Web Science

CourseWork2: Health Analytics & Geo-Localisation

StudentID: 2588922B
Student mail: 2588922b@student.gla.ac.uk

Question 1:

Research hypotheses: Analyse subreddit data to identify underlying mental health issues population is facing

Reddit is a "network of communities based on people's interest"

- Users can browse a variety of topics
- Submit their opinion or expertise to any topic they are interested in
    - Keeping in mind that other users might vote for or against it
- Links that gather the highest scores of user's votes will be presented in the home page
    - Users can comment on those links
    - Reply to other users' comments

Data will be collected by Reddit

- Subreddits of our interest are the ones regarding mental health problems such as PTSD, Suicide watch and more
- Only the text of posts that are shared publicly will be collected
    - People tend to share their thoughts and troubles; thus, the analysis of the text can provide us useful insight about them and allow us to propose solutions
- We need to pseudo-anonymize user identifiers (IDs)
- No demographic features of users will be collected
- Each post collected will be accompanied by metadata such as
    - Date of posting,
    - Number of (pseudo-anonymized) users replied
    - Number of responses author of post
    - Number of up-voted and down-voted replies
- Data set will also include
    - User roles (e.g. admin, moderator, simple user)
    - Dates that user joined and left the thread

Node in our case is a post, the size of the node is determined by the number of users replied (the more users the larger the node).
A graph is a set of nodes connected one to another with a line. Graphs can be directed, meaning there is a specific path to follow for each edge. To present our data we will use a User Interaction Graph, which will show each node and its connections.

- Node in our case is a post
    - Size of Node: the more replies a post gathers, the larger the node
    - Proxy for the interest of users, the more replies the higher the number of users interested in the particular post
        - Need to take care of spammers/marketers

        - Use the number of replies up-voted and down-voted, low quality posts will have more downvoted than upvoted replies
        - Use the number of users replied, low quality posts will have few people interested in, thus not many replies
        - Use the number of responses of post author, if the author is seeking for help, will respond to as many replies as possible to externalise the problem but also gather information
        - Use the dates user joined and left the thread and post, low quality users will probably join a post/thread just for a short period of time
- Edge in our case is a connection
    - Connection between two users is established when a user replies to the other
    - Directed Edge: User 1(V1) replies to User 2(V2)-> the edge has direction from V1 to V2

From the above graph,

- we can draw conclusions about the macro and micro dynamics of the thread community
- measure user engagement

Analysis:

Posting activity: We need to measure the posts per user over a constant period of time (e.g. a week) but also, we need to measure the overall number of posts in a thread. This will allow us to measure the new user engagement per thread over time. Moreover, we can track the posting frequency of users which will allow us to track the so-called super users.

Superusers are the users who are responsible for the majority of posting activity in online communities. They represent only a small percentage of the total users (maximum 5%), however, they are a key component as their activity is crucial for the functionality of thread. These users have a high number of connections; thus, they are essential for information dispersion in the whole community. Studies have shown that if we remove the top %5 of most connected users, the overall connectivity of the remaining users might fall up to 50%.

It is often the case that superusers tend to interact more with other superusers, forming stronger connections, which makes them a “closed functional hub”. The rich-club coefficient is a common score that is used to measure the degree of connectivity among superusers. The rich-club coefficient is calculated from the following formula

$$\varphi(k) = \frac{2 * E_{>k}}{N_{>k} * (N_{>k} - 1)}$$

High value of rich-club coefficient is often mentioned as rich-club effect, and refers to networks that pose a high degree of connectivity between important nodes, in our case superheroes. On the other hand, there is the opposite measure which is named anti-rich-club which suggests that superusers tend to avoid connecting with each other and prefer to connect with users that pose a low number of connections. This may suggest some kind of opposition between superusers.
Another issue we have to investigate into this section is the activity of superusers over time, more precisely to check if superusers activity was following a normal distribution e.g. it was mostly active in certain period of time, or it was uniformly distributed e.g. they keep on posting with the same frequency over the whole timespan.
Users expertise: we are going to use z-score to capture users "write new posts" and "write new threads" style. More precisely,

- If "write new replies" > "write new posts", then z-score is positive
- If "write new replies" < "write new posts", then z-score is negative
- If "write new replies" = "write new post", then z-score is zero

