(subjectivity)
```

Then we plot the distribution of these two scores and calculate the mean polarity and mean subjectivity.

Problem A. [Data Collection; 20 points] Pick an English keyword that interests your team. For example, iPhone, Vancouver, Trump, Exo, etc.[1] Using Twitter Streaming API, collect 10K tweets using the keyword and store the data into a JSON file.Question B- **[Preliminary Analysis; 20 points]** Using the collect tweets, please answer the following questions:

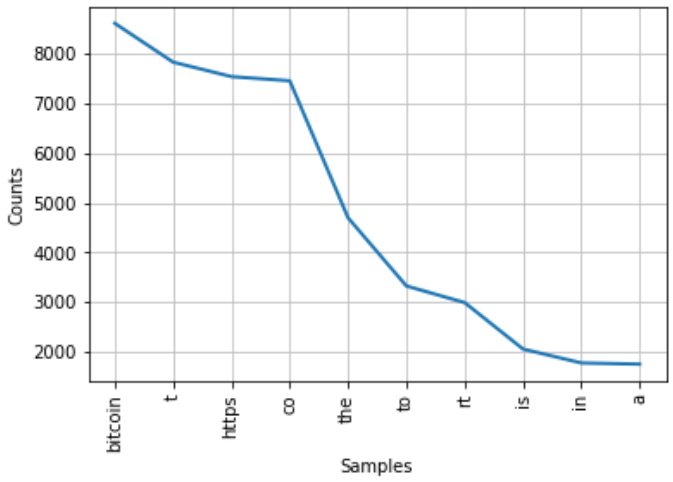
Keyword used: Bitcoin

JSON file containing 10K tweets has been attached with this file on CONNECT

Problem B. [Preliminary Analysis; 20 points] Using the collect tweets, please answer the following questions:

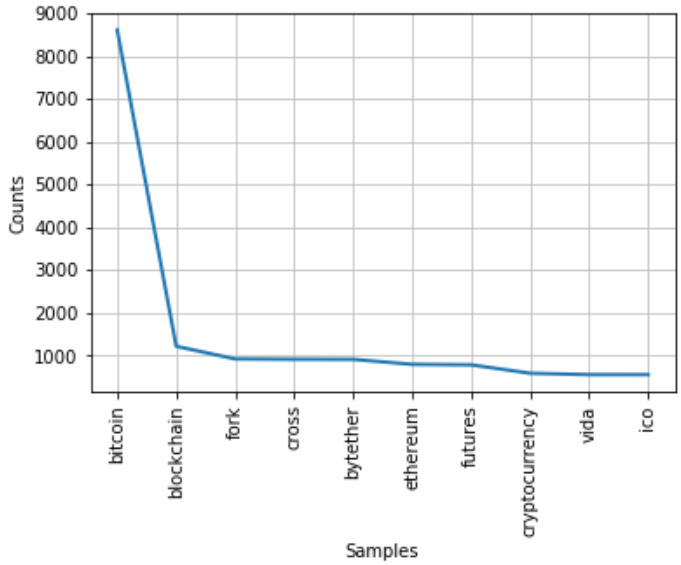
a. **What are the ten most popular keywords with and without stop words?**

Graph (including Stop Words)



Keywords: Bitcoin
Stopwords: t (tweet), https (url), co, the, to, rt (rtweet), is, in, a
We could not generate any relevant information from the above indicated stopwords. Hence, these words will be appended to the list of stopwords to reduce noise in the data.

Graph (without Stop Words)



It can be noticed by looking at the graph above that the frequency of word 'Bitcoin' is approx. 8500 in comparison to the expected value of 10,000. The primary reason for this could be that we only extracted the first 100 characters from each of the tweets. So in some cases, where occurrence of bitcoin was outside the first 100 character, those cases might not have accounted for the word 'Bitcoin'.

It is in a way self-explanatory that 'Bitcoin' indeed emerged as the most frequently occurring word, since our selected keyword was 'Bitcoin'. The frequency of the other top keywords ranged from 500-1200.

