ASSIGNMENT 2

Files included: cosine_based_retrieval.py, tf_idf_based_retrieval.py, index_construction_preprocessing.py, doc_id_doc_name.json, inverted_index.json, query_cache.json, query_cache_cosine.json, index.txt, wordcloud.png

Assumptions

1. Header is assumed to be everything above the first blank line.
2. Stop words present in the NLTK corpus are considered.
3. Document IDs are provided.

Pre-processing Steps

Removal of Header

(All the lines before the first blank line are removed)

Removal of Punctuation marks, comma, etc

(They are removed through regular expression)

Tokenization

(Tokens are formed using word_tokenize and special symbols are removed)

Removal of Stop Words

(Stop words are removed using NLTK stop words)

Normalization

(All token are converted into lower case)

Stemming

(Stemming is performed using Porter algorithm to get the root word)

Num2Words

(Numbers as token are converted into words)

Parsing of Title

Title of each Document file are parsed from index.html files using BeautifulSoup library and are written into index.txt files.

Index Construction

Inverted Index is constructed from term- doc pair and Inverted Index is dictionary with term as key and List as value. List contains the first element as Document Frequency and the subsequent elements as the List of DocID and Term Frequency.

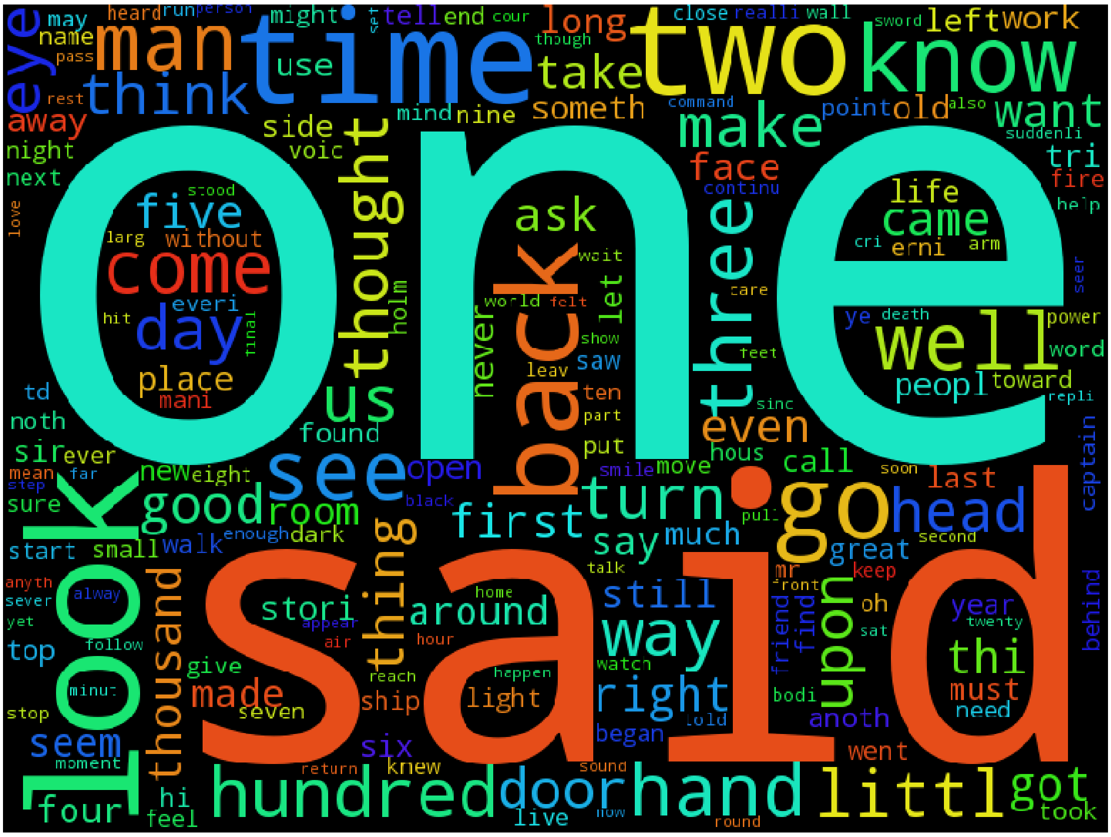
Another dictionary is also maintained to map the Document ID with the corresponding file.

Both the dictionary are dumped into json files.

Number of Documents: 485

Number of Tokens: 33679

WordCloud



1.Tf-Idf based retrieval

Query: one time thing

DocID: 468 Score: 0.977984812611

Document Name: **stories/vgilante.txt**

DocID: 258 Score: 0.971184597606

Document Name: **stories/hitch3.txt**

DocID: 450 Score: 0.953561101916

Document Name: **stories/timem.hac**

DocID: 246 Score: 0.948934823285

Document Name: **stories/gulliver.txt**

DocID: 257 Score: 0.943915212758

Document Name: **stories/hitch2.txt**

2.Cosine based retrieval

Query: one time thing

DocID: 0.00598972072898 Score: 302

Document Name: **stories/lmtchgrl.txt**

DocID: 0.00596658756469 Score: 394

Document Name: **stories/reality.txt**

DocID: 0.00583057046835 Score: 45

Document Name: **stories/3wishes.txt**

DocID: 0.00566365779724 Score: 263

Document Name: **stories/horsdonk.txt**

DocID: 0.00546667886177 Score: 338

Document Name: **stories/non2**

References

1. Introduction to Information Retrieval CS 221 Donald J. Patterson