**Student Name**

* **Pranav Vinod**

### *Notes:*

### *First, please re-save this document on your computer, RENAMING the file to contain your last name.*

### *Point values of each part are shown below; points allocated include the quality of your business writing (organization, clarity, grammar, etc.).*

### *Type or paste your responses into the boxes below. The boxes will expand to fit your answers.*

1. Obtain 3000 tweets using $GOOGL (GOOGL is the ticker symbol for Google). A tweet may be categorized in different groups such as original comment, retweet, feeling, sharing of article, Ad, etc. Looking at a few sample tweets, discuss strategies that you may use to categorize tweets into five to seven categories. Read 1st 30 tweets and classify them into categories that you have identified. Report results and provide comments. (20 pts).

Due to trouble getting tweets from twitter website, I used the csv files provided.

To categorize tweets in different groups we can look for certain keywords in the tweets and use that information. Such as:

1. Any tweets starting with the keyword **rt** means they are a retweet.
2. Any tweet with the keywords “I”, “think”, “observe” and such, would qualify as an original statement.
3. Any tweet with an html link could qualify as sharing of article.
4. Tweets containing the keyword “promoted” means that they are an advertisement.
5. Some tweets are quotes borrowed from others; we can identify them if the tweet is within quotation marks.

|  |  |
| --- | --- |
| Category | Number of tweets |
| Retweet | 17 |
| Original Statement | 0 |
| Sharing an article | 7 |
| Ad | 6 |
| Quotes | 0 |

We can see based on an elementary analysis; majority of tweets are retweets followed by sharing a link for an article or an ad.

There are very few original statements and quotes.

1. In light of your response to the first question, discuss some of the challenges one would encounter in creating an algorithm that would accurately classify these tweets into the proposed categories. (10 points)

There are challenges in creating an algorithm which could classify tweets in the proposed categories:

1. One challenge would be the fact that no tweet is not only one thing. A single tweet may be a quotation, followed by an original comment. A tweet might fall into multiple categories, so putting them in all the applicable categories is a challenge.
2. It is possible to miss keywords that classify tweets. For example, not all ads have the promoted tag at the bottom of the tweet. So, it would be difficult to accurately identify some categories.
3. Develop a frequency distribution of the most frequent word appearing in the tweets, *excluding* articles, conjunctions and obvious terms (e.g. “and”, “of”, “the”, “Google” etc.). Include your code AND the resulting bar plot (10 points):

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application, chat or text message

Description automatically generated

Chart, bar chart, histogram

Description automatically generated

1. Using your list of important words, create a Wordcloud. This wordcloud should be in gray-scale (10 points):

Text

Description automatically generated

1. Now experiment with alternative wordclouds, changing color, order, fonts, etc. to create a visualization that you think is more compelling or informative than the initial one. (10 points):

A picture containing circle

Description automatically generated

1. Carry out sentiment analysis of the tweets and summarize key insights. (10 points):

Chart, bar chart

Description automatically generated

The key insight about this analysis is that at the time of data collection there was positive anticipation about the company GOOGLE. The tweets score high on positive sentiment as well as anticipation.

Even though it has a high score on trust and positive sentiment, there is a significant negative feeling regarding the company as well.

1. Obtain 3000 tweets each using $AAPL and $AMZN, and carry out sentiment analysis of the tweets separately for the two companies. Comment on similarity or differences among the three companies. (20 points):

Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated

The three companies evoke similar sentiments among people on twitter. They score high on positive sentiment and there is high anticipation and trust for each of the three companies.

There are slight differences in the negative sentiment for each company. Also, there is a higher degree of anger for Apple and Amazon as compared to Google.

1. Provide code used for this assignment. (10 points):

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text, application, chat or text message

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text, application

Description automatically generated