**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

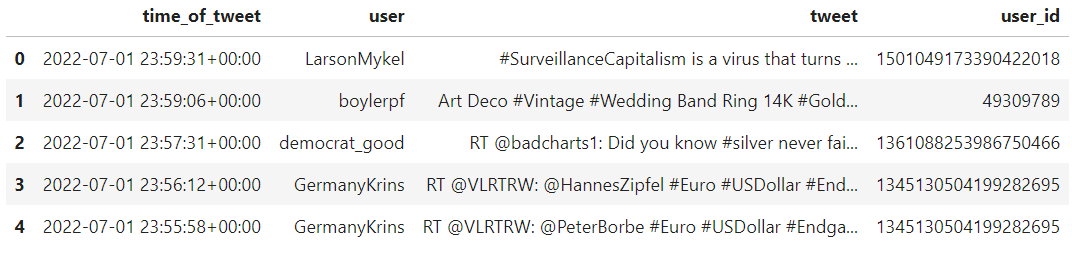
d. #Cryptocurrency

e. #Gold

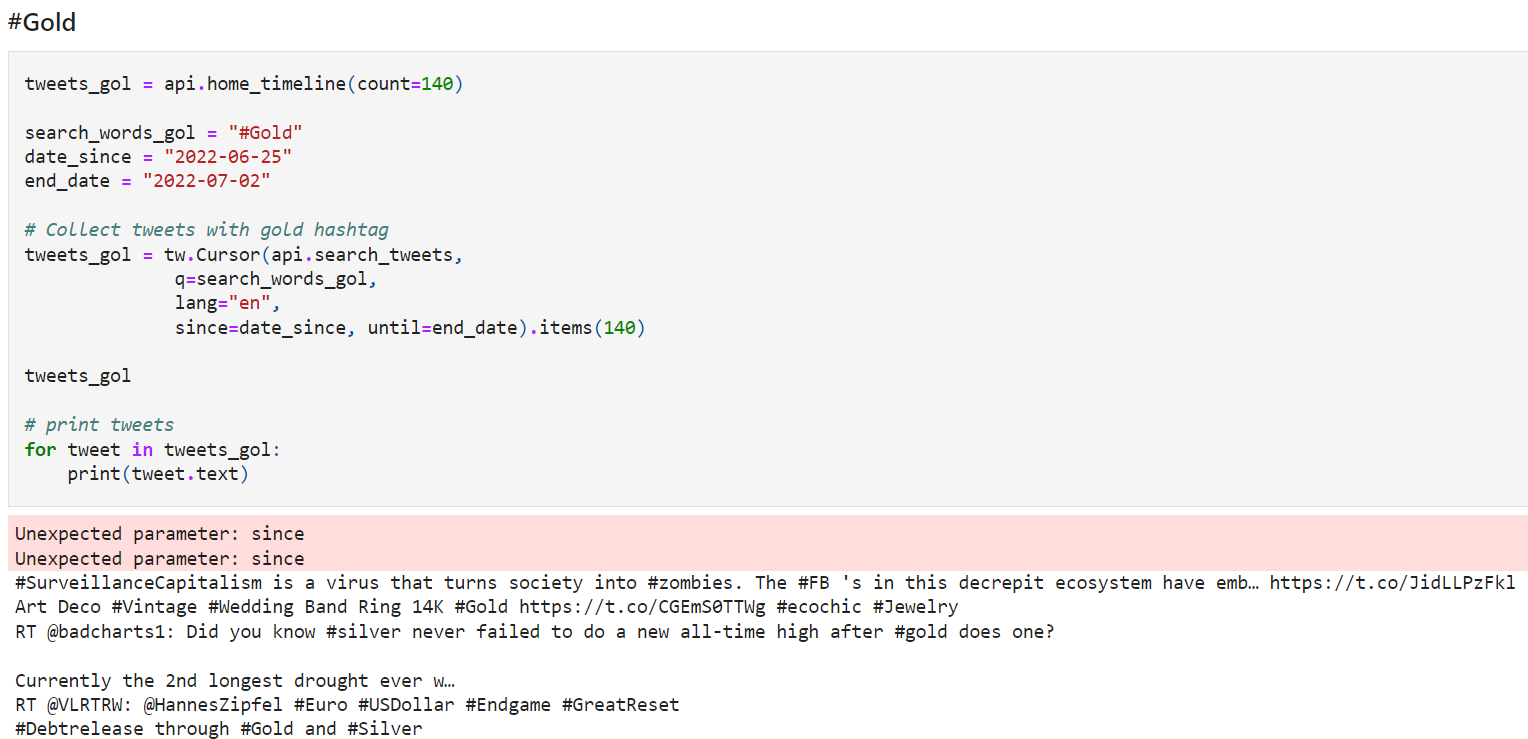
f. #APPL

g. #GOOG

h. #YHOO

­­­­

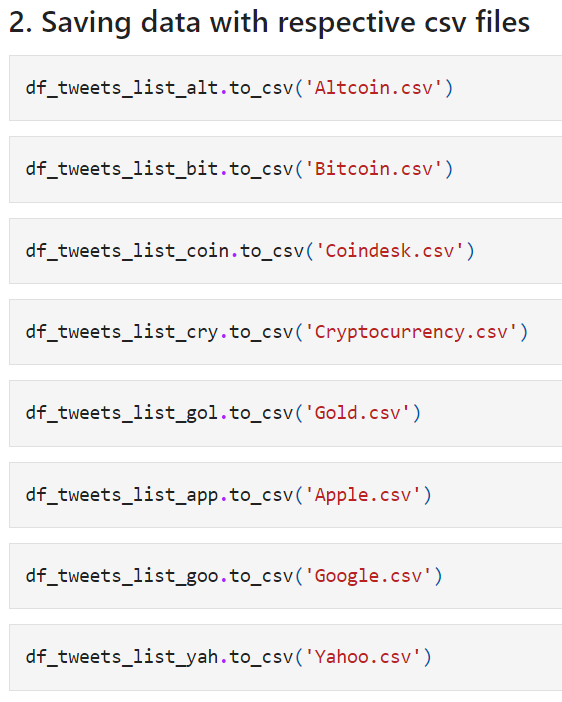
Sample Info for the data frame



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

1. **Saving data**

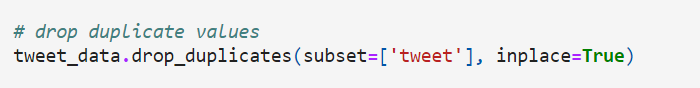
The collected information about the tweets will be in json format and then we are saving the json tweets data into 8 different CSV files where data related to each keyword is saved and each file consists of four columns with labels tweet id, time of the tweet, user id and text.



