CSE 7337 Somya Singh Spring 2017
Email: somyas@smu.edu
Student Id: 47304053

Term Project – Part 2

Table of Contents

Introduction and Key points

This part of the project takes the output from the first phase of the project and gives the user an ability to do search for a word or a term. There was no major change done to any of scripts in the first part of the project except adding a filter in crawler script to avoid downloading pptx (http://lyle.smu.edu/~fmoore/misc/poem-classification.pptx).

The key components of **search script** are:

- When you execute the script the first thing script displays is number of words it has in the dictionary (not including stop words and words from URL of the document).
- User is given an option to enter a word or the query he/she wants to do search.
- The search is case in-sensitive i.e. user can input in any case his or her search term.
- The script will keep executing unless user inputs the keyword "Stop" (case in-sensitive).
- The script will display maximum of 6 result for a search.
- The first thing the script displays the result found with time it took to do search
- The result displays the path, title, highlighted content (100 characters around the found search term) and cosine similarity score of the result.
- If the result found for the term is less than 3 then the script uses the dictionary that was provided to expand the query.
- If there are no results found the script will display zero results.
- The results are displayed in descending order based on score.
- The Score is increemented by 0.5 if there is a title hit for the query.

Python Packages

There is no new package compared to phase 1 of this project. The same packages as part 1 is used here. The only package needed for this part is "Whoosh". For installation you can refer to README file for details.

Functions Used

The script contains following functions:

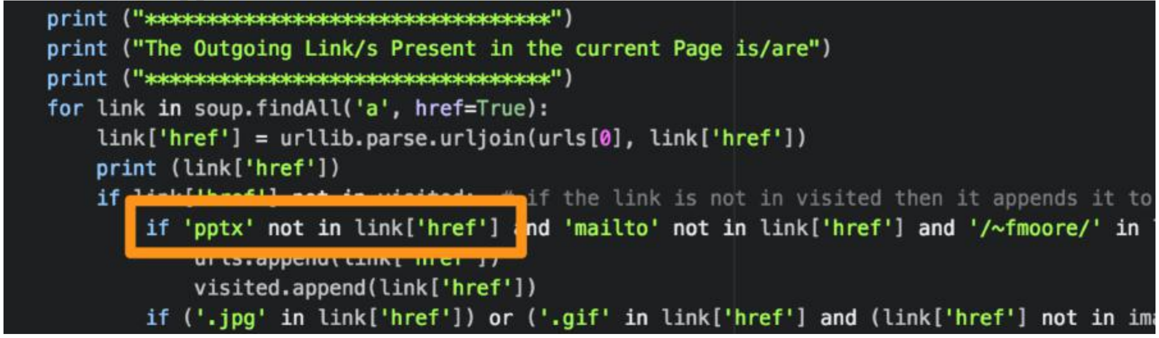
1. CountWords: The function used to count the words in dictionary
2. SearchAlgo: The function is doing actual searches
3. My_Dictionary: The function is to define the dictionary provided in project for expnsion of search results.
4. SearchLoop: The function keep calling Search unless user enters “Stop”.

Project Answers

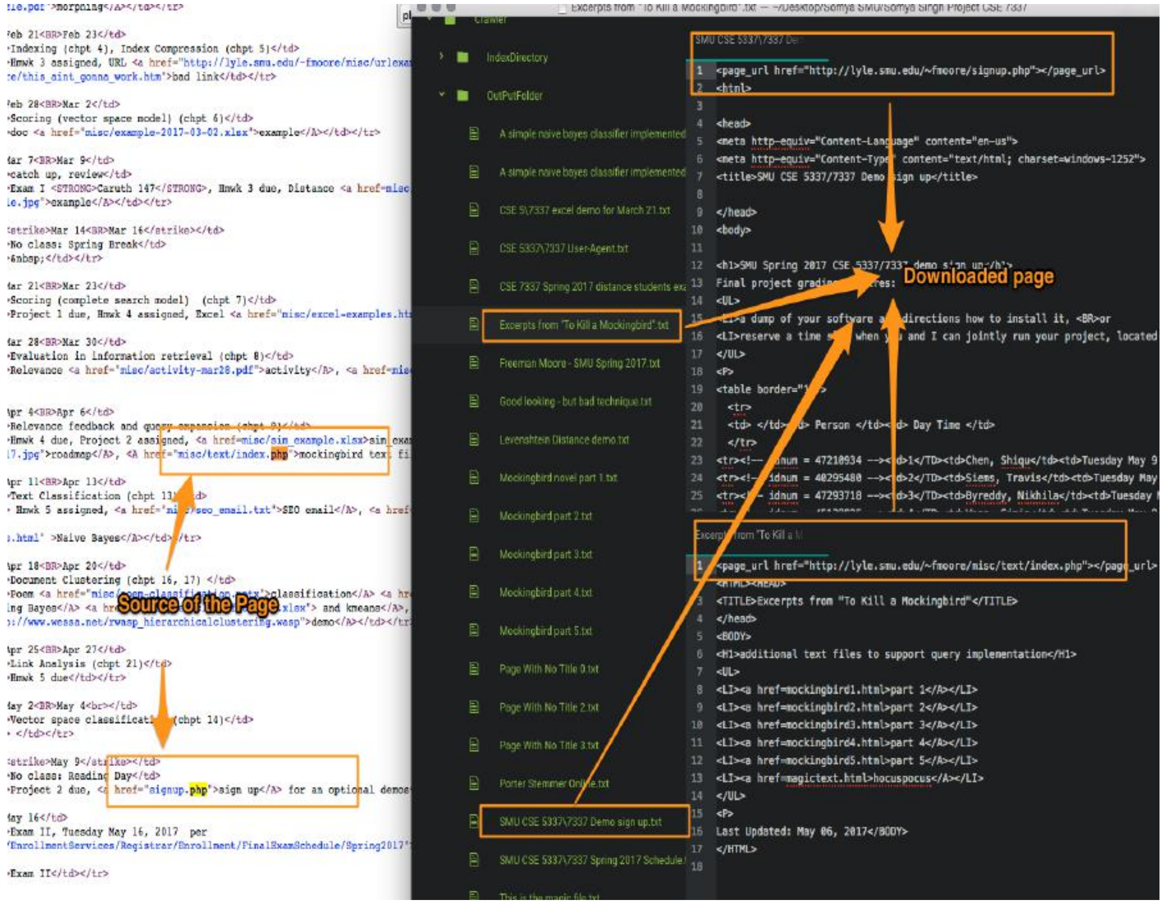
1. Use the web crawler you built in Project 1 that crawled a limited space, looking for text and html files ***and php files***. You may need to modify how you saved the words from the pages that you traversed to support the query engine. Describe in detail what you changed to support the second half of the project.

Answer:

I **did not have to do any major change** to my phase 1 project. The crawler script had only one update to avoid downloading pptx file (misc/poem-classification.pptx)



```
print ("**********************************")
print ("The Outgoing Link/s Present in the current Page is/are")
print ("**********************************")
for link in soup.findAll('a', href=True):
    link['href'] = urllib.parse.urljoin(urls[0], link['href'])
    print (link['href'])
    if link['href'] not in visited:  # if the link is not in visited then it appends it to
        if 'pptx' not in link['href'] and 'mailto' not in link['href'] and '/~fmoore/' in 
            urls.append(link['href'])
            visited.append(link['href'])
        if ('.jpg' in link['href']) or ('.gif' in link['href'] and (link['href'] not in ima
```


Crawler
IndexDirectory
OutPutFolder
A simple naive bayes classifier implemented
A simple naive bayes classifier implemented
CSE 5\7337 excel demo for March 21.txt
CSE 5337\7337 User-Agent.txt
CSE 7337 Spring 2017 distance students exa
Excerpts from 'To Kill a Mockingbird'.txt
Freeman Moore - SMU Spring 2017.txt
Good looking - but bad technique.txt
Levenshtein Distance demo.txt
Mockingbird novel part 1.txt
Mockingbird part 2.txt
Mockingbird part 3.txt
Mockingbird part 4.txt
Mockingbird part 5.txt
Page With No Title 0.txt
Page With No Title 2.txt
Page With No Title 3.txt
Porter Stemmer Online.txt
SMU CSE 5337\7337 Demo sign up.txt
SMU CSE 5337\7337 Spring 2017 Schedule
This is the magic file.txt
Source of the Page
Downloaded page
Excerpts from 'To Kill a Mockingbird'.txt — ~/Desktop/Somya SMU/Somya Singh Project CSE 7337

```
ile.pdf">morphing</A></td></tr>

Feb 21<BR>Feb 23</td>
Indexing (chpt 4), Index Compression (chpt 5)</td>
Hmwk 3 assigned, URL <a href="http://lyle.smu.edu/~fmoore/misc/urlexa
te/this_aint_gonna_work.htm">bad link</td></tr>

Feb 28<BR>Mar 2</td>
Scoring (vector space model) (chpt 6)</td>
doc <a href="misc/example-2017-03-02.xlsx">example</A></td></tr>

Mar 7<BR>Mar 9</td>
catch up, review</td>
Exam I <STRONG>Caruth 147</STRONG>, Hmwk 3 due, Distance <a href=misc
le.jpg">example</A></td></tr>

<strike>Mar 14<BR>Mar 16</strike></td>
No class: Spring Break</td>
&nbsp;</td></tr>

Mar 21<BR>Mar 23</td>
Scoring (complete search model)  (chpt 7)</td>
Project 1 due, Hmwk 4 assigned, Excel <a href="misc/excel-examples.ht

Mar 28<BR>Mar 30</td>
Evaluation in information retrieval (chpt 8)</td>
Relevance <a href="misc/activity-mar28.pdf">activity</A>, <a href=mis


Apr 4<BR>Apr 6</td>
Relevance feedback and query expansion (chpt 9)</td>
Hmwk 4 due, Project 2 assigned, <a href=misc/sim_example.xlsx>sim exa
17.jpg">roadmap</A>, <A href="misc/text/index.php">mockingbird text fi

Apr 11<BR>Apr 13</td>
Text Classification (chpt 13)</td>
 Hmwk 5 assigned, <a href='misc/seo_email.txt'>SEO email</A>, <a href

s.html' >Naive Bayes</A></td></tr>

Apr 18<BR>Apr 20</td>
Document Clustering (chpt 16, 17) </td>
Poem <a href="misc/poem-classification.pptx">classification</A> <a hr
ing Bayes</A> <a hr                    .xlsx'> and kmeans</A>,
://www.wessa.net/rwasp_hierarchicalclustering.wasp">demo</A></td></tr

Apr 25<BR>Apr 27</td>
Link Analysis (chpt 21)</td>
Hmwk 5 due</td></tr>

May 2<BR>May 4<br></td>
Vector space classification (chpt 14)</td>
 </td></tr>

<strike>May 9</strike></td>
No class: Reading Day</td>
Project 2 due, <a href="signup.php">sign up</A> for an optional demos

May 16</td>
Exam II, Tuesday May 16, 2017  per
/EnrollmentServices/Registrar/Enrollment/FinalExamSchedule/Spring2017"

Exam II</td></tr>
```

