University of Glasgow | School of Computing Science

Assessed Coursework

Course Name	Web Science (M/MSc) COMPSCI5107/COMPSCI5078			
Coursework Number	Reddit Data Analysis			
Deadline	Time:	4:30 PM	Date:	28th February 2025
% Contribution to final course mark	20%			
Solo or Group ✓	Solo	x	Group	
Anticipated Hours	20			
Submission Instructions	Submission on Moodle; link will be provided Code (zipped) and the report need to be submitted			
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

Individual Assessment: Reddit Data Analysis

COMPSCI5107/COMPSCI5078

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

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

All submissions are through Moodle.

Links to data from ***InvestmentClub*** (https://www.reddit.com/r/InvestmentClub/) subreddit is given.
Teams File area – File – Class Materials – coursework – Mlevel
Ps: use a small data set for the development and once you are happy with your software, apply the methods on full data!

Your task is to develop Network analysis on this data set.
Ps: Work with Jupyter notebook and submit code along with 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) You will be given two files on comments and submissions. This is a subset of Reddit data in **json** format. Process the data to conduct network analysis.

[5]

In the report –
Describe how did you aggregate data from two files and organised the data for further processing.
Explain the data organisation and rationale
Summarise the data, which will be useful when you interpret the results
– 5 marks

(ii) Use the data and create graphs and create visualisation.

[20]

In the report –
Describe the way you use the data for building graphs – what are origin nodes and what are destination nodes. Justify the approach (5 marks)

Graph Visualisation (15 marks)

- Visualisation of the graphs created; Graph for the entire data and zooming in on a part (10 marks – 5 marks each)
- How do you interpret this data? Can you make any observations about the data? – 5 marks

(iii) Make an analysis of the network and understand the important properties.

[45]

Your task is to conduct network analysis on the graph you created, (or the data given). You are free to conduct any analysis you may choose as long as you justify them. The best way is to identify a research question and conduct analysis and answer a research question as we discussed in the class. **Please have a look at for an example - https://pmc.ncbi.nlm.nih.gov/articles/PMC6060304/**

I am suggesting some sample research questions below.

3.1 Study the role of super users in the community. How central are they to the cohesiveness and functioning of the community?

- 3.2 How users on support communities, as a group behave over time?
- 3.3 How exclusive super users in the community in their behaviour?

In you report identify the research question you chose, and specify your solution (what metrics/measures used to answer the research question, provide a short description, pseudo-code to highlight the logic (the actual software you created will be on the Jupyter file), and the respective interpretation (15 marks each)

(iv) *[Open creativity tasks]*

[20]

Students are encouraged to explore further
It is up to the students to come up with solutions, though engaging in the class would help.

- **for example,** here you can explore additional data analysis, (Have a look at for an example - https://pmc.ncbi.nlm.nih.gov/articles/PMC6060304/)
- exploring the datasets further and discussing methods for modelling quality of data; …
- look at content analysis
- any other ideas you think useful

- You are free to do what you think appropriate. In you report identify the research question you chose or the reasons for the analysis you coinducted, and specify your solution (what metrics/measures used to answer the research question, provide a short description, pseudo-code to highlight the logic (the actual software you created will be on the Jupyter file), and the respective interpretation (15 marks each)

(v) *Report – 10 marks*

[10]

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

Report – Follow the tasks given above and answer them one by one.

Data File is in

Teams File area – File – Class Materials – coursework – Mlevel – Data

The interesting fields (as shown in red)

- url – Unique URL for the original post
  "url": "http://www.reddit.com/r/InvestmentClub/comments/p6vn4/goog_buy_when_at_or_around_575_short_when_it_is/",
- id – unique id
  "id": "c3mvgkg",   of the user
  "parent_id": "t3_p6dut",

Link to parent post or comment

id replied to parent_id

- created_utc

  date in utc format; convert this to python Date format to process "created_utc": "1328119996",

- name (use) – t3/t1

  "name": "t1_c3mvgkg", t1- comments; t3- submissions
- title – in case of submissions

  "title": "GOOG. Buy when at or around 575. Short when it is around 615.",
- author
- "author": "hobbitskill",
- body

  "body": "I think this is a great idea!",

  For example, in comments
- num_comments

  in case of submissions "num_comments": 2,

such as upvotes, downvotes, likes, and clicks.

- downs
- ups
- likes
- clicked
- score
- "score": 2,
  - Ups - downs

For your analysis it is better to create a csv file; I would let the students to organise them