4030ICT

Big Data Analytics and Social Media

Assignment Specifications

Course Convenor:

Sebastian Binnewies

Instructions

- **Structure**: This assignment is broken up into three milestones. You complete Milestone 1 first, then later Milestone 2, and finally Milestone 3. The milestones are aligned to your progress through the contents of the course.
- **Due**: See course site on Learning@Griffith for the due dates of each milestone.
- **Marks**: 30 marks total (= 30% of course grade).
  Milestone 1: 6 marks. Milestone 2: 9 marks. Milestone 3: 15 marks.
- **Late Submissions**: The standard penalty for delayed milestone submissions is the reduction of the mark allocated to the milestone by 5% of the maximum mark applicable for that milestone, for each working day or part working day that the milestone is late. Milestones submitted more than five working days after the due date are awarded zero marks.
- **Extensions**: If for any valid reason (e.g., being sick) you need an extension, you must apply for an extension by the due date of the milestone through this online form: https://www.griffith.edu.au/students/assessment-exams-grades/submitting-assignments/assignment-extension
- **Group Work**: You may complete this assignment in a group of maximum 2 students. If you work in a group, you must complete all steps below for two case studies (instead of only one if you are working by yourself).

Overview

In this assignment, you are required to think about a case study, in which you can apply social media analytics to gain insight about how a certain artist or band can improve their popularity. You will need to describe the setting for your case study, apply social media analytics using the tools introduced during the labs, and evaluate your findings and determine appropriate future actions.

You are required to use software for analysis and produce a written report. The report accounts for the majority of marks for each question and you will need to present your findings in the lab to your tutor to receive marks. Simply pasting screenshots of your analysis outputs will not give you full marks. Accuracy and reproducibility of your code will be checked.

- Choose data sources and data that are appropriate for your case study.
- Pay attention to how much data you retrieve and how frequently you retrieve data. If you try to get lots of data often, the APIs will impose a rate-limit on your account. However, you can still proceed after the rate-limit has ended.
- Use the software introduced in the labs (RStudio, Tableau…).
- Add headings in your R scripts so that we can easily find the code related to each question.
- Export all datasets as .RData files (so that we can re-run your code if needed).
- Add screenshots of your results.
- Make plots/visualisations wherever possible, using R functions, Gephi, and Tableau. (Most results can be displayed as a plot/visualisation!)

Milestone 1

- Choose a well-known artist or band. Assume you are the artist's/band's manager and want to help improve their popularity by using social media analytics.
- Your chosen artist/band should be well-known already so that there exists enough social media data that is somehow related to it. Otherwise, you may not be able to retrieve enough useful data for performing the analytical steps later.

Case Study Setting

1.1) Describe the artist/band you are managing by using data from the Spotify API. Add additional information from other sources (e.g., Wikipedia).

For example:

- How many years have they been active?
- How many albums & songs have they published?
- What are the prevalent features of their songs (e.g., valence)?

[1-2 paragraphs, 2 marks]

1.2) Describe the purpose of using social media analytics for your case study.

For example:

- How do you want to improve the popularity?
- How can social media analytics help you achieve it?
- What kind of social media data do you want to analyse?
- What is your hypothesis (expectation) about the analysis outcome?

[2-3 paragraphs]

Note: You need to complete Question 1.2) for Milestone 1 but will have a chance to improve your answer when you submit Milestone 3. Therefore, final marks for Question 1.2) will be allocated as part of your Milestone 3 submission.