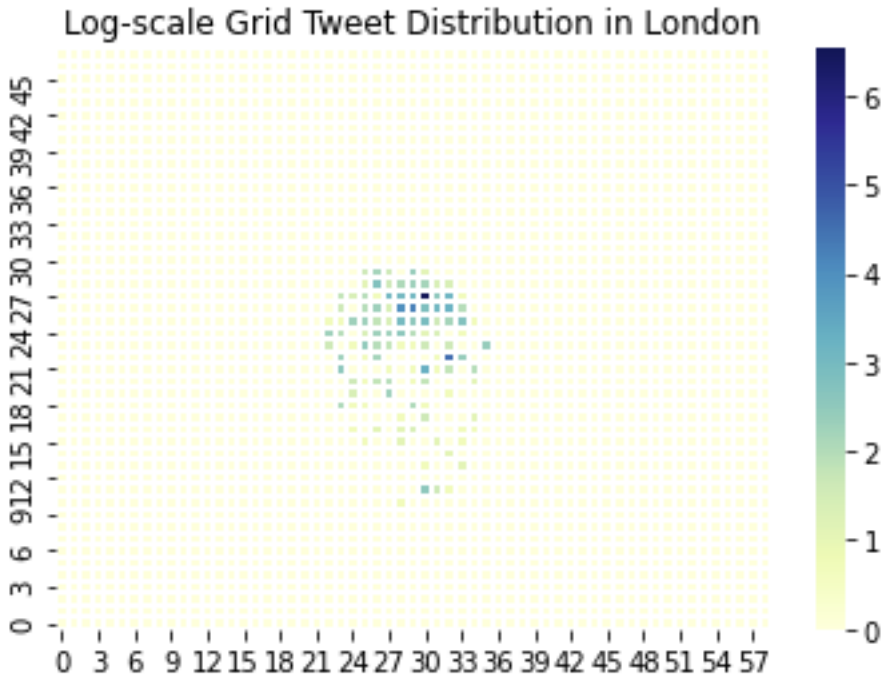
The more positive the z-score, the higher the engagement of users in the community. We need to pay more attention to superusers z-score, as it is probably going to be high, thus it will indicate that they have a broad field of knowledge and they can respond to various topics, thus it will be another reason they are crucial for our community.

Finally, we need to build a machine learning model, that will incorporate all the previous generated information, and present the most important data in a vector format. It will allow us to identify people that potentially need some kind of help, however, as there are is no collection data available, initially we will need human annotators that will observe the model's output and will assign a label, need or not need help, and as soon as we have a large enough sample we will be able to upgrade our model into supervised one.

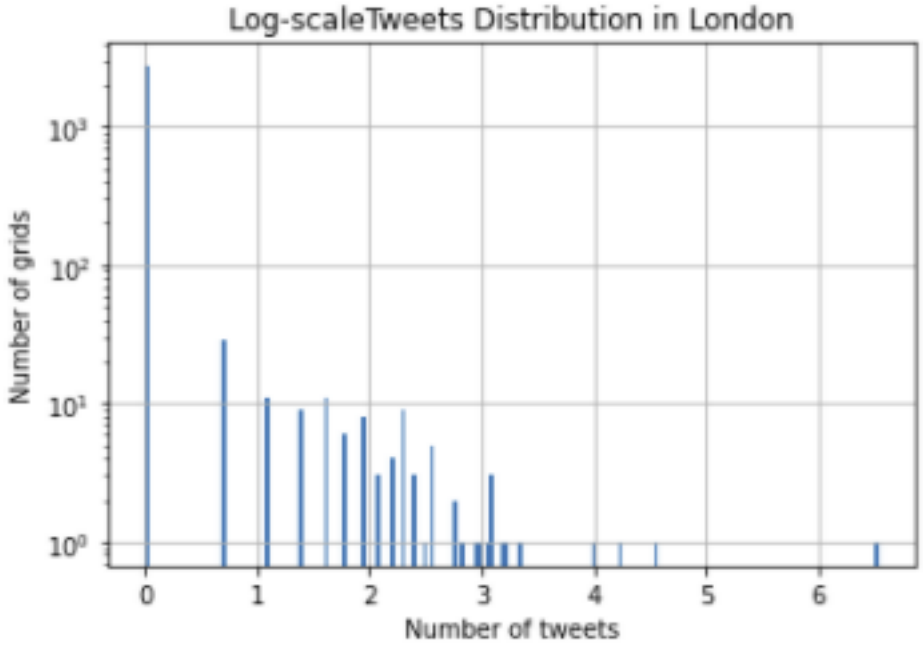
Question 2

I]