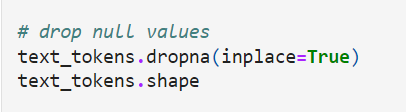
Saving the data frame as CSV files

1. **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.



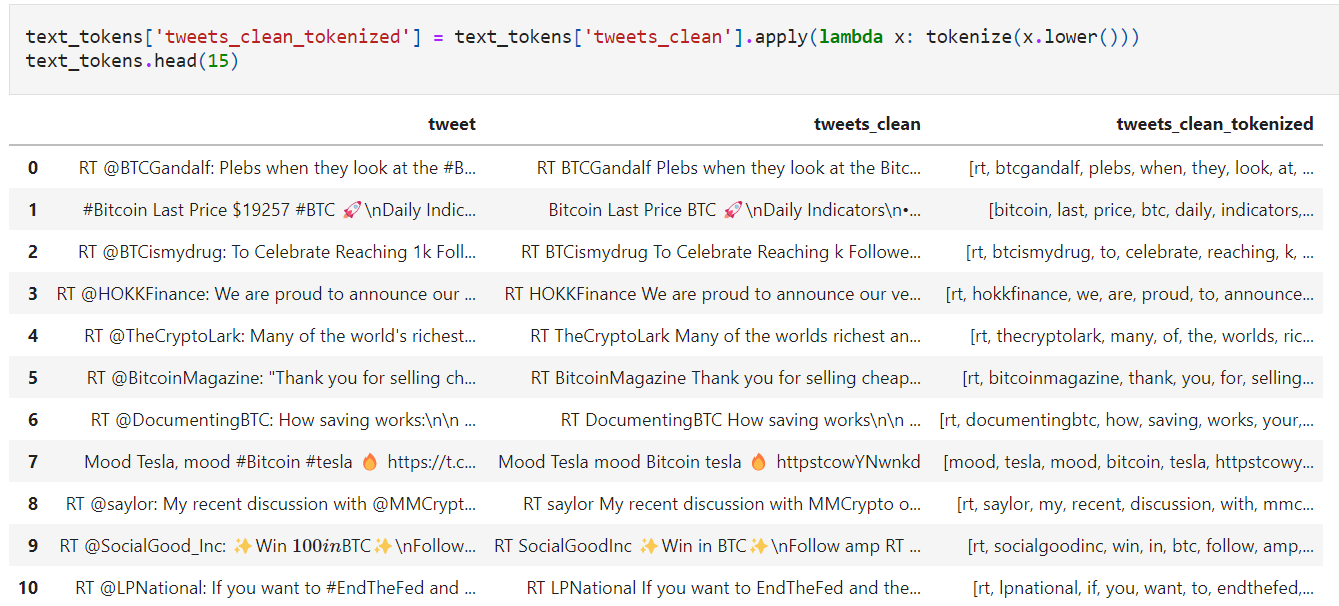
Dropping duplicates



Dropping the null values

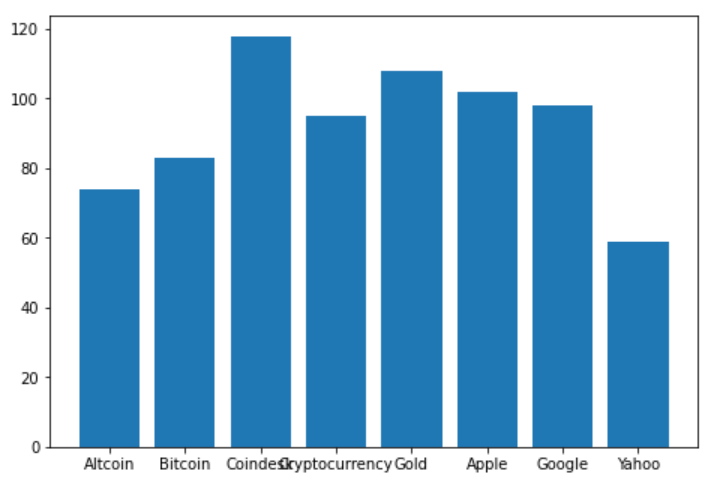


Removing Punctuations

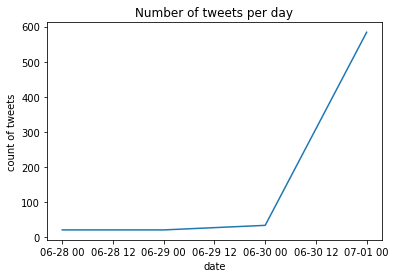


Cleaned and Tokenized Data Frame

1. **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://dev.twitter.com/overview/documentation%20)

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

<https://www.tweepy.org/>