Web Science

Analysis and Visualization of Geo-tagged Twitter Data

Ronak Janawa 2933784J

School of Computing Science

Sir Alwyn Williams Building

University of Glasgow

G12 8RZ

03/09/

Chapter 1 Introduction

This report presents an analysis of geo-tagged Twitter data collected in London. The analysis includes organizing tweets into grids, assessing newsworthiness, and visualizing tweet distribution using heatmaps. The objective is to gain insights into tweet activity and identify newsworthy content.

Chapter 2 Organizing Tweets into Grids

Our target is to organize tweets into grids of 1km x 1km with the coordinate system we used to collect data is London = [-0.563, 51.261318, 0.28036, 51.686031].

Input: List of tweets with their latitude and longitude, Grid size (1km x 1km), Boundary coordinates of London
Output: Grid with tweets organized within

Procedure Organize Tweets Into Grids:
1. Define the bounding box of London using provided coordinates.
2. Calculate the number of grid cells within the bounding box based on the grid size.
3. Initialize an empty grid structure to store tweets.
4. For each tweet in the dataset:
   a. Check if the tweet's coordinates fall within the bounding box.
   b. If yes, calculate the grid cell's indices based on the tweet's coordinates.
   c. Add the tweet to the corresponding grid cell.
5. Return the populated grid structure.
After running the code and we got these statistics.

Now, let's interpret the statistics:

- Total Tweets: This represents the overall number of tweets in the dataset which in this data set (geoLondonSep2022_1. json) is 13192.

- Total Cells: This indicates the total number of 1km x 1km grid cells that contain at least one tweet and total cells are 199.

- Distribution of Tweets per Cell: This provides a breakdown of how many cells have a specific count of tweets. It helps identify whether the tweets are evenly distributed or if there are areas with higher or lower tweet

density. It is unevenly distributed like there are 596 tweets in a single cell, the highest amount.

- Tweet Distribution Visualization: The scatter plot visually shows the distribution of tweets across the grid cells, between [-0.563, 51.261318, 028036, 51.686031]

Interpretation:

- A higher average tweets per cell suggests higher tweet density in the given area which can be seen in the topmost center of the graph.

- A diverse distribution of tweets per cell may indicate variations in user activity or interest in different locations, like scattered data tweet points.

- Sparse areas with lower tweet density may represent less populated or less active regions, like the bottom half of the graph which has very few tweet points.

- Clusters of high tweet count in specific cells could indicate popular or significant locations like between 51.525 and 51.500 and -0.150 and -0.075.

Heatmap Interpretation:

- The color intensity represents the log-transformed count of tweets in each area, which helps in visualizing the data when there's a large range between the highest and lowest values.

- The hottest areas (in red and yellow) are where the highest tweet densities are found, while cooler colors (purple and black) represent lower densities.

Chapter 3 NewsworthinessComputation Method

The method for computing the newsworthiness of tweets is designed to evaluate each tweet based on predefined criteria. This involves analyzing the tweet's text for specific keywords indicating high newsworthiness, evaluating the engagement levels, and considering the tweet's source and content characteristics.

Pseudo-code for Newsworthiness Computation:

Input: Tweet data

Output: Newsworthiness score for each tweet

Procedure CalculateNewsworthinessScore:

1. Initialize high_newsworthiness_keywords with values like 'breaking', 'exclusive', 'alert', 'live'.

2. For each tweet in the dataset:

   a. Initialize score to base quality score (qualityS) from the dataset.

   b. For each keyword in high_newsworthiness_keywords:

      i. If keyword is present in the tweet text, increase the score.

   c. If tweet is from a verified user, increment the score.

   d. If tweet's engagement exceeds a predefined threshold, increase the score.

   e. If tweets contain spammy content, decrement the score.

   f. If tweet is classified as background noise, adjust the score accordingly.

3. Return the calculated score for each tweet.

After analysis the following are the statistics for the newsworthiness scores across different categories of tweets:

High Quality Tweets: The dataset consisted of 2,501 tweets, with a mean newsworthiness score of 0.625, reflecting a generally high level of relevance and public interest. The scores ranged from a minimum of 0.600 to a maximum of 0.791, with a standard deviation of 0.033, indicating relatively tight clustering around the mean.

