PROJECT 2 (20%)

STQD6114 UNSTRUCTURED DATA ANALYTICS

SEMESTER 2 2023/2024

Data acquisition

Compile article news from any platform (e.g., New Straits Times, The Star, etc.) regarding **one** of the following themes:

i. Health
ii. Sport
iii. Financial issues
iv. Political views
v. Natural disasters
vi. Success story
vii. Science and technology

From the above,

Find at least **50 news** for your selected theme and save it to one folder. Extract the title and its content. Paste each news to different .txt files. You may refer to the folder "TextMining" under section Text Data Analysis in UKM Folio for your reference.

For the next analyses task, you may need to do some relevant data preprocessing, cleaning and converting to document term matrix.

Task 1: Perform topic modelling analysis using LDA.

1. Using the dataset created in **Data acquisition** section, create four topics, k=4.
2. Perform the relevant analysis such as
   i. Extract per-topic per word probabilities and visualize at least eight (8) terms that are most common within each topic.
   ii. Extract the relevant beta spread. You may want to understand the greatest difference between any topic (e.g., topic 1 and topic 3, etc.), depending on your preference.
   iii. Perform per-document per topic probabilities.
   iv. Other relevant analysis
3. Write a comprehensive report that is equipped with relevant outputs and interpretations. Your report must include the following:
   i. Introduction on the selected theme.
   ii. Discussion on the finding of the topic modelling analysis from part 2 above.
   iii. Conclusion.

Your report should be at least two pages long using times new roman, font 12 and spacing 1.5.

Task 2: Perform text clustering.

1. Using the dataset created in **Data acquisition** section, construct data clustering by using *k*-means, hierarchical and HDBScan algorithms.
2. Perform the relevant analysis (including the relevant visualization) on each of the clustering algorithms. Compare the results.
3. Write a summary to discuss the analysis obtained from part 2 above (and overall clustering results) together with your conclusion (the most appropriate clustering algorithm for your dataset along with its reason).

Task 3: Sentiment analysis.

1. Find any reviews data from Kaggle.com.
2. Perform the analysis such as obtaining sentiment scores using different lexicon, find the most common positive and negative words, perform emotion classification and other related analyses.
3. Write an essay that is equipped with relevant outputs and interpretations. Your essay must include the following:
   i. Introduction on the selected reviews.
   ii. Discussion on the analysis obtained from part 2 above.
   iii. Conclusion.

   Your essay should be at least 500 words long.

Due date: 9th June 2024, before 11 pm at UKMFolio.

Additional information:

Please attach your dataset (your folder in **data acquisition** section and txt/csv file in **Task 3**) and all the codes including the data preprocessing (in R script or attached as appendix).