Task Results

Description

1. **Definition of "word"**

   My definition a "word" is a string of non-space character(s), which begin with an alphabetic character and end with an alphabetic or numeric character. It can cantain special characters. And if a token does not begin with an alphabetic character, the program will remove those until it is alphabetic or null.

2. **Description of data**

   - 40 files
   - 1190 words
   - Data Structures of Crawler

     For the crawler, I use a **list** to hold objects of Class Page(information of page), and maintain a list of links as frontier queue which contains links will be crawled. Also has other lists for recording the links going out of the test data, broken links and non-text links. Besides, I use **set** that contain already seen content to detect exact duplicate content in crawled pages.

     For the tokenizaton part, I use **lists** to hold the tokenized words(tokens) and stemmed tokens. And use a **dictionary** to store the postings lists of crawled content.

     The structure of this dictionary is like:

     `Tokenpos{token1{doc2:[pos1,pos2...],doc2:[pos1,pos2...]...},token2{doc2:[pos1,pos2...],doc2:[pos1,pos2...]...}}`

   - Data Structures of Query Engine

     (What I changed to support the second part of this project)

     - Postingslists Dictionary
       - Gets data from `postingslists.xlsx`
       - Structure:

```
self.postingList = {
  'term1': {
    docId1: [pos1,pos2,...],
    docId2: [pos1,pos2,...],
    ...
  },
  ...
}
```

- Term Frequency Vector Dictionary
  - Gets data from `frequencyMatrix.xlsx`
  - Structure:

```
self.frequency = {
  'term1': array([tf1, tf2, ...]), # [0:40]
  'term2': array([tf1, tf2, ...]),
  ...
}
```

- Term Frequency Matrix Dictionary
  - Gets data from `frequencyMatrix.xlsx`
  - Structure:

```
self.termFrequency = {
  docId1: array([tf1, tf2, ...]), # [0:1190]
  docId2: array([tf1, tf2, ...]),
  ...
}
```

- Document Frequency Dictionary
  - Gets data from `frequencyMatrix.xlsx`
  - Structure:

```
self.df = {
  'term1': df,
  'term2': df,
  ...
}
```

- Pages Dictionary
  - Gets data from `all_pages.xlsx`
  - Structure:

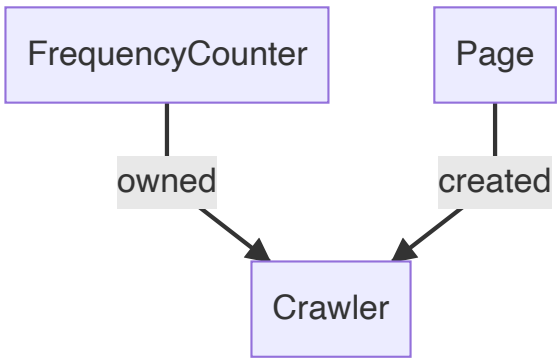
```
self.pages = {
  # Page(url,date,title,content_len,content)
  docId1: Page(),
  docId2: Page(),
  ...
}
```

- Theasurus Dictionary
  - Get data from `theasurus.xlsx`
  - Structure:

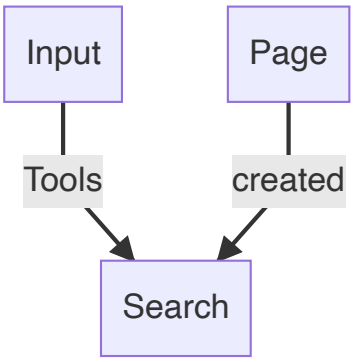
```
self.theasurus = {
  'term1': arrary(['theasurus1','theasurus2',...]),
  'term2': arrary(['theasurus1','theasurus2',...]),
  ...
}
```

3. **Description of the key architecture**

   My crawler consists of three classes. There are Crawler, FrequencyCounter and Page. A Crawler owns one frequency counter(FrequencyCounter) and can creates several objects of Page during crawling process.



