Northeastern University
CS6200 - Fall 2018

Assignment 1:

This is an individual assignment. If you get help from others you must write their names down on your submission and explain how they helped you. If you use external resources you must mention them explicitly. You may use third party libraries but you need to cite them, too.

Date posted: Friday September 21, 2018

Date Due: Monday October 01, 2018 by 11:59pm

Goal: Implementing your own web crawler. Performing focused crawling

Task 1: Crawling the documents:

A. Start with the following seed URL from Wikipedia:
   https://en.wikipedia.org/wiki/Time_zone
B. Your crawler has to respect the politeness policy by using a delay of at least one second between your HTTP requests.
C. Perform crawling as follows:
   Follow the links with the prefix **https://en.wikipedia.org/wiki** that lead to articles only (avoid administrative links containing ":"). Also, make sure to properly treat URLs with # which basically denotes a section within the (same) page and not a different one. Non-English articles, external links, main Wikipedia page, navigations and marginal/side links must not be followed. Only follow links that appear within the article itself (content block). You may ignore formulas, tables, images, and non-textual media.
D. Crawl no further than depth 6. The seed page is the first URL in your frontier and thus counts for depth 1. You should handle redirected pages to avoid duplicates. Compose the frontier such that your crawl stays as close to the seed URL as possible. Moreover, links encountered closer to the title within an article are deemed more important and hence the pages they link to should be acquired prior to ones occurring later in the article. You need to record the depth of each URL in the list.
E. Stop once you've crawled 1000 unique URLs in a crawl. Keep a list of these URLs in a text file. Keep the downloaded documents (raw html, in text format) with their respective URL for future tasks (transformation and indexing). Do not upload downloaded documents to Blackboard in this submission.
F. Repeat steps A-E using the following seeds (perform **one** run per URL)
   https://en.wikipedia.org/wiki/Electric_car
   https://en.wikipedia.org/wiki/Carbon_footprint

Task2: Merging Results:

Consolidate the lists obtained from all three crawls into one list of unique URLs. The list should follow the same order of link importance explained in part D is Task1. Explain your approach.

Task 3: Focused Crawling:

Your crawler should be able to consume two arguments: a URL and a keyword or a set of keywords to be matched against **anchor text** or text within a **URL**. Starting with the seed **https://en.wikipedia.org/wiki/Carbon_footprint**, crawl to depth 6 at most, using the keyword "green" (**not** case-sensitive). You should return at most 1000 URLS. Describe how you handled keyword variations.

What to hand in?

1- Your source code for solving this assignment.
2- A readme text file explaining in detail how to setup, compile, and run your program. Report maximum depth reached in all tasks
3- FIVE text files each containing at most 1000 URLs (three files for Task 1, one for Task 2, and one file for Task 3).
4- A text file with your explanation for Task 2.
5- A text file with your explanation for Task 3.
6- Compress your all files into one folder and name your folder using the following format: *FName_LName_HW1*