Northeastern University
CS6200 - Spring 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: Saturday January 27, 2018

Date Due: Monday February 05, 2018 by 11:59pm

Goal: Implementing your own web crawler. Performing focused crawling

Description:

Task 1: Crawling the documents:

A. Start with the following seed URL from Wikipedia: https://en.wikipedia.org/wiki/Solar_eclipse.
B. Your crawler has to respect the politeness policy by using a delay of at least one second between your HTTP requests.
C. Perform two crawling rounds:
    a. Following a breadth-first order (BFS)
    b. Following a depth-first order (DFS)
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, 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, images, and non-textual media.
E. 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.
F. Stop once you've crawled 1000 unique URLs in a crawl. Keep a list of these URLs in a text file. You should hence generate two separate files, one for each crawl (one for BFS and one for DFS). Keep the downloaded documents (raw html, in text format) with their respective URL from the BFS crawl only for future tasks (transformation and indexing). Do not upload downloaded documents to Blackboard in this submission.
G. Briefly compare the results obtained from the two crawls in terms of URL overlap, perceived quality and efficiency aspect, coverage of the crawl topic (i.e., solar eclipse).

Task 2: 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 same seed in Task 1, perform a BFS crawl to depth 6 at most, using the keywords “lunar” and “moon”. “Moon_landing”, “Moonlit”, “Lunar”, “honeymoons”, “LUNAR”, etc. should be considered as valid variations of these keywords.
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- THREE text files each containing at most 1000 URLs (two files for Task 1 and one file for Task 2).
4- A text file with your explanation for Task 1-G.
5- A text file with your explanation for Task 2.
6- Compress your all files into one folder and name your folder using the following format: *FName_LName_HW1*