MIE 223 Data Science
Lab and Assignment 05:
Data Science with Natural Language

In this lab and assignment, you will learn and gain hands-on experience on applying natural language processing (NLP) techniques to help analyse review data.

- Programming language: Python (Google Colab Environment)
- Due Date: Posted in Syllabus

Marking scheme and requirements: You are required to use the Jupyter notebook named `Assignment_05.ipynb` to provide appropriate answers to the questions in this assignment. All changes to the `Assignment_05.ipynb` notebook should be committed and pushed to your assignment repository on Github.

This assignment has *4 points* in total and the point allocation is shown below:

- Solution to Questions (1 point):
  - Q1: 0.1 point
  - Q2: 0.1 point
  - Q3(a): 0.1 point
  - Q3(b): 0.1 point
  - Q3(c): 0.1 point
  - Q3(d): 0.1 point
  - Q3(e): 0.1 point
  - Q4(a): 0.1 point
  - Q4(b): 0.1 point
  - Q4(c): 0.1 point
- Code review (1 point)
- Quiz (2 points)

What/how to submit your work:

- All your code should be included in the provided notebook named `Assignment_05.ipynb`.

- Commit and push your work to your github repository in order to submit it. Your last commit and push before the assignment deadline will be considered to be your submission. You can check your repository online to make sure that all required files have actually been committed and pushed to your repository.
- A link to create a personal repository for this assignment is posted on QUERCUS.

Notes that you should pay attention to

1. All your code should be written in the provided notebook named `Assignment_05.ipynb`.
2. All functions must *return* the specified return type.
3. Commit and push your work to your GitHub repository in order to submit it. **Your last commit and push will be considered to be your submission.** You can check your repository online to make sure that all required code has actually been committed and pushed to your repository.
4. **Your code should have output for all assignment questions unless otherwise stated**. If any of the assignment questions asks you to explain your outputs, please use markdown cells to clearly explain them.
5. As a backup plan, also upload a PDF copy of your assignment notebook to GitHub. This is useful especially when GitHub cannot render some images or outputs.
6. During the code review, TAs will go over your answer submissions to assess your understanding of the concepts relating to the assignment and the logic behind your codes. Please, ensure your codes are readable and well commented using python comments and markdown cells when necessary. Not have readable code may impact your code review score.
7. Please note the plagiarism policy in the syllabus.

1 Main Assignment

Please answer the questions below in the provided notebook named `Assignment_05.ipynb` using Google Colab. **Numpy** and **Pandas** functions should be used for all computations. Unless otherwise stated, the Pandas DataFrame '.plot()' function or **Matplotlib** should be used to generate plots. **Clearly give titles and labels to all plots**.

You are provided with a review dataset, scraped from TripAdvisor, for hotels in Kingston, ON. This dataset is similar to the one used in the lab. For each review, it contains: the hotel name, review text, rating, sentiment ground truth, datestamp, and the hotel address. As a data scientist looking to better understand the hotel scene in Kingston, you want to be able to extract useful insights from these reviews. Some questions you might want to consider are:

- How does the distribution of ratings look?
- Are ratings correlated with time?
- Can we better understand how certain words or phrases relate to sentiment and ratings?

Using the **"hotelDf"** DataFrame, complete the following questions.

Q1. Histogram

Complete the 'get_histogram' function provided to plot a histogram of the ratings in **"hotelDf"**. Ensure you label your axes and title the plot. Use the default parameters for bin size.

Q2. Time Series

Complete the 'plot_time_series' function to plot the number of reviews and rolling average rating score for 3 hotels over time. Plot these on separate axes, as shown in the lab. The 3 hotels you should use are provided in the Jupyter notebook. You should use a rolling average over the last 5 rating scores.

Q3. Frequency, MI, and PMI

(a) Complete the 'most_frequent_words' function to find the most common 500 words across all tokens in the dataset. You should return a list sorted in descending order by frequency. The words returned should be all in lowercase, with no stop words and have at least 3 characters. You should use NLTK's 'word_tokenize' function to extract the tokens. Print the top 10 most frequent words.

(b) Calculate the MI score of the top 500 most common words across all lowercase tokens in the dataset. Print a DataFrame showing the 5 words with the highest MI scores and their MI scores.

(c) Calculate the PMI for the words "great" and "dirty" across positive and negative sentiments and for the case that the word is present and not present. You should print 2 DataFrames summarizing the PMI results, similar to the ones in the lab.

(d) Create a DataFrame with the top 50 most common noun phrases across the reviews. Use the same grammar as in Lab 5. Print the top 10 most common noun phrases with their frequencies.

(e) Using only the hotels with at least 15 reviews, find the average rating for each of these hotels. Print the hotel and average rating of the hotel with the highest average rating and lowest average rating. Using only the reviews for each hotel (ignore reviews of all other hotels), determine the 5 noun phrases with the highest PMI for positive and negative sentiments separately. Do this for both the highest rated and lowest rated hotels. In total, you should print 4 DataFrames with 5 rows each and 2 columns (the noun phrase and PMI), plus a header row (PMI for positive and negative sentiment for each of the 2 hotels).

Q4. Noun Phrase Extraction

(a) Write the grammar RegEx for a noun phrase (named NP) that may or may not start with a determiner (DT tag), followed by any number of adjectives (including comparative and superlative adjectives using JJ, JJS, or JJR tags) or possibly none, followed by at least 1 singular or plural noun (NN or NNS tags).

(b) Calculate the PMI of each of the top 50 most common noun phrases with the positive sentiment. Print a DataFrame showing the 5 noun phrases with the highest PMI with positive sentiment. Repeat the same for negative sentiment.

(c) What does this PMI analysis tell you about what many people dislike about the hotels here?