ASSIGNMENT -1

LAMBTON COLLEGE

CBD 3335_2 DATA MINING AND ANALYSIS

Team Members

SAI VARUN KOLLIPARA – C0828403

BHANU PRAKASH MAHADEVUNI – C0850515

DEEKSHA NAIKAP – C0835440

PRAMOD REDDY GURRALA – C0850493

Instructor: Ali Nouhi

Date: 08-07-2022

Problem Statement:

The main objective is to explore and gain experience of collecting real-time data from Twitter using Twitter API and gain experience of saving data into different formats like csv and performing various data preprocessing and cleaning activities and finally to get some visualizations for the number of tweets per keywords and number of users versus the number of tweets.

Solution Approach:

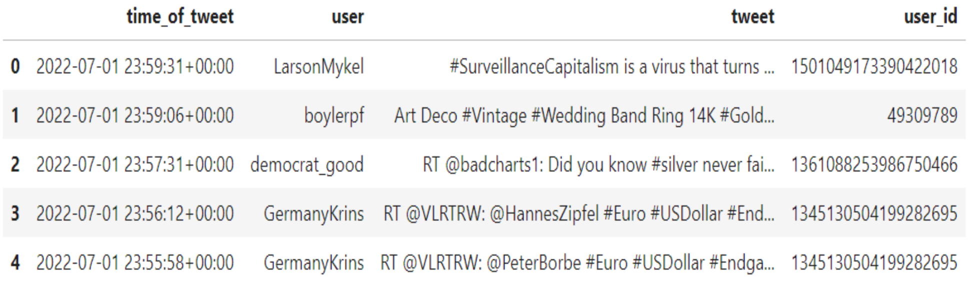
The approached solution is to make use of required python libraries and packages for data frame and data visualization by importing them and make use of twitter developer account for collection of required tweets. Upon collection of tweets based on keywords and saving it into 8 csv files, we performed some data preprocessing and cleaning steps and then performed data visualizations using python libraries like matplotlib.

Steps Involved:

- Collecting data
- Saving data
- Cleaning data
- Visualizing data

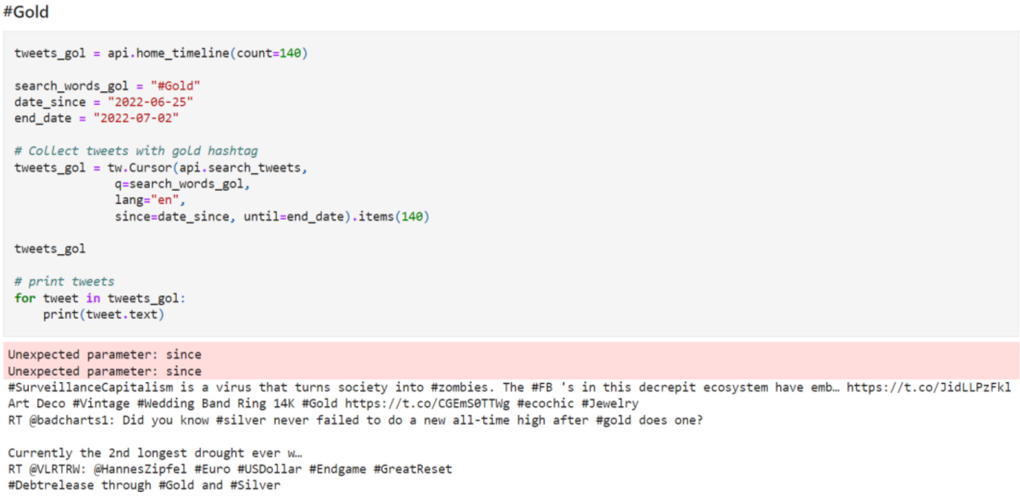
Methodology:

