Northeastern University
CS6200 - Fall 2017

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 24, 2017

Date Due: Wednesday October 04, 2017

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/Tropical_cyclone.
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 the earlier the hyperlink appears in a page, the more important it is (and hence must be crawled first) and that shallower depths are more important than deeper pages.
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. You may ignore formulas, images, and non-textual media.
E. Crawl to depth 6. 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) – do not upload downloaded documents. You should handle redirected pages to avoid duplicates.

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 6 at most, using the keyword "rain". "Falling_rain", "rain_fall", "rainband", "Rain", "rains", etc. should be considered as valid variations, whereas "grains", "Ukraine", etc. should not be considered as valid matches.
You should return at most 1000 URLS using the same crawler setup in the previous question. 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 both tasks
3- TWO text files each containing at most 1000 URLs (one file for Task 1-E and one file for Task 2).
4- A text file with your explanation for Task 2.
5- Compress your all files into one folder and name your folder using the following format:
FName_LName_TuTh_HW1 (for Tuesday & Thursday section)
FName_LName_Tu_HW1 (for Tuesday evening section)