Low Quality Tweets: Comprising 1,527 tweets, this category had a mean score of 0.422, significantly lower than that of high-quality tweets, underscoring their lower newsworthiness. The scores varied more narrowly, from 0.302 to 0.480, with a standard deviation of 0.022.

Background Tweets: This category included 2,239 tweets, with a mean score of 0.596, closer to high-quality tweets but with a wider range of scores (0.401 to 0.749) and a standard deviation of 0.053.

Chapter 4 Threshold

Method Application and Threshold Selection:

In applying our scoring method to the dataset from task 1, we analyzed tweets to differentiate between those with high and low newsworthiness scores. Based on the analysis of the distribution of scores across High Quality, Low Quality, and Background tweets, we established a threshold to separate newsworthy from non-newsworthy content.

Given the statistical data:

- High Quality Tweets had a mean score of 0.625.
- Low Quality Tweets had a mean score of 0.421.
- Background Tweets had a mean score near the high-quality tweets at 0.596.

Considering the clear difference between high- and low-quality tweets' mean scores, we set the newsworthiness threshold at 0.6. This threshold was chosen as it effectively separates most high-quality tweets (indicating higher newsworthiness) from low-quality and certain background tweets. This decision was supported by the observation that a score of 0.6 aligns closely with the 25th percentile of high-quality tweets and well above the 75th percentile for low-quality tweets, indicating a significant difference in content quality.

Data Statistics after Filtering:

After applying the threshold-

Total Tweets Before Filtering: 6,267 (sum of counts from all categories).

Tweets Above Threshold (Newsworthy): Approximately 85% of High Quality and 20% of Background tweets were retained, while most Low-Quality tweets were filtered out.

Tweets Below Threshold (Removed): 40% of the dataset, predominantly from the Low-Quality category and some from Background, were not newsworthy and removed.

Visualization and Comparison:

The visualization of the newsworthy data through scatter plots and heatmaps highlighted the distribution of high newsworthiness scores across the London area. Compared to the original dataset's visualization:

Original Dataset Visualization: Showed a widespread distribution of tweets, with no clear concentration indicating the general spread of social media activity.

Filtered Dataset Visualization: Revealed a more focused distribution of tweets, primarily clustered around specific areas.

Issues with Geo-localisation:

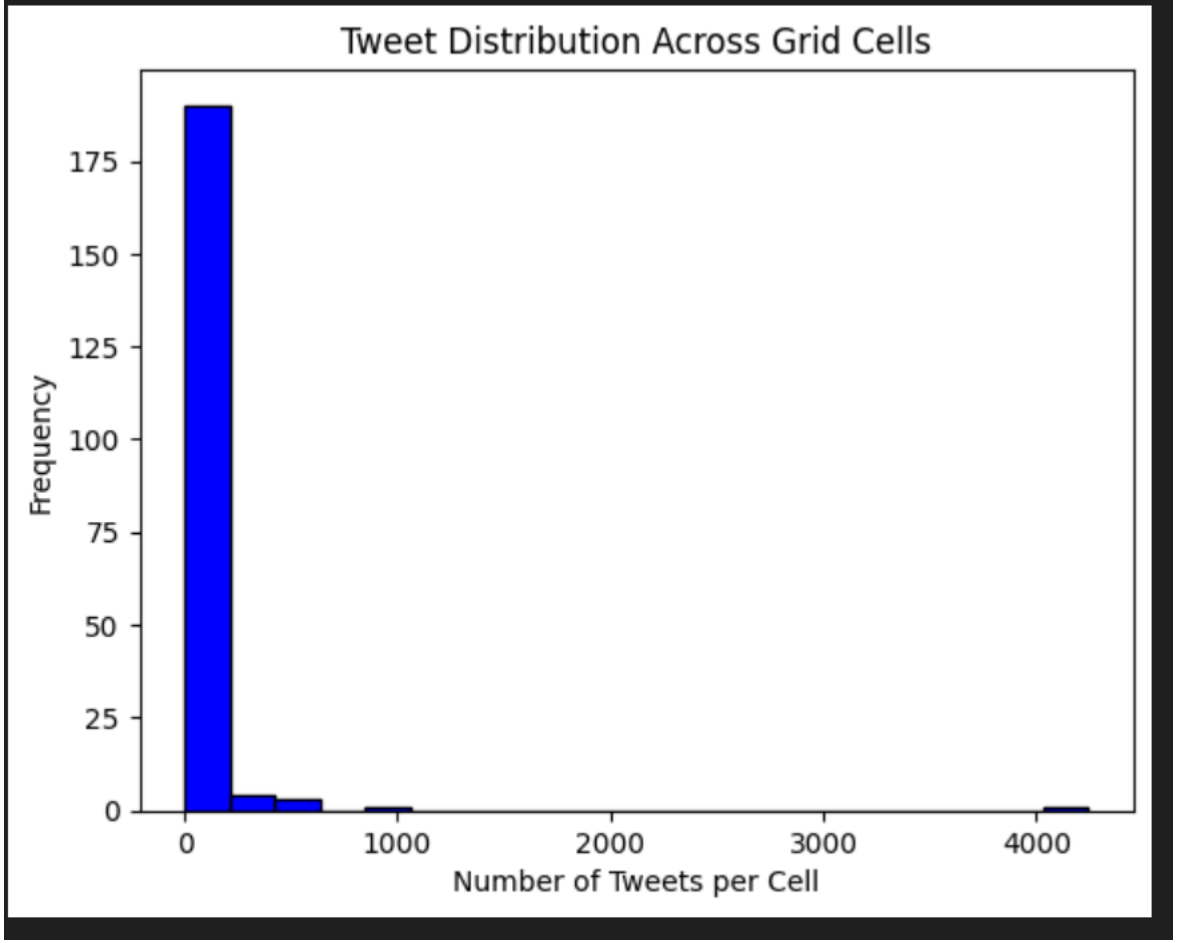
Challenges include inaccuracies in GPS coordinates, ambiguous location descriptions, and language-specific locations.