1. **Collecting Data**
   In this task, we have collected tweets from twitter for around one week using given keywords related to stock market. Here are the hashtags info we collected.
   a. #Altcoin
   b. #Bitcoin
   c. #Coindesk
   d. #Cryptocurrency
   e. #Gold
   f. #APPL
   g. #GOOG
   h. #YHOO



	time_of_tweet	user	tweet	user_id
0	2022-07-01 23:59:31+00:00	LarsonMykel	#SurveillanceCapitalism is a virus that turns ...	1501049173390422018
1	2022-07-01 23:59:06+00:00	boylerpf	Art Deco #Vintage #Wedding Band Ring 14K #Gold...	49309789
2	2022-07-01 23:57:31+00:00	democrat_good	RT @badcharts1: Did you know #silver never fai...	1361088253986750466
3	2022-07-01 23:56:12+00:00	GermanyKrins	RT @VLRTRW: @HannesZipfel #Euro #USDollar #End...	1345130504199282695
4	2022-07-01 23:55:58+00:00	GermanyKrins	RT @VLRTRW: @PeterBorbe #Euro #USDollar #Endga...	1345130504199282695

Sample Info for the data frame



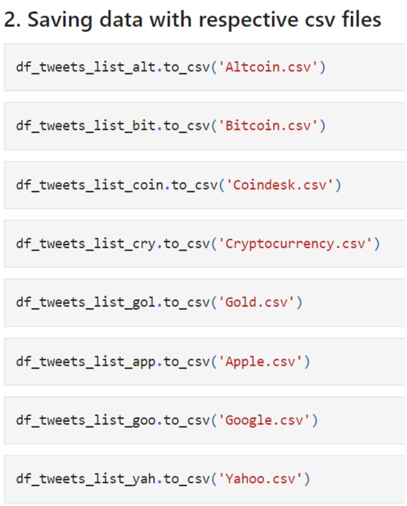
#Gold

```
tweets_gol = api.home_timeline(count=140)

search_words_gol = "#Gold"
date_since = "2022-06-25"
end_date = "2022-07-02"

# Collect tweets with gold hashtag
tweets_gol = tw.Cursor(api.search_tweets,
              q=search_words_gol,
              lang="en",
              since=date_since, until=end_date).items(140)

tweets_gol

# print tweets
for tweet in tweets_gol:
    print(tweet.text)
```

```
Unexpected parameter: since
Unexpected parameter: since
#SurveillanceCapitalism is a virus that turns society into #zombies. The #FB 's in this decrepit ecosystem have emb… https://t.co/JidLLPzFkl
Art Deco #Vintage #Wedding Band Ring 14K #Gold https://t.co/CGEmS0TTWg #ecochic #Jewelry
RT @badcharts1: Did you know #silver never failed to do a new all-time high after #gold does one?

Currently the 2nd longest drought ever w…
RT @VLRTRW: @HannesZipfel #Euro #USDollar #Endgame #GreatReset
#Debtrelease through #Gold and #Silver
```

Info on the timeline, hashtag and limit of tweets for #Gold Keyword

2. **Saving data**
   The collected information about the tweets will be in json format and then we are saving the json tweets data into 8 different CSV files where data related to each keyword is saved and each file consists of four columns with labels tweet id, time of the tweet, user id and text.



2. Saving data with respective csv files

```
df_tweets_list_alt.to_csv('Altcoin.csv')
```

```
df_tweets_list_bit.to_csv('Bitcoin.csv')
```

```
df_tweets_list_coin.to_csv('Coindesk.csv')
```

```
df_tweets_list_cry.to_csv('Cryptocurrency.csv')
```

```
df_tweets_list_gol.to_csv('Gold.csv')
```

```
df_tweets_list_app.to_csv('Apple.csv')
```

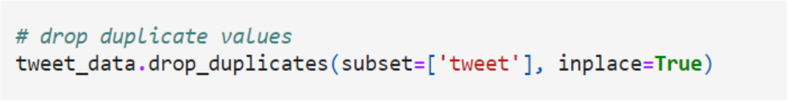
```
df_tweets_list_goo.to_csv('Google.csv')
```

```
df_tweets_list_yah.to_csv('Yahoo.csv')
```

Saving the data frame as CSV files

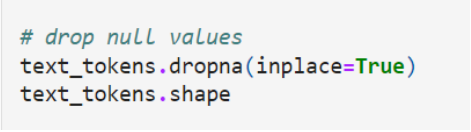
3. Cleaning data

As part of preprocessing and cleaning data we have dropped duplicate and null values, removed the punctuations, removed the numbers, checked the tweets with length less than 2 and removed them.



