Project 2: Query engine implementation

Can also find PDF version in this folder.

Used software

- Visual Studio Code
    - Version: 1.59.1
- macOS Monterey
    - Version: 12.3

Installation

1. Uncompressed archive(zip) named "Proj2", then you will get a file "Proj2"
2. Put this file under the path of your compiler(python3)
3. Open this file with Visual Studio Code(IDE)
4. Check uninstalled packages. Following is the list of packages used in this project:

    ```
    import requests
    from bs4 import BeautifulSoup
    from urllib import parse
    import nltk
    import pandas

    import math
    import heapq
    import os
    import sys
    import re
    import time
    import random
    ```

    You can use following shell command to install this package with pip3:

    ```
    pip3 install requests
    pip3 install bs4
    pip3 install urllib
    pip3 install nltk
    pip3 install pandas
    ```

Compilation

Python 3.9.7 64-bit

Execution instructions

This is a specialized web crawler, designed to crawler only on the data in http://freemanmoore.net. It will output all crawled data(information of pages) to `all_pages.xlsx` file, a term-document frequency matrix to `frequencyMatrix.xlsx`, a postings lists to `postingslists.xlsx`, and **the** list of special crawled links(going out links, non-text files links, broken links, duplicated links) to a 'otherList.txt' file.

After that, you will be able to enter multiple queries, consisting of one or more query words separately by space. Entering "stop" will cause this program to stop.

Can follow the following instructions to execute this program:

1. Simply run 'Main.py'
2. Waiting for the console print "Query?"
3. Then you can enter a query.
4. After the results are displayed, you can continue enter the next query, or enter "stop"