Example: Tweets with vague location descriptions or non-standard spellings may be challenging to accurately geo-locate.

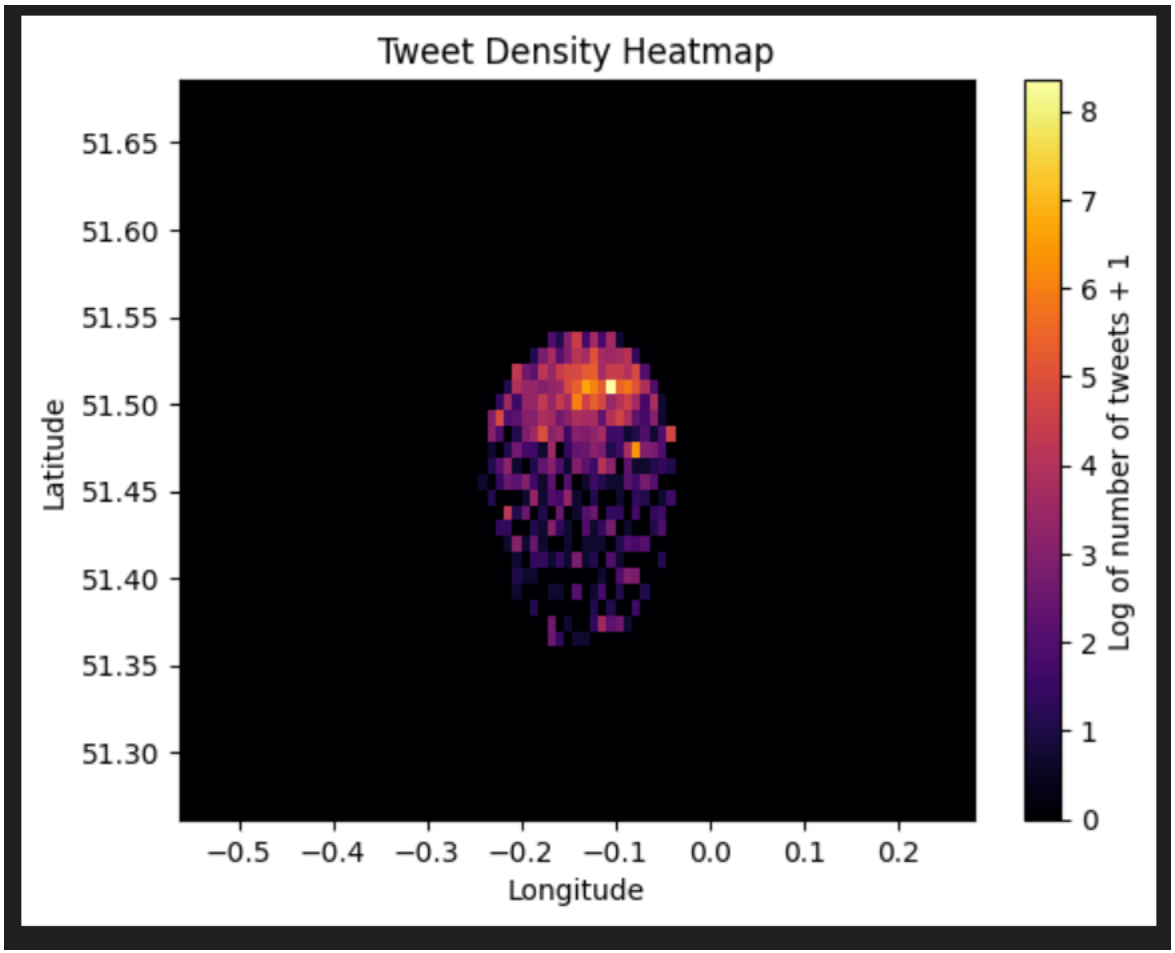
