Course work M – Geo Localisation
Web Science

COMPSCI4077/COMPSCI5107/COMPSCI5078

Joemon M Jose

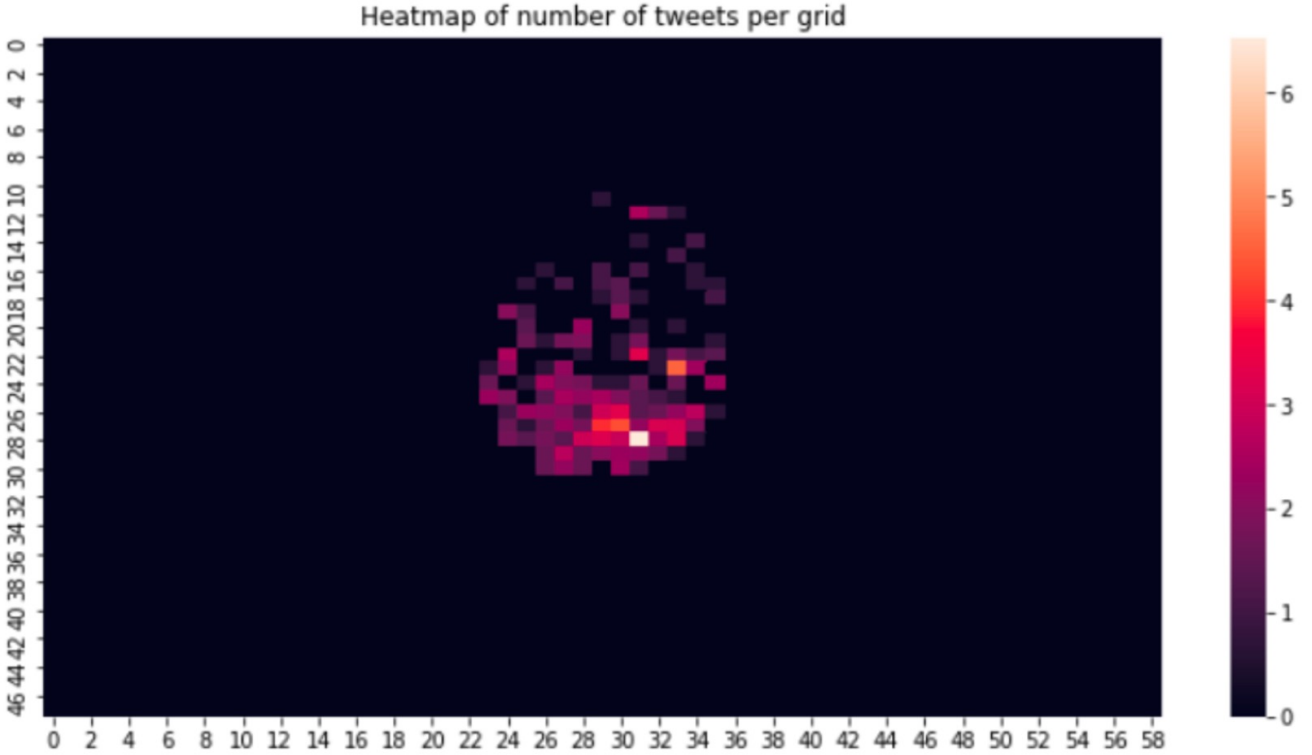
(i)

- A dataset will be given to you (In the Data folder teams). Write python code to organise tweets into grids of 1km x 1km. Draw charts and/or figures to analyse the distribution of data.
  The coordinate system we used to collect data is
  London = [-0.563, 51.261318, 0.28036, 51.686031]
- [25]

