Birla Institute of Technology and Science-Pilani, Hyderabad Campus

Second year First Semester
2020-2021

Information Retrieval (CS F469)

Design Document

Assignment 1

Domain Specific Information Retrieval System

Faishal Hussain siddiqui	2019H1030012H
Rojan Sudev	2019H1030008H
Arpit Roy	2019H1030118H
Suraj Abhiman Shinde	2019H1030507H

Under the guidance of
Dr. Aruna Malapati

Abstract:

This project means to construct a domain specific search engine based on a vector space model. Initially the documents to be searched are pre-processed to create terms. In the vector space model, documents are represented as vectors containing weights of these terms. At the point when a search query is given to the search engine, a vector is produced using its terms and similarity between the documents and query is found. Then the documents are returned in decreasing order of their relevance.

Architecture:

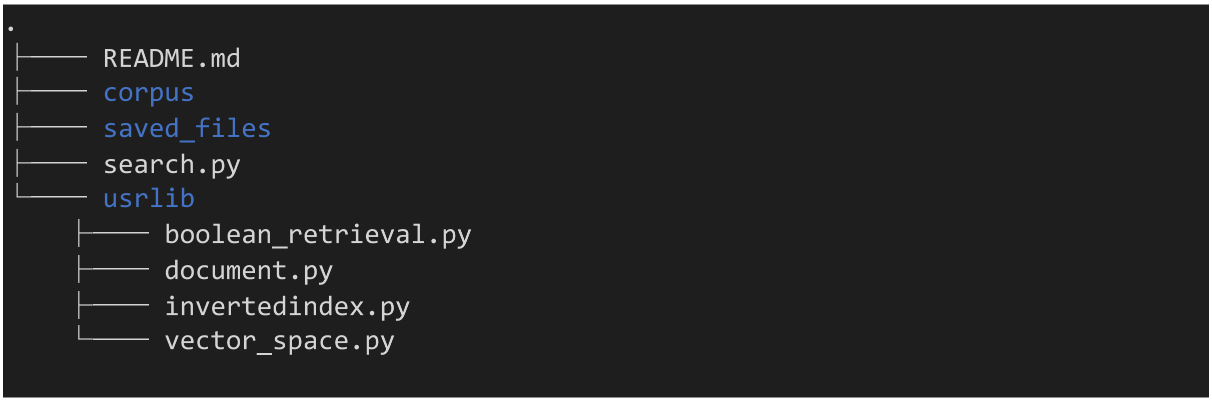
Programming Language: Python 3

Dataset: 42 text documents of Shakespeare's Play

Dataset Source: https://shakespeare.folger.edu/download-the-folger-shakespeare-complete-set/

Python libraries used: nltk, numpy, pandas, pickle

Folder Structure:



```
.
├─── README.md
├─── corpus
├─── saved_files
├─── search.py
└─── usrlib
     ├─── boolean_retrieval.py
     ├─── document.py
     ├─── invertedindex.py
     └─── vector_space.py
```

Classes used are

- Document
- InvertedIndex
- Tf_Idf

Working :

A document object is created from a text file. Pre-processing like case normalization, tokenization, removal of stop words and stemming is performed on the document object.

Unique words with their frequencies are stored in a dictionary.

Series of weights for each term in the corpus is represented as a document vector. The method used for calculating weights is tf-idf (term frequency - inverse document frequency).

To create a vector space model, a boolean retrieval model is used first. From boolean retrieval model results, the list of matched documents is given as input to a vector space model. This decreases the search space and improves the time efficiency.

