THE UNIVERSITY OF THE WEST INDIES
ST. AUGUSTINE

EXAMINATIONS OF MAY 2020

Code and Name of Course: COMP 3610 Big Data Analytics Paper: 1

Date and Time: 2nd June - 4th June 5pm Duration: 2 days

INSTRUCTIONS TO CANDIDATES: This paper has 3 pages and 2 questions.

Answer all questions.

Your answers must be submitted via the MyElearning platform as a zip folder containing files in PDF and Jupyter notebook formats.

PLEASE TURN TO THE NEXT PAGE

1 Question 1 [50 marks]

In this section you are required to use the dataset entitled “Emissions.csv”.

1. Using a Jupyter notebook, load and examine your dataset. [2 marks]

2. Clean the data and explain, using markdown code, your reason for each cleaning operation performed. [5 marks]

3. Use plots/graphs to visualize and examine the data in more detail. Describe, using markdown code, what you observed from each plot/graph. [8 marks]

4. Identify a machine learning problem that can be explored/solved using this dataset. Explain your problem definition in detail. [5 marks]

5. Based on your topic from question 4, perform a 300 word literature review. You are required to use at least 2 papers for your review. [8 marks]

6. Using your problem definition from above, answer all Heilmeier questions. If a question cannot be answered, state the reasons why. [6 marks]

7. Describe the steps you would take to solve your proposed problem. In addition, explain what methods (graphs, metrics, etc) you would use to evaluate your results. [10 marks]

8. Using at least one machine learning algorithm, solve piece of your machine learning problem. Provide details and explanations of the results. Note that the complete problem may require more analysis but you are not required to do this. [6 marks]

2 Quetion 2 [50 marks]

For this section you will use the dataset entitled “news.csv”.

1. Load and explore your dataset. [2 marks]

2. Perform any necessary cleaning and explain why each step is performed. [5 marks]

3. Perform topic modelling on the “title” column using NMF, LDA and SVD, identifying the top 15 topics for each. Indicate which algorithm gives the best topic clusters and why. Using the topics from this algorithm give each topic a category. [12 marks]

4. Repeat question 3 using the “text” column. [6 marks]

5. Compare the categories obtained from the “title” and “text” columns in order to determine how accurately the titles describe the text. [3 marks]

6. Depending on the topic being discussed, the author may choose to remain anonymous. Determine what topics do not have authors. Comment on these topics. [8 marks]

7. Plot the distribution of the topics by author. [3 marks]

8. Plot the frequency of articles by author. [3 marks]

9. Write a 300 word literature review on the applications of topic modelling. You are required to reference at least 2 papers in your review. [8 marks]