University of Glasgow | School of Computing Science

Assessed Coursework

Course Name	**Web Science (H) COMPSCI4077**			
Coursework Number	**Geo Localisation -**			
Deadline	**Time:**	**4:30PM**	**Date:**	**28th February 2025**
% Contribution to final course mark	**20%**			
Solo or Group ✓	**Solo**	**x**	**Group**	
Anticipated Hours	**20**			
Submission Instructions	**Submission through Moodle** **Report and Python codes**			
Please Note: This Coursework cannot be Re-Assessed				

Code of Assessment Rules for Coursework Submission

Deadlines for the submission of coursework which is to be formally assessed will be published in course documentation, and work which is submitted later than the deadline will be subject to penalty as set out below.

The primary grade and secondary band awarded for coursework which is submitted after the published deadline will be calculated as follows:

(i) in respect of work submitted not more than five working days after the deadline
   a. the work will be assessed in the usual way;
   b. the primary grade and secondary band so determined will then be reduced by two secondary bands for each working day (or part of a working day) the work was submitted late.

(ii) work submitted more than five working days after the deadline will be awarded Grade H.

Penalties for late submission of coursework will not be imposed if good cause is established for the late submission. You should submit documents supporting good cause via MyCampus.

Penalty for non-adherence to Submission Instructions is 2 bands

You must complete an "Own Work" form via https://studentltc.dcs.gla.ac.uk/ for all coursework - If you are not signing the declaration of originality, then your marks will not be released.

Individual Assessment: Geo-Localisation

Coursework is due on Friday, **28th February 2025**, 430 PM

CW is marked out of 100 marks & Weighted 20% to the final marks.

All submissions are through Moodle

Course work description is given at

Teams- File Area – Class Materials- coursework - Hlevel

Data is given at the data sub-directory

Around 10–20K tweets from London/UK are in the Data sub-directory. It is in json format.
Ps: Data will be uploaded by Wednesday
Ps: Please use Jupyter notebook and submit code along with cell outputs archived.
A written report should accompany the software- We are marking the report and using software to verify the facts in the report

(i) A dataset will be given to you (Teams File area). 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]

[30]

In the report:

"Write python code to organise tweets into grids of 1km x 1km." (10 marks, for the description of code and the actual code – (no need to reproduce the code in the report); correctness of the code.

Collect statistics of the data (total tweets, how many cells, how many are on the cells, and how it is distributed etc. – look at lecture slides to see examples of statistics reported) and interpret the statistics – what does this mean? (5 marks)

Provide heatmap and histograms for the visualisation of the grid data (10 marks –for code and description; and for outputs 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)

(ii) 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.

[25]

In the report:

Explain your newsworthiness computation method along with an algorithm/pseudo-code (10 marks).
Ps: Algorithm is given in the lecture slides; implement the algorithm;

Conduct data analysis & provide an analysis of various thresholds; data analysis may include for example, using or not using stop words, adapting thresholds etc. You use data in (i) for this (15 marks)

- Effect of using or removing stop words or not removing stop words (5)
- We use different thresholds in the algorithm; explore variations; demonstrate its appropriateness for the data given (10)

Ps: here you explore data and come up with a robust newsworthiness scoring scheme that work for the data given

(iii) Use the above newsworthy scoring techniques on the geo-tagged data set given (i) and discuss the potential effect on geo-localisation.

[15]

In the report:

In Task (ii) you have developed a scoring method and studied the effects of various thresholds. Apply your scoring method to data given above and remove tweets with low scores (no newsworthiness) and keep tweets with newsworthiness. Repeat task 1 on the remaining data - (5 marks)

and provide the statistics of remaining data; statistics of the data may include (total tweets, how many are with certain newsworthy scores, and how it is distributed etc. *how many removed*, see below); contrast with results in task (i) in terms of statistics and heatmap/histograms. What can we say about the difference? (10 marks)

(iv) [Open tasks]

Identify and discuss, with examples, issues for geo-localisation due to the nature of tweets or sources

[20]

It is up to the students to come up with solutions, though engaging in the class would help.

Example

- ***Variations in the newswothiness scoring***
- ***Study the bias in data and its visualisations***
- ***Study the distance Formulae variations***
- ***Propose methods to improve the effectiveness***

(v) Report – 10 marks

[10]

a. *Structuring and formatting - 3*
b. *Articulation of ideas - 3*
c. *Creativity in addressing the tasks- 4*