Data Selection & Exploration

1.3) Select social media platforms (Twitter, Spotify) and retrieve data. Make sure to choose keywords for data retrieval that are most relevant to your artist/band. However, try not to be too narrow. As a rough guide, you should retrieve at least 1000 data items (e.g., tweets from Twitter). Explain what you have done. *(=> Labs 2, 3)*

[0.5 mark]

1.4) List the top 5 most influential users for your artist/band. Find out what other interests/characteristics they have besides those related to your artist/band. Do these 5

have something in common? *(=> Lab 2)*

[1 mark]

1.5) List the top 10 most important terms that appear together with your keyword(s) related to your artist/band. Explain the results. *(=> Lab 2)*

[0.5 mark]

1.6) For your Twitter dataset, calculate how many of your retrieved tweets are retweets. Alternatively, if you filtered out retweets in your query, calculate how many unique user accounts there are in your dataset. What do the results tell you? *(=> Labs 1, 2)*

[1 mark]

1.7) Find related artists/bands on Spotify and create a network graph. Did you find any interesting relationships? *(=> Lab 3)*

[1 mark]

Milestone 2

Data Selection & Exploration (continued)

2.1) Retrieve data relevant to your artist/band from YouTube. Which videos have the highest number of views and likes? Do you see a correlation between views and likes? (Your dataset may contain hundreds of videos, so it's OK if you choose only a subset of those to get their statistics, in order to avoid hitting the rate-limit. However, you should get statistics for at least 5 videos.) *(=> Lab 6)*
[1 mark]

Text Pre-Processing

2.2) Perform text pre-processing and create a Term-Document Matrix. What are the 10 terms occurring with the highest frequency? How are they different to your answer for 1.5) above? *(=> Lab 4)*
[1 mark]

Social Network Analysis

2.3) Perform centrality analysis by detecting degree centrality, betweenness centrality, and closeness centrality. Explain how relevant the results are to your artist/band. What are the actual degree, betweenness, and centrality scores for your artist/band node in the network? Compare these scores to the scores for related artists. *(=> Lab 5)*
[3.5 marks]

2.4) Perform community analysis with the infomap, Girvan-Newman (edge betweenness) and Louvain methods. Explain how relevant the results are to your artist/band. Perform the community analysis also for related artists. Is their community structure similar? *(=> Labs 5, 6)*
[3.5 marks]

Milestone 3

Case Study Setting (continued)

Revisit Question 1.2) from Milestone 1 and improve your answer based on your new knowledge that you gained throughout the last weeks.
[2-3 paragraphs, 1 mark]

Machine Learning Models

3.1) Use k-means clustering to classify a user's friends (following) and followers. You have to identify one influential user related to your artist/band and analyse his/her friends and followers. Justify why he/she is an influential user. Explain the results. *(=> Lab 7)*
[1.5 marks]

3.2) Build a decision tree and evaluate its performance in predicting whether a song is by your artist/band. *(=> Lab 7)*
[1.5 marks]

3.3) Use sentiment analysis to identify how the public reacts to events and/or topics related to your artist/band. Provide a summary of public opinions (emotions, reactions). *(=> Lab 8)*
[2 marks]

3.4) Use LDA topic modelling to identify some terms that are closely related to your artist/band. Find at least 3 significant groups of words that can be meaningful to your analysis. Explain your findings. *(=> Lab 8)*
[2 marks]

Visualisation

3.5) Plot the location of tweets from your Twitter dataset on a map using RStudio. Explain your findings. (If you do not have enough tweets with location data in your Twitter dataset that you have been using for the previous questions, you can run a new search to get a dataset with more location data.) *(=> Lab 5)*
[0.5 mark]

3.6) Plot the location of tweets from your Twitter dataset on a map using Tableau. Add other attributes as additional marks to your map (e.g., as colour, size). Explain your choice of attributes and visual marks. *(=> Labs 9, 10)*
[1 mark]

3.7) Create at least two other charts from your datasets using Tableau and combine them together with your plot from the previous question into a dashboard. Explain the functionality of your dashboard. *(=> Labs 9, 10)*
[1.5 marks]

Evaluation

3.8) What are the findings of your social media analytics?
[2-4 paragraphs, 2 marks]

3.9) What actions for improving the popularity of your artist/band do you suggest based on your findings?
[1-2 paragraphs, 1 mark]

3.10) How could you refine your social media analytics?
For example:

- Could you use different data sources?
- Could you choose different parameters?
- Can you think of ways to obtain more relevant data?

[1-2 paragraphs, 1 mark]