**Name: Zeyan Liang**

**Student Number: 2627888L**

**1. A data set will be given to you (In the Data folder teams). Develop software to organize 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] [40]**

Step 1: define a function to calculate the actual distance using latitude and longitude coordinates in KM. Use this defined function to count number of grids in London by given coordinate.

begin

point1 = input ask the first point you want to calculate

point2 = input ask the second point you want to calculate

R = 6373.0

lat1 = lat of point 1

long1 = long of point 1

lat2 = lat of point 2

long2 = long of point 2

phi1 = lat1\*(math.pi/180)

phi2 = lat2\*(math.pi/180)

delta1 = (lat2-lat1)\*(math.pi/180)

delta2 = (long2-long1)\*(math.pi/180)

a = math.sin(delta1/2)\*\*2+math.cos(phi1)\*math.cos(phi2)\*math.sin(delta2/2)\*\*2

c = 2\*math.atan2(math.sqrt(a),math.sqrt(1-a))

d = R\*c

d is distance from point 1 o point 2

end

Use this formula to calculate the distance between southwest and southeast London coordinates, and then calculate the distance between southeast London and northeast coordinates. The former is the horizontal distance and the latter is the vertical distance.

58.157106147216226

47.24075616044219

According to the distance we calculated, there are 58\*47 grids on London map.

Step 2: calculate distance from everyone twitter’s coordinates, calculate density/frequency of number of twitters in each grid.

Use similar method as previous to calculate the distance from southwest London coordinates to twitter’s place coordinates of each data.

Then use sns.heatmap to draw the heatmap.

A picture containing graphical user interface

Description automatically generated

It can be seen from the data that there are a total of **4212** posts, and their locations are distributed in more central areas of London. they mainly come from 'City of London, London', 'London, England', 'Camberwell, London' 'Kensington, London' and 'Camden Town, London' …... I think the main reason why most social media data on transport, community, activity and city related issues are related to these specific areas is that these areas are located in central London where both transport and the economy are important, so these areas get more attention.

Text

Description automatically generated

Step 3: Sort the data from step 2 to draw distribution bar chart.

A picture containing background pattern

Description automatically generated

This data distribution graph is very long, I only intercepted the X-axis before 700. It can be seen that there are very few grids with 100 or even more than 50 posts. Most grids have no data, and the data distribution is very concentrated.

This data is also reasonable and easy to understand, since the transportation system in the suburbs of the city is usually underdeveloped and the population density is much lower than that of the city center. This has resulted in few posts discussing these remote areas.

**2. You will be given a set of high-quality tweets and low-quality tweets. Develop newsworthy scoring method based on this dataset.[30]**

Step 1: posts are processed

Mainly removing stop words, spaces and punctuation. Labeled to distinguish between high-quality posts and low-quality posts

Create a pipeline with space then use NLP functions to remove all stop words punctuation and spaces

Define another a function to calculate the frequency of term t in respective model.

The main method judges whether the sub-term exists in a specific model, if it does not exist, it is assigned to 0, and if it exists, the frequency of the corresponding term is taken from the established data dictionary.

Define a function called calculate\_cntd to calculate CNTD. This method has two parameters, a list of all terms in a post and a threshold. The function of the threshold is to adjust the parameters to achieve the maximum accuracy of the model.

This function is more complicated. First, use the frequency of the term to be calculated in the current model to get FT\_HQ and FT\_LQ, and then use len() to calculate the raw frequency of all terms in the model. Then use (FT\_HQ/F\_HQ)/(FT\_LQ/F\_LQ) to calculate likelihood ratio for each term. Then use the threshold to determine whether the CNTD is available, if not, set it to 0.

The last step is to calculate Newsworthiness

The calculation method is to calculate the sum of the CNTD scores obtained in the previous step, and add 1 to the base value to prevent the log2 from being unable to be calculated. The SHQ sum of the high-quality model is divided by 1 and the SHQ sum of the low-quality model is added by 1.

The threshold is used as a parameter to adjust the parameters to obtain models with different accuracy. The parameter adjustment method I use is to input the threshold of different values. Using the current model to predict the label of the input data, for example, the input data is from a high-quality model, but the final predicted label indicates that the data is a worthless post, which is a misjudgment.

I define the initial threshold to be 1.0 and increase the original data by 0.1 each time, input all the data of high-quality posts and low-quality posts for prediction, and then compare the predicted results with the actual results, and take the two data sets of posts. The sum of the number of errors to determine the accuracy of the model.

The results show that the minimum number of errors has been reached when the threshold is 1.8. Because I use 1.8 as the best threshold to do predictions on other datasets.

Text

Description automatically generated

I also tried not text processing the content of the post, keeping stop words, punctuation and spaces. I found the results to be unsatisfactory. I think the main reason for this is that commas, single quotes and prepositions appear very frequently. If they are put into the model as terms, these have no specific meaning and are not important. The data of the information will affect the accuracy of the model to a certain extent.

After processing the posts, the number of terms in high quality posts, low quality posts and the sum is 14396,13196,27592 respectively.

Graphical user interface, text

Description automatically generated

**3. Use the above newsworthy scoring techniques to analyse the geo-tagged data set given above and discuss the results[30]**

I printed out posts with scores greater than 0 and posts with scores less than 0 respectively. After analyzing the content of posts, I found that posts with scores greater than 0 almost did not contain any valuable information in the text. Even a large part of the data are pictures, because the data is read in csv format, so the pictures cannot be recognized. However, in the posts with scores greater than 0, these contents have a similar point. The text structure has two themes, one is the event and the other is the location. Usually, there is a specific and easy-to-understand text in the text. For example, the Canadian police banned incomplete Vaccinated truck drivers across provinces. Or Liverpool News, which broadcasts real-time weather conditions in Liverpool, and various events in the city. Therefore, it is a good choice to set the newsworthiness threshold to 0 to judge whether a news is valuable or not.

The above methods are all implemented in the form of def, so in the third question, I use the geo data set analyzed in the first question as input to predict the label of each post. There are two types of labels, and those with scores greater than 0 are valuable posts. The text continues to remain in the pandas dataframe, and the low-value posts with a score less than 0 are removed from the pandas dataframe.

The result after filtering is 889 high-value posts and 3253 low-value posts.

889 rows × 15 columns

3253 rows × 15 columns

Then I made a heat map and a distribution map of the data results.

A picture containing graphical user interface

Description automatically generated

A picture containing application

Description automatically generated

Comparing these figures with the results of the first question, the number of posts plummeted, with high-value posts accounting for only 21.1% of all posts. The distribution of these high-quality posts is uneven but regular. The main origins of high-quality posts are similar to those in the first question, and they are mainly from several important places in London (the number of posts is distributed in a specific area and the picture is below).

As can be seen from the histogram, there are still more than 50 high-quality posts in very few squares, which is similar to the result of the first question.

Graphical user interface, text

Description automatically generated

(all posts distribution by area)

Text

Description automatically generated

(high quality posts distribution by area)

In my opinion, whether it is high-quality posts or low-quality posts, the number of posts they post is closely related to the region, and popular regions such as 'City of London, London', 'London, England', 'Camberwell, London' and 'Camden Town, London' always have an amazing number of posts compared to other regions. The grid they are on will also have this larger number. But the quality of the posts and the regions or grids can't find a logical relationship to prove that there is a connection between them. All grids or regions have high and low quality posts. And it can be seen from the available data that the number of regional Twitter distribution rankings are similar for both times.