7230ICT

Big Data Analytics and Social Media

**Assignment Specifications**

# 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 music 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. In the report, you need to explain your analysis findings. The report accounts for the majority of marks for each question. Simply pasting screenshots of your analysis outputs will not give you full marks. Accuracy and reproducibility of your code will be checked. You will also need to present your findings in a recorded video presentation.

- 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 (e.g., RStudio, Gephi).
- 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 in the report. Include in the screenshot the code you wrote to produce that result. (Usually this would be the last line of code.)
- Make plots/visualisations wherever possible. (Most results can be displayed as a plot/visualisation!)

# Advice

1. Make sure to read through all the information in detail & before you start (everything in this document, the marking rubric, and the submission instructions on the course site).
2. Make sure to address all parts of each Assignment question and use the marking rubric to guide you.
3. Start answering the Assignment questions by going back to the lab scripts and altering them to fit your case study based on the specifications. Then, use the rubric to improve your answer incrementally.
4. Once you are ready to submit, make sure to follow the submission requirements, otherwise you will lose marks.

# Instructions

- 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) Describe the artist/band you are managing. Make sure to reference your sources properly (don't plagiarise). Use APA referencing style.
   For example:
   - How many years have they been active?
   - How many albums & songs have they published?
   [1-2 paragraphs, 1 mark]

**Data Selection & Exploration**

2) Collect data about your artist/band from YouTube and/or Reddit. Make sure to choose keywords/videos/subreddits/threads for data retrieval that are most relevant to your artist/band. However, try not to be too narrow. You should retrieve 3000-8000 data points. List the keywords/videos/subreddits/threads and explain your search strategy, choice of data sources, and how much data you have collected.
   [1.5 marks]

3) Create actor networks from your data and list the top 5 most influential actors for your artist/band according to page rank. Explain the results.
   [2 marks]

4) Calculate how many unique actors there are in your datasets. Explain the code you have used for the calculation. What do the results tell you?
   [2 marks]

5) Use the Spotify API to extract data about your artist/band.

   For example:

   ○ How many years have they been active?

   ○ How many albums & songs have they published?

   ○ With whom have they often collaborated?

   ○ What are the prevalent features of their songs (e.g., valence)?

   How does the Spotify data compare to the information you collected from other sources in Question 1)?

   [1.5 marks]

## Text Pre-Processing

6) After performing text pre-processing, create Term-Document Matrices for your data. What are the 10 terms occurring with the highest frequency? Explain the results.
   [2 marks]

7) After performing text pre-processing, create semantic (bigram) networks from your data and list the top 10 most important terms according to page rank. Explain how and why they differ from your results for the question above.
   [2 marks]

## Social Network Analysis

8) 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 closeness centrality scores for your artist/band node in the network? Compare these scores to the scores for other artists that are related to your artist/band.
   [3.5 marks]

9) Perform community analysis with the 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?
   [3.5 marks]

**Machine Learning Models**

10) 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).
[2 marks]

11) Build a decision tree and evaluate its performance in predicting whether a song is by your artist/band.
[2 marks]

12) 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.
[2 marks]

**Visualisation**

13) Create two dashboards (pages), each with at least three charts, from your datasets using Power BI. Describe each chart in your dashboard and why you chose to include it. Explain the functionality of your dashboard and what insights you can obtain from it.
[3 marks]

**Analysis Review**

14) Research and review other methods/algorithms for network analysis, machine learning models, or visualisation. Compare them to the methods you used in this Assignment. Did you find a method that could give you better insights or more promising results for your social media analytics? Explain why you think so.
[2-4 paragraphs, 2 marks]

# Video Presentation

To complete your Assignment submission, you will need to record a video presentation of **maximum 5 minutes** duration. You should use PowerPoint slides or similar to show the results from your report.

**You will also need to record yourself while you are presenting and show your student ID (or driver licence) at the beginning.** When you show your ID, make sure that your video is recording in full screen, and <u>not</u> minimised over the presentation (see images below). You must check the recording to verify that your ID is legible.



In the video, you should answer the following questions:

**Evaluation**

- Briefly introduce yourself (show your student ID) and your artist/band. What data have you collected (data sources, search terms, search parameters, amount of data)?
  [2.5 marks]

- What are the findings of your social media analytics? How did you obtain your findings?
  [5 marks]

- How could you refine your social media analytics?
  For example:
  - Could you choose different parameters?
  - Could you choose different methods for analysis?
  - Can you think of ways to obtain more relevant data?
  [2.5 marks]