UNIVERSITÄT DES SAARLANDES

Department of Language Science and Technology

Muhammad Umer Butt

Introduction to Python Programming

Search Engine

Project Report

27th March - 2023

Introduction:

In this project, I have created a complete search engine using tfidf from scratch. The code, data and tf,idf files are also uploaded on github repository [1] (it will be a private repository till the deadline, and then i will make it public, to avoid any cheating. Please let me know if its not allowed and I will remove it).

I followed the instructions from the project instructions file and was able to achieve the exact results given in the file. Below I discuss some of aspects of the project.

The tf and idf output files are located in the project home directory. I wont be able to submit the nyt199501.tf file because its too big to be uploaded so i will upload its zip file. Please extract it to use it. I wont able to upload the full data file (nyt199501.xml) because of the same reason.

Environment:

Since the search engine was created from scratch so there arent many packages requirements. The only requirement is of “xml” package containing “xml.etree.ElementTree” to read the data.

Code Structure:

I have divided the project code files into 3 separate class. SearchEngine, Data_Handler and TFIDF_custom_class. All the classes and functions are completely modular and has detailed documentation on each and every function for how to use that function, what arguments it takes, their data types and what data to they return. Comments are also added inside the functions to explain the working of the code and the approach that i have taken. Since I have created the tfidf class separately so it can be used with any data as long as its structure is what is expected by the class (and that structure is also explained in the code). I have also added the clear step by step instructions on how to use the tfidf class, its written at the start of the class.

SearchEngine class is the the class whose function arguments were given in the assignment instructions.
Data_Handler class handles the data for example reading the data, processing the data and query, saving or loading the the tf or idf.
TFIDF_custom_class implements the tfidf algorithm from scratch.

UML diagram of the classes is given below in fig 1.

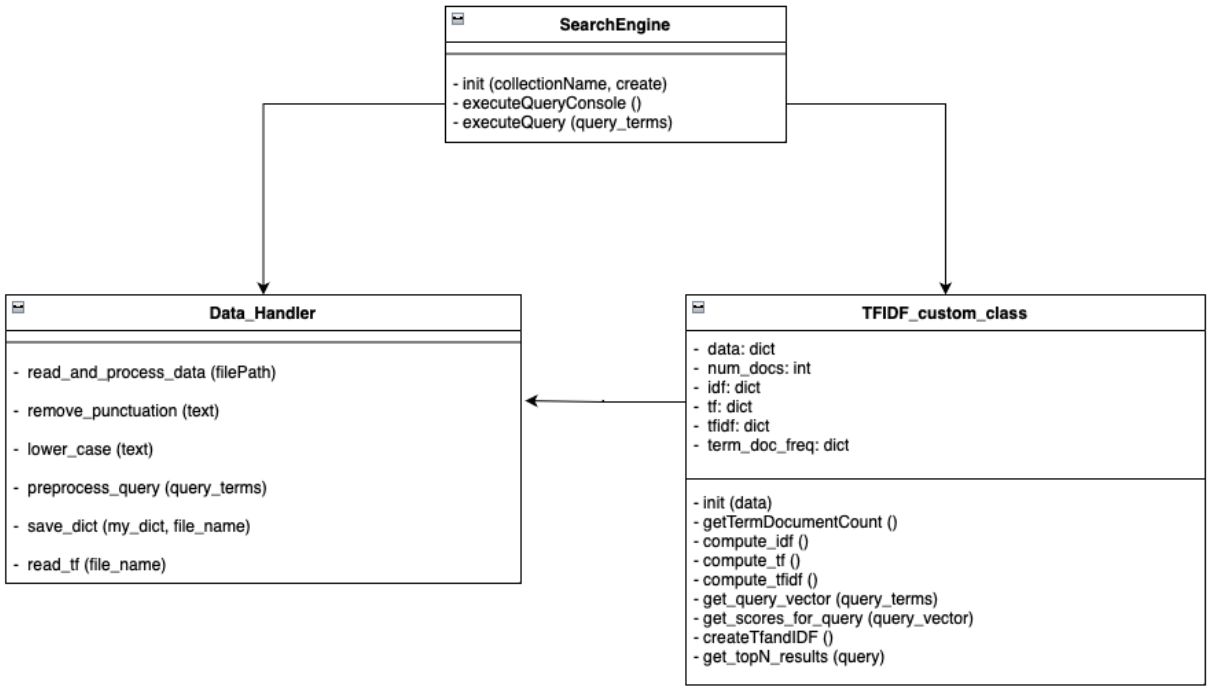

Figure 1. UML diagram of the class

Project Directory Structure:

