Web Search Engine

Final Project for Course CS582 - Information Retrieval

Submitted by : Sheetal S Prasad (sprasa22)

UIN: 677918305

Files and Folder:

- IRHW1_Project.zip: Zip folder containing the python files , data saved as pickle files, images folder, dictionary text file, and readme file.
- README.md(this file): Information about environment setup to run the Python file and information about the libraries used.
- spider.py: Python file that crawls the web and scrapes the web contents and processes textual data.
- pr_tfidf.py: Python file where for each token is assigned a tfidf weight and inlinks for each URL is calculated.
- QDPR.py - Python file that calculated the query dependent page rank value for each file.
- Final_calc.py - Python file that is called by the User interface when a query is obtained and calculated cosine ranking and query dependent page rank.
- preprocessing.py - Python files containing the methods used for cleaning the web content as well as the query.
- pickle_functions.py - Python file wit methods to open pickles and save pickles.
- spellchecker.py- Python file containing the code to correct misspelled query.
- UIC_app.py- User interface for the web search engine project created using Tk GUI from python.
- dictionary.txt : A text file containing a large collection of words used in English language used to calculate the probability in the spellchecker.
- Images directory - Includes all the images used in the project report.
- Pickle files - 7 Pickle files are included in the the IR_Project.zip folder , these save data collected from the web pages as well as results of the calculations performed such as the inverted index.

Code and how to run it:

Method 1: semi-dynamic

- The source code for this assignment has been written in Python 3.7, and Windows operating system.
- The file that needs to be executed to see the results of the search engine is : UIC_app.py.
- Extract the zip file on to your C drive and execute the source code from the terminal.
- To execute the python file run the code below:

```
python UIC_app.py
```

- Make sure the source code and the dataset directory are both within the same directory, as the program is written in such a way that it fetches a pickle files, text files, images and other python programs in the same directory.

- The user interface includes an image, the console needs to be set up to work with the image.

Method2 : Dynamic

If you choose to run the web search engine in real time follow the step below:

- Run the python file name spider.py with the search limit equal to the number of web pages that need to be crawled, by default the value in the python file is 3000. this will create 4 pickle files.
- Next run pr_tfidf to create 2 more pickle files containing inlinks and tfidf scores.
- Run the QDPR to create the final pickle used in query-dependent PageRank.
- Finally run the user interface UIC_app.py.

Setup and Prerequisites:

The source code uses the below modules and suite of libraries for processing the dataset .

```
import os, nltk, string, re, pickle, requests
import math
import tkinter as tk
from PIL import Image,ImageTk
from bs4 import BeautifulSoup
from urllib.parse import urlparse,urljoin
from urllib.request import urlopen
from nltk.tokenize import RegexpTokenizer
from nltk.stem import PorterStemmer
```

- If any of the above is not available install them using :

  ```
  pip install os
  ```