
You will need to write a report with multiple chapters, one for each of the following task.

Task 1 (12 points)

This task is about data scrapping, data wrangling and natural language processing.

(a) (2 points) Formulate one clearly-defined question to be answered by studying tweets.

(b) (1 point) Scrap as many data points and as many twitter users as twitter allows.

(c) Process the data and answer the question in (a). Your analysis must include

i. (2 points) Forming a rectangular tidy data

ii. (1 point) Analysing word frequency

iii. (2 points) Sentiment analysis

iv. (1 point) Visualisation

(d) (3 points) Present your findings in a report format. Submit your R-code on canvas.

Task 2 (10 points)

This task is about model formulation and prediction.

The goal is to use past movie ratings to predict how users will rate movies they haven't watched yet. This type of prediction algorithm forms the underpinning of recommendation engines, such as the one used by many streaming providers. We have two sources of data,

Original	Additional	
1_movie_names.tsv	2_credits.csv	2_movies_metadata.csv
1_movie_ratings.csv	2_keywords.csv	2_ratings_small.csv
1_users.csv	2_links_small.csv	2_ratings.csv
1_predict.csv	2_links.csv	

The dataset 1_predict.csv contains ratings for you to predict, with all ratings set initially to 0, every rating needs to be between 1 and 5. You are restricted to use only the given datasets. But you are free to choose any approach/model to make your rating predictions. For example, the choice between using integer ratings and allowing any real number rating is yours. Present your findings in a report format. Submit your R-code on canvas.

Task 3 (8 points)

This task is about working with a big dataset. You need to form a *group of 3-4 for the task*.

The Million Song Dataset (280GB)

provides a million contemporary popular music tracks. The dataset consists the feature analysis and metadata for the songs. It does not include any audio, only the derived features. We will use only a subset of it, and revisit the whole dataset using Hadoop later.

(a) (2 points) Load the subset of the Million Song Dataset into R.

(b) (5 points) Use three of your computers to form a cluster using the H2O package to perform k-means for $k = 1, 2, 3, 4, 5$ to partition the 10,000 songs in the subset. Use as much information in the subset as possible. You are recommended to use data.table to wrangle the data so that it is appropriate for k-means.

(c) (1 point) Present your findings *individually*. Submit your R-code on canvas.