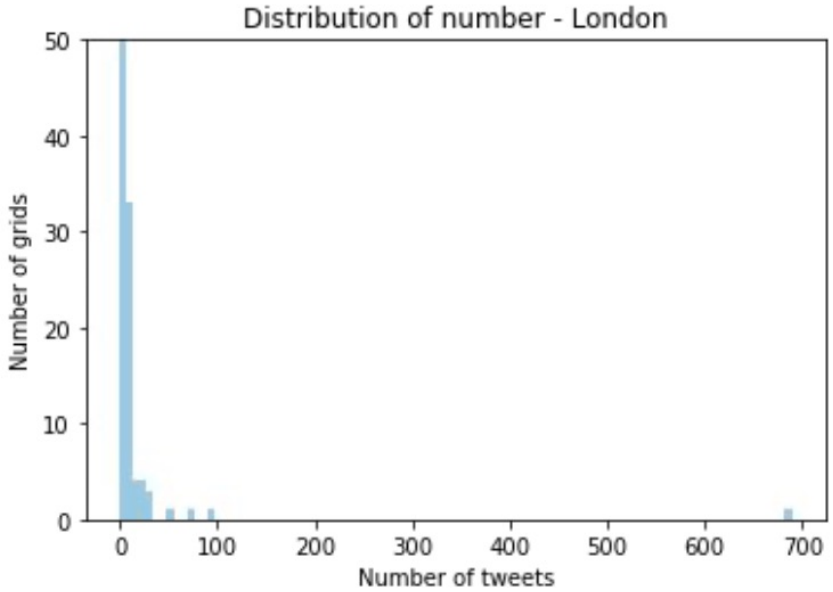
solution

- Describe the algorithm with a pseudo-code
  - (5 marks)
- Provide statistics of the data (total tweets, how many are on the cells, and how it is distributed etc.) and interpret the statistics – what does this mean?
  - (5 marks)
- Provide charts and/or figures for the visualisation of the grid data
  - (10 marks – 5 for code and description; 5 for output shown)
- Describe your views/interpretation on the data (and the resulting visualisation) - you may want to highlight issues with any potential geo-localisation techniques with this data.
  - (5 marks)

Heat map – how do you interpret?



Data distribution? How do you interpret



(ii)

2. You will be given a set of high-quality, low-quality and background tweets. Develop newsworthy scoring method based on this dataset. Empirically adjust the thresholds to modify newsworthiness and discuss the results.

- [30]

solution

- *In the report:*
- Lecture slides on newsworthiness scoring
- Straight forward implementation
  - Explain your newsworthiness computation method along with an algorithm/pseudo-code
    - (15 marks)
  - Conduct data analysis & provide an analysis of various thresholds; data analysis may include for example, using or not using stop words, adapting thresholds etc.
- Critical Discussion –
  - Scoring method quality
  - Your high, low and background quality dataset
  - (15 marks)

(iii)

(i) Use the above newsworthy scoring techniques to analyse the geo-tagged data set given (i) and discuss the results

- [25]

solution

- Take each tweet in the data set
- Score its newsworthiness
- Remove the one is not newsworthy?
- Keep the one that is newsworthy

- Apply your scoring method to data in task 1 -
  - Investigate tweets with low scores and high scores; find an appropriate threshold to separate them and remove tweets with low newsworthy scores
  - Justify the threshold used with any supportive information you can collect
- (10 marks)

- Provide statistics of the data (total tweets, how many are with certain newsworthy scores, and how it is distributed etc. *how many removed*, see below)
- (5 marks)

Apply the visualization you created on newsworthy data; Draw the figures/charts and compare them with results in (1). What can we say about the difference?

- Redraw the heatmap and histogram
  - 10 marks

[Open tasks] – 10 marks

- *Identify and discuss, with examples, issues for geo-localisation due to the nature of tweets or sources*
- *The idea here is to demonstrate your skillset; apply your knowledge to potentially a real-life task*
- *Explore newsworthiness score and variations*
- *I may have mentioned some of these things in the class at various points*
  - *May not be in the slides*

- This part is for rewarding initiatives

Report – 10 marks

a. *Structuring and formatting*
b. *Articulation of ideas*
    a. *New insights drawn*
c. *Creativity in addressing the tasks*
    a. *Going further than just copying bullet points from sliodes*