2023 Sem 2 IT2391 Assignment

Assignment (40%) **(Total = 60 marks)**

INSTRUCTIONS:

1. This assignment will be done in Jupyter Notebook with Python programming. When you finished editing, re-run all the cells to make sure they work.
2. Submit .ipynb file for each part of the assignment.
3. This assignment contains 2 parts.
4. Name your files as follows: IT2391_<admin_no>_Part_1. ipynb and IT2391_<admin_no>_Part_2. ipynb.
5. Zip up all the files and submit to Brightspace.
6. Your submission to Brightspace is **final**. Please be sure to check all files carefully before submitting to Brightspace.

SUBMISSION DATE: 15 Dec (Friday), 2359hrs

OBJECTIVES:

On successful completion of this practical, the students should be able to:

- Apply the different text pre-processing techniques needed for text analysis
- Create a text analytic model using classification method
- Elicit a good understanding of the data
- Perform good coding practices by describing your code / task / findings in code comments and markdown text

Dataset:

Please download the zipped file (data.zip) from Brightspace and unzip the contents. There is two files in the folder: data.csv and data_test.csv. This dataset will be used in your Practical Assignment Part 1 and Part 2.

Practical Assignment Part 1 – Data Understanding and Text Preprocessing

Tasks (your tasks should include but not limited to):

- Get a general understanding of the data (for example, how many data points in total, the percentage of each category, the distribution of character lengths for the category by using visualization, etc.)
- Apply pre-processing steps/techniques to the text (for example, remove all words that contain numbers, make all the text lowercase, etc.)
- After going through these pre-processing steps, demonstrate your understanding of the data again (for example, what are the most common words now? Do they make more sense? and etc.)
- Save your cleaned data to cleaned_data.csv

Practical Assignment Part 2 – Text Classification

Tasks (your tasks should include but not limited to):

- Create a text classification system A
  - Split the data into training and test sets
  - Apply feature engineering
  - Use Logistic Regression / Naïve Bayes model to classify review as positive or negative
  - Measure model performance
- Create a text classification system B or more using different features and/or different model (Logistic Regression / Naïve Bayes)
- Evaluate the performance of system A vs. system B (or more systems)
- Choose the best model and apply it to data_test.csv. Evaluate the result.

Marking Scheme

Part 1 (30 marks)	Initial data understanding	5 marks
	Pre-processing techniques	20 marks
	Data understanding after text pre-processing, save as a new dataset	5 marks
Part 2 (30 marks)	Multiple text classification system creation and performance comparison	18 marks
	Best model to predict unseen data	12 marks

-End-