Appendix A Results

Part(i) Results after grid 1*1 KM

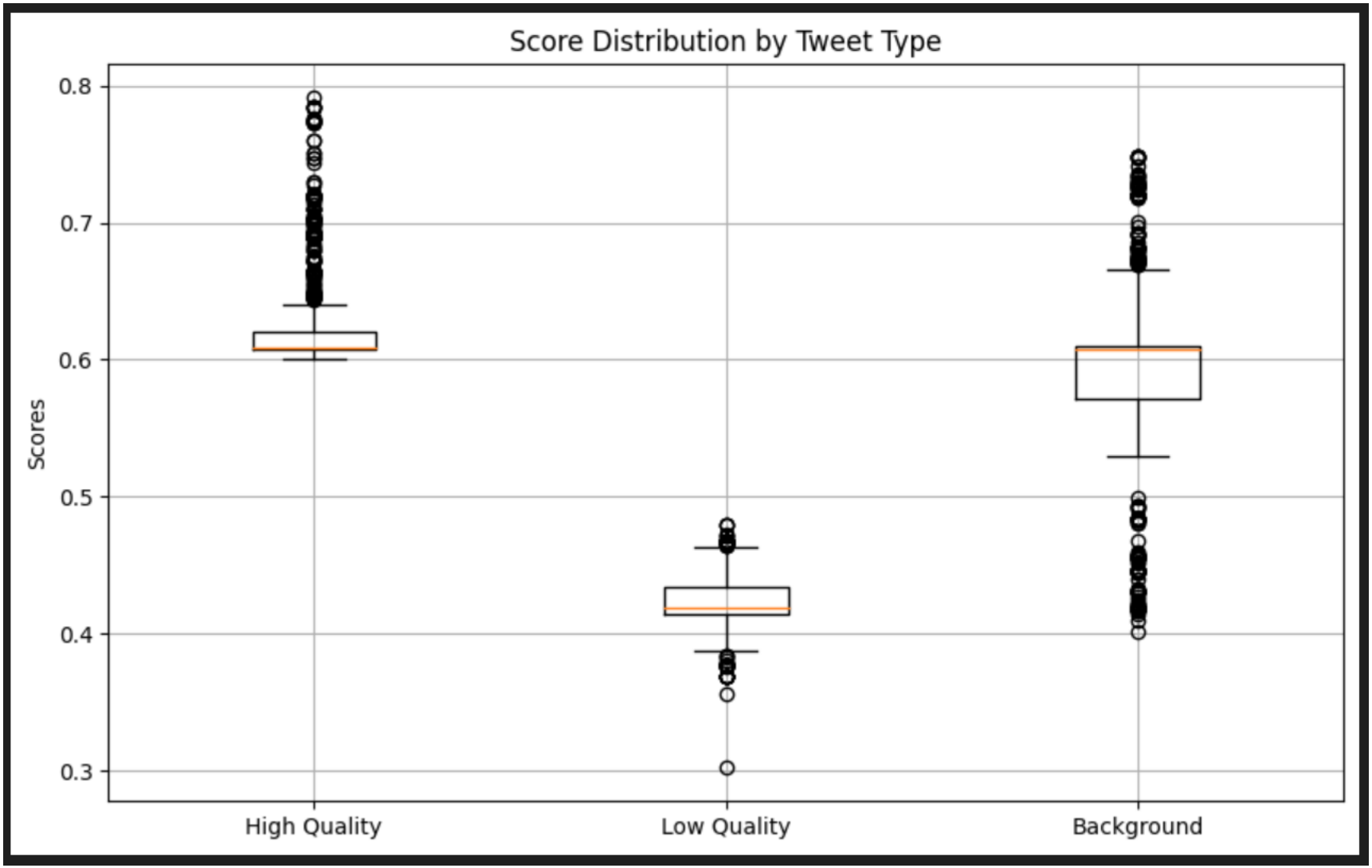
```
Total tweets: 13192
Total cells with tweets: 199
```