SMU CSE 5337\7337 Dem

```
<page_url href="http://lyle.smu.edu/~fmoore/signup.php"></page_url>
<html>

<head>
<meta http-equiv="Content-Language" content="en-us">
<meta http-equiv="Content-Type" content="text/html; charset=windows-1252">
<title>SMU CSE 5337/7337 Demo sign up</title>

</head>
<body>

<h1>SMU Spring 2017 CSE 5337/7337 demo sign up</h1>
Final project grading       res:
<UL>
<LI>a dump of your software and directions how to install it, <BR>or
<LI>reserve a time s    when you and I can jointly run your project, located
</UL>
<P>
<table border="1">
  <tr>
  <td> </td><td> Person </td><td> Day Time </td>
  </tr>
<tr><!-- idnum = 47210934 --><td>1</TD><td>Chen, Shiqu</td><td>Tuesday May 9
<tr><!-- idnum = 40295480 --><td>2</TD><td>Siems, Travis</td><td>Tuesday May
<tr><!-- idnum = 47293718 --><td>3</TD><td>Byreddy, Nikhila</td><td>Tuesday
```

Excerpts from 'To Kill a M

```
<page_url href="http://lyle.smu.edu/~fmoore/misc/text/index.php"></page_url>
<HTML><HEAD>
<TITLE>Excerpts from "To Kill a Mockingbird"</TITLE>
</head>
<BODY>
<H1>additional text files to support query implementation</H1>
<UL>
<LI><a href=mockingbird1.html>part 1</A></LI>
<LI><a href=mockingbird2.html>part 2</A></LI>
<LI><a href=mockingbird3.html>part 3</A></LI>
<LI><a href=mockingbird4.html>part 4</A></LI>
<LI><a href=mockingbird5.html>part 5</A></LI>
<LI><a href=magictext.html>hocuspocus</A></LI>
</UL>
<P>
Last Updated: May 06, 2017</BODY>
</HTML>
```

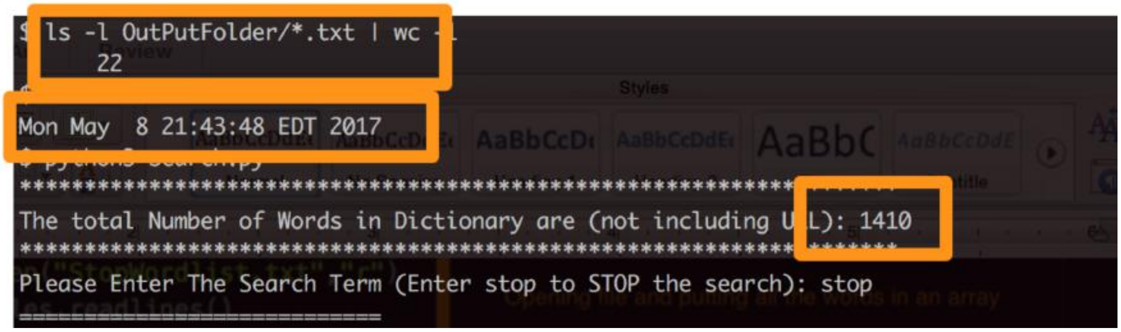
2. You will need a dictionary of words.
   a. What is your definition of "word"? Did it change from project1?

   Answer: In my project the word definition is the token after performing normalization like stop word elimination, punctuation elimination and stemming that is the stemmed words (include both number and words) stored in index in lower case to allow case insensitive searching. **There was no change to my word definition** from the Phase 1 of the project. Same indexing script is used for second phase also without any changes.

   b. How many words are in your dictionary?

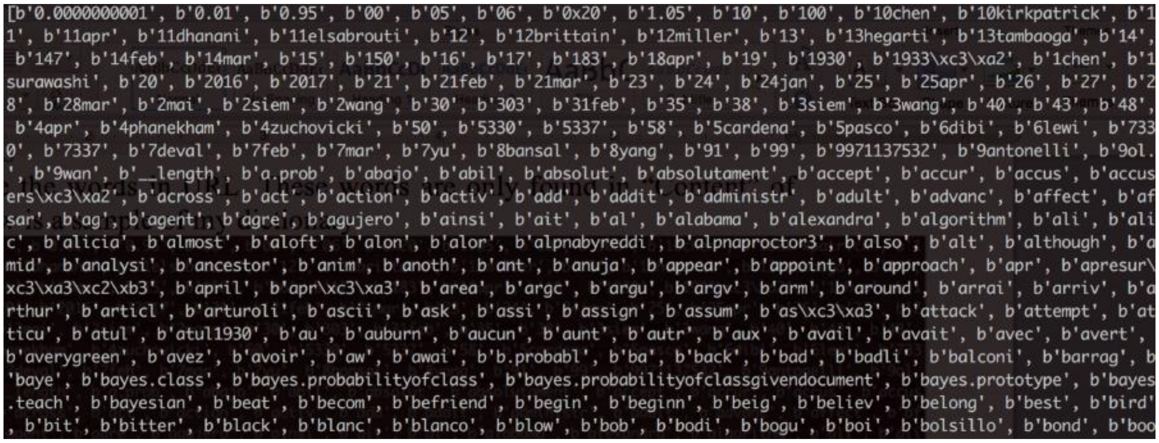
   Answer: At the time of this document Sept May 8 10 pm ET my script downloaded and indexed 22 documents with total unique words of **1,410** which do not include the words in URL. These words are only found in "Content" of files.

Email: somyas@smu.edu
Student Id: 47304053



```
$ ls -l OutPutFolder/*.txt | wc -l
      22

Mon May  8 21:43:48 EDT 2017
************************************************************
The total Number of Words in Dictionary are (not including URL): 1410
************************************************************
Please Enter The Search Term (Enter stop to STOP the search): stop
============================
```

Below is a sample of my dictionary



```
[b'0.0000000001', b'0.01', b'0.95', b'00', b'05', b'06', b'0x20', b'1.05', b'10', b'100', b'10chen', b'10kirkpatrick', b'1
1', b'11apr', b'11dhanani', b'11elsabrouti', b'12', b'12brittain', b'12miller', b'13', b'13hegarti', b'13tambaoga', b'14',
 b'147', b'14feb', b'14mar', b'15', b'150', b'16', b'17', b'183', b'18apr', b'19', b'1930', b'1933\xc3\xa2', b'1chen', b'1
surawashi', b'20', b'2016', b'2017', b'21', b'21feb', b'21mar', b'23', b'24', b'24jan', b'25', b'25apr', b'26', b'27', b'2
8', b'28mar', b'2mai', b'2siem', b'2wang', b'30', b'303', b'31feb', b'35', b'38', b'3siem', b'3wang', b'40', b'43', b'48',
 b'4apr', b'4phanekham', b'4zuchovicki', b'50', b'5330', b'5337', b'58', b'5cardena', b'5pasco', b'6dibi', b'6lewi', b'733
0', b'7337', b'7deval', b'7feb', b'7mar', b'7yu', b'8bansal', b'8yang', b'91', b'99', b'9971137532', b'9antonelli', b'9ol.
', b'9wan', b'__length', b'a.prob', b'abajo', b'abil', b'absolut', b'absolutament', b'accept', b'accur', b'accus', b'accus
ers\xc3\xa2', b'across', b'act', b'action', b'activ', b'add', b'addit', b'administr', b'adult', b'advanc', b'affect', b'af
sar', b'ag', b'agent', b'agre', b'agujero', b'ainsi', b'ait', b'al', b'alabama', b'alexandra', b'algorithm', b'ali', b'ali
c', b'alicia', b'almost', b'aloft', b'alon', b'alor', b'alpnabyreddi', b'alpnaproctor3', b'also', b'alt', b'although', b'a
mid', b'analysi', b'ancestor', b'anim', b'anoth', b'ant', b'anuja', b'appear', b'appoint', b'approach', b'apr', b'apresur\
xc3\xa3\xc2\xb3', b'april', b'apr\xc3\xa3', b'area', b'argc', b'argu', b'argv', b'arm', b'around', b'arrai', b'arriv', b'a
rthur', b'articl', b'arturoli', b'ascii', b'ask', b'assi', b'assign', b'assum', b'as\xc3\xa3', b'attack', b'attempt', b'at
ticu', b'atul', b'atul1930', b'au', b'auburn', b'aucun', b'aunt', b'autr', b'aux', b'avail', b'avait', b'avec', b'avert',
b'averygreen', b'avez', b'avoir', b'aw', b'awai', b'b.probabl', b'ba', b'back', b'bad', b'badli', b'balconi', b'barrag', b
'baye', b'bayes.class', b'bayes.probabilityofclass', b'bayes.probabilityofclassgivendocument', b'bayes.prototype', b'bayes
.teach', b'bayesian', b'beat', b'becom', b'befriend', b'begin', b'beginn', b'beig', b'believ', b'belong', b'best', b'bird'
, b'bit', b'bitter', b'black', b'blanc', b'blanco', b'blow', b'bob', b'bodi', b'bogu', b'boi', b'bolsillo', b'bond', b'boo
```

c. What technique did you use to store your dictionary (fixed size, string array, one-large-string)?

Answer: I used “string array” to store my dictionary. The words are stored in lower case in stem form after removing stop words.

---

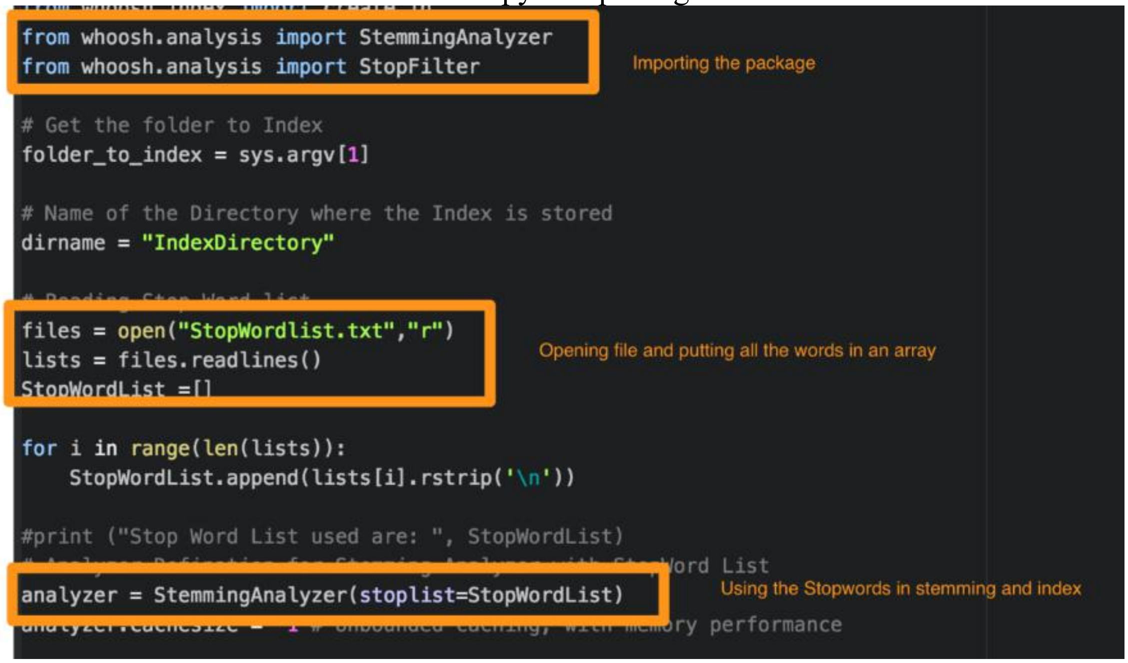
3. For the purpose of this project, you may assume a maximum of 50 documents. You will need to create a word/document frequency matrix to support item 5

a) Remove documents if the content has already been seen.

