Group Project Specification

300958 Social Web Analytics

Due Date: Friday of Week 13

1 Aim

The Group Project provides you with a chance to analyse the Social Web using knowledge obtained from this unit and a computer based-statistical package. For this project, we will focus on identifying a chosen company's Twitter image.

2 Method

To complete this project:

1. Read through this specification.

2. Form a group and register the members in your group using the Project Groups section of vUWS.
3. Choose a company that is active on Twitter, check that it is not already on the list of Group Project Twitter Handles. Then submit the Twitter handle of the company using the same link. Note that a given company cannot be allocated to more than one group. If duplicate company names are found on the list, the group with the later time stamp will be asked to find a new company.

4. Complete the data analysis required by the specification.

5. Write up your analysis using your favourite word processing/typesetting program, making sure that all of the working is shown and presented well.

6. Include the student declaration text on the front page of your report. Please make sure that the names and student numbers of each group member are clearly displayed on the front page. If a group member did not contribute to any part of the project, state their contribution as 0% (no contribution means 0 mark).

7. Submit the report as a PDF file by the due date using the Submit Group Project link
   All code and the outputs must be shown in the project, also include comments in the code to explain what you tried to do. Put all the code in the text (not to the Appendix). Any submissions other than a PDF file will not be marked.

3 Group Size and Organisation

Students in groups of size 4 or 5 are to work together to complete this project. One project report is to be submitted per group.

The group must be formed by signing-up to a group within the Project section of 300958 in vUWS. Zero marks will be awarded to lone submissions.
Groups must be formed by week 7. Once the group is formed, one person should be nominated within the group to be responsible for submitting the report.

4 Due date and Submission

The project report is due by 11:59 p.m. on the Friday of week 13. The report must be submitted as a PDF file using the assignment submission facilities in the Assessment 1 section of 300958 in vUWS. Only one student from each group needs to submit the assignment, once submitted all group members will be able to see the submission. It is your responsibility to know who has submitted on behalf of the group. No special permissions will be entertained.

5 Report Format

Once the required analysis is performed by the group, the members of the group are to write up the analysis as a report. Remember that the assessor will only see the group's report and will be marking the group's analysis based on your report. Therefore, the report should contain a clear and concise description of the procedures carried out, comments on the code, explanations of what you tried to do, the analysis of results and any conclusions reached from the analysis.

The required analysis in this specification covers the material presented in lectures and labs. Students should use the computer software R to carry out the required analysis and then present the results from the analysis in the report.

6 Marks

This project is worth 30% of your final grade, and so the project will be marked out of 30. The project consists of three investigations and will be marked using the following criteria:

Marks	**Criteria Satisfied**
10 marks	First section completed correctly.
11 marks	Second section completed correctly.
7 marks	Third section completed correctly.

There are also two marks allocated for presentation (based on the report formatting, style, grammar, clarity and mathematical notation). If the report looks professional enough to be submitted to an employer, then full two marks will be awarded.

If a report is submitted late, the maximum mark it can achieve will be reduced by 10% per day.

7 Declaration

The following declaration must be included in a clearly visible and readable place on the first page of the report.

"Names and Student IDs of all group members who contributed the project"

Student Name	Student Number	Contribution(%)

By including this statement, we the authors of this work, verify that:

- We hold a copy of this assignment that we can produce if the original is lost or damaged.
- We hereby certify that no part of this assignment/product has been copied from any other student's work or from any other source except where due acknowledgement is made in the assignment.
- No part of this assignment/product has been written/produced for us by another person except where such collaboration has been authorised by the subject lecturer/tutor concerned.
- We are aware that this work may be reproduced and submitted to plagiarism detection software programs for the purpose of detecting possible plagiarism (which may retain a copy on its database for future plagiarism checking).
- We hereby certify that we have read and understand what the School of Computing and Mathematics defines as minor and substantial breaches of misconduct as outlined in the learning guide for this unit.

Note: An examiner or lecturer/tutor has the right not to mark this project report if the above declaration has not been added to the cover of the report.

8 Project Description

Due Week 13, Friday 11:59 pm

The company "Progressive Business Private Ltd", also known as PBP wants to start using social media to promote its business. They have approached your team with a request to find what other businesses have done successfully using social media. PBP are particularly interested in using Twitter and so have asked your group to perform the following analysis on Twitter. To begin, find a company (say X) that has a Twitter handle with over 10,000 followers and 1500 tweets, then perform the following tasks using the chosen Twitter handle. Note, no two groups should use the Twitter handle for the same company X.

8.1 Analysing relationship between friends count and followers count from tweets

Company X wants to know what kind of relationship if any exists between follower count and friend count in general. Hence complete the following activities.