We have also included **btc** in stoppers because it is the value of bitcoin, in other words it is a representation of bitcoin itself (like USD for dollar)

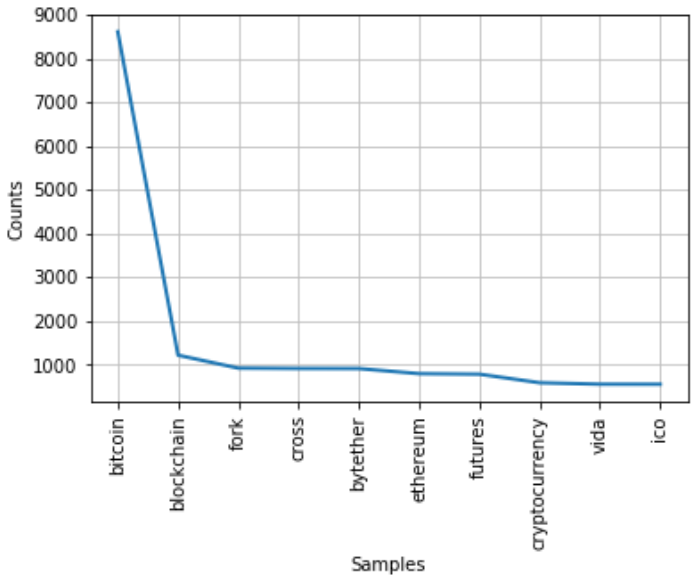
We can see that the keywords in the graph make sense as these terminologies are closely related to bitcoin/blockchain:

1. Blockchain: is a technology that makes bitcoin possible. Since bitcoin is a use case of blockchain technology, the two are often used interchangeably, which explains why blockchain follows next in the most frequently occurring words
2. Fork: A fork is a technical event that occurs when a blockchain diverges into two potential paths
3. Cross: A term associated with cross-blockchain transactions
4. Bytether: A new bitcoin planned to launch in Feb 2018
5. Ethereum: Ethereum is the 3rd most valuable form of digital money after bitcoin
6. Futures: A term related to futures trading in bitcoin
7. Cryptocurrency: A digital currency, of which bitcoin is the first decentralized example

8. ICO: stands for Initial Coin Offering, bitcoin's equivalent of IPO

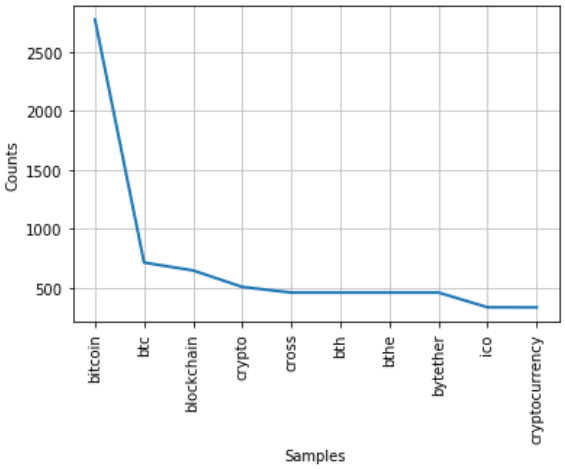
NOTE: We couldn't get an explanation for the association of "vida" in relation to bitcoin – we couldn't infer anything significant from this one as it's a Spanish word

Data only including noun, adjective and verb:



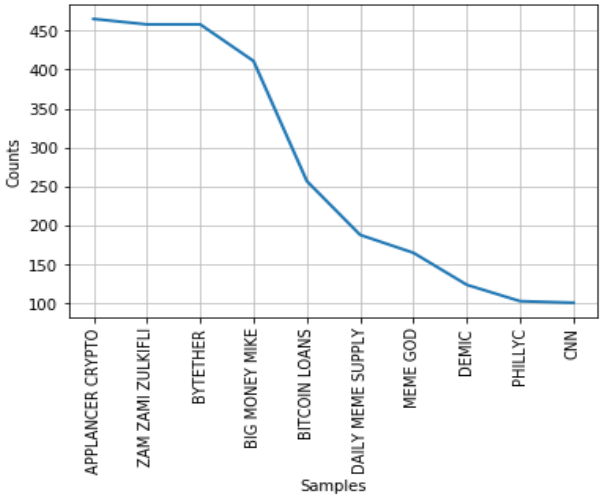
The top ten frequent words that we obtained with only including noun, adjective and verb is in line with the top ten frequent occurring words that we received after removing the stop words, since most of the keywords that we have received are actually nouns.

b. What are the ten most popular hashtags (#hashtag)?



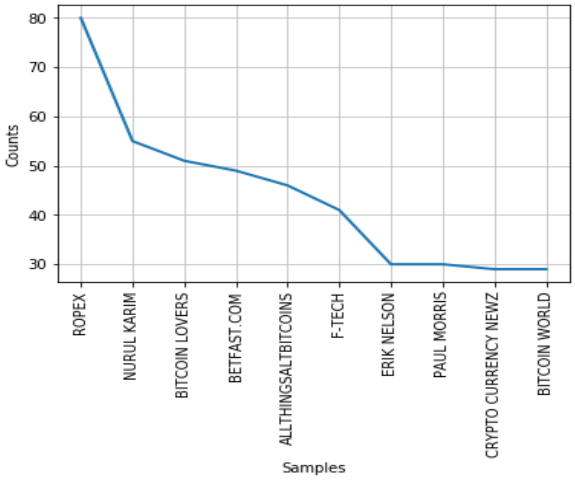
The set of results is similar to the keywords that we've listed above. As for the new additions to the list: *"bth"* is a forking (branching) of original currency btc, *"bthe"* implies bth exchange

c. What are the ten most frequently appearing usernames (@username)?