Answer: I am performing near duplicate detection using k-shingling (k=5) and calculating Jaccard coefficient .
(threshold > 80(duplicate data then discard))
Every Page is checked with all other pages downloaded to make sure the same content is not there.

b) Remove stop words from documents. What list did you use?

Answer: I created a list of words in a file “StopWordlist.txt” which I used as stop words. While storing my words in dictionary I used this file to remove stop words. In order to achieve this I used the python package Whoosh.



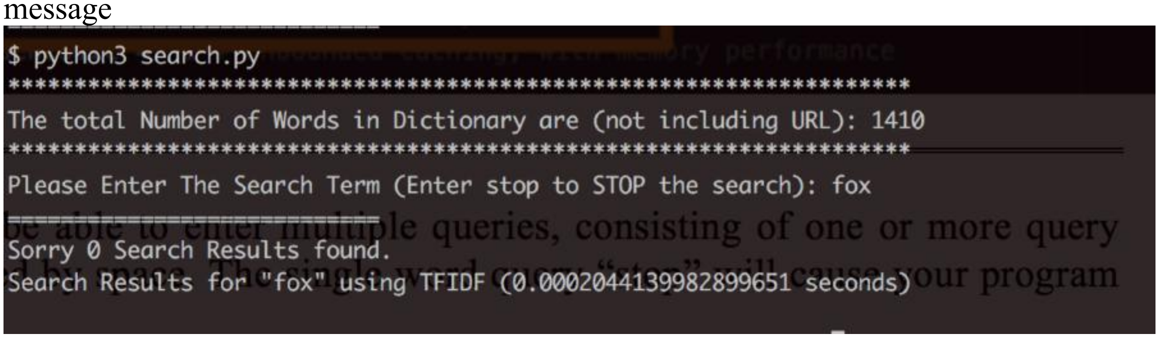
```
from whoosh.analysis import StemmingAnalyzer
from whoosh.analysis import StopFilter

# Get the folder to Index
folder_to_index = sys.argv[1]

# Name of the Directory where the Index is stored
dirname = "IndexDirectory"

files = open("StopWordlist.txt","r")
lists = files.readlines()
StopWordList =[]

for i in range(len(lists)):
    StopWordList.append(lists[i].rstrip('\n'))

#print ("Stop Word List used are: ", StopWordList)
analyzer = StemmingAnalyzer(stoplist=StopWordList)
```

Importing the package

Opening file and putting all the words in an array

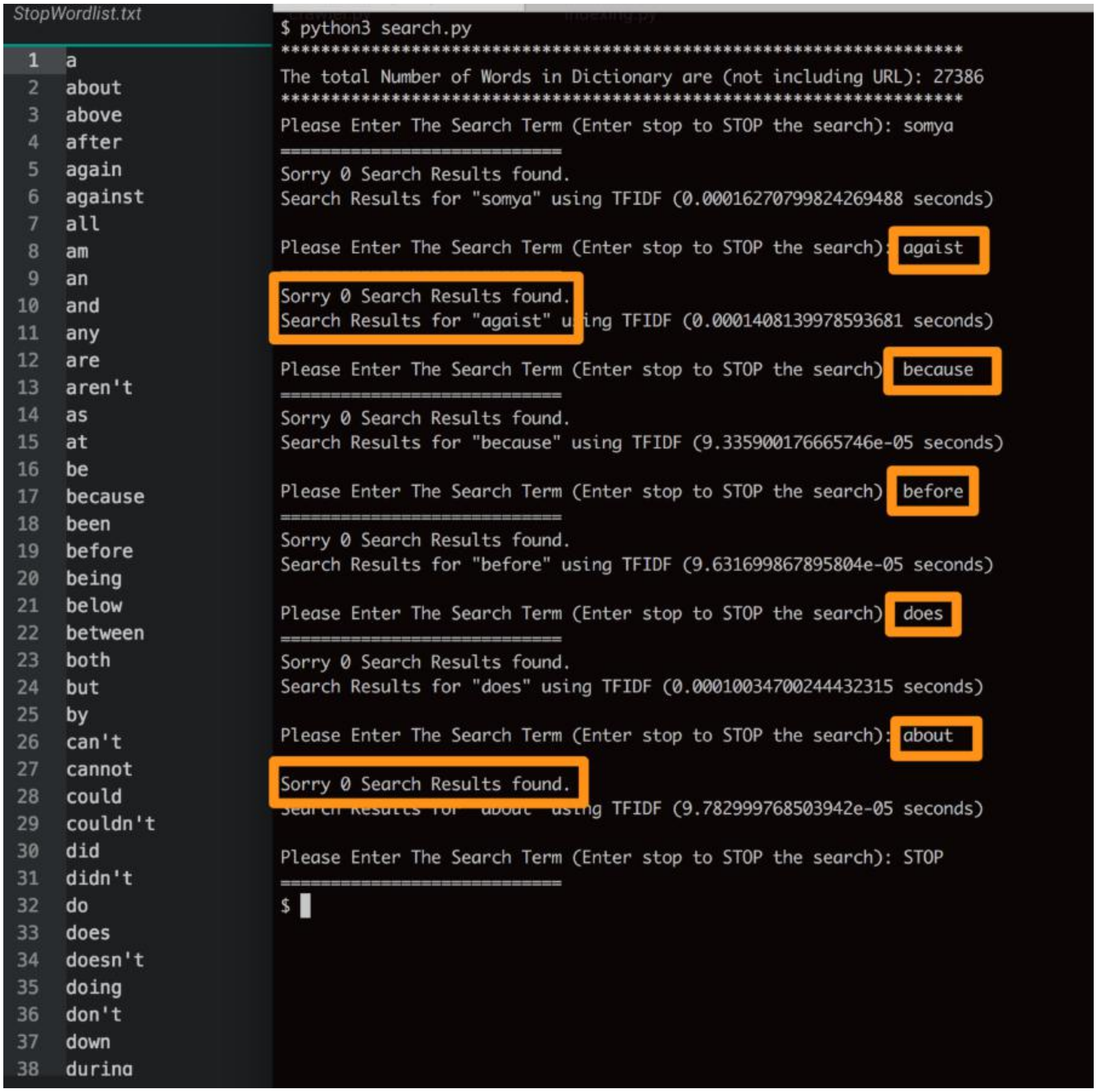
Using the Stopwords in stemming and index

4. The user will be able to enter multiple queries, consisting of one or more query words separated by space. The single word query “stop” will cause your program to stop.
   a) What happens if a user enters a word that is not in the dictionary?
   Answer: If the word/term user entered is not found the user will see the zero result message



```
$ python3 search.py
*****************************************************************
The total Number of Words in Dictionary are (not including URL): 1410
*****************************************************************
Please Enter The Search Term (Enter stop to STOP the search): fox

Sorry 0 Search Results found.
Search Results for "fox" using TFIDF (0.0002044139982899651 seconds)
```

b) What happens if a user enters a stop word?
Answer: If the user enters a stop word that is defined in my “StopWordlist.txt” the user will not get any result returned.
Also if user enter the term “STOP” the script stop execution.



StopWordlist.txt

```
a
about
above
after
again
against
all
am
an
and
any
are
aren't
as
at
be
because
been
before
being
below
between
both
but
by
can't
cannot
could
couldn't
did
didn't
do
does
doesn't
doing
don't
down
during
```

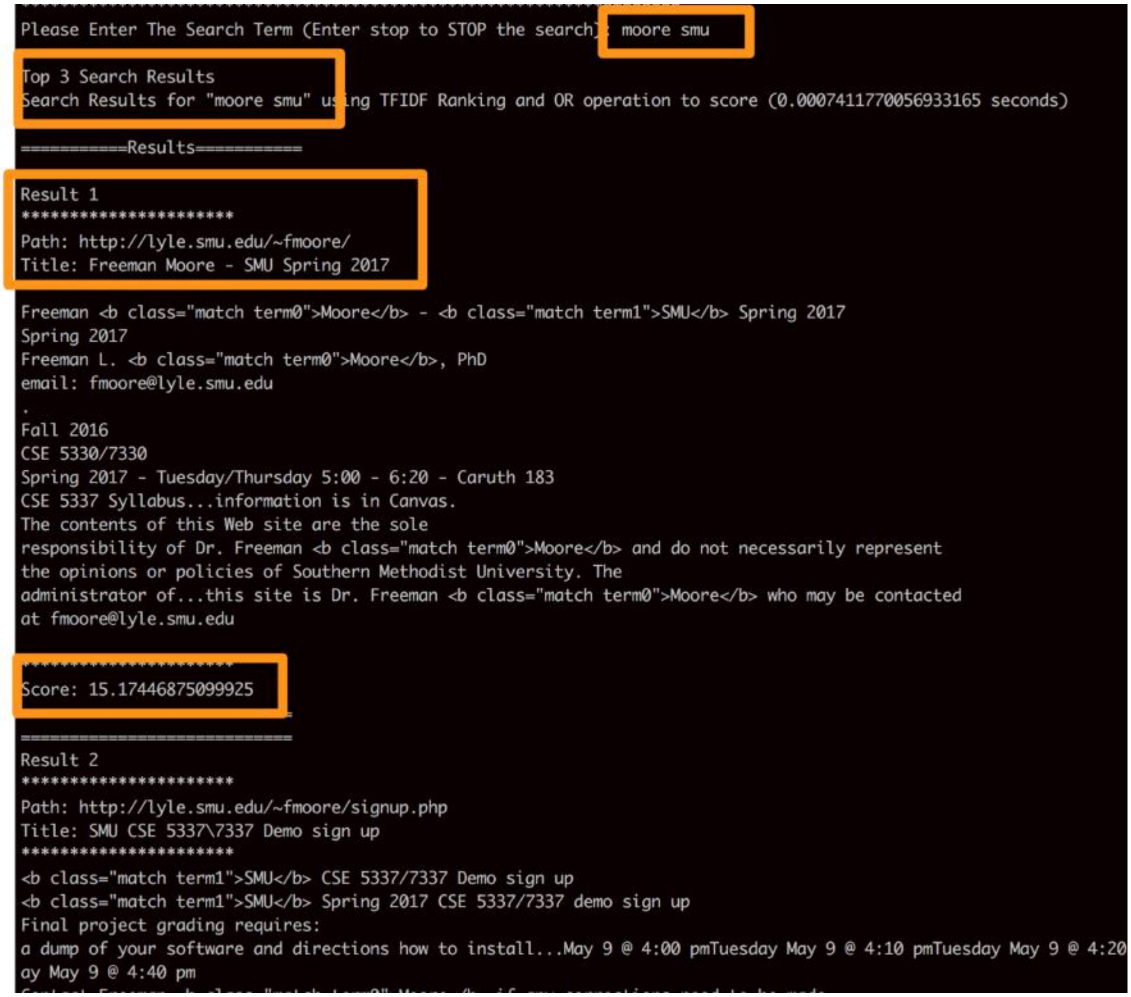
```
$ python3 search.py
*******************************************************************
The total Number of Words in Dictionary are (not including URL): 27386
*******************************************************************
Please Enter The Search Term (Enter stop to STOP the search): somya
============================
Sorry 0 Search Results found.
Search Results for "somya" using TFIDF (0.00016270799824269488 seconds)

Please Enter The Search Term (Enter stop to STOP the search): agaist
============================
Sorry 0 Search Results found.
Search Results for "agaist" using TFIDF (0.0001408139978593681 seconds)

Please Enter The Search Term (Enter stop to STOP the search): because
============================
Sorry 0 Search Results found.
Search Results for "because" using TFIDF (9.335900176665746e-05 seconds)

Please Enter The Search Term (Enter stop to STOP the search): before
============================
Sorry 0 Search Results found.
Search Results for "before" using TFIDF (9.631699867895804e-05 seconds)

Please Enter The Search Term (Enter stop to STOP the search): does
============================
Sorry 0 Search Results found.
Search Results for "does" using TFIDF (0.00010034700244432315 seconds)

Please Enter The Search Term (Enter stop to STOP the search): about
============================
Sorry 0 Search Results found.
Search Results for "about" using TFIDF (9.782999768503942e-05 seconds)

Please Enter The Search Term (Enter stop to STOP the search): STOP
============================
$ 
```

c) A set of queries will be provided.

