Project has 2 different Pyton files:

1-SearchEngineWithoutTfIdf.py

This file is able to load all documents given, and make search on them. The missing part is; search results is not in the order of their Tf-Idf values.

2-SearchEngine.py

This file still needs some work on the part of Tf-Idf calculation. As it requires calculation throughout the document, some performance issues needs to be solved. But it is working on small sized documents (e.g. Sample.txt).

Parts of the Project:

Main Function:
In this part, required functions are called. To load a file, 'LoadDocument' function, for searching in loaded files 'Search' function is called.

Load File Step:

- Parsing file with *BeatifulSoap* html parser
- Eliminating stop words and punctuation with *EliminateStopWords* function
- Splitting and storing data in *Dictionary* data structure
  Key: string data type, name of the word
  Value: list data type, tfidf, url, title
- If a word is already inserted into wordDict(Dictionary), only Value part of the object is updated with inserting new tf-idf, url, and title values
- If a word is a new word, word is inserted into the wordDict(Dictionary).While checking if the word is in wordDict, a new variable in *Set* data type is used, to gain the advantage of constant time reach to searched word.

Tf-Idf Calculation Step:
First TF value is calculated as below:

TF = Number of times word occurs in the text / Total number of words in text

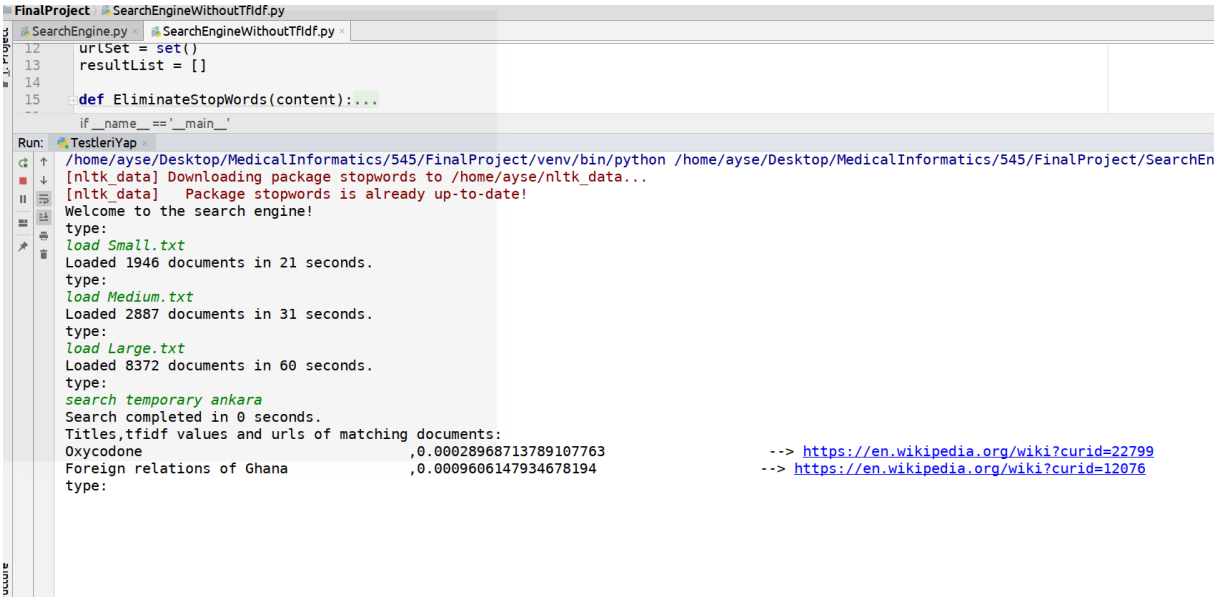
Secondly IDF values is calculates as below:

IDF = Total number of documents / Number of documents with word in it

Lastly, TF and IDF values are multiplied to get the Tf-Idf value.

Search Step:
In this part I took search keys in a variable called searchKeys, which is a *List* data type
I used the *intersection property of Set* data structure, to get the list of documents which has *ALL* the searched words.
I returned the search results in a *list* typed variable to the main function, to be able to print the results.

Outputs:


FinalProject
SearchEngineWithoutTfIdf.py
SearchEngine.py
SearchEngineWithoutTfIdf.py
urlSet = set()
resultList = []
def EliminateStopWords(content):...
if __name__=='__main__'
Run:
TestleriYap
/home/ayse/Desktop/MedicalInformatics/545/FinalProject/venv/bin/python /home/ayse/Desktop/MedicalInformatics/545/FinalProject/SearchEn
[nltk_data] Downloading package stopwords to /home/ayse/nltk_data...
[nltk_data] Package stopwords is already up-to-date!
Welcome to the search engine!
type:
load Small.txt
Loaded 1946 documents in 21 seconds.
type:
load Medium.txt
Loaded 2887 documents in 31 seconds.
type:
load Large.txt
Loaded 8372 documents in 60 seconds.
type:
search temporary ankara
Search completed in 0 seconds.
Titles,tfidf values and urls of matching documents:
Oxycodone ,0.00028968713789107763 --> https://en.wikipedia.org/wiki?curid=22799
Foreign relations of Ghana ,0.0009606147934678194 --> https://en.wikipedia.org/wiki?curid=12076
type: