Name:Prathusha JS Naidu

Information Retrieval
Project Phase 3

1. Introduction

The objective of this assignment is to build an index for the documents in the collection. In this sense, an index is also called an inverted file: instead of looking up a file to see what terms occur in it, we're looking up a term to see what files that term occurs in.
I have used Python3 to code the program.

To execute the program, type the following:
python3 index.py <input_directory> <output_directory>
where, the input-directory contains the original files (unprocessed) and the output-directory contains all generated output files

2. Input

A file folder with all the unprocessed HTML documents.

3. Output

The output folder contains two files
a) Dictionary.txt : Sorted alphabetically with 3 fields for each term
  The word
  the number of documents that contain that word
  the location of the first record for that word in the postings file
b)Postings.txt : contains two fields:
  the document id
  the normalized weight of the word in the document

4. Preprocessing

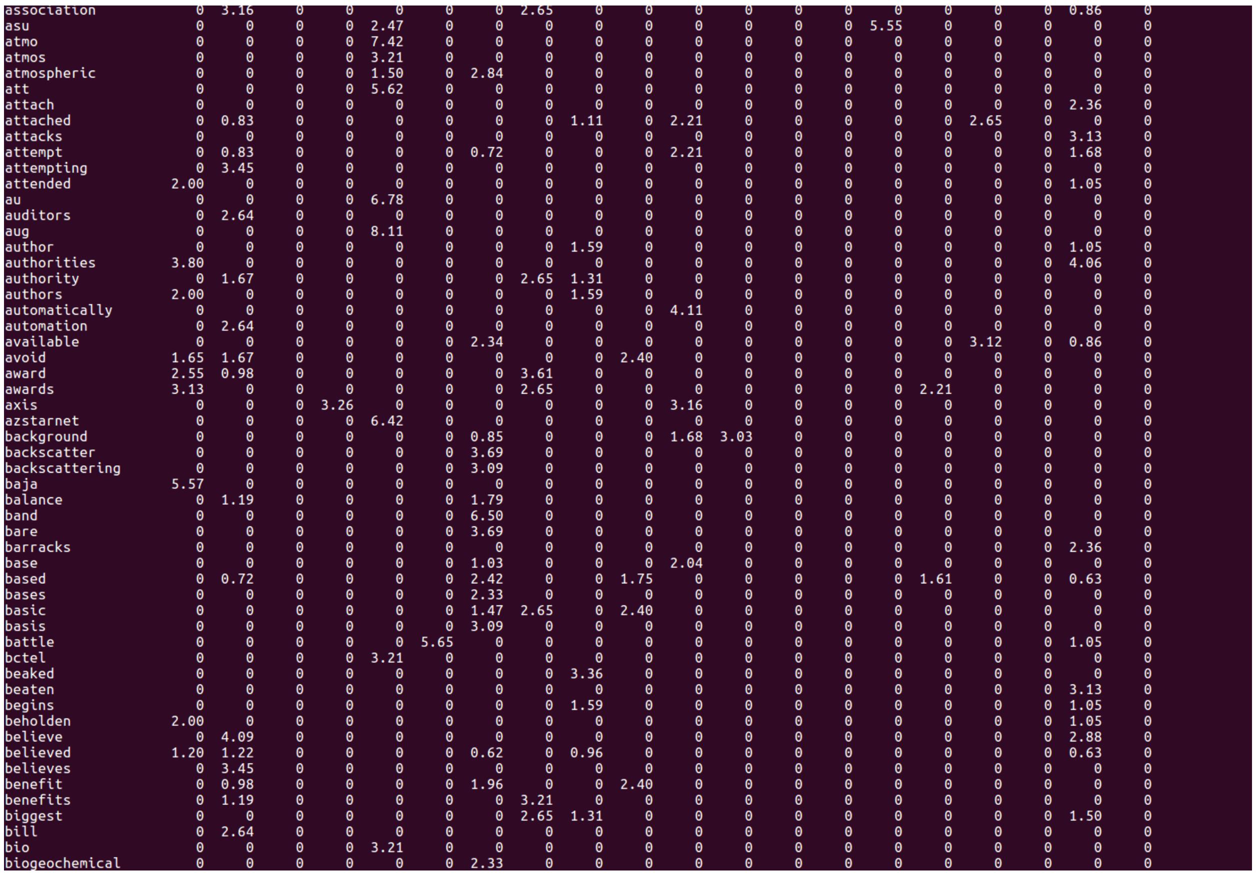
All the preprocessing steps like removal of stop words and tokenization have been used from previous stages of the Project. The following are the changes made:

a) Instead of storing the results of tokenization in temporary files as done in Phase 2, I used a nested dictionary to store the tokenized information from all the input files. Thus eliminating the need for reading and writing into files twice.
b)I have removed the part where term weights for each document are written into output files.