What we see here are some usernames related to currency and finance: appliance crypto, byether, bitcoin loans; the others are news/entertainment-related usernames: daily meme supply, meme god, big money mike, phillyc, cnn. So this indicates that bitcoin's popularity transcends the boundaries of finance enthusiasts and has an established reputation in the meme world!

d. Who is the most frequently tweeting person about the keyword?



The above graph indicates the user names that created the buzz around the keyword 'Bitcoin' at the time when tweets were extracted. Ropex takes the lead in this, with some 80 counts. Followed by a significant gap are Nurul Karim and Bitcoin lovers, with close to 50 counts each. The interesting thing to note here is that Ropex has very recently joined

twitter and managed to build a good reach (with 3,961 followers). Since we have captured tweets over a very small timeframe of 2-3 hours, we didn't get significant data on retweets from users' tweets and can't infer conclusively about their influence among their followers in relation to the keywords.

Snapshot of Ropex's profile on twitter

e. Which is the most influential tweet? (Let's define that influence is the sum of retweet count, reply count, and quote count.)

i. <u>The most influential tweet (that our sample retweeted from) based on sum of counts:</u>

This calculation is done on the basis of stats available from the 'retweeted_status' section of the extracted data – i.e., it represents the sum of counts till date (time of extraction) of the tweet's retweet, reply and quote. Note that this tweet selected may not be available in our data except in the form of a retweet.

Bitcoin Giveaway: Once Bitcoin officially passes the $10,000 milestone, we will be giving away 1 BTC to a random follower. <u>https://t.co/XyRc9cF6d4</u>

Influence was found out to be: 20,048 (i.e. sum of retweet count, reply count, and quote count)

This is a post by @btc handle on twitter, which runs the website Coinbase: a secure online platform for buying, selling, transferring, and storing digital currency. The post was posted on early hours of Nov 28, 2017.

<u>Market scenario before tweet:</u>
The following is an <u>excerpt</u> from a news site dated Nov 27, when the price was $9,531:

> **27 November, Swissquote** -Talk of Bitcoin reaching $10,000 before the end of 2017 had unquestionably done its rounds, still, to be holding at $9,531.97 before the end of November, $10,000 is going to be a predictable conclusion, hoping the bubble doesn't burst.

Market scenario after tweet:
The price hit $10,000 by Tuesday, 28 Nov 2017.

(Click here for image source)

The tweet's popularity can be attributed to the following:

i. Timing of the situation – see market price of bitcoin before and after the tweet
ii. Interest in the speculation around bitcoin, since bitcoin mainly derives its valuation from speculation in futures market
iii. Contest of retweeting that promised an award of 1 bitcoin generated traction contributing to rising popularity (see tweet for context)

ii. The most influential tweet (present in our extracted sample):

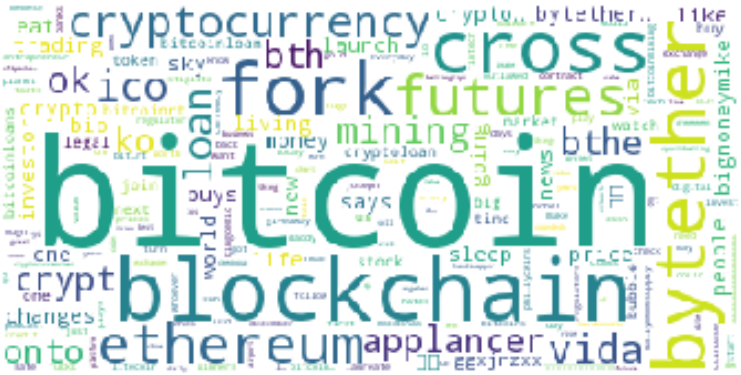
Nobel laureate Joseph Stiglitz says bitcoin "ought to be outlawed"
https://t.co/9sO0qa6ys3

Count: 215

This post released on Dec 01, 2017 talks about the opinion of noble winners Joseph Stiglitz and Robert Shiller towards bitcoin. Both people are notable for their contribution to the fields of finance and economics, and serve as influencers to their followers. With the bitcoin valuation surging beyond $10,000 on Nov 28, it drew attention from these people who criticized it for being deregulated, and anticipating a bubble burst.

