Collecting tweets related to stock market

Objectives

- Gain experience of collecting data from Twitter using Twitter API
- Gain experience of data storage to store the data and query
- Gain experience of collecting real-time data
- Gain experience of data cleaning

Important Notes

- Work in groups of 4 students
- All reports must be submitted as a PDF, with source code as an archive (e.g. zip, tar.gz)
- Save the submission as “Assignment-1_Group-#.zip”

Assignment Details

This assignment consists of four parts:

1. **Collecting data:** In this assignment you need to collect data related to stock market from Twitter for one week. In Twitter, ticker symbols like #gold are used for stocks and companies. You are requested to collect the tweets with some specific keywords and store them in different files. The following keywords should be used:

a. Altcoin
b. Bitcoin
c. Coindesk
d. Cryptocurrency
e. Gold
f. APPL
g. GOOG
h. YHOO

Each tweet is a json file with the following format:

```
{"created_at":”……….”,
"id":”………..”,
"text":" Time to buy some ether!\n#ethereum #investing #cryptocurrency”
“user_id”:”………..”
…
}
```

2. **Saving data:** You need to save the requested data into csv format of 8 files where data related to each keyword is saved. Each file consist of four columns: tweet id, time of tweet, user id and text.

3. **Cleaning data:** remove duplication, remove punctuations, remove numbers in tweets, and remove words with length less than 2.

4. **Visualizing data:** You need to present the daily number of tweets for each keyword as well as the daily number of users.

Use Clustering of similar tweets if feasible and applicable.

Recourses

It is recommended for this assignment to use Python for programming with Tweepy.

The following resources are helpful.
https://dev.twitter.com/overview/documentation
https://www.python.org/doc/