Comp 479/6791 # Project 3 Final Project **Fall 2019**

Objectives: Web crawling, scrape and index a set of web documents, rank documents, answer extraction.

Due Date: This project has to be demonstrated December 2, 3, 4. The complete project has to be submitted by December 2, 2019. This project is an individual and a team project, team size is two.

Individual Work Description:

1. starting from page `https://www.concordia.ca/research.html`, crawl for links (you may use crawling tools such as Websphinx but you may also find other tools, such as NYUcrawl). To extract the text from web pages, consider Boilerpipe. Describe and attribute any tools used. Make sure you obey the standard for robot exclusion. 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. The final index (ConcordiaAI) should cover as many documents as possible. (5 pts, Attrib 5)
2. using the index you compiled, design and test queries to try to satisfy the following information needs:

   (a) Which departments have AI research?

   (b) Which researchers are working on AI research?

   (c) What AI research is being conducted at Concordia?

   You should compare and report on at least 3 queries for each information need. While you can test as many as you'd like, pick the three most interesting ones for your report. (3 pts, Attrib 5, 6)
3. test and compare two ranking schemes for your queries: tf-idf and BMI as implemented in Project 2. Report on the usefulness of your results @10, @50, @100 (2 pts, Attrib 5, 6)
4. for more useful results, create an index (AIindex) in the same way as above starting from root `https://aitopics.org/search`. Restrict the vocabulary of AIindex to AI research specific terms. Compute df for all final terms in AIindex and report your vocabulary with df values. (1 pts, Attrib 5, 6)
5. Use the df weights of AIindex in your ranking for your queries. Compile any any additional resources (list of department names, for instance) that would help you to turn your ranked retrieval into answers to the information needs above. Try to be as comprehensive as possible. Report your findings. (3pts, Attrib 5, 6)

Final report: no points will be given, if the respective question is not addressed in the report. The report has to be carefully written, it should address trouble encountered, issues resolved, and findings.

Team Work: In teams of two, run each other's queries and compare results. Do this throughout the development of your individual systems and compare your approaches (without copying code). Report your insights in your individual reports. (1pt for graduate students, 3pts for undergraduate students, Attrib 6)

Deliverable in Moodle:

- code
- the indeces
- one file each for the nine queries and their top 100 returns
- one file each with your final answer to 2(a)-(c) (answers, not documents!)
- the final report, documenting the design of the code and report on the experiments, the results, and answer the questions specified above. Individual contributions of each team member have to be specified at the end of the report. Make sure it is clear from the report what work was done manually and what was done automatically.