Answer:

Search Term 1: **“moore smu” fetched 3 results**

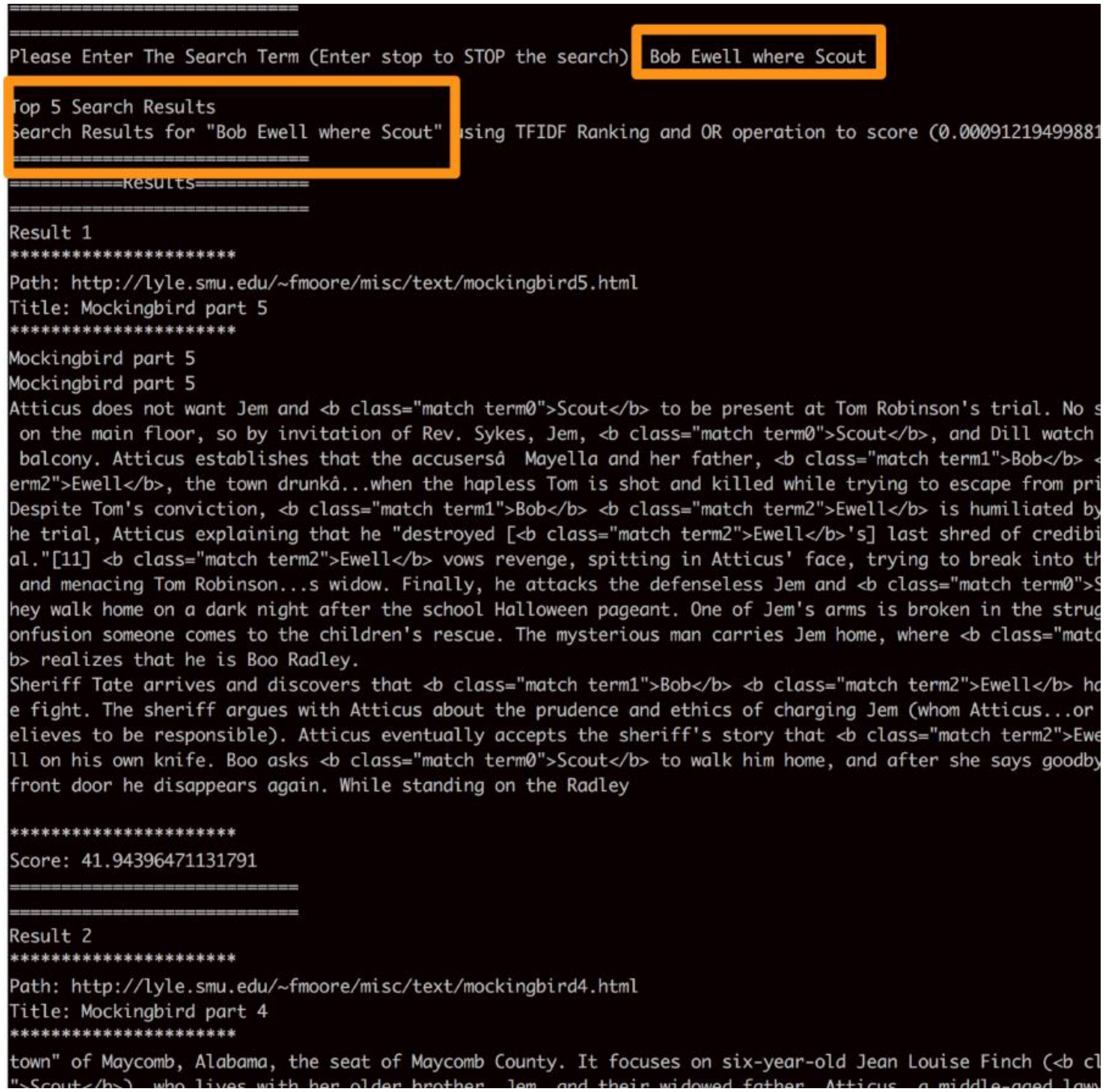
Result 1: http://lyle.smu.edu/~fmoore/
Result 2: http://lyle.smu.edu/~fmoore/signup.php
Result 3: http://lyle.smu.edu/~fmoore/schedule.htm



```
Please Enter The Search Term (Enter stop to STOP the search): moore smu

Top 3 Search Results
Search Results for "moore smu" using TFIDF Ranking and OR operation to score (0.0007411770056933165 seconds)

===========Results===========

Result 1
**********************
Path: http://lyle.smu.edu/~fmoore/
Title: Freeman Moore - SMU Spring 2017

Freeman <b class="match term0">Moore</b> - <b class="match term1">SMU</b> Spring 2017
Spring 2017
Freeman L. <b class="match term0">Moore</b>, PhD
email: fmoore@lyle.smu.edu
.
Fall 2016
CSE 5330/7330
Spring 2017 - Tuesday/Thursday 5:00 - 6:20 - Caruth 183
CSE 5337 Syllabus...information is in Canvas.
The contents of this Web site are the sole
responsibility of Dr. Freeman <b class="match term0">Moore</b> and do not necessarily represent
the opinions or policies of Southern Methodist University. The
administrator of...this site is Dr. Freeman <b class="match term0">Moore</b> who may be contacted
at fmoore@lyle.smu.edu

**********************
Score: 15.17446875099925
=============================

Result 2
**********************
Path: http://lyle.smu.edu/~fmoore/signup.php
Title: SMU CSE 5337\7337 Demo sign up
**********************
<b class="match term1">SMU</b> CSE 5337/7337 Demo sign up
<b class="match term1">SMU</b> Spring 2017 CSE 5337/7337 demo sign up
Final project grading requires:
a dump of your software and directions how to install...May 9 @ 4:00 pmTuesday May 9 @ 4:10 pmTuesday May 9 @ 4:20
ay May 9 @ 4:40 pm
Contact Freeman <b class="match term0">Moore</b> if any corrections need to be made
```

Search Term 2: **“Bob Ewell where Scout” fetched 5 results**

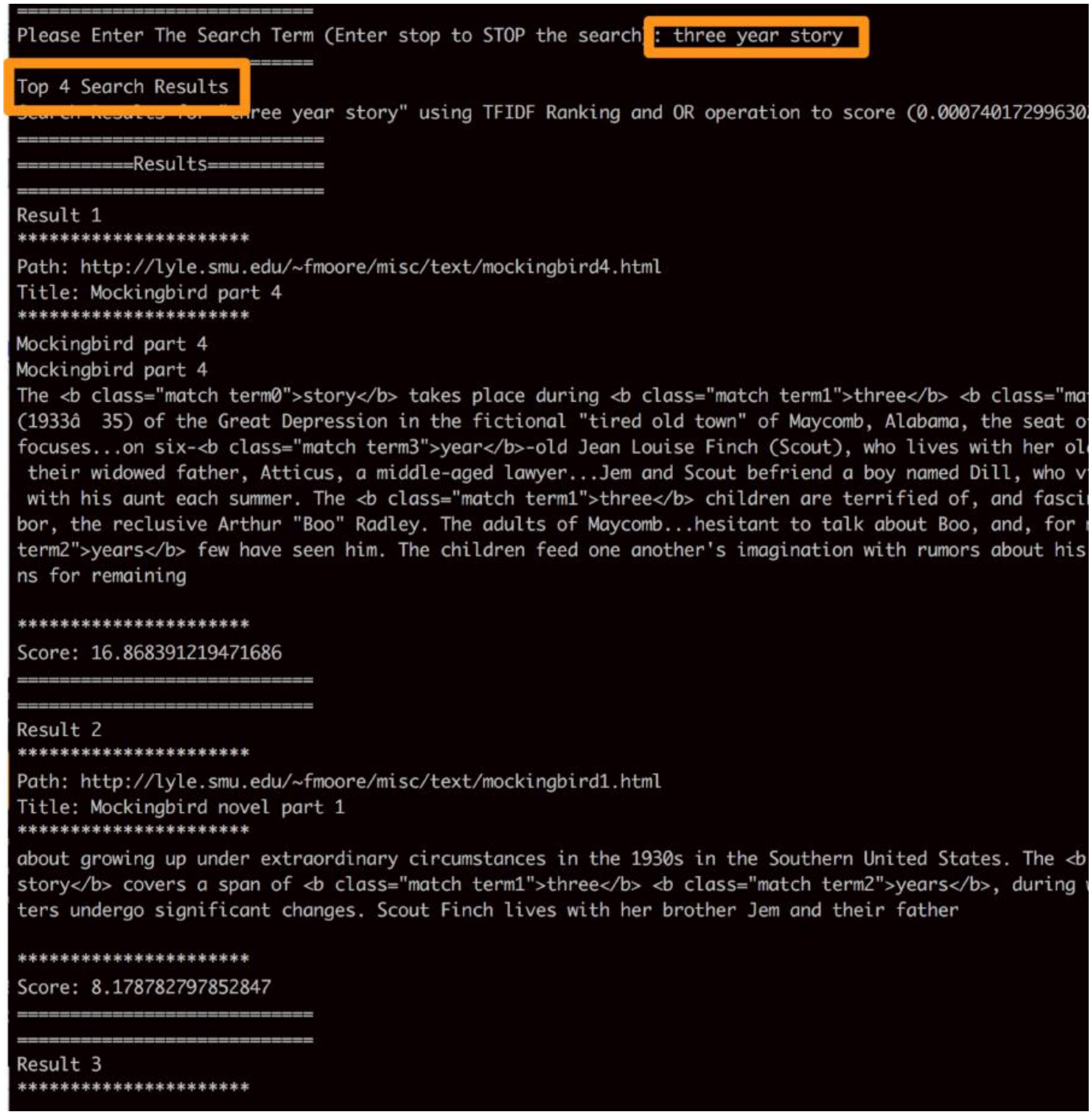
Result 1: http://lyle.smu.edu/~fmoore/misc/text/mockingbird5.html
Result 2: http://lyle.smu.edu/~fmoore/misc/text/mockingbird4.html
Result 3: http://lyle.smu.edu/~fmoore/misc/text/mockingbird3.html
Result 4: http://lyle.smu.edu/~fmoore/misc/text/mockingbird2.html
Result 5: http://lyle.smu.edu/~fmoore/misc/text/mockingbird1.html



```
==============================
==============================
Please Enter The Search Term (Enter stop to STOP the search) Bob Ewell where Scout

Top 5 Search Results
Search Results for "Bob Ewell where Scout" using TFIDF Ranking and OR operation to score (0.00091219499881
==============================
===========Results===========
==============================
Result 1
*********************
Path: http://lyle.smu.edu/~fmoore/misc/text/mockingbird5.html
Title: Mockingbird part 5
*********************
Mockingbird part 5
Mockingbird part 5
Atticus does not want Jem and <b class="match term0">Scout</b> to be present at Tom Robinson's trial. No s
 on the main floor, so by invitation of Rev. Sykes, Jem, <b class="match term0">Scout</b>, and Dill watch
 balcony. Atticus establishes that the accusersâ  Mayella and her father, <b class="match term1">Bob</b> <
erm2">Ewell</b>, the town drunkâ...when the hapless Tom is shot and killed while trying to escape from pri
Despite Tom's conviction, <b class="match term1">Bob</b> <b class="match term2">Ewell</b> is humiliated by
he trial, Atticus explaining that he "destroyed [<b class="match term2">Ewell</b>'s] last shred of credibi
al."[11] <b class="match term2">Ewell</b> vows revenge, spitting in Atticus' face, trying to break into th
 and menacing Tom Robinson...s widow. Finally, he attacks the defenseless Jem and <b class="match term0">S
hey walk home on a dark night after the school Halloween pageant. One of Jem's arms is broken in the strug
onfusion someone comes to the children's rescue. The mysterious man carries Jem home, where <b class="matc
b> realizes that he is Boo Radley.
Sheriff Tate arrives and discovers that <b class="match term1">Bob</b> <b class="match term2">Ewell</b> ha
e fight. The sheriff argues with Atticus about the prudence and ethics of charging Jem (whom Atticus...or
elieves to be responsible). Atticus eventually accepts the sheriff's story that <b class="match term2">Ewe
ll on his own knife. Boo asks <b class="match term0">Scout</b> to walk him home, and after she says goodby
front door he disappears again. While standing on the Radley

*********************
Score: 41.94396471131791
==============================
==============================
Result 2
*********************
Path: http://lyle.smu.edu/~fmoore/misc/text/mockingbird4.html
Title: Mockingbird part 4
*********************
town" of Maycomb, Alabama, the seat of Maycomb County. It focuses on six-year-old Jean Louise Finch (<b cl
">Scout</b>), who lives with her older brother, Jem, and their widowed father, Atticus, a middle-aged lawy
```