   My search engine consists of three classes. There are Search, Input and Page. Object "Search" uses "Input" as a tool class to get the data that has been crawled and stored, and converted into a data structure that supports search engines. Objects of Page is part of the data structure. An Object of Page stores all the information about a page.



Demonstration and explanation

1. Start

```
(base) evelynxu@jos-MacBook-Pro-2 workplace_vscode %  /usr/bin/env /Users/evelynxu/opt/anaconda3/bin/python /Users/evelynxu/.vscode/extensions/ms-python.python-2021.10.1365161279
/pythonFiles/lib/python/debugpy/launcher 49510 -- /Users/evelynxu/Documents/workplace_vscode/WebSearch/Hw/Proj2.1/Main.py
CS 7337 project written by Yiwen Xu
Crawling complete, indexed 40 files. Stemmed dictionary has 1190 words

Query? 
```

2. If user enters a word that is not in the dictionary

   A notice will pop up and ask to re-enter the new query.

```
Query? aaa
No matched document. Please Try other query.

Query? 
```

3. Entering "words" per my definition

   Accroding to my definition, A word can cantain special characters. And if a token does not begin with an alphabetic character, the program will remove those until it is alphabetic or null.

```
Query? $$moore
1 documents match, displaying top K=1
0.6596012101862708 http://freemanmoore.net/ Freeman Moore - SMU Spring 2020
spring freeman moor phd email fmoor lyle.smu.edu fall cs keep look at the cours calendar in for the latest inform

Query? moore
1 documents match, displaying top K=1
0.6596012101862708 http://freemanmoore.net/ Freeman Moore - SMU Spring 2020
spring freeman moor phd email fmoor lyle.smu.edu fall cs keep look at the cours calendar in for the latest inform

Query? 
```

4. Thesaurus expansion

```
Query? Moore smu story
Query expanded to: Moore smu story novel book
8 documents match, displaying top K=5
0.5025415378775715 http://freemanmoore.net/ Freeman Moore - SMU Spring 2020
spring freeman moor phd email fmoor lyle.smu.edu fall cs keep look at the cours calendar in for the latest inform
0.37094979063704936 http://freemanmoore.net/textfiles/extratextfiles/mockingbird1.html Mockingbird novel part 1
mockingbird part to kill mockingbird is primarili novel about grow up under extraordinari circumst in the in the southern unit
0.21637037003889348 http://freemanmoore.net/schedule.htm SMU CS 5337/7337 Spring 2020 Schedule
smu cs spring preliminari schedul thi page is maintain as the latest schedul of content and activ date topic activ
0.1376759667419815 http://freemanmoore.net/textfiles/index.html SMU CSE 5/7337 Spring 2018 Textfiles
textfil for cluster golf golf golf golf golf basketbal basketbal basketbal basketbal basketbal basebal basebal basebal basebal basebal footbal footbal
0.13605231641490023 http://freemanmoore.net/textfiles/extratextfiles/index.php SMU CS 5/7337 Spring 2020 text files"
addit text file to support queri implement part part part part part hocuspocu word file for test exact match same
```

5. Stop

```
Query? stop
12 query processed
```

6. Test queries

   1. moore southern

```
```

```
Query? moore southern
2 documents match, displaying top K=2
0.615596355280391 http://freemanmoore.net/ Freeman Moore - SMU Spring 2020
spring freeman moor phd email fmoor lyle.smu.edu fall cs keep look at the cours calendar in for the latest inform
0.16105633296029973 http://freemanmoore.net/textfiles/extratextfiles/mockingbird1.html Mockingbird novel part 1
mockingbird part to kill mockingbird is primarili novel about grow up under extraordinari circumst in the in the southern unit

Query?
```

According to the query results, it can be seen that the first result has a much higher rating than the second. This means that the first result is much more consistent compared to the second.

