PROJECT 1 (20%)

STQD6114 UNSTRUCTURED DATA ANALYTICS

SEMESTER 2 2023/2024

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 two pages long using times new roman, font 12 and spacing 1.5.

Task 3

Find three 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.

Due date: 28th April 2024, 5 pm and submit to UKMFolio

Additional information:

1. This is individual project and you have about 3 weeks to complete this project.
2. Please be noted that you cannot copy/plagiarism other members project.
3. Please sent all the supplementary materials used in this project, e.g. csv files of your task 3, etc.
4. Please also sent the R codes and I will randomly check later that the submission is original from you. You may share the R script or copy the codes and attached as appendix or show one by one along the results, that's no format on the codes submission as long as it's attached together with your submission.

"ALL THE BEST"