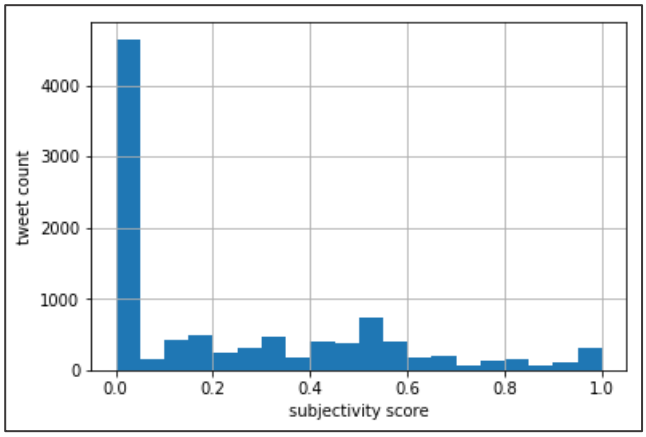
The influence of this tweet indicates the perception about the bitcoin's value is not clear.

Question C - [Word Cloud; 20 points] Create a word cloud from the collected 10K tweets. Depending on the needs, you may want to remove stop words and do stemming before feeding into the word cloud module.

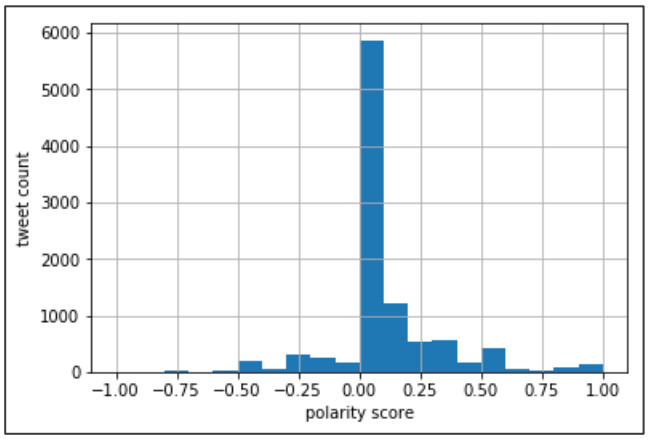


As anticipated, the WordCloud is consistent with our observations on the most frequently occurring words. The size of the word indicates its relative frequency. We can also notice less frequent but significant words such as mining, investors, stock, price which indicate bitcoin's association with financial market, especially stock market.

Question D - [Sentiment Analysis; 20 points] Using TextBlob, calculate the polarity and subjectivity scores for each tweet in the 10K tweet corpus. Summarize the calculated scores with histograms using Matplotlib, where X-axis is the score and Y-axis is the tweet count in the score bin. Also, provide the average of the polarity and subjectivity scores.



The subjectivity score indicates that there is strong opinion associated with it as most tweets gravitate towards zero score. However other tweets (avg. frequency around 300) display other range of subjectivity scores.
We can interpret this as that since the market over time has matured for bitcoin and it has a strong follower base, which is why majority of the *tweeple* have evolved to have a strong view on bitcoin. However a smaller chunk of population perceives it as subjective, and this could be tied with the speculation and uncertainity around bitcoin in the market



The average polarity score gravitates towards zero, indicating a neutral sentiment towards the concept.
For the tweets that were opinionated, the positive score dominated the negative; correlating with the market sentiment at that time – with the bitcoin valuation surging beyond $10,000 for the first time.
The negative sentiments can be attributed to uncertainty given a deregulated market and the fear of "bubble burst"

Mean polarity score= 0.0879182
Mean subjectivity score= 0.240587

Question E - [Insights; 20 points] At the end of the day, what we want from all these analyses is the insights. Please describe the insights you gained from the analyses. I look forward to seeing your unique perspectives. Good luck!

1. The most frequent keywords are practically buzzwords around bitcoin right now. Some of the words like forking and bytether are very recent developments and are already gathering attention.
2. For the most frequently occurring keywords on bitcoin, we can run a further sentiment analysis to see how positive/negative is the sentiment around them. For example, bitcoin itself may have a negative sentiment associated with it, but blockchain technology could have a positive sentiment. In this case, someone addressing audience on bitcoin might prefer to promote it as a blockchain product than specifically bitcoin
3. We can identify other words that aren't important and can be removed to keep the context relevant. E.g. seeing the WordCloud we know that OK, onto are general words but they have pretty much similar size as ICO – which is significant keyword. So their presence can be distracting and with no value add
4. Code effectiveness, and suggestions: Pulling more characters (more data consumed) to see if there are patterns by using more related keywords would provide more insight into the field. This is because the first 100 may not necessarily contain the relevant and complete point. For e.g. the first part could be a statement/observation, followed by a negation/ counterargument in the second part – in this case the latter is the one being emphasized and should be taken into account.