- Project home directory
  - |
  - ——— softwareAssignment.py
  - ——— TFIDF.py
  - ——— dataHandler.py
  - ——— nyt199501.tf
  - ——— nyt199501.idf
  - ——— nytsmall.tf
  - ——— nytsmall.idf
  - ——— sample_output.png
  - ——— uml_diagram.png
  - ——— data
    - |
    - ——— [xml data files]
  - ——— stemming
    - |
    - ——— [stemming files provided by instructor]

How To Run :

```
>> python softwareAssignment.py
```

To run the program you just have to run the softwareAssignment.py and it will the start the program.

At the moment default create argument is TRUE. which means index will be created. To change this and load the tf and idf matrices from the file, change the value of the constant CREATE_INDEX in the softwareAssignment.py.

There are other constants also which might have to be tweaked if you want to run the program on your machine.
All the constants are mentioned at the top of softwareAssignment.py file (above the class).

You have to change only these constants if you want to run the program on your machine, nothing else.

Note: you will have to also change atleast the BASE_DIR constant if you want to the program on your machine.

Descriptions of the module:

I have written detailed doc strings for each class and each function of every class. I have also mentioned the basic structure above in the Introduction section. So, I am not going to rewrite the same thing here as it will prolong the report unnecessarily.

Sample Run:

I tests on multiple queries and it seems to be working fine. But to show here that its working exactly how it should be, i am attaching a picture, fig 2 below, of the output for sample queries that were given in the project instructions.pdf

```
(base) umer@Umers-MacBook-Pro-2 project % python softwareAssignment.py
Reading index from file...
Done.
Please enter query, terms separated by whitespace: philadelphia hurricane
I found the following documents:
NYT_ENG_19950101.0001    (0.15211094925838722)
NYT_ENG_19950101.0056    (0.02912567342257731)
NYT_ENG_19950101.0022    (0.028534637840005445)
NYT_ENG_19950101.0017    (0.01673266417666542)
Please enter query, terms separated by whitespace: cleveland american football conference
I found the following documents:
NYT_ENG_19950101.0064    (0.23973042239309542)
NYT_ENG_19950101.0074    (0.11484316681019026)
NYT_ENG_19950101.0103    (0.036440292369455056)
NYT_ENG_19950101.0107    (0.033323052050357106)
NYT_ENG_19950101.0045    (0.026175315954348147)
NYT_ENG_19950101.0019    (0.02500709946390535)
NYT_ENG_19950101.0116    (0.021793107603788366)
NYT_ENG_19950101.0022    (0.019889466582403995)
NYT_ENG_19950101.0095    (0.013848647710351597)
NYT_ENG_19950101.0035    (0.010800429071449616)
Please enter query, terms separated by whitespace: aTermNotInVocab
Sorry, I didn't find any documents for this term.
Please enter query, terms separated by whitespace: america
I found the following documents:
NYT_ENG_19950101.0048    (0.06637072573296174)
NYT_ENG_19950101.0005    (0.05462572110399223)
NYT_ENG_19950101.0043    (0.0500462746676232)
NYT_ENG_19950101.0068    (0.04299818023490763)
NYT_ENG_19950101.0016    (0.04166078791910162)
NYT_ENG_19950101.0019    (0.040636516837853257)
NYT_ENG_19950101.0028    (0.03886407767894348)
NYT_ENG_19950101.0070    (0.03279405236958527)
NYT_ENG_19950101.0094    (0.02481293665098882)
NYT_ENG_19950101.0032    (0.021362490855110723)
Please enter query, terms separated by whitespace:
(base) umer@Umers-MacBook-Pro-2 project %
```

Figure 2. Sample output

References:

1) https://github.com/UmerTariq1/TF-IDF