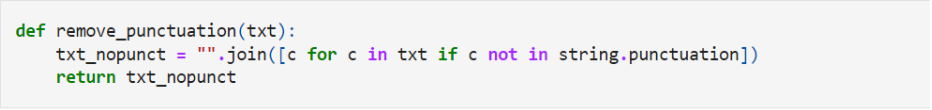
```
# drop duplicate values
tweet_data.drop_duplicates(subset=['tweet'], inplace=True)
```

Dropping duplicates



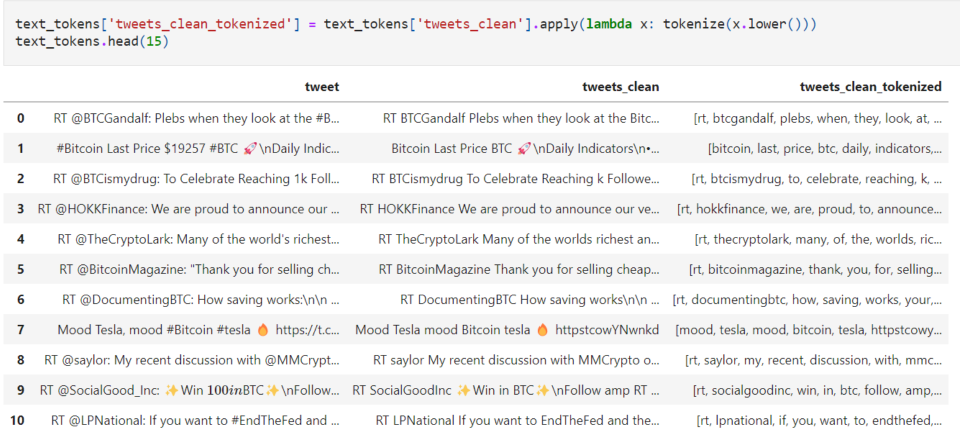
```
# drop null values
text_tokens.dropna(inplace=True)
text_tokens.shape
```

Dropping the null values



```
def remove_punctuation(txt):
    txt_nopunct = "".join([c for c in txt if c not in string.punctuation])
    return txt_nopunct
```

Removing Punctuations

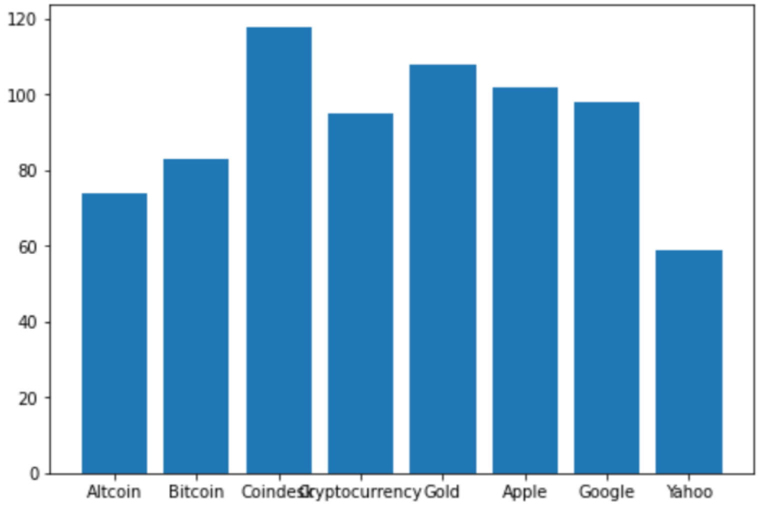


```
text_tokens['tweets_clean_tokenized'] = text_tokens['tweets_clean'].apply(lambda x: tokenize(x.lower()))
text_tokens.head(15)
```