BM25 is used to calculate term weights in Phase 3 as well

5. Building Inverted Index

A term document matrix was first created from the nested dictionary. This gives details of the term frequency of each word in every document.
Fig 1 shows the term document matrix for 20 documents.

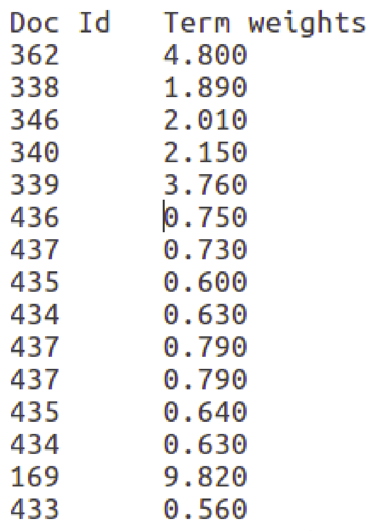


```
association         0  3.16     0     0     0     0     0  2.65     0     0     0     0     0     0     0     0     0     0  0.86     0
asu                 0     0     0     0  2.47     0     0     0     0     0     0     0     0     0  5.55     0     0     0     0     0
atmo                0     0     0     0  7.42     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0
atmos               0     0     0     0  3.21     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0
atmospheric         0     0     0     0  1.50     0  2.84     0     0     0     0     0     0     0     0     0     0     0     0     0
att                 0     0     0     0  5.62     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0
attach              0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0  2.36     0
attached            0  0.83     0     0     0     0     0     0  1.11     0  2.21     0     0     0     0     0  2.65     0     0     0
attacks             0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0  3.13     0
attempt             0  0.83     0     0     0     0  0.72     0     0     0  2.21     0     0     0     0     0     0     0  1.68     0
attempting          0  3.45     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0
attended         2.00     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0  1.05     0
au                  0     0     0     0  6.78     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0
auditors            0  2.64     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0
aug                 0     0     0     0  8.11     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0
author              0     0     0     0     0     0     0     0  1.59     0     0     0     0     0     0     0     0     0  1.05     0
authorities      3.80     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0  4.06     0
authority           0  1.67     0     0     0     0     0  2.65  1.31     0     0     0     0     0     0     0     0     0     0     0
authors          2.00     0     0     0     0     0     0     0  1.59     0     0     0     0     0     0     0     0     0     0     0
automatically       0     0     0     0     0     0     0     0     0     0  4.11     0     0     0     0     0     0     0     0     0
automation          0  2.64     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0
available           0     0     0     0     0     0  2.34     0     0     0     0     0     0     0     0     0  3.12     0  0.86     0
avoid            1.65  1.67     0     0     0     0     0     0     0  2.40     0     0     0     0     0     0     0     0     0     0
award            2.55  0.98     0     0     0     0     0  3.61     0     0     0     0     0     0     0     0     0     0     0     0
awards           3.13     0     0     0     0     0     0  2.65     0     0     0     0     0     0     0  2.21     0     0     0     0
axis                0     0     0  3.26     0     0     0     0     0     0  3.16     0     0     0     0     0     0     0     0     0
azstarnet           0     0     0     0  6.42     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0
background          0     0     0     0     0     0  0.85     0     0     0  1.68  3.03     0     0     0     0     0     0     0     0
backscatter         0     0     0     0     0     0  3.69     0     0     0     0     0     0     0     0     0     0     0     0     0
backscattering      0     0     0     0     0     0  3.09     0     0     0     0     0     0     0     0     0     0     0     0     0
baja             5.57     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0
balance             0  1.19     0     0     0     0  1.79     0     0     0     0     0     0     0     0     0     0     0     0     0
band                0     0     0     0     0     0  6.50     0     0     0     0     0     0     0     0     0     0     0     0     0
bare                0     0     0     0     0     0  3.69     0     0     0     0     0     0     0     0     0     0     0     0     0
barracks            0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0  2.36     0
base                0     0     0     0     0     0  1.03     0     0     0  2.04     0     0     0     0     0     0     0     0     0
based               0  0.72     0     0     0     0  2.42     0     0  1.75     0     0     0     0     0  1.61     0     0  0.63     0
bases               0     0     0     0     0     0  2.33     0     0     0     0     0     0     0     0     0     0     0     0     0
basic               0     0     0     0     0     0  1.47  2.65     0  2.40     0     0     0     0     0     0     0     0     0     0
basis               0     0     0     0     0     0  3.09     0     0     0     0     0     0     0     0     0     0     0     0     0
battle              0     0     0     0     0  5.65     0     0     0     0     0     0     0     0     0     0     0     0  1.05     0
bctel               0     0     0     0  3.21     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0
beaked              0     0     0     0     0     0     0     0  3.36     0     0     0     0     0     0     0     0     0     0     0
beaten              0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0  3.13     0
begins              0     0     0     0     0     0     0     0  1.59     0     0     0     0     0     0     0     0     0  1.05     0
beholden         2.00     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0  1.05     0
believe             0  4.09     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0  2.88     0
believed         1.20  1.22     0     0     0     0  0.62     0  0.96     0     0     0     0     0     0     0     0     0  0.63     0
believes            0  3.45     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0
benefit             0  0.98     0     0     0     0  1.96     0     0  2.40     0     0     0     0     0     0     0     0     0     0
benefits            0  1.19     0     0     0     0     0  3.21     0     0     0     0     0     0     0     0     0     0     0     0
biggest             0     0     0     0     0     0     0  2.65  1.31     0     0     0     0     0     0     0     0     0  1.50     0
bill                0  2.64     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0
bio                 0     0     0     0  3.21     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0
biogeochemical      0     0     0     0     0     0  2.33     0     0     0     0     0     0     0     0     0     0     0     0     0
```