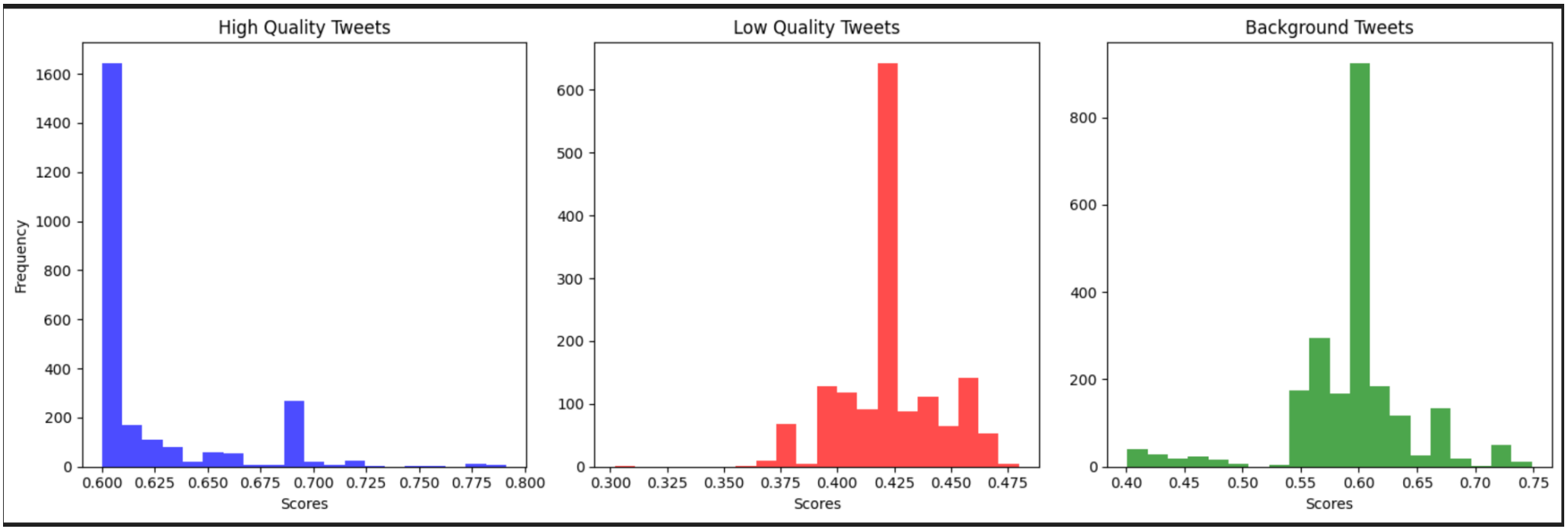


Heat Map:

