PROJECT 1 (30%)

STQD6114 UNSTRUCTURED DATA ANALYTICS

SEMESTER 2 2024/2025

PART 1 – Unstructured Text Mining

Task 1

1. Pick one example of raw data. Explain how this raw data turns into wisdom using DIKW pyramid.
2. Based on your opinion, explain why unstructured data gains its popularity nowadays?

Answer the following questions based on video https://youtu.be/dK4aGzeBPkk

3. Two popular examples of big data technology are Nextflix and credit-card. Explain how big data is used in these examples.
4. For EVERY application of big data found in different fields (e.g., banking telecommunication, healthcare, etc.) mentioned in the video, give two real-life examples of each field.
5. Construct the table to differentiate between structured, unstructured and semi-structured. Give at least 3 points, including sources of the data.

Task 2

Find any website that have multiple pages regarding **one** of the following:

1. Online purchase website (example, Amazon, lazada, etc). Select two different products with the same categories (example, Sneekers and high heels).
2. Movies of two different genres.
3. Songs from two different artists. The artists must have produced at least 20 songs.
4. Providers of two different services/industries.

From the above,

a. Extract the information for the first three pages.
b. Build a data set for the information you gained. Your data set should consist of the profiles/ characteristics of the items you have chosen with at least four variables (example, for a movie, it may comprise of ratings, year, title, director, production company etc. and for a product, it may comprise of name of item, description, prices etc).
c. Perform simple analysis to compare the two different groups in (1).
c. Write a short article on your findings which includes the following:
-Introduction (what have been chosen and why was it chosen)
-Compare these two different groups.
-Conclusion

Your short article should be at least one full pages using times new roman, font 12 and spacing 1.5.

Task 3

Find five lyrics of different themes and save it to csv file.

1. Perform data cleaning/preprocessing such as remove punctuation, remove stop words, etc.
2. Convert to document term matrix and find the frequency. Tabulate the frequency and its corresponding terms (at least five terms).
3. Represent your terms in the form of word cloud of any shape
4. Write short essay which includes the following:
   -Introduction of the lyrics
   -Discussion on the overall finding in part 2 and 3 above.
   -Conclusion

Your short essay must be at least 300 words.

PART 2 – Text Data Analysis

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 three topics, k=3.
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. Based on question 2 above, try different number of topics and perform related analyses. What can you comment?
4. 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 question 2 and 3 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 by selecting relevant number of clusters.
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. Your summary must be at least 500 words.

Task 3: Sentiment analysis.

1. Find any reviews data from any trusted sources.
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: 22th June 2025, before 11.59 pm at UKMFolio.

Additional information:

1. This is individual project and you have about six weeks to complete this project.
2. Please be noted that you cannot copy/plagiarism other members project.
3. Please submit all the supplementary materials used in this project for each question.
4. Please submit the R script for each questions.
5. Similarity check will be conducted at random, to ensure the originality of your submission. Penalty will be imposed for those who been detected with high plagiarism and highly usage of AI assistance.

"ALL THE BEST"