**Comp 479/6791** # Project 4 **Fall 2020**

---

**Objectives:** Experiment with web crawling, scrape and index a set of web documents, use the sentiment dictionary aFinn to associate sentiment values to the index, make document ranking reflect sentiment.

**Due Date:** 11 December 2020

**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 Websphinx but you may also find other tools, such as NYUcrawl). Describe and attribute any tools used in your Report (2pt, Attrib 5)
- make sure you obey the standard for robot exclusion (`https://www.robotstxt.org`) (-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 Boilerpipe (1pt, Attrib 5)
- create an inverted index, encoding `tf` with the docIDs in the postings lists and `df` with the vocabulary terms in the index dictionary. You should order the postings lists by `tf`. You may limit your postings lists to the 50 pages with the highest `tf` (1pt bonus for implementing a better "goodness" function). The final index should index as many documents as possible (2pt, Attrib 5)
- formulate two different queries for each of the information needs listed below (2pts, Attrib 5)
- for each of your 4 queries, retrieve and rank documents and return the top 15 results for both of the rankings, BM25 and simple tf/idf weighting (2pts, Attrib 5)
  For the bonus point to be given, you must also retrieve, rank, and return the top 15 results for your personal ranking. Make sure to describe it clearly in your Report.
- report on the different behaviour of the ranking schemes, on any issues with the tf-ranked postings lists, the top-15 return functionality, and your experience with crawling and scraping of web pages. Limit your Report to 5 pages (2pts, Attrib 6)

**Information needs:**

1. which researchers at Concordia worked on COVID 19-related research?
2. which departments at Concordia have research in environmental issues, sustainability, energy and water conservation?

**Challenge queries:** I will post challenge queries on 9.12.2020 that you have to run and include your top-15 returns (1pt, Attrib 5)

**Deliverables:**

- code for the markers to rerun
- the inverted index
- one *Returns* file with the (clearly annotated) top 15 returns for all your queries and for the Challenge queries
- one .pdf file of no more than 5 pages called *Report*, including your findings

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