Fig 1: Term Document Matrix

A postings file was then constructed by iterating through the term document matrix. Each entry in the postings file represents a non-zero entry in the term

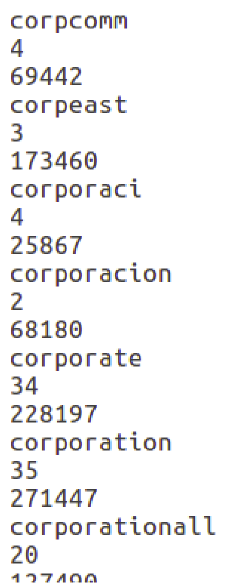
document matrix. A counter variable called postings_id was used to keep track of the number of entries a particular word has in the file
Fig 2 shows a screenshot of a part of the postings file.



```
Doc Id    Term weights
362       4.800
338       1.890
346       2.010
340       2.150
339       3.760
436       0.750
437       0.730
435       0.600
434       0.630
437       0.790
437       0.790
435       0.640
434       0.630
169       9.820
433       0.560
```

Fig 2: Screenshot of Postings file

Finally a dictionary was created using the entries in the postings file for each word and has 3 fields for every entry and is sorted alphabetically: The word , number of entries in postings file and location of first entry in the postings file. Fig 3 shows a screenshot of the part of dictionary file



```
corpcomm
4
69442
corpeast
3
173460
corporaci
4
25867
corporacion
2
68180
corporate
34
228197
corporation
35
271447
corporationall
20
```

Fig 3: Screenshot of dictionary file

6. Evaluation

a) Size:

Each of the output file sizes (Postings and Dictionary) is plotted against the size of input files for varying number. In every case , the size of the dictionary is significantly smaller than that of a postings file and the input file
Fig 4 shows the graph of sizes for output files and input files

