Northeastern University
CS6200 - Fall 2016

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: Sunday September 25, 2016

Date Due: Wednesday October 5, 2016

Goal: Implementing your own web crawler. Performing focused crawling

Description:

Task 1: Crawling the documents:

A. Start with the following seed URL: https://en.wikipedia.org/wiki/Sustainable_energy; a Wikipedia article about green energy.
B. Your crawler has to respect the politeness policy by using a delay of at least one second between your HTTP requests.
C. Your crawler must assume that pages in a shallower depth are more important than deeper ones, also, within each individual webpage, hyperlinks appearing earlier on in the webpage should be crawled first.
D. 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 and external links must not be followed.
E. Crawl to depth 5. The seed page is the first URL in your frontier and thus counts for depth 1.
F. Stop once you've crawled 1000 unique URLs. Keep a list of these URLs in a text file. Also, keep the downloaded documents (raw html, in text format) with their respective URL for future tasks (transformation and indexing)

Task 2: Focused Crawling:

Your crawler should be able to consume two arguments: a URL and a keyword to be matched against anchor text or text within a URL. Starting with the same seed in Task 1, crawl to depth 5 at most, using the keyword "solar". You should return at most 1000 URLS for each of the following:

A. Breadth first crawling
B. Depth first crawling
C. In a few sentences compare and explain the approaches above. Briefly compare the results obtained in A & B in this task in terms of the total number of URLs crawled, and the top 5 URLs (topical content).

Task 3: Combined Results

Repeat Task 1 this time using the seed https://en.wikipedia.org/wiki/Solar_power. Assume you were asked to combine the results of these two independent runs into one file with 1000 links at most. Describe briefly (in a few steps) how you would approach the merging process with minimal loss of information about the deemed importance of the hyperlinks (reflected by the order in which they are crawled – as described in Task 1).

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.
3- FOUR text files each containing 1000 URLs at most (one file for Task 1-E, two files for Task 2- A & B, and one for Task 3).
4- A text file with your explanation for Task 2-C.
5- A text file with your explanation for Task 3.