Data Science for Business

Assignment 2: Text Scraping & Classification

Marking Guidelines (100%):

Data collection: 30%

Marks awarded for:

- Correctly scraping and storing the title and body text of **all** news articles, and filtering irrelevant content.
- Correctly retrieving and storing the category labels for all news articles.

Text classification: 50%

Marks awarded for:

- Loading the data into an appropriate representation for analysis.
- Constructing a document-term matrix, using appropriate text preprocessing and term weighting steps.
- Building **two or more multi-class** classification models on the data, with **two or more** different classification algorithms and tuning them appropriately to achieve the highest performance as possible for each selected classifier.
- Running an appropriate experimental evaluation to compare the performance of the two classification models, using one or more performance criteria.
- Presenting, visualizing, and discussing the results of the experimental evaluation.

Code quality & explanation text: 20%

Marks awarded for:

- Use of Markdown cells to explain each step of the process, from data collection to classifier evaluation.
- Readability and clarity of the code. This includes use of comments, though they should not be used excessively.
- Complexity of the code, and appropriate use of Python packages. Note that code performance/speed will not be taken into account.

Note: For this assignment, **only** these third-party packages can be used: NumPy, Pandas, Scikit-learn, NLTK, SciPy, Requests, BeautifulSoup, Matplotlib, Seaborn.

Penalties for Late Submission:

- 1-5 days late: 20% deduction from overall mark
- 6-10 days late: 40% deduction from overall mark
- Assignments will not be accepted after 10 days without an extenuating circumstances form and/or a medical certificate.