

## Inverted Index and Positional Index

### Approach and Methodologies:-

In my approach to creating an inverted index and positional index, I have created eight files. In these eight files, I have distributed the task of preprocessing text files present in the 'text\_files' folder to the process of asking questions by the user. These are the names of the .py files I have made, along with their functionalities and the sequence in which they are used.

### Python Files:-

Preprocess.py : In this file I have used NLTK, BeautifulSoup and String library the main working of this file is to perform the pre-processing steps of Tokenization, Lowering the letters Removing stop words , Removing the blank space tokens, and Removing punctuations.

AssignmentIR.py : In this file I have imported the above preprocess.py module and , This file loops through the files1.txt to file999.txt and calls the method in Preprocess to and stored the result back in the folder named as preprocess\_text\_files with same name.

Inverted\_Index.py : In this file I have created the inverted index by using the text file present in preprocess\_text\_files folder and this gives us the inverted index in the form of dictionary where the keys are the unique terms in the document(file) and the values are the posting list which contains the document(file) name where that term is present.

Positional\_Index.py : In this file I have created the Positional\_index using the text\_file present in the preprocess\_text\_files folder and this gives us the positional\_index in form of dictionary where the term is the key and the value is dictionary in itself , The dictionary which is present as a value contains the key as the document(file) name where that term is present and the value is the list which contains the positions where that words is present in the document(file).

QueryInputQ1.py : In this file what we do is we are basically taking the input than we are performing preprocessing on it by performing all the 5 steps as mentioned it starts with i) Lowering the Text, ii)Performing the Tokenization , iii)removing the stop words iv) removing the stopwords v)Removing the blank space tokens , This file has all function for all the above operation we only have to provide the file number and it will give us the preprocessed file.txt and store it in preprocessed\_files folder.

QueryInputQ2.py : It takes the input from the user as a sentence and it also takes the operations as well which can be AND, NOT,AND NOT,OR NOT, than it extracts the posting list of each term in the sentence and performs the above operations on them , Make sure the number of operation is one less than the number of terms in the list.

QueryInputQ3.py : It takes the input from the user as a sentence and it performs the phrase search on the sentence , when it first takes the input it preprocess it by performing all the operation's on them like Tokenization, Lowering the letters Removing stop words , Removing the blank space tokens, and Removing punctuations, Then it finds all the document which contains all these terms then we return it.

printInvertedIndex.py: There is a python file separately which helps you in printing the inverted index.

### Library Used:-

The library which I have used for this assignment is NLTK and BeautifulSoup:-

**NLTK** : Provides tools for natural language processing tasks such as Tokenization Stemming , Tagging , Parsing and Semantic Reasoning.