Calculating Tf-Idf :

Term Frequency = $f_{t,d}$

$$tf_{t,d} = log\left(1 + f_{t,d}\right)$$

where t = term t and d = document

Document Frequency = df_t

Number of documents containing the term t

Inverse Document frequency = idf_t

$$idf_t = log\left(\frac{N}{df_t}\right)$$

where N = number of documents

Tf-idf weight = $tf_{t,d} * idf_t$

$$tf_{t,d} * idf_t = log\left(1 + f_{t,d}\right) * log\left(\frac{N}{df_t}\right)$$

for a term t, in document d

Similarity Metrics:

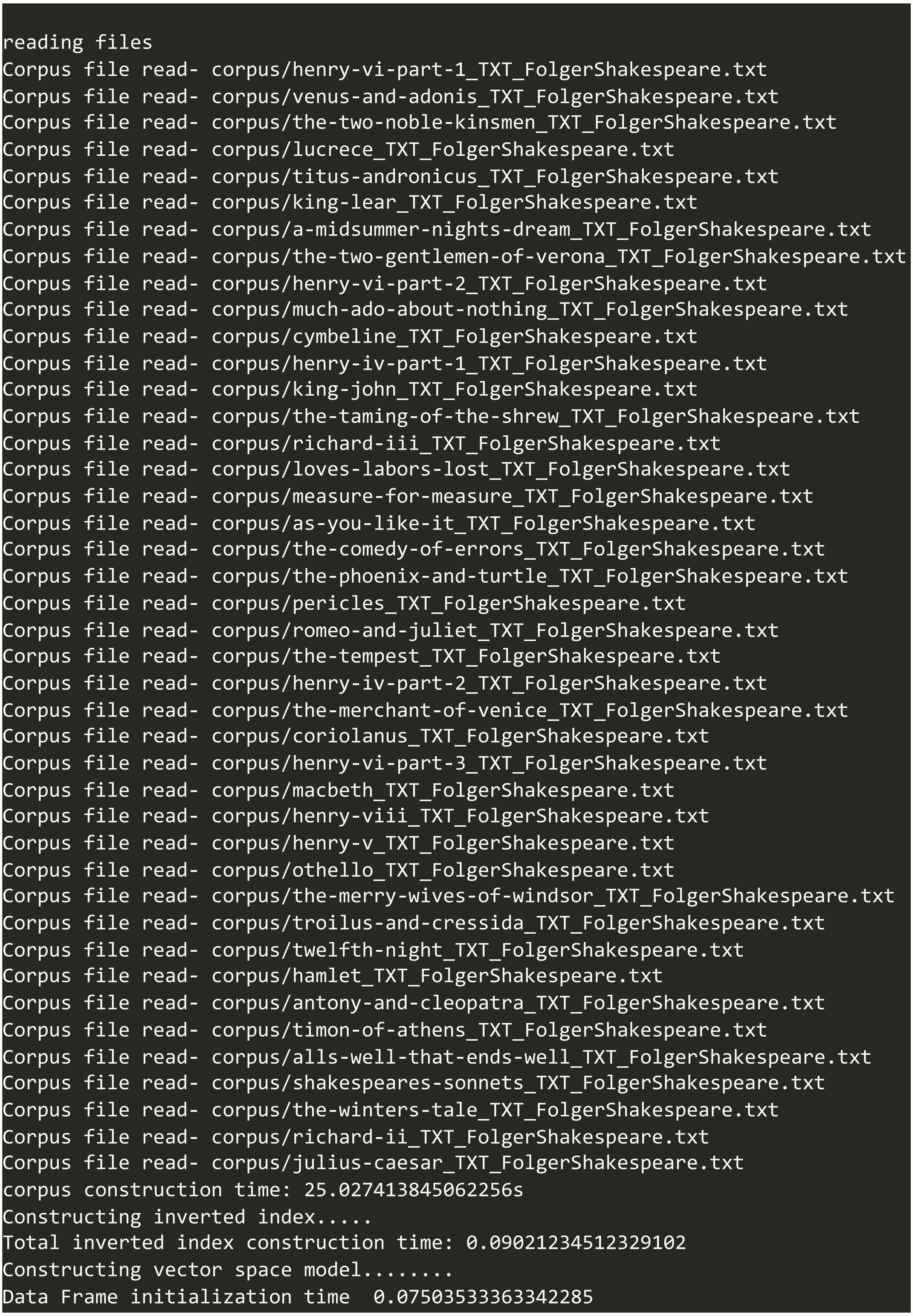
A vector is generated from a query using tf-idf weights. Cosine similarity is used to calculate the similarity between the above query vector and all the document vectors and they are returned in order of decreasing similarity. The top ten results are returned to the user.