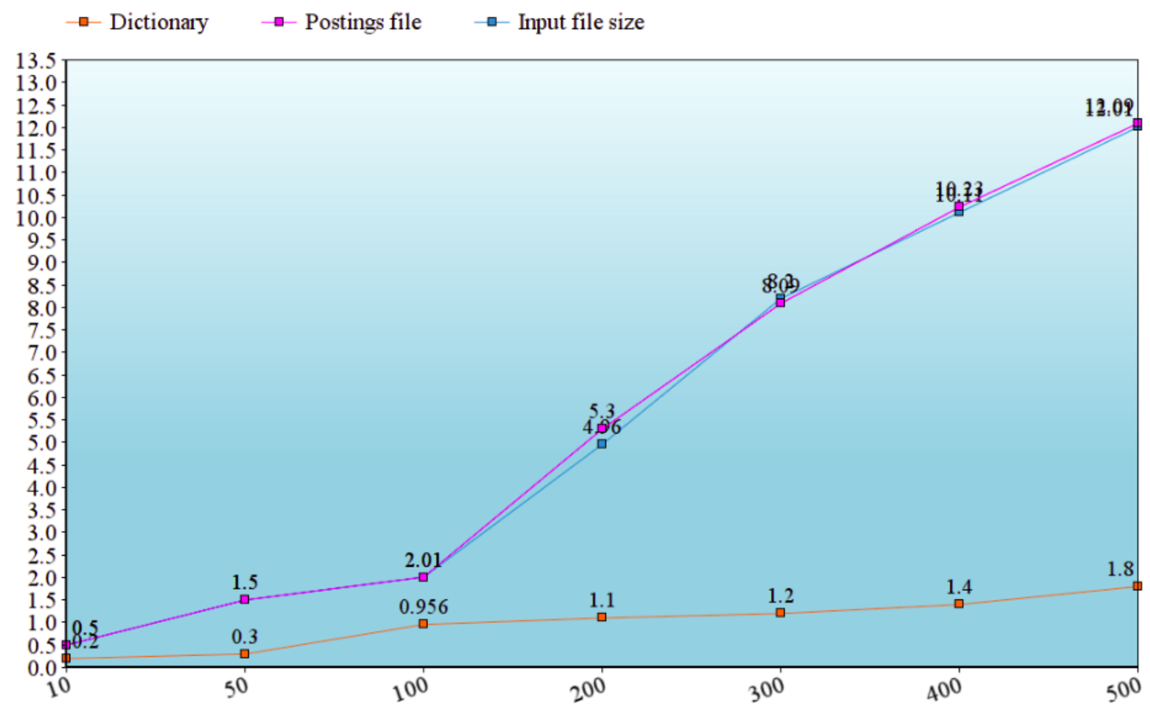


Fig 4 : graph of sizes for output files and input files

b) Time

Evaluation is also done on the time required for executing the program. Both the metrics “System time” and “Elapsed Time” have been plotted for varying number of input files. The total time required for processing all 503 files was just around 20s , which is a significant improvement from Phase 2 of the project. I believe this was due to the use of nested dictionaries and minimizing the number of loops used.
Fig 5 shows time plotted against varying number of input files

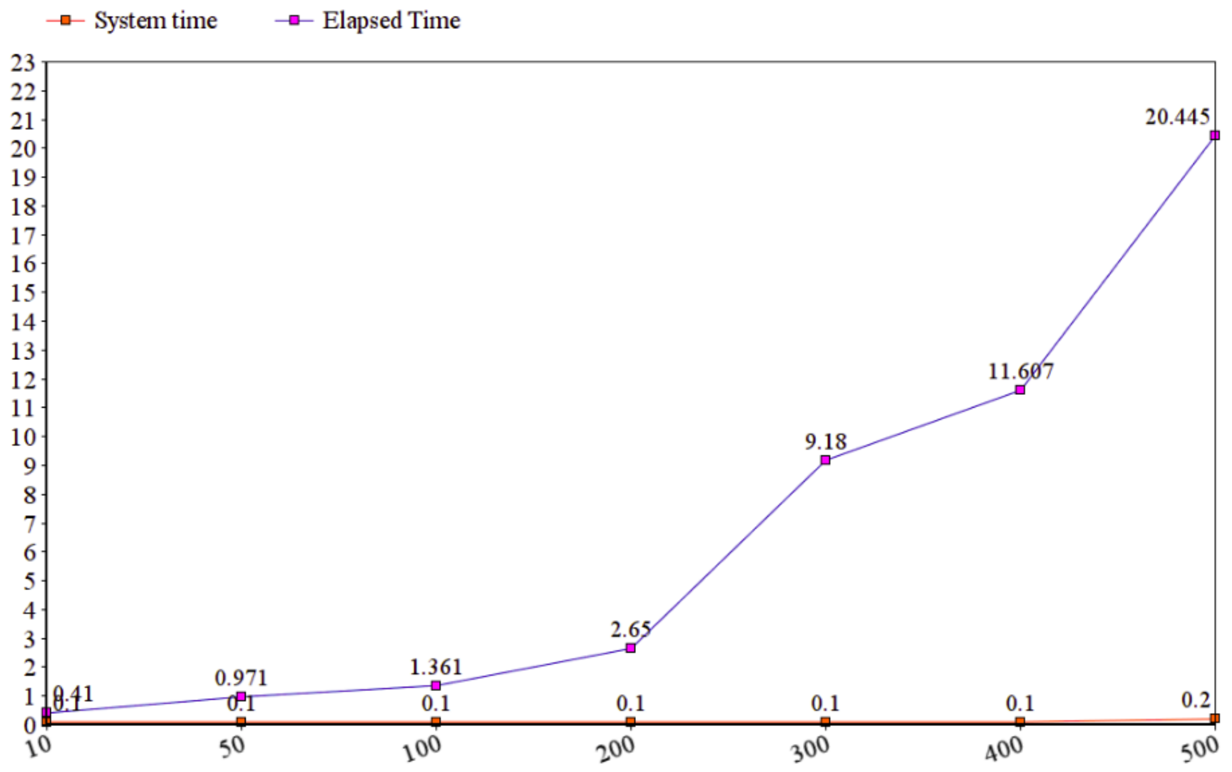


Fig 5 : Time vs Input files