**Mini Project Report2**

Sijin Zhou IST652 M001

1. **The data and source**

My data come from Twitter website. For all the questions below, I extracted 500 tweets from tweets include keywords, or each 500 tweets from different users.

Here are the API\_KEY and Token information:

**Graphical user interface, text, application, email

Description automatically generated**

1. **Data preparation**

Firstly, I searched for 500 tweets that contain key word "Biden" since 2020-1-1 and changed all those json format tweets into dataframe using pd.DataFrame function to clearly see the fields of those tweets.

Then I used nltk package to analyze hashtags of those tweets, look at the word count of the texts and hashtags.

Finally, I did sentiment analysis of those tweets using textblob package to explore the polarity.

1. **Unit of analysis**

**Q1:** Process one collection of data and summarize information from several fields.

**Q1.a:** select specific fields of tweets, they are the time of this tweet created, the user name of this tweet, the location of this user and the tweet content. The head of this dataframe is like:

A picture containing graphical user interface, application

Description automatically generated

**Q1.b:** Analyze the hashtag of those tweets

1. Find how many tweets include hashtags of those 500 tweets
2. Select the most 10 common hashtags and convert the hashtag word and word count into a dataframe. This step will help us understand what hashtag usually be when users posed tweets about ‘Biden’

Table

Description automatically generated

**Q1.C:** Analyze the text of those tweets

1. Use NLTK package to analysis what text usually be in those tweets, most people post the election and Trump with Biden

Table

Description automatically generated

**Q1.D:** Polarity in tweets

1. Polarity ranges from -1 to 1, it usually represents the degree of emotion. We can see the result below which represents the polarity of each tweet

Table

Description automatically generated

**Q2:**  Process one collection of data and separate it into different categories and provided summary statistics on those categories.

1. Through this analysis, we can bin how many tweets Biden posted by hour and by day.
   1. By hour

Table

Description automatically generated

* 1. By day

A close up of a clock

Description automatically generated

Q3: Process two or more collections of data and compare some summary data about the two collections together

1. Two users I selected are Biden and Trump. I fetch 500 @JoeBiden tweets and 500 @realDonaldTrump tweets and compare the number of their followers, friends, public lists, favorites, tweets been retweeted and favorited.

Table

Description automatically generated

1. **Program**

In read data step, I used Cursor function from tweepy package. In data processing and cleaning step, I chose all the fields I need from each tweet and restore information in a list using loop. Then I convert those into data frame using Pandas package.

For the text analysis step, I use alpha\_filter function to take a word and return true if it consists only of non-alphabetic characters and filter those characters. I also use nltk package to drop stopwords in English.

For the sentiment analysis, I installed the textblob package to define polarity in each tweet. In this step, I used remove\_url function to drop urls in tweet.

1. **Output files**
2. Tweet\_text contains 500 records and 4 columns: created\_at, user\_name, location,text. Each row represents the tweet information of when did it post, who post it, where did it post, and the tweet content
3. hashtag\_wordcount contain the most common hashtags of those tweets and how many those hashtags appear.
4. wordcount\_df contain two columns word and score, representing the major key information of each tweets.
5. sentiment\_df contains two cloumns: polarity and tweet, representing the emotion of each tweet. Positive, neural, or negative.
6. biden\_groupby\_hour contains two columns, summary how many tweets Biden post in each hour.
7. biden\_groupby\_day suggests how many tweets Biden post in each day of a month
8. Biden\_trump\_summary contain 2 columns and 6 rows, summary data about Biden and Trump followers, friends, public lists, favorites, times of their tweets been retweets and favorited