Search Term 3: **"three year story" fetched 4 results**

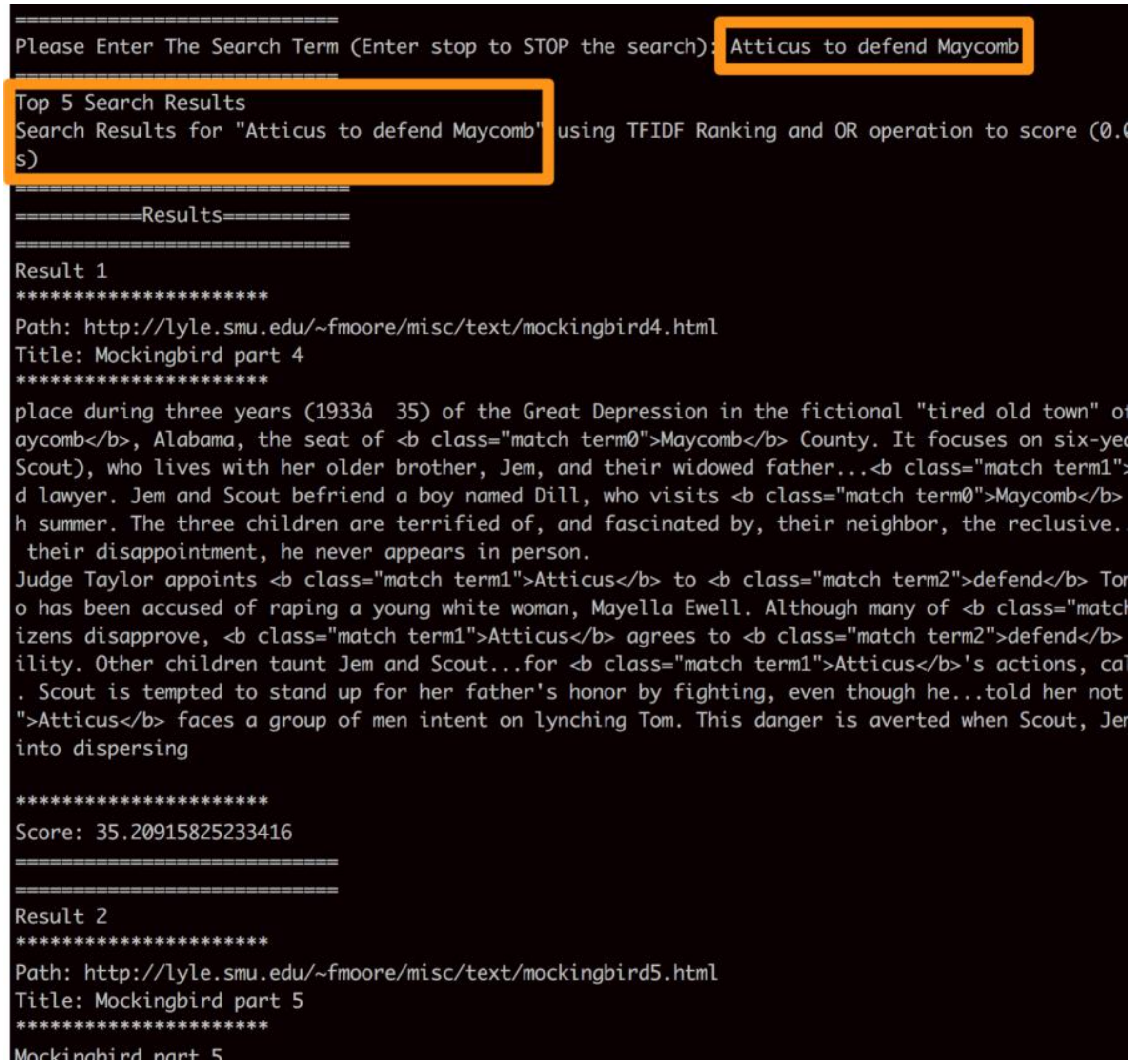
Result 1: http://lyle.smu.edu/~fmoore/misc/text/mockingbird4.html
Result 2: http://lyle.smu.edu/~fmoore/misc/text/mockingbird1.html
Result 3: http://lyle.smu.edu/~fmoore/misc/text/mockingbird2.html
Result 4: http://lyle.smu.edu/~fmoore/misc/text/mockingbird5.html



```
============================
Please Enter The Search Term (Enter stop to STOP the search : three year story
============================
Top 4 Search Results
Search Results for "three year story" using TFIDF Ranking and OR operation to score (0.00074017299630
=============================
===========Results===========
=============================
Result 1
**********************
Path: http://lyle.smu.edu/~fmoore/misc/text/mockingbird4.html
Title: Mockingbird part 4
**********************
Mockingbird part 4
Mockingbird part 4
The <b class="match term0">story</b> takes place during <b class="match term1">three</b> <b class="ma
(1933â  35) of the Great Depression in the fictional "tired old town" of Maycomb, Alabama, the seat o
focuses...on six-<b class="match term3">year</b>-old Jean Louise Finch (Scout), who lives with her ol
 their widowed father, Atticus, a middle-aged lawyer...Jem and Scout befriend a boy named Dill, who v
 with his aunt each summer. The <b class="match term1">three</b> children are terrified of, and fasci
bor, the reclusive Arthur "Boo" Radley. The adults of Maycomb...hesitant to talk about Boo, and, for 
term2">years</b> few have seen him. The children feed one another's imagination with rumors about his
ns for remaining

**********************
Score: 16.868391219471686
============================
============================
Result 2
**********************
Path: http://lyle.smu.edu/~fmoore/misc/text/mockingbird1.html
Title: Mockingbird novel part 1
**********************
about growing up under extraordinary circumstances in the 1930s in the Southern United States. The <b
story</b> covers a span of <b class="match term1">three</b> <b class="match term2">years</b>, during 
ters undergo significant changes. Scout Finch lives with her brother Jem and their father

**********************
Score: 8.178782797852847
============================
============================
Result 3
**********************
```