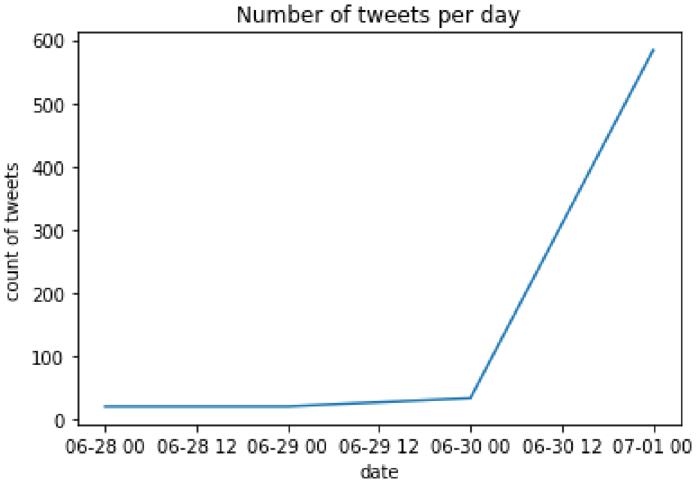
	tweet	tweets_clean	tweets_clean_tokenized
0	RT @BTCGandalf: Plebs when they look at the #B...	RT BTCGandalf Plebs when they look at the Bitc...	[rt, btcgandalf, plebs, when, they, look, at, ...
1	#Bitcoin Last Price $19257 #BTC 🚀\nDaily Indic...	Bitcoin Last Price BTC 🚀\nDaily Indicators\n•...	[bitcoin, last, price, btc, daily, indicators,...
2	RT @BTCismydrug: To Celebrate Reaching 1k Foll...	RT BTCismydrug To Celebrate Reaching k Followe...	[rt, btcismydrug, to, celebrate, reaching, k, ...
3	RT @HOKKFinance: We are proud to announce our ...	RT HOKKFinance We are proud to announce our ve...	[rt, hokkfinance, we, are, proud, to, announce...
4	RT @TheCryptoLark: Many of the world's richest...	RT TheCryptoLark Many of the worlds richest an...	[rt, thecryptolark, many, of, the, worlds, ric...
5	RT @BitcoinMagazine: "Thank you for selling ch...	RT BitcoinMagazine Thank you for selling cheap...	[rt, bitcoinmagazine, thank, you, for, selling...
6	RT @DocumentingBTC: How saving works:\n\n ...	RT DocumentingBTC How saving works\n\n ...	[rt, documentingbtc, how, saving, works, your,...
7	Mood Tesla, mood #Bitcoin #tesla 🔥 https://t.c...	Mood Tesla mood Bitcoin tesla 🔥 httpstcowYNwnkd	[mood, tesla, mood, bitcoin, tesla, httpstcowy...
8	RT @saylor: My recent discussion with @MMCrypt...	RT saylor My recent discussion with MMCrypto o...	[rt, saylor, my, recent, discussion, with, mmc...
9	RT @SocialGood_Inc: ✨Win 100inBTC✨\nFollow...	RT SocialGoodInc ✨Win in BTC✨\nFollow amp RT ...	[rt, socialgoodinc, win, in, btc, follow, amp,...
10	RT @LPNational: If you want to #EndTheFed and ...	RT LPNational If you want to EndTheFed and the...	[rt, lpnational, if, you, want, to, endthefed,...

Cleaned and Tokenized Data Frame

4. Visualizing data



Bar Graph representation for number of tweets per hashtag



Graph representation for number of tweets per day

Results and Conclusion:

- From the analysis we learnt how to use the tweetpy API to collect the tweets.
- We learnt how to handle the tweets and paraments to collect the tweets based on dates, hashtags and number of tweets.
- We also performed cleaning the tweets text and tokenizing them.
- We also plotted the graphs and charts on number of tweets per hashtag and number of tweets per day.

References:

https://dev.twitter.com/overview/documentation

https://www.python.org/doc/

https://www.tweepy.org/