WEB SCIENCE
COURSE WORK 2

Matriculation Number - 2576183S

1. Large number of serious mental health issues originated during COVID. Reddit channels (known as subreddits) discuss mental health issues. You have been recruited by the NHS to collect and analyse such data and identify mental health issues. Design an experimental study where you can identify mental health issues in the population. [Hint: identify research hypotheses, data collection method; analytics techniques- what kind of graphs you create and what kind of measures you may use.]

Research Hypotheses

The research hypothesis is to identify mental health issues originated during COVID in the population by analysing the reddit data.

Data collection method

Many people who suffer from mental illness have found sense of community and support on reddit platforms as these are safe space to express and discuss their issues online. As we know, Reddit is composed of subreddits. subreddits are forums dedicated to specific topic.

We can collect public reddit data using API's, we can perform filtering techniques to select subreddits that are covid and mental health illness related, users in mental health related subreddits and use keywords to search for mental health related words. The reddit posts should also be dated after February 2020, because that's where the actual problem started.

Using NLP, we can analyse linguistic features like (word counts and structural features) and sentiment analysis of the posts and compare between various disorders and group posts together that belong to same mental disorder.

Also, most of the affected people are connected to each other, since they belong to same group. As we know majority of data is generated by only fewer percentage of people also known as super-users. Since interactions is held by superusers, identifying their posts and replies on their posts is important. Most of the users don't generate new posts but are likely to reply to the posts of superusers, so their network is important to reach to other users. They are also responsible for creating threads of posts.

We also need to take gender biasing and demographic biasing in account.

Data Analytics Techniques

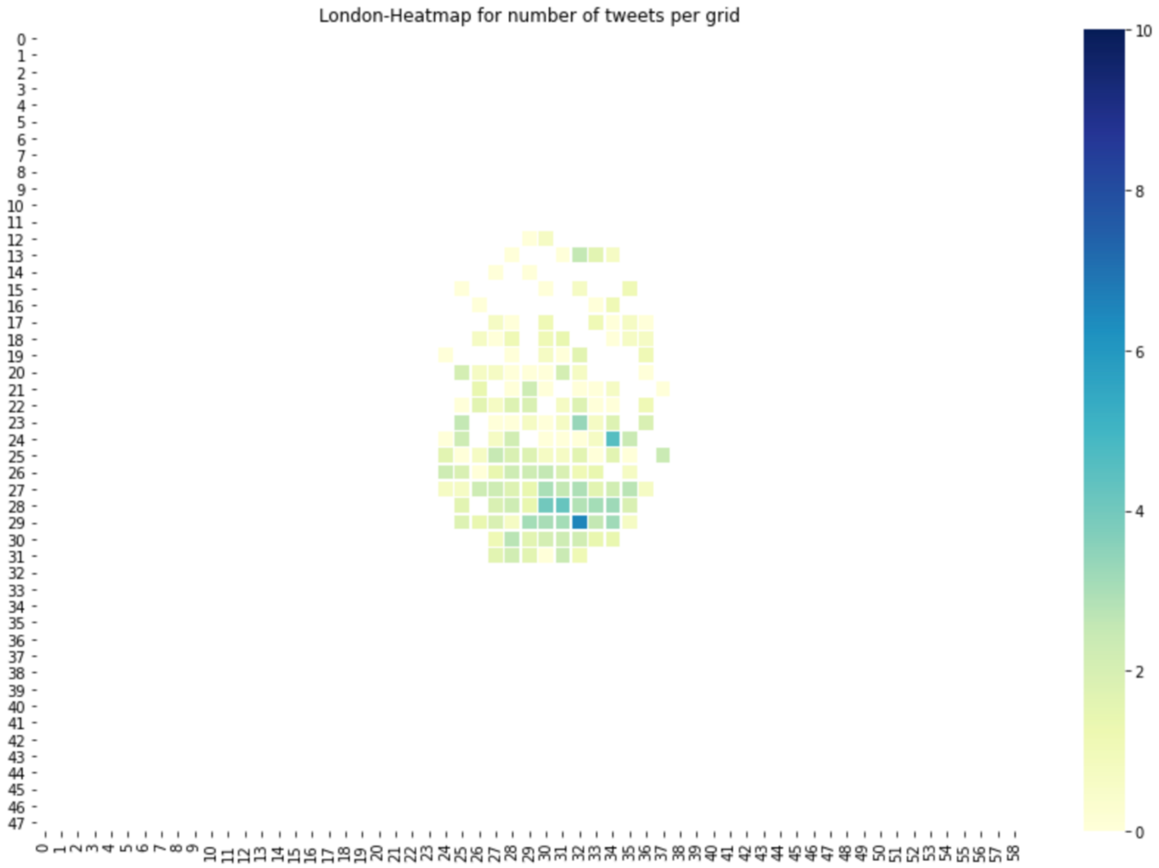
Looking at number of users, number of posts and connections per user and posting frequency related to one mental disorder can help analyse statistics related to one disorder. This is social network analysis. Since superusers are very less in number, so analysing the data based on mean and standard deviations won't be sufficient, we need to consider median values too. We can also use time patterns in posting activity and also how superusers are connected to each other. Measures used can be z-score for user-expertise (if he initiates a posts or replies). Various graphs can be plotted using the above information, we can identify no. of users affected and grouped by mental health disorders, if a condition is more specific to a gender and a demography, patterns in posting can tell if the number of patients are increasing or decreasing, also analyse if mental health problems are pre covid or post covid affects.