Results:

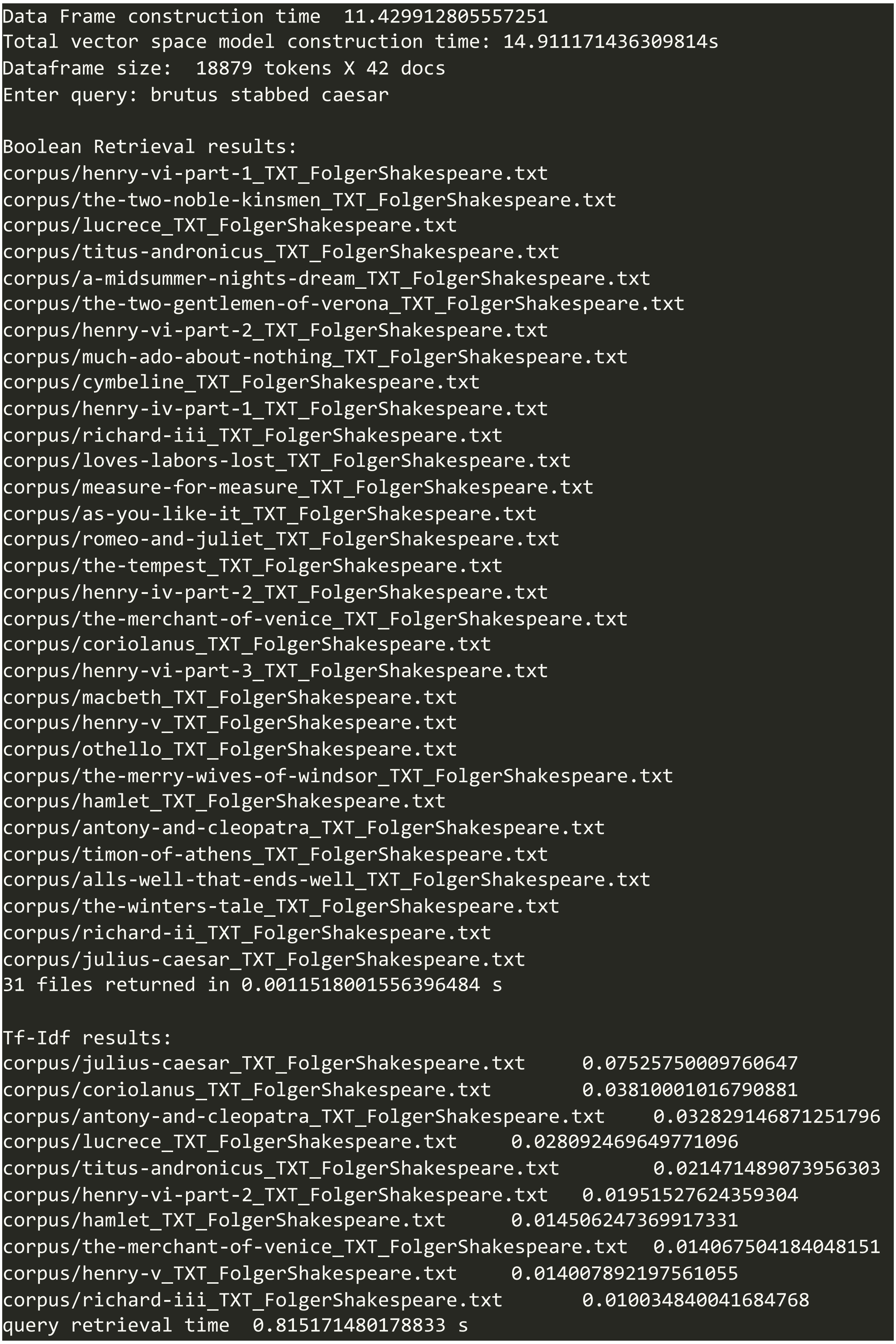
Running times:

● corpus construction time	25.027413845062256s
● inverted index construction time:	0.09021234512329102s
● Data Frame initialization time:	0.07503533363342285s
● Data Frame construction time:	11.429912805557251
● Total vector space model construction time:	14.911171436309814s
● Size of dataframe:	18879 tokens X 42 docs
● query retrieval time:	0.815171480178833 s

Output:



```
reading files
Corpus file read- corpus/henry-vi-part-1_TXT_FolgerShakespeare.txt
Corpus file read- corpus/venus-and-adonis_TXT_FolgerShakespeare.txt
Corpus file read- corpus/the-two-noble-kinsmen_TXT_FolgerShakespeare.txt
Corpus file read- corpus/lucrece_TXT_FolgerShakespeare.txt
Corpus file read- corpus/titus-andronicus_TXT_FolgerShakespeare.txt
Corpus file read- corpus/king-lear_TXT_FolgerShakespeare.txt
Corpus file read- corpus/a-midsummer-nights-dream_TXT_FolgerShakespeare.txt
Corpus file read- corpus/the-two-gentlemen-of-verona_TXT_FolgerShakespeare.txt
Corpus file read- corpus/henry-vi-part-2_TXT_FolgerShakespeare.txt
Corpus file read- corpus/much-ado-about-nothing_TXT_FolgerShakespeare.txt
Corpus file read- corpus/cymbeline_TXT_FolgerShakespeare.txt
Corpus file read- corpus/henry-iv-part-1_TXT_FolgerShakespeare.txt
Corpus file read- corpus/king-john_TXT_FolgerShakespeare.txt
Corpus file read- corpus/the-taming-of-the-shrew_TXT_FolgerShakespeare.txt
Corpus file read- corpus/richard-iii_TXT_FolgerShakespeare.txt
Corpus file read- corpus/loves-labors-lost_TXT_FolgerShakespeare.txt
Corpus file read- corpus/measure-for-measure_TXT_FolgerShakespeare.txt
Corpus file read- corpus/as-you-like-it_TXT_FolgerShakespeare.txt
Corpus file read- corpus/the-comedy-of-errors_TXT_FolgerShakespeare.txt
Corpus file read- corpus/the-phoenix-and-turtle_TXT_FolgerShakespeare.txt
Corpus file read- corpus/pericles_TXT_FolgerShakespeare.txt
Corpus file read- corpus/romeo-and-juliet_TXT_FolgerShakespeare.txt
Corpus file read- corpus/the-tempest_TXT_FolgerShakespeare.txt
Corpus file read- corpus/henry-iv-part-2_TXT_FolgerShakespeare.txt
Corpus file read- corpus/the-merchant-of-venice_TXT_FolgerShakespeare.txt
Corpus file read- corpus/coriolanus_TXT_FolgerShakespeare.txt
Corpus file read- corpus/henry-vi-part-3_TXT_FolgerShakespeare.txt
Corpus file read- corpus/macbeth_TXT_FolgerShakespeare.txt
Corpus file read- corpus/henry-viii_TXT_FolgerShakespeare.txt
Corpus file read- corpus/henry-v_TXT_FolgerShakespeare.txt
Corpus file read- corpus/othello_TXT_FolgerShakespeare.txt
Corpus file read- corpus/the-merry-wives-of-windsor_TXT_FolgerShakespeare.txt
Corpus file read- corpus/troilus-and-cressida_TXT_FolgerShakespeare.txt
Corpus file read- corpus/twelfth-night_TXT_FolgerShakespeare.txt
Corpus file read- corpus/hamlet_TXT_FolgerShakespeare.txt
Corpus file read- corpus/antony-and-cleopatra_TXT_FolgerShakespeare.txt
Corpus file read- corpus/timon-of-athens_TXT_FolgerShakespeare.txt
Corpus file read- corpus/alls-well-that-ends-well_TXT_FolgerShakespeare.txt
Corpus file read- corpus/shakespeares-sonnets_TXT_FolgerShakespeare.txt