Search Term 4: **"Atticus to defend Maycomb" fetched 5 results**

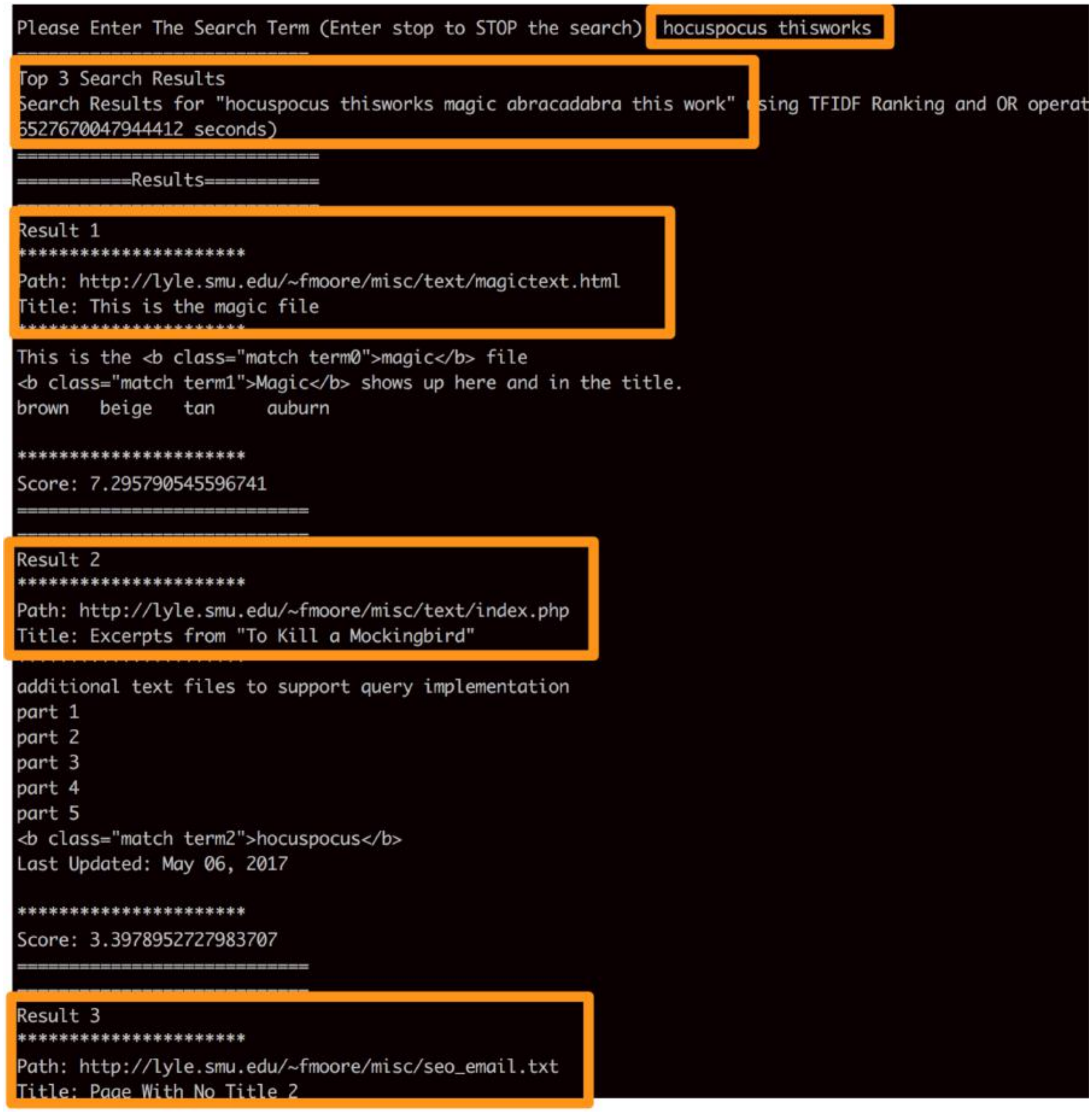
Result 1: http://lyle.smu.edu/~fmoore/misc/text/mockingbird4.html
Result 2: http://lyle.smu.edu/~fmoore/misc/text/mockingbird5.html
Result 3: http://lyle.smu.edu/~fmoore/misc/text/mockingbird1.html
Result 4: http://lyle.smu.edu/~fmoore/misc/text/mockingbird2.html
Result 5: http://lyle.smu.edu/~fmoore/misc/text/mockingbird3.html



```
============================
Please Enter The Search Term (Enter stop to STOP the search): Atticus to defend Maycomb
============================
Top 5 Search Results
Search Results for "Atticus to defend Maycomb" using TFIDF Ranking and OR operation to score (0.
s)
=============================
===========Results===========
=============================
Result 1
**********************
Path: http://lyle.smu.edu/~fmoore/misc/text/mockingbird4.html
Title: Mockingbird part 4
**********************
place during three years (1933â  35) of the Great Depression in the fictional "tired old town" o
aycomb</b>, Alabama, the seat of <b class="match term0">Maycomb</b> County. It focuses on six-ye
Scout), who lives with her older brother, Jem, and their widowed father...<b class="match term1"
d lawyer. Jem and Scout befriend a boy named Dill, who visits <b class="match term0">Maycomb</b>
h summer. The three children are terrified of, and fascinated by, their neighbor, the reclusive.
 their disappointment, he never appears in person.
Judge Taylor appoints <b class="match term1">Atticus</b> to <b class="match term2">defend</b> To
o has been accused of raping a young white woman, Mayella Ewell. Although many of <b class="matc
izens disapprove, <b class="match term1">Atticus</b> agrees to <b class="match term2">defend</b>
ility. Other children taunt Jem and Scout...for <b class="match term1">Atticus</b>'s actions, ca
. Scout is tempted to stand up for her father's honor by fighting, even though he...told her not
">Atticus</b> faces a group of men intent on lynching Tom. This danger is averted when Scout, Je
into dispersing

**********************
Score: 35.20915825233416
============================
============================
Result 2
**********************
Path: http://lyle.smu.edu/~fmoore/misc/text/mockingbird5.html
Title: Mockingbird part 5
**********************
Mockingbird part 5
```

Search Term 5: “**hocuspocus thisworks**” fetched 3 results. The search term “hocus-pocus thisworks” fetched only 1 result so using dictionary the term was expanded and it fetched total of 3 records

Result 1: http://lyle.smu.edu/~fmoore/misc/text/magictext.html
Result 2: http://lyle.smu.edu/~fmoore/misc/text/index.php
Result 3: http://lyle.smu.edu/~fmoore/misc/seo_email.txt



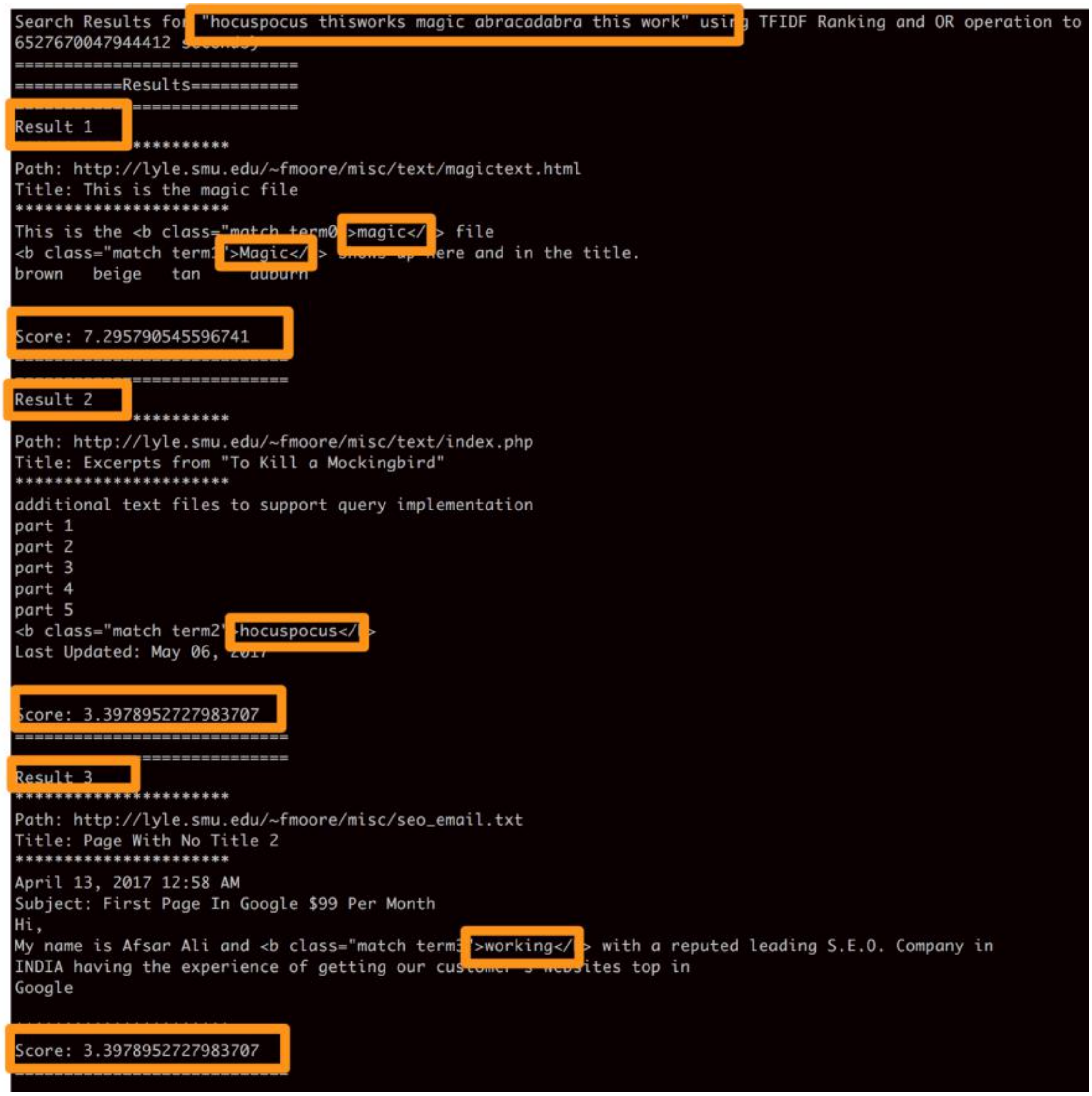
```
Please Enter The Search Term (Enter stop to STOP the search) hocuspocus thisworks
----------------------------
Top 3 Search Results
Search Results for "hocuspocus thisworks magic abracadabra this work" using TFIDF Ranking and OR operat
6527670047944412 seconds)
=============================
===========Results===========
-----------------------------
Result 1
**********************
Path: http://lyle.smu.edu/~fmoore/misc/text/magictext.html
Title: This is the magic file
**********************
This is the <b class="match term0">magic</b> file
<b class="match term1">Magic</b> shows up here and in the title.
brown    beige    tan      auburn

**********************
Score: 7.295790545596741
============================
----------------------------
Result 2
**********************
Path: http://lyle.smu.edu/~fmoore/misc/text/index.php
Title: Excerpts from "To Kill a Mockingbird"
**********************
additional text files to support query implementation
part 1
part 2
part 3
part 4
part 5
<b class="match term2">hocuspocus</b>
Last Updated: May 06, 2017

**********************
Score: 3.3978952727983707
============================
----------------------------
Result 3
**********************
Path: http://lyle.smu.edu/~fmoore/misc/seo_email.txt
Title: Page With No Title 2
```

5. Implement the cosine similarity of the query against all documents.
a) If any of the query words appear in the <title>, add 0.5 to the query score.
Answer:
I have used whoosh **scoring** package and imported **TF_IDF** class scoring for this project. The class calculates cosine similarity of the query against all documents it has indexed. I added the logic to add the additional score of 0.5 if the query is found in title of the document.