This conclusion is very reasonable. This can be seen by looking at the printed title and the first 20 words. The query word (moore) appears in the title of the first page and in the first 20 words. On the second page, the word "southern" appears only in the first 20 words.

We can also click on the link further to view the content of the page and see if it meets expectations. Following is the page with the two results.

We can see that the first page has "moore" and "southern" very many times besides what was mentioned before. The second page has no additional matching section.

SMU | BOBBY B. LYLE SCHOOL OF ENGINEERING

Spring 2020

Freeman L. Moore, PhD Keep looking at the Course calendar in for the latest information.

email:
fmoore@lyle.smu.edu

Spring 2020 - T/Th 5:00 - 6:20 pm Junkins 113

5337 Syllabus

Fall 2019
CS 5330/7330

7337 Syllabus

The contents of this Web site are the sole responsibility of Dr. Freeman Moore and do not necessarily represent the opinions or policies of Southern Methodist University. The administrator of this site is Dr. Freeman Moore who may be contacted at fmoore@lyle.smu.edu.

Page of result 1

Mockingbird part 1

To Kill a Mockingbird is primarily a novel about growing up under extraordinary circumstances in the 1930s in the Southern United States. The story covers a span of three years, during which the main characters undergo significant changes. Scout Finch lives with her brother Jem and their father Atticus in the fictitious town of Maycomb, Alabama. Maycomb is a small, close-knit town, and every family has its social station depending on where they live, who their parents are, and how long their ancestors have lived in Maycomb. A widower, Atticus raises his children by himself, with the help of kindly neighbors and a black housekeeper named Calpurnia. Scout and Jem almost instinctively understand the complexities and machinations of their neighborhood and town.

Page of result 2

2. what is the score of this page

```
Query? what is the score of this page
33 documents match, displaying top K=5
1.2 http://freemanmoore.net/simplescorepage.html what is the score of this page
what is the score of thi page what is the score of thi page
0.3113095297131505 http://freemanmoore.net/textfiles/cow3.txt
what cow is brown cow
0.22522003573637922 http://freemanmoore.net/textfiles/extratextfiles/magictext.html This is the magic file
magic show up here and in the titl brown beig tan auburn thi is the magic file
0.22430498350255224 http://freemanmoore.net/textfiles/basketball5.txt
basketbal is limited-contact sport play on rectangular court while most often play as team sport with five player on each
0.18331921516264627 http://freemanmoore.net/schedule.htm SMU CS 5337/7337 Spring 2020 Schedule
smu cs spring preliminari schedul thi page is maintain as the latest schedul of content and activ date topic activ
```

According to the query results, it can be seen that the first result matched the query exactly (reaching a maximum score of 1.2), because both the topic and the content part were exactly the same as the query. The scores of the second result also match better than the other results. Because the content has only 4 words in total, two of them are matched with the query.

3. three year story

```
Query? three year story
Query expanded to: three year story novel book
14 documents match, displaying top K=5
0.44831731988057 http://freemanmoore.net/textfiles/extratextfiles/mockingbird1.html Mockingbird novel part 1
mockingbird part to kill mockingbird is primarili novel about grow up under extraordinari circumst in the in the southern unit
0.16032678119874944 http://freemanmoore.net/textfiles/extratextfiles/mockingbird4.html Mockingbird part 4
mockingbird part the stori take place dure three year of the great depress in the fiction tire old town of
0.1311962910377378 http://freemanmoore.net/textfiles/basketball1.txt
under normal circumst the news that dirk nowitzki is done for the year would be devast but thursday announc wa
0.04310027134998584 http://freemanmoore.net/textfiles/extratextfiles/mockingbird2.html Mockingbird part 2
mockingbird part the onli neighbor who puzzl them is the mysteri arthur radley nicknam boo who never come outsid when
0.026682857807890197 http://freemanmoore.net/textfiles/golf2.txt
jay haa nearli shot hi age friday to take the lead into hole saturday finish in the profession golf associ
```

The query was expanded to "three year story novel book". The first result has matches word in both title and the first 20 words (“novel“). While the second one has only "three year" in the first 20 words.