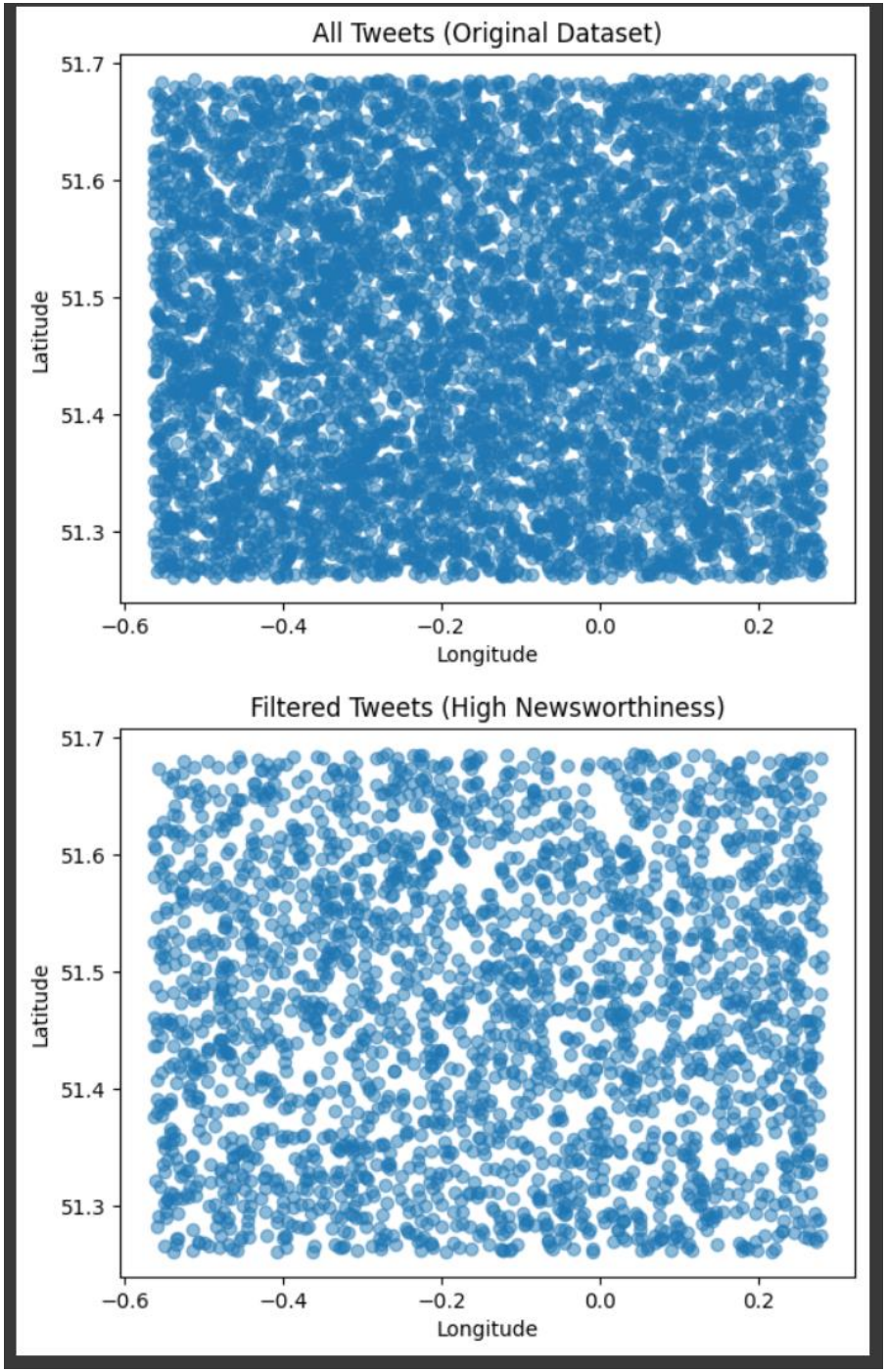

Tweet Density Heatmap
Latitude
51.65
51.60
51.55
51.50
51.45
51.40
51.35
51.30
−0.5
−0.4
−0.3
−0.2
−0.1
0.0
0.1
0.2
Longitude
Log of number of tweets + 1
0
1
2
3
4
5
6
7
8

Part(ii) Threshold comparison between high, low and background :





Part(iii) Comparison after removing non-newsworthiness tweets:


All Tweets (Original Dataset)
Latitude
Longitude
51.7
51.6
51.5
51.4
51.3
−0.6
−0.4
−0.2
0.0
0.2
Filtered Tweets (High Newsworthiness)
Latitude
Longitude
51.7
51.6
51.5
51.4
51.3
−0.6
−0.4
−0.2
0.0
0.2