COMP 479/6791

Lab Assignment 3

Fall 2023

Objectives: Web crawling

Getting started: Study Chapter 20 in the forthcoming edition of Information Retrieval (https://nlp.stanford.edu/IR-book/pdf/20crawl.pdf)

1. investigate 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)

2. develop your own crawler, that takes as a parameter a starting web page and crawls for links. Set as a second parameter a list of domains, to filter links. When that parameter is not 'nil', it should only crawl to links within that list of permissible domains

3. make sure you obey the standard for robot exclusion (https://www.robotstxt.org covers robots.txt and meta tags)

4. your crawler must accept a third parameter, which sets 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

5. extract the text from the web pages, consider using **BeautifulSoup** (https://www.crummy.com/software/BeautifulSoup/)