**Beautiful Soup**: A Library for parsing HTML and XML documents and extracting data from Them.

**Sent\_Tokenize**: (from nltk.tokenize) : Tokenizes text into sentences.

**Word\_Tokenize**: (from nltk.tokenize) : Tokenizes text into words.

**stopwords**: (from nltk.corpus) Provides a collection of common stop words for english languages.

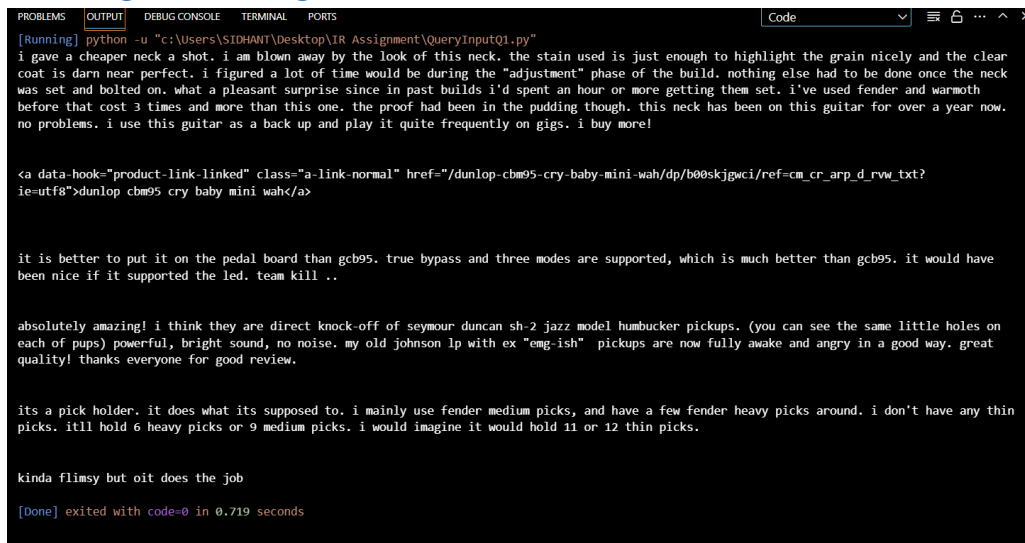
**String**: provides constant and function for string manipulation.

## WORKING RESULTS OF EACH PROBLEM (with screen shots)

Q1. Here we have to print the contents of five sample file before and after performing each operation's ?.

Ans:-

### Lowering and Printing it :-



```
PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS
[Running] python -u "c:\Users\SIDHANT\Desktop\IR Assignment\QueryInputQ1.py"
i gave a cheaper neck a shot. i am blown away by the look of this neck. the stain used is just enough to highlight the grain nicely and the clear coat is darn near perfect. i figured a lot of time would be during the "adjustment" phase of the build. nothing else had to be done once the neck was set and bolted on. what a pleasant surprise since in past builds i'd spent an hour or more getting them set. i've used fender and warmoth before that cost 3 times and more than this one. the proof had been in the pudding though. this neck has been on this guitar for over a year now. no problems. i use this guitar as a back up and play it quite frequently on gigs. i buy more!

<a data-hook="product-link-linked" class="a-link-normal" href="/dunlop-cbm95-cry-baby-mini-wah/dp/B00SKJGWCI/ref=cm_cr_ar_p_d_rvw_txt?ie=utf8">dunlop cbm95 cry baby mini wah</a>

it is better to put it on the pedal board than gcb95. true bypass and three modes are supported, which is much better than gcb95. it would have been nice if it supported the led. team kill ..

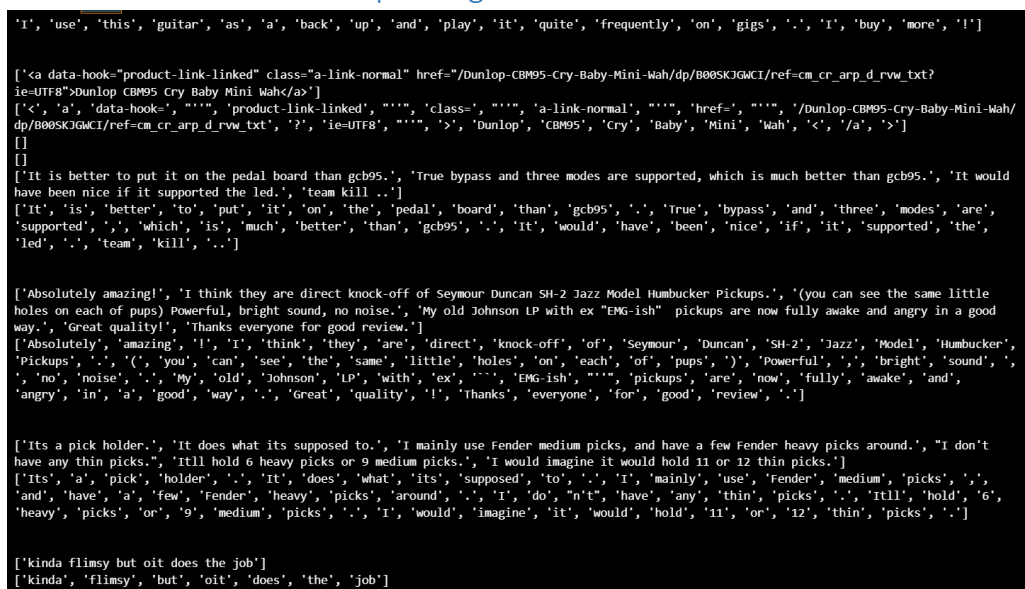
absolutely amazing! i think they are direct knock-off of seymour duncan sh-2 jazz model humbucker pickups. (you can see the same little holes on each of pups) powerful, bright sound, no noise. my old johnson lp with ex "emg-ish" pickups are now fully awake and angry in a good way. great quality! thanks everyone for good review.

its a pick holder. it does what its supposed to. i mainly use fender medium picks, and have a few fender heavy picks around. i don't have any thin picks. itll hold 6 heavy picks or 9 medium picks. i would imagine it would hold 11 or 12 thin picks.

kinda flimsy but oit does the job

[Done] exited with code=0 in 0.719 seconds
```

