**Documentation**

**Py files**

**BeautifulSoup.py**:

All the documents are scrapped from https://www.medicalnewstoday.com/popular using Beautiful Soup the text is grepped from the class: ”article\_body” .

**T\_N.py**:

It takes files\_list(list of documents) from lod.py as the input argument and perform tokenization and normalization and stores tokenized words in T\_N.txt.(Tokenized using NLTK)

**stop\_words.py**:

The file T\_N.txt is opened and stop word are removed from this file and the words are the stored in another file stop\_word.txt

**stemmed\_words.py**:

The file stop\_words.txt is opened and the words are stemmed using Porters Stemmer and are stored in stemmed.txt file.

**lod.py**:

Contains two file lists and the definitions for Tokenization, Normalization, removing stop words, stemming and unique\_list.

**Functions implemented in lod**:

**T\_N(file\_content):** Tokenizes and Normalizes the given input “file\_content”

**remove\_stopwords(file)**: Removes stop words from the input “file”

**stem\_words(files)**: Stems the input “files”

**unique\_words(text)**: Removes duplicates of a word

**inverted\_index.py**:

In this module documents are assigned their id’s inverted index, term frequency table for each word, idf table is built

and the tables are stores in tf\_document\_data.csv, idf\_table.csv.

**document\_frequency.py**:

The function “idf” in this module calculates the idf of each word.

**stem\_documents.py**:

Stems each document in files\_list(imported from lod.py) ,creates files listed in stem\_document\_list(lod.py) and stores the stemmed document.

**query.py**:

Takes the input from the user and writes documents names and cosines scores in the descending order into total\_scores.csv file.