Corpus file read- corpus/the-winters-tale_TXT_FolgerShakespeare.txt
Corpus file read- corpus/richard-ii_TXT_FolgerShakespeare.txt
Corpus file read- corpus/julius-caesar_TXT_FolgerShakespeare.txt
corpus construction time: 25.027413845062256s
Constructing inverted index.....
Total inverted index construction time: 0.09021234512329102
Constructing vector space model........
Data Frame initialization time  0.07503533363342285
```



```
Data Frame construction time  11.429912805557251
Total vector space model construction time: 14.911171436309814s
Dataframe size:  18879 tokens X 42 docs
Enter query: brutus stabbed caesar

Boolean Retrieval results:
corpus/henry-vi-part-1_TXT_FolgerShakespeare.txt
corpus/the-two-noble-kinsmen_TXT_FolgerShakespeare.txt
corpus/lucrece_TXT_FolgerShakespeare.txt
corpus/titus-andronicus_TXT_FolgerShakespeare.txt
corpus/a-midsummer-nights-dream_TXT_FolgerShakespeare.txt
corpus/the-two-gentlemen-of-verona_TXT_FolgerShakespeare.txt
corpus/henry-vi-part-2_TXT_FolgerShakespeare.txt
corpus/much-ado-about-nothing_TXT_FolgerShakespeare.txt
corpus/cymbeline_TXT_FolgerShakespeare.txt
corpus/henry-iv-part-1_TXT_FolgerShakespeare.txt
corpus/richard-iii_TXT_FolgerShakespeare.txt
corpus/loves-labors-lost_TXT_FolgerShakespeare.txt
corpus/measure-for-measure_TXT_FolgerShakespeare.txt
corpus/as-you-like-it_TXT_FolgerShakespeare.txt
corpus/romeo-and-juliet_TXT_FolgerShakespeare.txt
corpus/the-tempest_TXT_FolgerShakespeare.txt
corpus/henry-iv-part-2_TXT_FolgerShakespeare.txt
corpus/the-merchant-of-venice_TXT_FolgerShakespeare.txt
corpus/coriolanus_TXT_FolgerShakespeare.txt
corpus/henry-vi-part-3_TXT_FolgerShakespeare.txt
corpus/macbeth_TXT_FolgerShakespeare.txt
corpus/henry-v_TXT_FolgerShakespeare.txt
corpus/othello_TXT_FolgerShakespeare.txt
corpus/the-merry-wives-of-windsor_TXT_FolgerShakespeare.txt
corpus/hamlet_TXT_FolgerShakespeare.txt
corpus/antony-and-cleopatra_TXT_FolgerShakespeare.txt
corpus/timon-of-athens_TXT_FolgerShakespeare.txt
corpus/alls-well-that-ends-well_TXT_FolgerShakespeare.txt
corpus/the-winters-tale_TXT_FolgerShakespeare.txt
corpus/richard-ii_TXT_FolgerShakespeare.txt
corpus/julius-caesar_TXT_FolgerShakespeare.txt
31 files returned in 0.0011518001556396484 s

Tf-Idf results:
corpus/julius-caesar_TXT_FolgerShakespeare.txt     0.07525750009760647
corpus/coriolanus_TXT_FolgerShakespeare.txt        0.03810001016790881
corpus/antony-and-cleopatra_TXT_FolgerShakespeare.txt    0.032829146871251796
corpus/lucrece_TXT_FolgerShakespeare.txt      0.028092469649771096
corpus/titus-andronicus_TXT_FolgerShakespeare.txt         0.021471489073956303
corpus/henry-vi-part-2_TXT_FolgerShakespeare.txt    0.01951527624359304
corpus/hamlet_TXT_FolgerShakespeare.txt    0.014506247369917331
corpus/the-merchant-of-venice_TXT_FolgerShakespeare.txt    0.014067504184048151
corpus/henry-v_TXT_FolgerShakespeare.txt   0.014007892197561055
corpus/richard-iii_TXT_FolgerShakespeare.txt     0.010034840041684768
query retrieval time  0.815171480178833 s
```