COMP41680 Assignment 2:
Text Scraping and Classification

Deadline: 5pm, Friday 4th May 2018

Overview:

The objective of this assignment is to scrape a corpus of news articles from a set of web pages, pre-process the corpus, and evaluate the performance of automated classification of these articles in a supervised learning context.

The assignment should be implemented as a single IPython Notebook (not a script). Your code should be clearly documented, using comments and Markdown cells to explain your code and results.

Part 1. Data Collection

The goal here is to collect a labelled news corpus. Tasks to be completed:

1. Identify the URLs and category labels for **all** news articles listed on the website: http://mlg.ucd.ie/modules/COMP41680/archive/index.html
2. Retrieve all web pages corresponding to these article URLs. From the web pages, extract the main body text containing the content of each news article. Save the body of each article as plain text.
3. Save the category labels for all articles in a separate file.

Part 2. Text Classification

The goal here is to analyse the corpus of documents from Part 1 in a text classification context. Tasks to be completed:

1. From the files created in Part 1, load the set of raw documents into your notebook. Ensure that each document has a class label, based on the original category label that you identified.
2. From the raw documents, create a document-term matrix, using appropriate text pre-processing and term weighting steps.
3. Build two multi-class classification models using two different classifiers of your choice.
4. Compare the predictions of the two classification models using an appropriate evaluation strategy. Report and discuss the evaluation results in your notebook.

Guidelines:

- For the assignment, **only** these third-party packages can be used: NumPy, Pandas, Scikit-learn, NLTK, SciPy, Requests, BeautifulSoup, Matplotlib, Seaborn.
- Submit your assignment via the COMP41680 Moodle page. Your submission should clearly state your full name and student ID number. Include your full dataset with your submission. Your submission should be in the form of a single ZIP file containing the IPython notebook and the data.

- This assignment should be completed individually. Any evidence of plagiarism will result in a 0 grade.
- Hard deadline: Submit before 5pm on Friday 4th May 2018
    - 1-5 days late: 10% deduction from overall mark
    - 6-10 days late: 20% deduction from overall mark
    - Assignments will not be accepted after 10 days without an extenuating circumstances form and/or a medical certificate.