### Perform Tokenization and printing it :-



```
['I', 'use', 'this', 'guitar', 'as', 'a', 'back', 'up', 'and', 'play', 'it', 'quite', 'frequently', 'on', 'gigs', '.', 'I', 'buy', 'more', '!']

['<a data-hook="product-link-linked" class="a-link-normal" href="/Dunlop-CBM95-Cry-Baby-Mini-Wah/dp/B00SKJGWCI/ref=cm_cr_ar_p_d_rvw_txt?ie=utf8">Dunlop CBM95 Cry Baby Mini Wah</a>']
['<', 'a', 'data-hook=', '""', 'product-link-linked', '""', 'class=', '""', 'a-link-normal', '""', 'href=', '""', '/Dunlop-CBM95-Cry-Baby-Mini-Wah/dp/B00SKJGWCI/ref=cm_cr_ar_p_d_rvw_txt', '?', 'ie=utf8', '""', '>', 'Dunlop', 'CBM95', 'Cry', 'Baby', 'Mini', 'Wah', '<', '/a', '>']
[]
[]
['It is better to put it on the pedal board than gcb95.', 'True bypass and three modes are supported, which is much better than gcb95.', 'It would have been nice if it supported the led.', 'team kill ...']
['It', 'is', 'better', 'to', 'put', 'it', 'on', 'the', 'pedal', 'board', 'than', 'gcb95', '.', 'True', 'bypass', 'and', 'three', 'modes', 'are', 'supported', '.', 'which', 'is', 'much', 'better', 'than', 'gcb95', '.', 'It', 'would', 'have', 'been', 'nice', 'if', 'it', 'supported', 'the', 'led', '.', 'team', 'kill', '...']

['Absolutely amazing!', 'I think they are direct knock-off of Seymour Duncan SH-2 Jazz Model Humbucker Pickups.', '(you can see the same little holes on each of pups) Powerful, bright sound, no noise.', 'My old Johnson LP with ex "EMG-ish" pickups are now fully awake and angry in a good way.', 'Great quality!', 'Thanks everyone for good review.']
['Absolutely', 'amazing', '!', 'I', 'think', 'they', 'are', 'direct', 'knock-off', 'of', 'Seymour', 'Duncan', 'SH-2', 'Jazz', 'Model', 'Humbucker', 'Pickups', '.', '(', 'you', 'can', 'see', 'the', 'same', 'little', 'holes', 'on', 'each', 'of', 'pups', ')', 'Powerful', '!', 'bright', 'sound', '!', 'no', 'noise', '!', 'My', 'old', 'Johnson', 'LP', 'with', 'ex', '""', 'EMG-ish', '""', 'pickups', 'are', 'now', 'fully', 'awake', 'and', 'angry', 'in', 'a', 'good', 'way', '.', 'Great', 'quality', '!', 'Thanks', 'everyone', 'for', 'good', 'review', '.']

['Its a pick holder.', 'It does what its supposed to.', 'I mainly use Fender medium picks, and have a few Fender heavy picks around.', 'I don't have any thin picks.', 'Itll hold 6 heavy picks or 9 medium picks.', 'I would imagine it would hold 11 or 12 thin picks.']
['Its', 'a', 'pick', 'holder', '.', 'It', 'does', 'what', 'its', 'supposed', 'to', '.', 'I', 'mainly', 'use', 'Fender', 'medium', 'picks', '.', 'and', 'have', 'a', 'few', 'Fender', 'heavy', 'picks', 'around', '.', 'I', 'do', 'n't', 'have', 'any', 'thin', 'picks', '.', 'Itll', 'hold', '6', 'heavy', 'picks', 'or', '9', 'medium', 'picks', '.', 'I', 'would', 'imagine', 'it', 'would', 'hold', '11', 'or', '12', 'thin', 'picks', '.']

['kinda flimsy but oit does the job']
['kinda', 'flimsy', 'but', 'oit', 'does', 'the', 'job']
```

