COMP41680 Assignment 2: Text Scraping & Clustering

Deadline: Friday 5th May 2017

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 then to apply unsupervised clustering algorithms to explore and summarise the contents of the corpus.

Part 1. Text Data Scraping (40% of marks)

This part of the assignment should be implemented as a Python script, which includes comments to explain your work.

Tasks to be completed in your script:

1. Identify the URLs for all news articles listed on the website: http://mlg.ucd.ie/modules/COMP41680/news/index.html
2. Retrieve all web pages corresponding to these article URLs.
3. 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.

Part 2. Corpus Exploration (60% of marks)

This part of the assignment should be implemented as an IPython Notebook. Include Markdown cells in your notebook to explain your work.

Tasks to be completed in your notebook:

1. Load the text corpus generated in Part 1. Apply any appropriate pre-processing steps and construct a document-term matrix representation of the corpus.
2. Summarise the overall corpus by identifying the most characteristic terms and phrases in the corpus.
3. Apply two alternative clustering algorithms of your choice to the document-term matrix to produce clusters of related documents. This might require applying each algorithm several times with different parameter values.
4. For each clustering generated in Step 3, summarise the contents of the clusters. Based on your summary, suggest a topic/theme for each cluster.

Note: For the assignment you can use any of the following packages: NumPy, Pandas, Scikit-learn, BeautifulSoup, Requests, NLTK, SciPy, Matplotlib, Seaborn.

Guidelines:

- Submit your assignment via the COMP41680 CS Moodle page. Include your full name and student ID number with your submission.
- Your submission should be in the form of a single ZIP file containing:
    1. Implementation of Part 1 described above, as a Python script
    2. The corpus plain text data files created in Part 1 above.
    3. Implementation of Part 2 described above, as an IPython Notebook.

- The assignment should be completed individually. Any evidence of plagiarism will result in a 0 grade.
- Hard deadline: Submit before 5pm on Friday 5th May 2017
    - 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.