Practical Specification

Social Web Analytics

There a chance to analyse the Social Web using knowledge obtained from this unit with assistance from a computer based statistical package. For this project, we will focus on identifying a chosen companies Twitter image.

1 Method

To complete this project:

1. Read through this specification.
2. Complete the data analysis required by the specification.
3. Write up your analysis using your favourite word processing/typesetting program, making sure that all of the working is shown and that is it presented well.

Report Format

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.

2 Marks

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

Marks	Criteria Satisfied
0-5	One of the project parts have been completed correctly.
6-10	Two of the project parts have been completed correctly.
11-15	Three of the project parts have been completed correctly.
16-20	All of the project parts have been completed correctly.
21-25	The required work has been completed correctly and the company questions have been answered based on the results.

There are also five marks allocated to presentation (based on the report formatting, style, grammar and mathematical notation). If the report looks like something that would be submitted to an employer, then the full five marks will be awarded.

If a report is submitted late, the maximum mark it can achieve will be reduced by 10% (3 marks) per day. E.g., if a report is submitted five days late, it can receive at most 15 marks.

3 Declaration

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

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.

4 Project Description

A company is investigating its public image and has approached your team to identify what the public associates with the company name. The company wants the four pieces of analysis to be performed.

8.1 Analysis of Twitter language

In this section, we want to examine the language used in tweets.

1. Download the random sample of tweets randomSample2016.csv , and load using `randomSample = read.csv(″randomSample2016.csv″)`
2. Use the `tm` library to construct a document-term matrix of term frequencies.
3. Sum the columns to obtain a vector of term frequencies summed over all tweets.
4. Compute the proportion of each term in the random sample, from the vector of term frequencies.
5. List the top 10 words and their proportion.

What do these words tell us about the random sample?

8.2 Analysis of the company image on Twitter

Important: Choose @steam_games

In this section we want to examine if the tweets about the company are of random topics, or if they contain specific information.

1. Use the `searchTwitter` function from the `twitteR` library to search for 1000 tweets about the company.
2. Combine the company tweets with the random sample of tweets from part 1 and construct the document-term matrix of term frequencies over all tweets.
3. Sum the rows associated to the company tweets to obtain a vector of term frequencies over all company tweets, and do the same for the random tweets.
4. Combine the two vectors (of company tweets and random tweets) to create a $2 \times M$ table (where M is the number of terms, and also the length of the vectors), then perform a χ^2 test for independence on this table.

What do the results of the test tell us about the company tweets and random sample of tweets?

8.3 Connection between public and company.

In this section, we want to determine what the public and company are tweeting about, when tweeting about the company.

1. Use the function `userTimeline` from the `twitteR` library to download the last 1000 tweets from the company.
2. Combine the tweets about the company from part 2 with the tweets from the company and cluster the tweets using an appropriate clustering method.
3. List the set of terms associated to each cluster.
4. Compute the proportion of company tweets in each cluster.

What do these results tell us about the tweet topics from the company and public about the company?

8.4 Company Image

By combining all of this information, what can we say about the company's image on Twitter? Also identify any problems with the analytical process used in each part and how the results may have been effected by these problem (do not include programming problems).

The company want the above four parts of analysis to be written up in a professional report. Each part should have its own section of the report and all questions should have thoughtful answers. Any code that is used should be included and clearly explained (include comments in the code).