## RemoveStopwords and printing it :-

We might find that some stop words are present because the stop words which nltk corpus gives are all in lower case letter so we should perform this step after lowering only, just for the sake of question I am doing this.

```
PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS
[Running] python -u "c:\Users\SIDHANT\Desktop\IR Assignment\QueryInputQ1.py"
I gave cheaper neck shot. I blown away look neck. The stain used enough highlight grain nicely clear coat darn near perfect. I figured lot time would "adjustment" phase build. Nothing else done neck set bolted on. What pleasant surprise since past builds I'd spent hour getting set. I've used Fender Warmoth cost 3 times one. The proof pudding though. This neck guitar year now. No problems. I use guitar back play quite frequently gigs. I buy more!

<a data-hook="product-link-linked" class="a-link-normal" href="/Dunlop-CBM95-Cry-Baby-Mini-Wah/dp/B00SKJGWC1/ref=cm_cr_arp_d_rvw_txt?ie=UTF8">Dunlop CBM95 Cry Baby Mini Wah</a>

It better put pedal board gcb95. True bypass three modes supported, much better gcb95. It would nice supported led. team kill ..

Absolutely amazing! I think direct knock-off Seymour Duncan SH-2 Jazz Model Humbucker Pickups. (you see little holes pups) Powerful, bright sound, noise. My old Johnson LP ex "EMG-ish" pickups fully awake angry good way. Great quality! Thanks everyone good review.

Its pick holder. It supposed to. I mainly use Fender medium picks, Fender heavy picks around. I thin picks. Itll hold 6 heavy picks 9 medium picks. I would imagine would hold 11 12 thin picks.

kinda flimsy oit job
[Done] exited with code=0 in 0.759 seconds
```

## Remove Punctuations and printing it :-

```
PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS
[Running] python -u "c:\Users\SIDHANT\Desktop\IR Assignment\QueryInputQ1.py"
I gave a cheaper neck a shot I am blown away by the look of this neck The stain used is just enough to highlight the grain nicely and the clear coat is darn near perfect I figured a lot of time would be during the adjustment phase of the build Nothing else had to be done once the neck was set and bolted on What a pleasant surprise since in past builds Id spent an hour or more getting them set Ive used Fender and Warmoth before that cost 3 times and more than this one The proof had been in the pudding though This neck has been on this guitar for over a year now No problems I use this guitar as a back up and play it quite frequently on gigs I buy more

a datahookproductlinklinked classalinknormal hrefDunlopCBM95CryBabyMiniWahdpB00SKJGWC1refcmcrarpdrvwtxtieUTF8Dunlop CBM95 Cry Baby Mini Waha

It is better to put it on the pedal board than gcb95 True bypass and three modes are supported which is much better than gcb95 It would have been nice if it supported the led team kill

Absolutely amazing I think they are direct knockoff of Seymour Duncan SH2 Jazz Model Humbucker Pickups you can see the same little holes on each of pups Powerful bright sound no noise My old Johnson LP with ex EMGish pickups are now fully awake and angry in a good way Great quality Thanks everyone for good review

Its a pick holder It does what its supposed to I mainly use Fender medium picks and have a few Fender heavy picks around I dont have any thin picks Itll hold 6 heavy picks or 9 medium picks I would imagine it would hold 11 or 12 thin picks

kinda flimsy but oit does the job
```