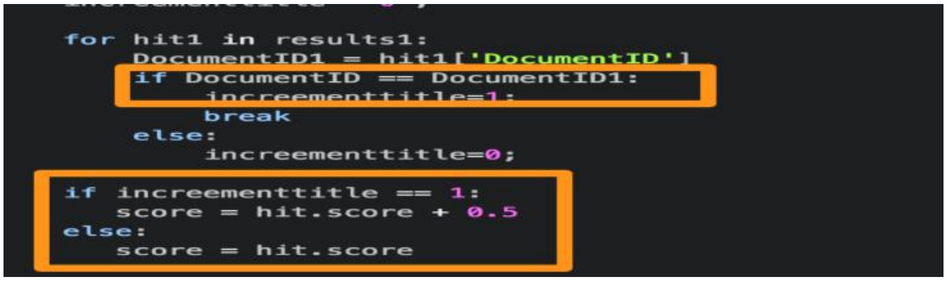
In order to verify this I did the search for "hocuspocus thisworks". The query fetches three matched document:

Result 1: http://lyle.smu.edu/~fmoore/misc/text/magictext.html
Result 2: http://lyle.smu.edu/~fmoore/misc/text/index.php
Result 3: http://lyle.smu.edu/~fmoore/misc/seo_email.txt

The Result 2 and 3 have hocuspocus and work respectfully as hit so each document got a score of 3.397. The Result 1 has two matches for magic in the page so its score was 3.397*2 which is 6.795. One of the match for the Result 1 is in the title of the document "This is the magic file" so the score was updated to 7.295 (6.795 + 0.5 = 7.295)



```
Search Results for "hocuspocus thisworks magic abracadabra this work" using TFIDF Ranking and OR operation to
6527670047944412 seconds)
==============================
===========Results===========
==============================
Result 1
**********************
Path: http://lyle.smu.edu/~fmoore/misc/text/magictext.html
Title: This is the magic file
**********************
This is the <b class="match term0">magic</b> file
<b class="match term1">Magic</b> shows up here and in the title.
brown   beige   tan     auburn

Score: 7.295790545596741
==============================
==============================
Result 2
**********************
Path: http://lyle.smu.edu/~fmoore/misc/text/index.php
Title: Excerpts from "To Kill a Mockingbird"
**********************
additional text files to support query implementation
part 1
part 2
part 3
part 4
part 5
<b class="match term2">hocuspocus</b>
Last Updated: May 06, 2017

Score: 3.3978952727983707
==============================
==============================
Result 3
**********************
Path: http://lyle.smu.edu/~fmoore/misc/seo_email.txt
Title: Page With No Title 2
**********************
April 13, 2017 12:58 AM
Subject: First Page In Google $99 Per Month
Hi,
My name is Afsar Ali and <b class="match term3">working</b> with a reputed leading S.E.O. Company in
INDIA having the experience of getting our customer's websites top in
Google

**********************
Score: 3.3978952727983707
==============================
```



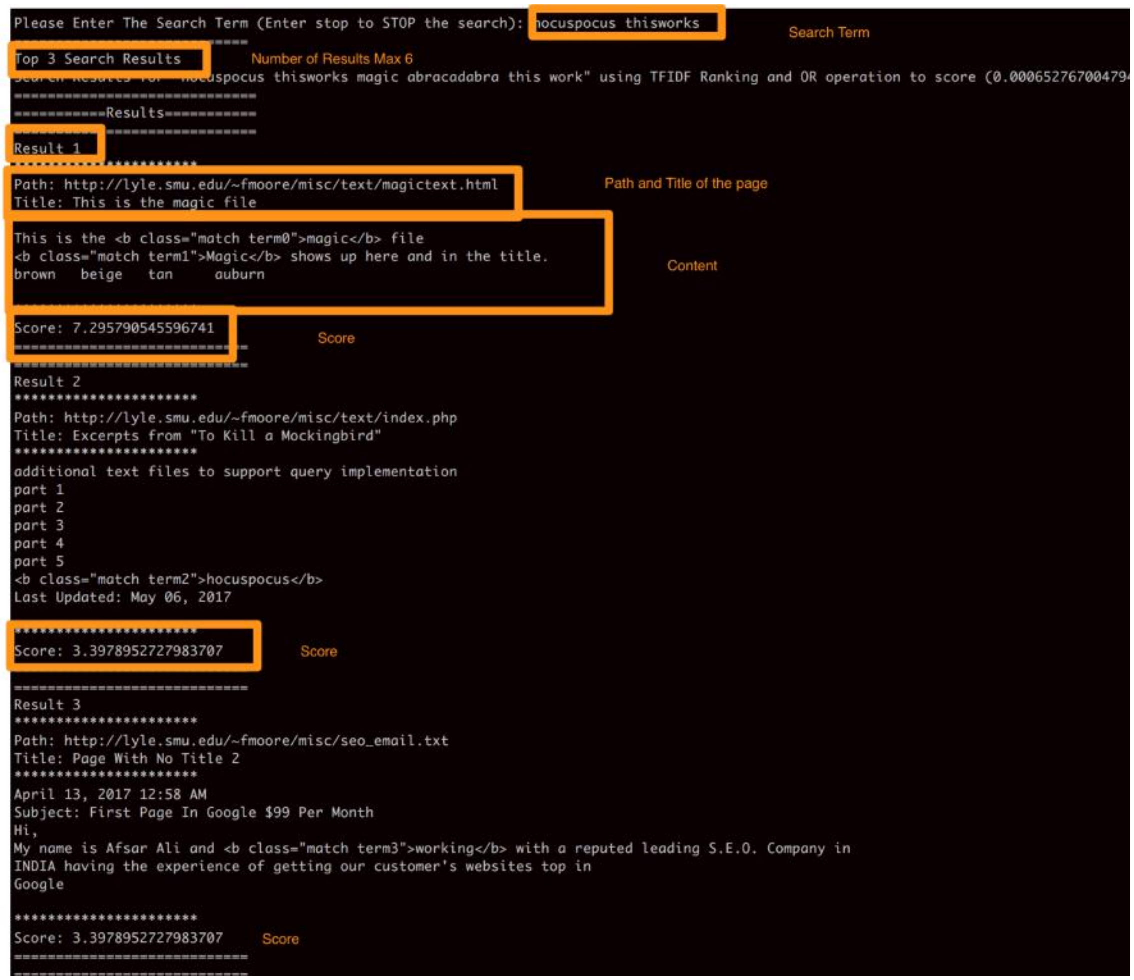
```
    increementtitle = 0;

    for hit1 in results1:
        DocumentID1 = hit1['DocumentID']
        if DocumentID == DocumentID1:
            increementtitle=1;
            break
        else:
            increementtitle=0;

    if increementtitle == 1:
        score = hit.score + 0.5
    else:
        score = hit.score
```

b) Display the similarity measure, document URL, and document title in descending numerical order for the top 6 results.

Answer:

For each result found I do display:

1. Path
2. Title
3. Matched Content
4. Score (similarity measure)

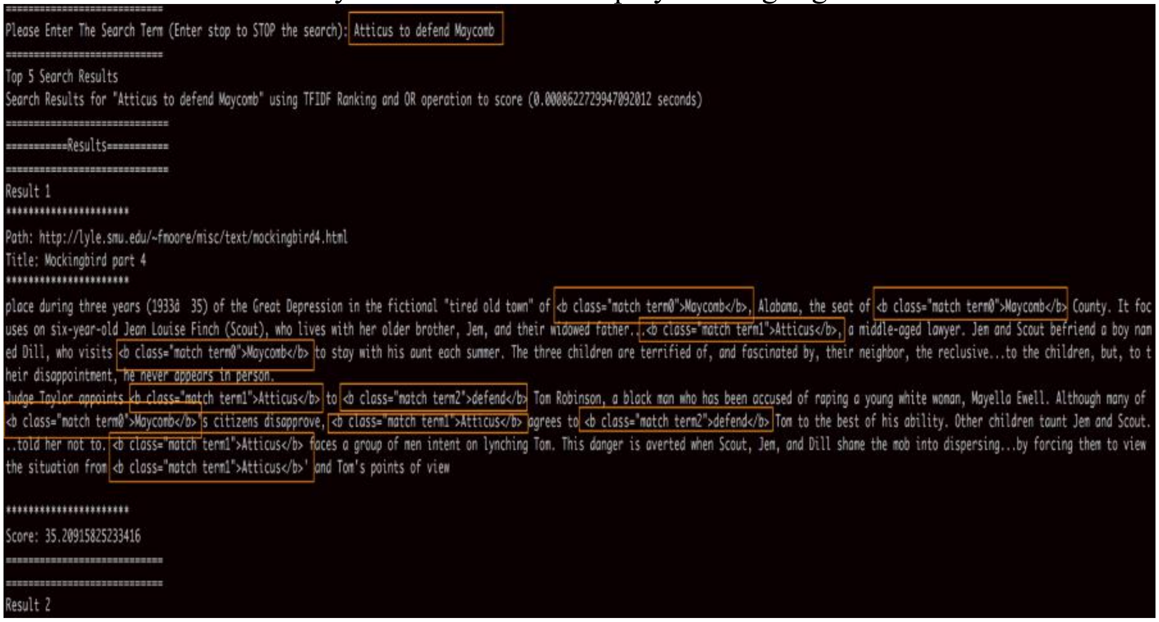


```
Please Enter The Search Term (Enter stop to STOP the search): hocuspocus thisworks
============================
Top 3 Search Results
Search Results for "hocuspocus thisworks magic abracadabra this work" using TFIDF Ranking and OR operation to score (0.00065276700479
==============================
===========Results===========
==============================
Result 1
**********************
Path: http://lyle.smu.edu/~fmoore/misc/text/magictext.html
Title: This is the magic file
**********************
This is the <b class="match term0">magic</b> file
<b class="match term1">Magic</b> shows up here and in the title.
brown   beige   tan     auburn

**********************
Score: 7.295790545596741
============================
============================
Result 2
**********************
Path: http://lyle.smu.edu/~fmoore/misc/text/index.php
Title: Excerpts from "To Kill a Mockingbird"
**********************
additional text files to support query implementation
part 1
part 2
part 3
part 4
part 5
<b class="match term2">hocuspocus</b>
Last Updated: May 06, 2017

**********************
Score: 3.3978952727983707

============================
Result 3
**********************
Path: http://lyle.smu.edu/~fmoore/misc/seo_email.txt
Title: Page With No Title 2
**********************
April 13, 2017 12:58 AM
Subject: First Page In Google $99 Per Month
Hi,
My name is Afsar Ali and <b class="match term3">working</b> with a reputed leading S.E.O. Company in
INDIA having the experience of getting our customer's websites top in
Google

**********************
Score: 3.3978952727983707
============================
============================
```

6. Include in the display, the first 20 words of the document

Answer:

Each result that I get I highlight the hit using whoosh package class “highlight”. Once the script finds a matched word of the term it list the 100 characters (assuming average of 5 character per word) around it. For e.g. in below screen shot I searched for “Atticus to defend Maycomb” and result displays the highlighted words.

