Milestone 1 - Twitter Sensitivity Analysis

Objective

The objective of my project is to analyse the sentiment of various 'tweets' from Twitter, and observe trends in the tweets sentiments.

Data

My data will come from Twitter and will be obtained via the R package `rtweet` which contains functions such as `search_tweets()`, `stream_tweets()` or `search_users()` to find tweets and their attributes. Furthermore, the sentiment data will come from the `textdata` package which contains lexicons for word-to-emotion data. The main attributes from Twitter will include the tweet *text*, which is the words of the tweet itself. Additionally, information such as the tweet time stored in the *created_at* variable and the user's *location* will be valuable attributes to determine various relationships. Other useful attributes includes *reply_count*, *retweet_count* and *favourite_count* (number of replies, retweets and likes respectively) as these could give an indication of sentiment or controversy. The lexicon data provides *sentiment* and *score* attributes for each word in the data set which will be used to analyse the words within tweets.

Some issues with the data will include missing entries (i.e. N/A) for some attributes, for example not everyone enables location for their tweets. Furthermore, the tweet text data will be faily small due to limited characters per tweet and may require filtering in order to accurately analyse and predict the sentiment.

Exploratory Ideas

- As per the objective above, the main goal is to analyse the sentiment of tweets. More specifically if we are given a tweet, we can use the resulting project to specify whether the tweet is positive or negative, and even give a score as to how positive/negative the tweet is.
- Once the main objective is working, can we see where tweets are coming from in the world and then compare the sentiment of tweets in different locations?
- Do certain topics tend towards a particular sentiment? For example, what percentage of tweets on politics or maybe corona virus are positive and negative?
- Finally, it would be interesting to see whether tweets are more positive or negative on average depending on when they were posted - eg. time of day or day of the week etc.

Approach

To start with, I will gather a large data set of tweets. Through data wrangling and tidying, I can extract the useful attributes into an easy to read format. I will likely need to filter out unnecessary words in the tweets message and tidy the lexicon data set so it is easy to work with. Next, by comparing the tweets to the word sentiment repository, I can use a scoring method to give the tweet a sentiment value. The method to score a tweet's sentiment may take some thought, but could be along the lines of averageing out the score of each unique/relevant word in the tweet.

To achieve the other goals outlined above, I can group the data into locations, times and topics etc. Note that certain topics can be filtered when using the `rtweet` functions.

Challenges

- Tweets are relatively small in terms of text size due to a limited number of characters per tweet. This could make it harder to get an accurate representation of the sentiment of the tweet.
- Furthermore, tweets often use images, videos or emojis, which could aid in the sentiment analysis, but will require more difficult methods to extract the emotion of the tweet.