## Remove blank space tokens and printing it :-

```
PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS Code
[Running] python -u "c:\Users\SIDHANT\Desktop\IR Assignment\QueryInput01.py"
I gave a cheaper neck a shot I am blown away by the look of this neck The stain used is just enough to highlight the grain nicely and the clear coat is darn near perfect I figured a lot of time would be during the adjustment phase of the build Nothing else had to be done once the neck was set and bolted on What a pleasant surprise since in past builds Id spent an hour or more getting them set Ive used Fender and Warmoth before that cost 3 times and more than this one The proof had been in the pudding though This neck has been on this guitar for over a year now No problems I use this guitar as a back up and play it quite frequently on gigs I buy more

a datahookproductlinklinked classalinknormal hrefDunlopCBM95CryBabyMiniWahdpB00SKJGWCirefcrcrarpdrwtxttieUTF8Dunlop CBM95 Cry Baby Mini Waha

It is better to put it on the pedal board than gcb95 True bypass and three modes are supported which is much better than gcb95 It would have been nice if it supported the led team kill

Absolutely amazing I think they are direct knockoff of Seymour Duncan SH2 Jazz Model Humbucker Pickups you can see the same little holes on each of pups Powerful bright sound no noise My old Johnson LP with ex EMGish pickups are now fully awake and angry in a good way Great quality Thanks everyone for good review

Its a pick holder It does what its supposed to I mainly use Fender medium picks and have a few Fender heavy picks around I dont have any thin picks Itll hold 6 heavy picks or 9 medium picks I would imagine it would hold 11 or 12 thin picks

kinda flimsy but oit does the job
```

## Q2. Unigram Inverted Index and Boolean Queries

Input format:-

Input format:

- a. The first line contains N denoting the number of queries to execute
- b. The next 2N lines contain queries in the following format:
  - i. Input sequence
  - ii. Operations separated by comma

```
PS C:\Users\SIDHANT\Desktop\IR Assignment> python .\QueryInputQ2.py
1
guitar is better than car
AND,OR
Query: guitar AND better OR car

QUERY 1 :No of documents retrived for query: 31
Names of document retrived for Query 1 is: file641.txt ,file514.txt ,file68.txt ,file264.txt ,file893.txt ,file271.txt ,file591.txt ,file978.txt ,file277.txt ,file469.txt ,file982.txt ,file24.txt ,file541.txt ,file413.txt ,file542.txt ,file801.txt ,file356.txt ,file37.txt ,file484.txt ,file743.txt ,file168.txt ,file166.txt ,file746.txt ,file628.txt ,file174.txt ,file239.txt ,file245.txt ,file54.txt ,file886.txt ,file314.txt ,file61.txt ,

PS C:\Users\SIDHANT\Desktop\IR Assignment> |
```

### Q3. Positional index and queries

a. The first line contains N denoting the number of queries to execute

b. The next N lines contain phrase queries

4. Output Format:

a. 2N lines consisting of the results in the following format:

i. Number of documents retrieved for query X using positional index

ii. Names of documents retrieved for query X using positional index

```
PS C:\Users\SIDHANT\Desktop\IR Assignment> python .\QueryInputQ3.py
1
i broke my guitar
Number of document retrived for query 1 using positional index: 5
Names of document retrived for query 1 using positional index: File68.txt ,
File841.txt ,
File649.txt ,
File782.txt ,
File123.txt ,
PS C:\Users\SIDHANT\Desktop\IR Assignment> █
```