The third result also has "year" in the first 20 words.

7. Atticus to defend Maycomb

```
Query? Atticus to defend Maycomb
28 documents match, displaying top K=5
0.34911768170608487 http://freemanmoore.net/textfiles/extratextfiles/mockingbird4.html Mockingbird part 4
mockingbird part the stori take place dure three year of the great depress in the fiction tire old town of
0.329806900334718 http://freemanmoore.net/textfiles/extratextfiles/mockingbird1.html Mockingbird novel part 1
mockingbird part to kill mockingbird is primarili novel about grow up under extraordinari circumst in the in the southern unit
0.2591797894659972 http://freemanmoore.net/textfiles/extratextfiles/mockingbird5.html Mockingbird part 5
mockingbird part atticu doe not want jem and scout to be present at tom robinson trial no seat is avail
0.15145601601128644 http://freemanmoore.net/textfiles/baseball2.txt
it open day and for one day our team is go to win the whole bleep thing it open day
0.14222810525392027 http://freemanmoore.net/textfiles/extratextfiles/mockingbird3.html Mockingbird part 3
mockingbird part suddenli scout and jem have to toler barrag of racial slur and insult becaus of atticu role in
```

The matching part may not be obvious from the title and the first 20 words of the output. But as before, by clicking on the link to view the page, you can see that the rest of the content has a very large number of words that match the query.

8. hocuspocus thisworks

```
Query? hocuspocus thisworks
Query expanded to: hocuspocus magic abracadabra thisworks this work
16 documents match, displaying top K=5
0.5613491301196275 http://freemanmoore.net/textfiles/extratextfiles/magictext.html This is the magic file
magic show up here and in the titl brown beig tan auburn thi is the magic file
0.24434697199730954 http://freemanmoore.net/simplescorepage.html what is the score of this page
what is the score of thi page what is the score of thi page
0.21727508926582234 http://freemanmoore.net/textfiles/extratextfiles/index.php SMU CS 5/7337 Spring 2020 text files"
addit text file to support queri implement part part part part part hocuspocu word file for test exact match same
0.1124626291479635 http://freemanmoore.net/useragent/useragent.php CSE 5337/7337 User-Agent
thi is the user-ag inform receiv if you are crawler did you defin uniqu name http_accept http_client_ip server_protocol http/1.1 request_metho
0.10836528727746436 http://freemanmoore.net/textfiles/football5.txt
the dalla cowboy overhaul their group of assist coach thi off-season richard is proud to join the dalla cowboy footbal
```

The query was expanded to "hocuspocus magic abracadabra thisworks this work". The first result contains the lots of expanded words, such as "magic" and "this". And query words also show in title and first 20 words. Thus, it has the highest score. In the third result, we can see that the content contains "hocuspocus".

additional text files to support query implementation

- part 1
- part 2
- part 3
- part 4
- part 5
- hocuspocus
- 3 word file for testing exact match
- same 3 words as building1.txt but in reverse order, words match but files don't match

Last Updated: April 06, 2020

9. Brown cow

```
Query? brown cow
Query expanded to: brown beige tan auburn cow
7 documents match, displaying top K=5
0.6943520096102751 http://freemanmoore.net/textfiles/extratextfiles/magictext.html This is the magic file
magic show up here and in the titl brown beig tan auburn thi is the magic file
0.48716079976709825 http://freemanmoore.net/textfiles/cow4.txt
brown cow
0.43371638165596527 http://freemanmoore.net/textfiles/cow1.txt
how brown is the brown cow
0.4286649545598271 http://freemanmoore.net/textfiles/cow3.txt
what cow is brown cow
0.3588811922310531 http://freemanmoore.net/textfiles/cow2.txt
how now brown cow
```

The query was expanded to "brown begie tan auburn cow". For the first result pretty match the expanded query, it has higher score than the second result, which exactly match the origin query.