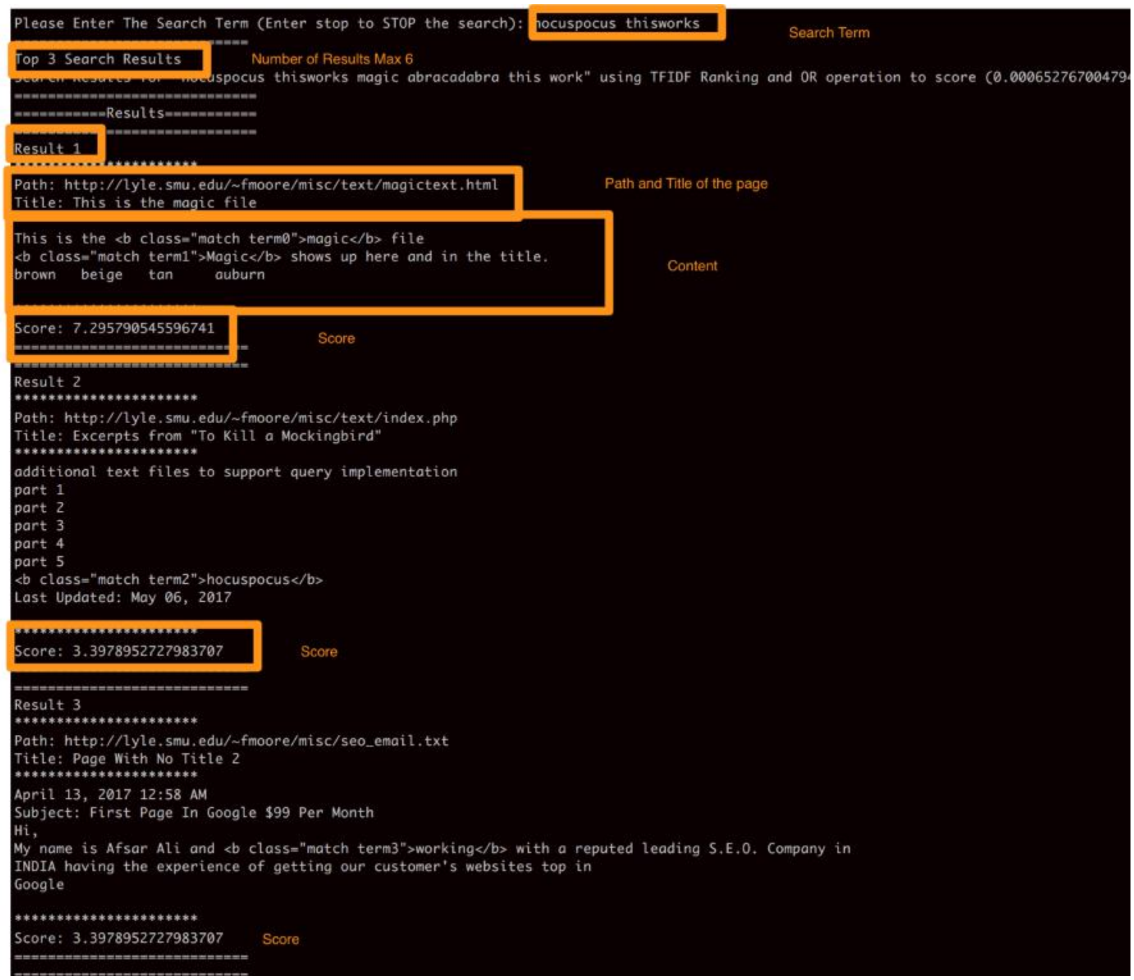


```
============================
Please Enter The Search Term (Enter stop to STOP the search): Atticus to defend Maycomb
============================
Top 5 Search Results
Search Results for "Atticus to defend Maycomb" using TFIDF Ranking and OR operation to score (0.0008622729947092012 seconds)
=============================
===========Results===========
=============================
Result 1
*********************
Path: http://lyle.smu.edu/~fmoore/misc/text/mockingbird4.html
Title: Mockingbird part 4
*********************
place during three years (1933â  35) of the Great Depression in the fictional "tired old town" of <b class="match term0">Maycomb</b>, Alabama, the seat of <b class="match term0">Maycomb</b> County. It foc
uses on six-year-old Jean Louise Finch (Scout), who lives with her older brother, Jem, and their widowed father...<b class="match term1">Atticus</b>, a middle-aged lawyer. Jem and Scout befriend a boy nam
ed Dill, who visits <b class="match term0">Maycomb</b> to stay with his aunt each summer. The three children are terrified of, and fascinated by, their neighbor, the reclusive...to the children, but, to t
heir disappointment, he never appears in person.
Judge Taylor appoints <b class="match term1">Atticus</b> to <b class="match term2">defend</b> Tom Robinson, a black man who has been accused of raping a young white woman, Mayella Ewell. Although many of
<b class="match term0">Maycomb</b>'s citizens disapprove, <b class="match term1">Atticus</b> agrees to <b class="match term2">defend</b> Tom to the best of his ability. Other children taunt Jem and Scout.
..told her not to. <b class="match term1">Atticus</b> faces a group of men intent on lynching Tom. This danger is averted when Scout, Jem, and Dill shame the mob into dispersing...by forcing them to view
the situation from <b class="match term1">Atticus</b>' and Tom's points of view

*********************
Score: 35.20915825233416
============================
============================
Result 2
```

7. If less than N/2 documents are returned for a query, rerun the query using thesaurus expansion. A list of words, along with 1 – 3 synonyms will be provided.

Answer:

If the query of user fetches less than 3 records then I use the dictionary that is provided. The script expands the query to include all the words and alternate provided. For e.g. in below for query the “hocuspocus thisworks” there are only 1 results but using dictionary we get three result back.



```
Please Enter The Search Term (Enter stop to STOP the search): hocuspocus thisworks
============================
Top 3 Search Results
Search Results for "hocuspocus thisworks magic abracadabra this work" using TFIDF Ranking and OR operation to score (0.00065276700479
============================
===========Results===========
============================
Result 1
**********************
Path: http://lyle.smu.edu/~fmoore/misc/text/magictext.html
Title: This is the magic file
**********************
This is the <b class="match term0">magic</b> file
<b class="match term1">Magic</b> shows up here and in the title.
brown   beige   tan     auburn

**********************
Score: 7.295790545596741
============================
============================
Result 2
**********************
Path: http://lyle.smu.edu/~fmoore/misc/text/index.php
Title: Excerpts from "To Kill a Mockingbird"
**********************
additional text files to support query implementation
part 1
part 2
part 3
part 4
part 5
<b class="match term2">hocuspocus</b>
Last Updated: May 06, 2017

**********************
Score: 3.3978952727983707
============================
============================
Result 3
**********************
Path: http://lyle.smu.edu/~fmoore/misc/seo_email.txt
Title: Page With No Title 2
**********************
April 13, 2017 12:58 AM
Subject: First Page In Google $99 Per Month
Hi,
My name is Afsar Ali and <b class="match term3">working</b> with a reputed leading S.E.O. Company in
INDIA having the experience of getting our customer's websites top in
Google

**********************
Score: 3.3978952727983707
============================
============================
```

Search Term

Number of Results Max 6

Path and Title of the page

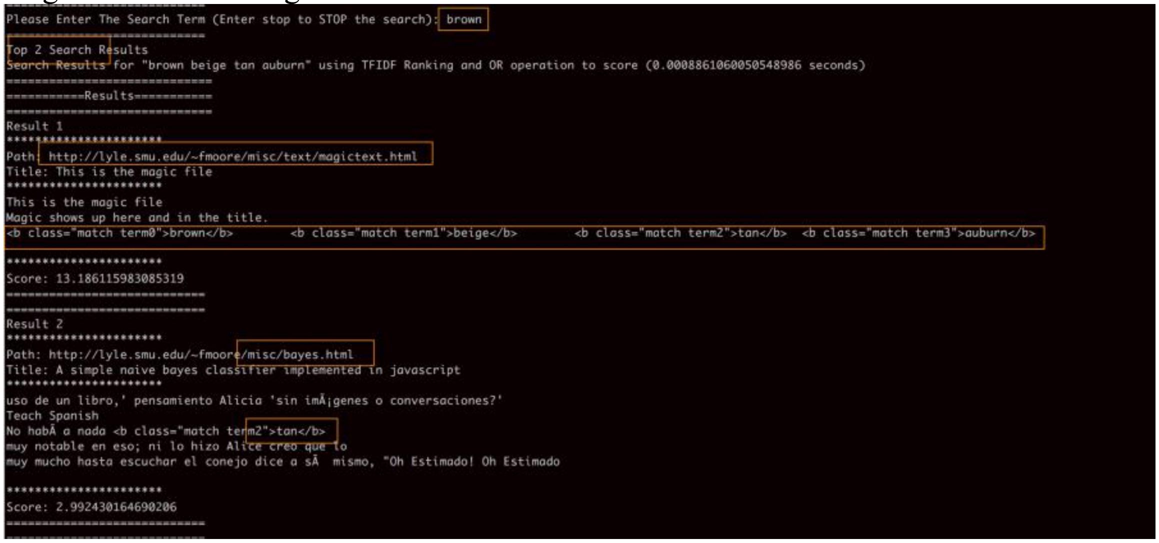
Content

Score

Score

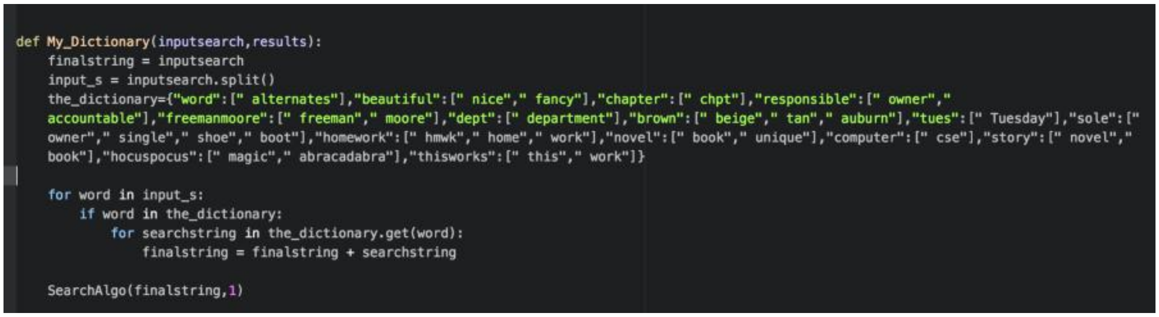
Score

For query “brown” there is one result but when the query is expanded to “brown beige tan auburn” we get two hit.



```
============================
Please Enter The Search Term (Enter stop to STOP the search): brown
============================
Top 2 Search Results
Search Results for "brown beige tan auburn" using TFIDF Ranking and OR operation to score (0.0008861060050548986 seconds)
============================
===========Results===========
============================
Result 1
**********************
Path: http://lyle.smu.edu/~fmoore/misc/text/magictext.html
Title: This is the magic file
**********************
This is the magic file
Magic shows up here and in the title.
<b class="match term0">brown</b>        <b class="match term1">beige</b>        <b class="match term2">tan</b>  <b class="match term3">auburn</b>

**********************
Score: 13.186115983085319
============================
============================
Result 2
**********************
Path: http://lyle.smu.edu/~fmoore/misc/bayes.html
Title: A simple naive bayes classifier implemented in javascript
**********************
uso de un libro,' pensamiento Alicia 'sin imÃ¡genes o conversaciones?'
Teach Spanish
No habÃ a nada <b class="match term2">tan</b>
muy notable en eso; ni lo hizo Alice creo que lo
muy mucho hasta escuchar el conejo dice a sÃ  mismo, "Oh Estimado! Oh Estimado

**********************
Score: 2.992430164690206
============================
============================
```

Below is defination for my dictionary for query expansion.



```
def My_Dictionary(inputsearch,results):
    finalstring = inputsearch
    input_s = inputsearch.split()
    the_dictionary={"word":[" alternates"],"beautiful":[" nice"," fancy"],"chapter":[" chpt"],"responsible":[" owner","
    accountable"],"freemanmoore":[" freeman"," moore"],"dept":[" department"],"brown":[" beige"," tan"," auburn"],"tues":[" Tuesday"],"sole":["
    owner"," single"," shoe"," boot"],"homework":[" hmwk"," home"," work"],"novel":[" book"," unique"],"computer":[" cse"],"story":[" novel","
    book"],"hocuspocus":[" magic"," abracadabra"],"thisworks":[" this"," work"]}

    for word in input_s:
        if word in the_dictionary:
            for searchstring in the_dictionary.get(word):
                finalstring = finalstring + searchstring

    SearchAlgo(finalstring,1)
```