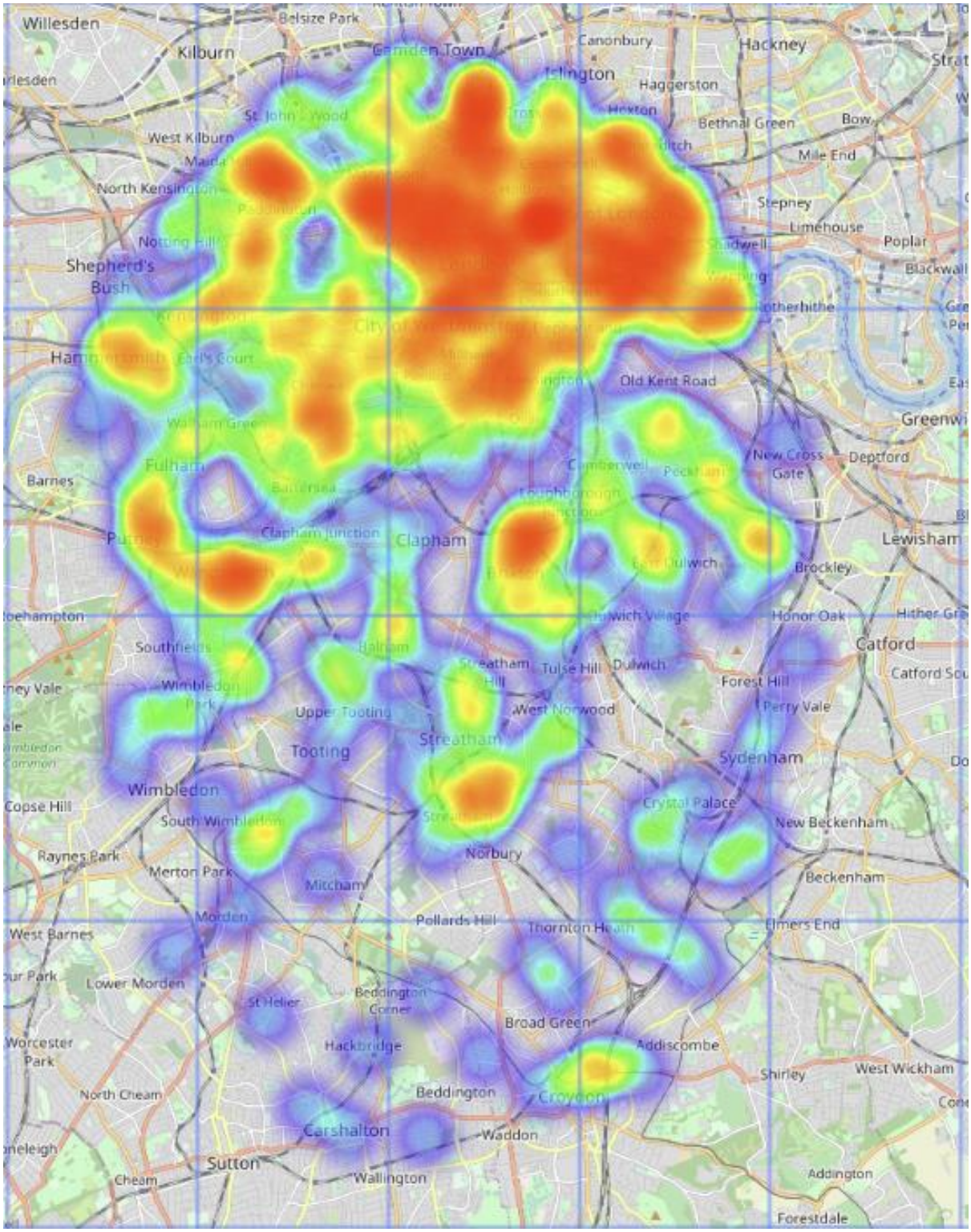
Heatmap



Histogram



Heatmap overlay of London



II]

Location bias: Urban locations are more densely populated than rural ones, thus there is a bias towards urban locations as there will inevitably exist many more people, thus the number of tweets into these areas is going to be higher. This can is verified in our case by the heatmap overlay of London shown above, where the north side of London is way more active. Thus, we

face an imbalance, so if we use accuracy as a metric it will be biased too and not reflective of the whole map, as it will be dominated by the urban areas.

- A potential solution would be to normalize by the maximum number of people per grid and also make use of precision, recall and f1-score metrics.

Activity Bias: Only a small percentage of tweets are geotagged, thus the dataset is not representative of the distribution of tweets on map, which may be proven to lead to a model that poses poor performance when deployed to evaluate its generalisability.

- A potential solution would be a to apply a mechanism to infer the location of the non-geotagged tweets
- Another approach would be to apply a weighting scheme

Population bias: Areas that are sparsely populated will not get enough local news, as they will be dominated by the densely populated ones.

- A potential solution would be to apply over/under-sampling to rural-urban areas, more precisely either over-sample rural ones or under-sample urban ones.

Statistical Bias: Modifiable Areal Unit Problem (MAUP), the way the grid will be designed is another source of bias, as data will be aggregated in different ways, depending on the way that spatial partition will take place. Thus, outcomes of data aggregation are highly dependent on the experimenter's choice of modifiable areal unit.

- To address this problem, we should implement our solution with various spatial partition techniques and compare them, if there is significant divergence, then we need to apply a different scale and reevaluate, until all methods will converge.

III]

Get the Top-N content based most similar geotagged tweets to a non-geotagged tweet. Make use of a weighted majority voting algorithm: select the location with the highest frequency in the Top-N tweets as the inferred location for the non-geotagged tweet. The inferred geo-location will be weighted accordingly based on user's credibility and similarity score.
After that, use a train-validation-test split, where train and validation consist of tweets that we have location and test set consists of tweets we have inferred locations for. Then we train a classifier on the train set, apply it on the validation set in order to predict their location by measuring the average distance from the predicted location to the actual location. Finally, if we have sufficient performance, apply the classifier on the test set and predict the locations of non-geotagged tweets. After that we can compare the predictions our classifier produced with the inferred locations from the previous geo-location inference process. This will provide us with a double validation that our predictions are as accurate as possible.