COMP3020 Group Project 2023

Click on a question number to see how your answers were marked and, where available, full solutions.

Question Number

Project Tasks
Question 1

Performance Summary

Exam Name:	COMP3020 Group Project 2023
Session ID:	013048887
Student's Name:	Long Nguyen (5736751264634f2cb31bc12a2e47e34c)
Exam Start:	Fri Sep 22 2023 22:15:40
Exam Stop:	Fri Sep 22 2023 22:19:17
Time Spent:	0:03:37

Question 1

Project Tasks

To get started, pick a topic that interests you, something you'd like to explore throughout the entire project.

Make sure each group selects a unique topic, and once you've settled on one, go ahead and add it to the `Group Project Topics` (https://vuws.westernsydney.edu.au/webapps/blogs-journals/execute/viewBlog?

course_id=_46416_1&blog_id=_426167_1&type=blogs) list.

If, as you begin working on the project, you find that there aren't many active users or discussions related to your chosen topic, feel free to change it. Just remember to update your choice on the topic list when you do.

Advice: After downloading data from the required Social Web App, it is recommended to save it to your working directory. Otherwise, each time you run your code, you will have a different dataset and the solutions you provided for the previous dataset will be incorrect for the new one.

Note that:

- The relevant terminology may differ between Reddit and Mastodon APIs, spending time reading about APIs and R packages and relevant functions, is expected.
- Labs featured Reddit and Mastodon APIs are introductory, you are required to investigate more on the relevant R functions and the data structure.
- Some research/preparation into downloaded data organisation is required, as data frames usually contain one or more columns of interest to select and filter on.

Question 1

Using **Reddit API,** identify the relevant thread URL's for your chosen topic. Focus on either weekly or monthly timeframe.

Find the top three threads with the highest number of comments.

Retrieve and display the main post from from each of these three threads. Generate a word cloud for the comments and replies within each of these threads, resulting in a word cloud for each thread.

Comment on your word clouds. Explain the key discussions and topics being addressed in each of the three thread.

5 marks

Question 2

Combine all the threads you collected in question 1 and create a column to label them with their thread number.

For example, label the first thread comments as "1", label the second thread comments as "2", label the third thread comments as "3".

Next, apply K-means clustering to cluster the combined threads.

Visualise the results of your clustering in two-dimensional vector space. Ensure that your visualisation includes both the clusters and the original labels of the documents.

Comment on your findings. Assess the performance of K-means clustering in terms of correctly identifying clusters. Did it effectively identify the clusters?

6 marks

Question 3

Use all thread URLs on Reddit that you identified in question 1.

Test if there exists a linear relationship between the number of comments in threads and their corresponding dates.

Note: You may need to convert dates to a date format.

3 marks

Question 4

Retrieve the content of the top thread you identified in question 1.

Test whether the number of comments on a thread is equally likely on each day.

2 marks

Question 5

Using **Mastodon,** identify users who are related to your chosen topic.

Identify the top five most active users, based on the highest number of statuses they've posted.

Download 50 followers and 50 friends (these are the users that your user follows) of these 5 top users.

Create a graph and visualize the relationships among these users.

Note 1: You can select any five users from the top 10 users for which you can successfully obtain friend and follower data. Keep in mind that you might encounter limitations in downloading friends, such as discoverability issues or some users having no friends.

Note 2: If you can only download fewer than 50 followers or friends for certain users that's acceptable

5 marks

Question 6

Find the most central users in your graph using all centrality measures you learned in this subject. Comment on your findings.

2 marks

Presentation

Submission Requirements (2 marks for design and readability):

Cover Sheet: Ensure your submission includes a cover sheet.

Group Member Information: Clearly list the names and student IDs of all group members. Identify the group leader.

Question Labeling: Clearly label the question number for each section of your submission.

Inclusion of Question Statement: Include the original question statement in your submission.

Code Inclusion: Include all the code you used in your analysis. This is necessary for us to verify your output.

Output Presentation: When required, display the output of your code. You don't need to print the entire dataset but include the relevant portions that are essential for understanding your solution to the question. We need to confirm you are on the right track.

Comments on Results: Provide comments on your findings. For instance, the conclusion of a hypothesis test, interpreting a plot,...

Following these guidelines will help ensure the clarity and quality of your work, earning you the full 2 marks for design and readability.

2 marks