CS 4364/5364: INFORMATION RETRIEVAL AND INFORMATION VISUALIZATION
DEPARTMENT OF COMPUTER SCIENCE
THE UNIVERSITY OF TEXAS AT EL PASO
FALL 2021

Assignment# 1: Crawling and Indexing

Due: September 21, 2021, in Blackboard

1. Write a program that will crawl a given link (seed URL) to generate the link structure of the website (starting from the seed URL).
    a. For each URL, scrape the content of the page and remove stop words, hyphens, punctuation, etc.
        i. Use the stop words from the given stopwords file provided with the assignment.

    b. Using your program, write each URL in a new line in a text file, with the content of the page listed as <word::frequency>. For example,
        <URL_1> → word1:: f1 word2:: f2
        indicates f1 is the frequency of word1 in URL_1, f2 is the frequency of word2 in URL_1

    c. If there is more than one webpage in a URL, use indentation (or, four spaces) for listing those subsequent URLs.

    Thus, if URL_1 has two URLs listed in that page, then the text file will include:
    < URL_1>→ word1:: f1 word2:: f2
    <URL_a> → word3::f3 word4::f4 word8::f8
    <URL_b> → word1::f1(in URL_b) word4::f4 (in URL_b)

2. Run your program with this URL: http://www.cs.utep.edu/makbar/A3/A2.html

3. The program should respect the rules stated in robots.txt.

Sample output:

- A sample output for https://zenhabits.net/ "**could**" partially look like this:

https://zenhabits.net/→ many::1 power::16....
/ → many::1 power::16 ....
https://zenhabits.net/behind/ → …
http://leobabauta.com → …
/archives/ → …….
https://feeds.feedburner.com/zenhabits → …….
/subscribe/ →…….

https://twitter.com/zen_habits →…….
https://www.facebook.com/officialzenhabits/ →…….
/about/ →…….
/uncopyright/ →…….

- There could be relative URLs (e.g., /, /about/).
- There could be loops in the URLs.

Deliverables: 2 files

1. The source code of the program. The name of the file should be lastname_firstname.xyz (replace xyz with proper extension).
2. The inverted index, containing the output of your program, in a txt file. The filename should be lastname_firstname.txt

Resources:

- If you are using Python, you can use Scrapy or Beautiful Soup. You can use Heritrix in java.