Introduction

Quora is an American question and answer website that contains questions that are posted and answered by registered site users. Quora questions are automatically categorised using AI techniques.

For your second CA you will build an AI system that can automatically detect the correct topic for any inputted question. You will use the Quora dataset which contains a sample of over 800,000 questions. The relevant category is not tagged with each question in this dataset.

The dataset is called **Quora questions** and is available for download under the Assignments & Tests section in Blackboard. **You must use this dataset and not one from the internet.**

This assessment is worth 35% of the total marks for the Artificial Intelligence 2 module. You must submit your work through Blackboard in the **NLP CA 2 submission** link that can be found under the **Assignments & Tests** link on the left of the Blackboard screen.

Q1 Text Classification

Choose 200,000 records from this quora dataset and discuss the processes needed to load and pre-process it. Then critique and evaluate the most accurate and suitable text classifier for your data. Build the text classification system that fits best with your dataset. Justify all your choices in detail.

Categorise your text classification outputs using suitable descriptors for each topic. Here is a sample of descriptors used by Quora. **Please note: The sample dataset may not contain all these topics so choose wisely.**

- Politics, Law, Government, and Judiciary
- Humanities
- Life, Relationships, and Self
- Science, Technology, Engineering, and Mathematics
- Recreation, Sports, Travel, and Activities
- Literature, Languages, and Communication
- Business, Work, and Careers
- Art, Design, and Style
- Medicine and Healthcare
- Food, Cuisines, and Cooking
- Education, Schools, and Learning
- News, Entertainment, and Pop Culture
- Major Concepts
- Honours and Recognition

Assign a topic description to each topic and update your quora dataset with this information.

Store the newly categorised dataset into a new csv file with the name **quora_supervised**

Q2 Supervised Learning

Using your modified and categorised csv dataset called quora_supervised, evaluate and develop the most accurate supervised learning model for your data. Critically analyse and discuss each of the steps required to choose the most suitable model.

Finally prove that the trained supervised model works by choosing 10 random questions from your dataset and demonstrate whether the selected model produces output as expected. Discuss and evaluate the output of your model.

Important Information

Plagiarism will not be accepted and will result in an automatic mark of zero.

If you use references, the Harvard referencing must be adopted. Please use the following link which might help you create the references required: http://www.neilstoolbox.com/bibliography-creator/.

Due Date: Wednesday 15th April at 23:59. You must submit your work through Blackboard using the relevant link. Submit all files you used including your Python code as a Jupyter notebook file. A cover sheet must also be submitted with your jupyter notebook file. You may submit all your work in a compressed file.