Project Proposal

Students: ***Triguna Bangalore Narasaraj*** ***StudentId: r0648544***

Swarnalata Patra ***StudentId: r0729319***

Anthoula Mountzouri ***StudentId: r0736452***

The general idea:

We are going to use some Twitter Data, in order to analyse and visualise them in a graph representation format.

Data:

We will get the tweet data using Twitter API for hashtags such as #Brexit, #Ukip, etc. Retweets are filtered. Only tweets tweeted in English language will be considered. Due to the restrictions from Twitter API, only the tweets that are tweeted in the past 7 days are available.

We are attaching some initial data. There are two tables. The first table contains tweet data for a particular hashtag. The second table contains the actual social network information i.e., followers, following. Here, only the users who follow each other are considered.

The features for table 1 are:

- **created_at**: (Timestamp) that indicates date & time of creation of the tweet.
  Eg: 2019-03-29 23:59:58
- **screen_name**: (Text) is the twitter account id name of a user.
  Eg: screen_name of Donald Trump is 'realDonaldTrump'.
- **location**: (Text) is geographic location of users. This is set by users themselves.
- **followers_count**: (Int) indicates number of twitter users who are following the given user.
- **friends_count**: (Int) indicates how many twitter users the given user follow.
- **retweet_count**: (Int) indicates the number of times the given tweet is retweeted by others.
- **text**: (Text) contains the actual tweet.
  Eg: '*#May has failed in delivering #Brexit #EU*'
- **tags**: (List of Strings) indicates other hashtags present in tweet (text).
  Eg: For the tweet, '*#May has failed in delivering #Brexit #EU*', the 'tags' attribute contains the value ['#May','#Brexit','#EU']
- **mentions**: (List of Strings) indicates other twitter users who are mentioned in tweet (text).
  Eg: For the tweet, '@theresa_may has failed in delivering #Brexit. @Nigel_Farage is to blame', the 'mentions' attribute contains the value ['@theresa_may','@Nigel_Farage']

The data from Table 1 can be filtered. Only the twitter users whose tweets are retweeted more than N times (Eg, 100) in the given time period can be considered.

Nodes are the unique '**screen_name**'.

Table 2 is used to find the edges between nodes. The features for table 2 are:

- **source_screen_name**: (Text) is the twitter account id name of a user 1.
- **destination_screen_name**: (Text)) is the twitter account id name of a user 2.
- **has_mutual_following**: (Bool) indicates if both the source and destination users follow each other.
- **source_follow_dest**: (Bool) indicates if the source user follow destination user.
- **dest_follow_source**: (Bool) indicates if the destination user follow source user.

We will compute edges between nodes (source_screen_name & destination_screen_names) using either **has_mutual_following, source_follow_dest,** or **dest_follow_source. (Probably we will keep only the case of has_mutual_following)**

Tasks:

- ***Task 1:***

  Get the required data from Twitter using Twitter API according to specific hashtag (Eg: #Brexit). We are planning to get the data for this specific hashtag for the whole month of April, meaning we need to mine the twitter data every 7 days. The data will be in the format described in the above section.

- ***Task 2:***

  Perform the sentiment classification on the tweets. Aggregate the sentiment of nodes for each week. Only consider the subset of nodes who have tweeted about the specific hashtag for each of the four weeks of April. This task will be implanted using Python.

- ***Task 3:***

  Build a social network graph of this nodes subset for each of the 4 weeks of April (meaning 4 visualized graphs). The edges in the graph between 2 nodes indicate that both the users are following each other. The graph will be interactive. Some clustering will be represented on it, according to the findings of the previous tasks, entities such as size and neighborhood will also be able to be recognised easily and many other characteristics similar to the functionalities of Neo4j. The size of the node can indicate the number of times the given user has tweeted using the given hashtag (#Brexit). This task will be implemented using the scripting language JavaScript and the library D3.js.

- ***Task 4:***

  Analyse the measurements of this social network like centrality, degree etc, as well as relations such as friendship and so on. Analyse, also, how the sentiment has changed for these node subset for each of the four weeks.