Comp 479/6791 # Project 4 **Fall 2023**

Objectives: Experiment with web crawling, scrape and index a set of web documents. Cluster the documents using k-means. Use the AFINN sentiment analysis script to assign a sentiment score to each cluster.

Due Date: 5 December 2022

Description and marking:

- starting from page `https://www.concordia.ca`, crawl for links within the `concordia.ca` domain **(you may use crawling tools such as spidy (`https://github.com/rivermont/spidy`), but avoid tools that scrape the pages directly. You may find inspiration on `https://github.com/BruceDone/awesome-crawler`)**. Describe and attribute any tools used in your Report (3pt, Attrib 5)
- make sure you obey the standard for robot exclusion (`https://www.robotstxt.org` covers robots.txt and meta tags) (-1pt, if not implemented)
- your crawler must accept as part of its input an upper bound on the total number of files to be downloaded. In developing, testing, and debugging, this number should be kept as small as possible. Develop your own closed test set of HTML files for testing and debugging. (-1pt, if parameter not implemented)
- extract the text from the web pages, consider using **BeautifulSoup (`https://www.crummy.com/software/BeautifulSoup/`)** (1pts, Attrib 5)
- use scikit learn to cluster the resulting document collection (see `https://scikit-learn.org/stable/auto_examples/text/plot_document_clustering.html#sphx-glr-auto-examples-text-plot-document-clustering-py`) (3pt, Attrib 5)
- run two different clustering runs, one for k=3 and one for k=6. Try to assess the clusters and see what 'name' you would give them (2pt, Attrib 6)
- use the AFINN sentiment lexicon (`http://corpustext.com/reference/sentiment_afinn.html`) and the afin 0.1 script (`https://pypi.org/project/afinn/`) for sentiment analysis and determine sentiment values for the clusters. You have to determine, describe, and motivate your own formula to derive cluster sentiment scores (3pts, Attrib 5)
- report on the different behaviour of the two clusterings, the usefulness of the sentiment values, and your experience with crawling and scraping of web pages. Limit your Report to 5 pages (2pts, Attrib 6)
- compile a demo file, where you present walkthroughs for the markers highlighting good and bad example clusters (1pt, Attrib 6)

Deliverables:

- 'code' for the markers to rerun. Give the sequence of calls to the different packages with all parameters in a separate file called README. Make sure that urls to the package version you use are part of your in-line comments
- two files, one each for 3 clusters and 6 clusters. Print out 20 index terms for each cluster that you find most informative
- one .pdf file called *Report*, including your design choices, findings, and anything interesting that the marker should pay attention to
- one demo file called *Demo* that showcases your efforts and insights

Note: you may disable crawling a branch of the Concordia html tree, if necessary, but be careful that you don't omit links to relevant pages