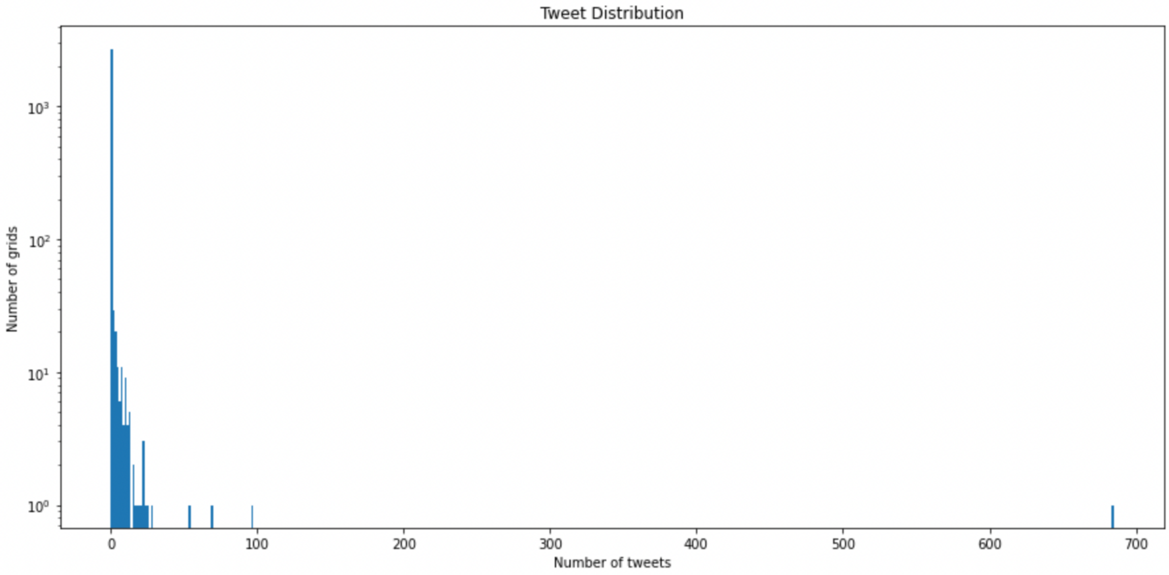
2. Assume that you are hired by the Transport for London (TfL). TfL manages the London's buses, London Underground, Docklands Light Railway, London Overground and London Trams. Your first task is to create a transport alert system exploiting Twitter data. It is assumed that there are enough social media data discussing transport related issues. Given this context, answer the following questions.

(i) A data set will be given to you (In the Data folder teams). Develop software to organise tweets into grids of 1km x 1km. Draw charts or figures to analyse the distribution of data. The coordinate system we used is

London = [-0.563, 51.261318, 0.28036, 51.686031]

<u>Analyse Distribution of Data</u>





(ii) If you were to use a grid-based system, identify the potential biases you may encounter and how are you going to address them?

In real the shape of the area of a city or a state is not a perfect square, its either square or rectangle, thus if we use grid-based system the shape of area and grid will be different. This difference in shape affects the accuracy of the result.

Instead of using Grid based system, using choropleth maps will be more beneficial, because choropleth maps are great way to show spatial patterns.

(iii) Majority of the data collected from an area/region is non-geo tagged. How do you test the effectiveness of the model for non-geo tagged data set?

As majority of tweets are non-geo tagged, and only 1-6% of the total tweets are geotagged. So, we can make some estimations based on this information. For testing effectiveness of model for non-geo tagged dataset, we can assign geolocation to a given tweet. We estimate the location of the tweets, so that we have adequate data to train, predict and test for both geo-tagged and non-geo tagged posts.

To assign geolocation to a given tweet-

Collect and store all the tweets to be processed. Before the actual location estimation all the tweets should be pre-processed (removal of all unwanted noise in the tweet). This includes URL, stop words, special characters and hashtag removal, repeated characters or words removal and tweet normalization. After the tweets are pre-processed, the tweets are separated based on, if it is geo-tagged or not. All non-geo-tagged tweets are stored in a separate database for further processing, as these are the tweets on which location estimation is performed. Considering tweet's three attributes namely, 'tweet content 'because location can be sometime mentioned in the content, 'user's activity' and 'user's network'; the locations are estimated. These estimated locations are assigned to each tweet and now all tweets has its own geolocation, therefore these tweets can be further used for Information Retrieval tasks if required.