1. Use rtweet library to search and download 1000 tweets written in English mention your Company’s name. Ignore retweets while searching. Save these tweets as
   “tweets.company”.
   Save this information in a file named **Download_1.Data**
2. Find the follower count (save in variable x) and friend count (save in variable y) of each twitter user in (1). Be aware there might be common users.
3. A **statistic** is any summary number, like an average or percentage, that describes the sample. Calculate the average followers_count (save as xbar) and the average friends_count (save as. ybar) of your sample.
4. Find the proportion of the followers count that are higher than the average. Save this as px_hat. Similarly find the proportion of the friends count that are higher than the average. Save this as py_hat.
5. A **population** is any large collection of objects or individuals, such as about all tweets for which some information is desired. A **parameter** is any summary number, like an average or percentage, that describes the entire population (that is, not just the sample).
   a First generate a bootstrap distribution of the followers counts in the sample and use this distribution for the following
   b Plot a histogram of the bootstrap distribution of the followers counts
   c Estimate a 97% confidence interval of the average of the followers count in the population
   d Estimate a 97% confidence interval of the average of the friends count in the population
6. Use a 97% confidence to estimate the proportion of the users in the population who have higher friends_count than the average count.
7. Company X also wants to find if there is evidence to suggest that followers_count and friends_count are **independent**. Do the following to answer this question.
   a First group the number of tweets regarding the followers count as follows:
   followers_count <100 as “tens”,
   100<= followers_count < 1000 as “hundreds”
   1000<= followers_count < 2000 as “1thousands”
   2000<= followers_count < 3000 as “2thousands”
   3000<= followers_count < 4000 as “3thousands”
   4000<= followers_count < 5000 as “4thousands”
   followers count >= 5000 as “5thousandsOrMore”
   and find the frequency of the followers count under each group.
   b Similarly find the frequency of the friends count under each group, use same grouping in (a). Show your grouping result in a table.
   c Find the expected counts under each group and use a suitable statistical test in R to test the independence.
   d What is the conclusion of your test? Interpret your results.

8.2 Finding Themes in Tweets

Company X wants its twitter users to have more friends and followers. Its plan is to provide “useful information” to those users who have lower friend count or followers count. So they want to find the themes of messages posted by those users who have higher friends count than the average friends count. Those themes could then be treated as “useful information” to be disseminated to those who have lower friend count or followers count. Hence:

8. find unique users who have higher friends count than the average friends count in the sample.
9. find unique users who have less than or equal to the average friends count in the sample.
10. find the tweets of those users identified in (8) and (9), combine them and save them in a variable **tweets**
11. clean and pre-process the tweet text data in tweets
12. display the first two tweets before and after the cleaning/processing
13. create a term document matrix. Use TFIDF weights in your analysis and find how many documents were empty following the TFIDF process.
14. use the elbow method to find the appropriate number of clusters of themes among the combined tweets using cosine distance. Assume the range of clusters can be from 1 to 15.
15. Find the number of tweets in each cluster assuming there are 2 to 6 clusters.
16. Visualize your clustering in 2-dimensional vector space. Show each cluster in a different colour. Use different symbols for “the tweets that are greater than average friend count” and for “the tweets that are less than the friends count”.
17. Comment on your visualization.
18. Which cluster has the highest proportion of tweets that are greater than the average friends_count.
19. Display five samples of the tweets in clusters that you found in 18.
20. Find the important themes in the cluster with the highest friends_count proportion and the cluster with lowest friends_count proportion.
21. Use word cloud to display the themes of the clusters identified by question (20).
22. Use dendrogram to display the themes of the clusters identified by question (20).
23. What is the conclusion regarding the themes from the analysis in this section? Interpret your findings.

8.3 Building Networks

Company X wants you to find some of the influential users in the network of friends. Hence you are required to create a network to find them.

24. Find the 10 most popular friends of the chosen Twitter handle.
25. Obtain a **2-degree egocentric** graph centred at the chosen Twitter handle and plot the graph. The egocentric graph should contain the most popular 10 friends of the chosen Twitter handle.
26. Compute the **closeness** centrality score for each Twitter handle in your graph.
27. List the top 3 most central people in your graph according to the **closeness** centrality.
28. Comment on your results.

Important notes:

Note that in Section 8.3, depending on the friends of the chosen twitter handle, you possibly will reach the rate limit of the Twitter API. I strongly recommend that you save your objects as an RData file once you download friends - so you can continue downloading friends the following day or with a different authentication key. For more information on how to save your objects see:
https://stackoverflow.com/questions/19967478/how-to-save-data-file-into-rdata
(https://stackoverflow.com/questions/19967478/how-to-save-data-file-into-rdata) .

See this https://developer.twitter.com/en/docs/basics/rate-limiting.html
(https://developer.twitter.com/en/docs/basics/rate-limiting.html) for more information on the rate limit.

The company wants the above three-part analysis to be written up as a professional report. Each part should have its own section of the report and all questions should have thoughtful answers. Include all the R code along with its output in your assignment. Output without the code, or code without the output will result zero marks for the relevant section.