COMP47670 Assignment 2

Deadline: Friday 23rd April 2021

Overview:

The objective of this assignment is to scrape a corpus of news stories from a set of web pages, pre-process the data, and evaluate the performance of both binary and multi-label text classification algorithms on the data. The news stories are archived by month at the link below and each story has been assigned one of 9 news categories.

http://mlg.ucd.ie/modules/COMP41680/assignment2/index.html

Based on this data, you should complete the three tasks listed below. The assignment should be implemented as a single Jupyter Notebook (not a script file). Your notebook should be clearly documented, using comments and Markdown cells to explain the code and results.

Task 1. Data Collection

1. Select **three** of the 9 news categories: [Books, Business, Film, Life-and-Style, Music, Politics, Sport, UK-News, US-News]
2. From the link above, retrieve details regarding all stories corresponding to your three selected categories, covering all months January to December 2020. For each story you will need to parse the HTML to extract the following information:
   i) The *title* of the news story.
   ii) The short *text snippet* for the story which represents the start of the complete news article.
   iii) The *category label* assigned to the story.

   Note: You **do not** have to retrieve the full linked article from The Guardian, only the data on the mlg.ucd.ie website.
3. Store the parsed data that you have collected in an appropriate format.

Task 2. Binary Text Classification

1. Load the data from Task 1 and create a set of documents, one per news story. Each document should consist of the concatenation of the story's title and text snippet. Each document should also have a class label, based on the story's news category.
2. For each unique pair of categories (A,B) from the three that you selected:
   i) Apply appropriate preprocessing steps to create a numeric representation of the documents from these two categories, suitable for classification.
   ii) Train a classification model using a **binary classifier** of your choice, which can distinguish documents in category A from documents in category B.

iii) Test the predictions of the classification model using an appropriate evaluation strategy. Report and discuss the evaluation results.

Task 3. Multi-Class Text Classification

1. Using all three categories (A,B,C) that you have selected:

iv) Apply appropriate preprocessing steps to create a numeric representation of the documents for these three categories, suitable for classification.

v) Train a classification model using a **multi-class classifier** of your choice, which can distinguish documents from the categories A, B, and C.

vi) Test the predictions of the classification model using an appropriate evaluation strategy. Report and discuss the evaluation results.

Guidelines:

- The assignment should be completed individually. Any evidence of plagiarism will result in a 0 grade.
- The grade awarded will depend on the complexity of the analysis and level of detail, i.e. the need for data cleaning, data preparation etc.
- Submit your assignment via the COMP47670 Brightspace page. Your submission should be in the form of a single ZIP file containing the notebook (i.e. IPYNB file) and your data as stored in Task 1.
- Hard deadline: Submit by end of **23rd April 2021**
    - 1-5 days late: 1 grade point deduction, e.g. B to B-
    - 6-10 days late: 2 grade point deduction, e.g. B to C+
    - Assignments will not be accepted after 10 working days without Extenuating Circumstances formally approved by UCD.