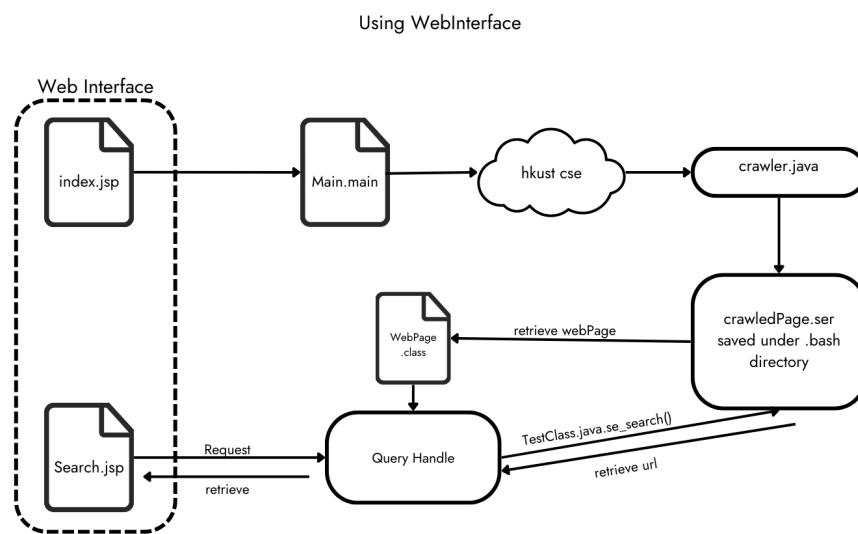


## 1. Introduction

This course is intended to have the whole picture of a search engine that has been used around the world. This project is one of the practical ways to perform the outcome of the courses. We establish the structure of the search by initiating the Crawler with Main.java as the main search that retrieves the crawled information, processes the index, connects the query input, etc. The programming language that we used is Java, and this project is built and run with IntelliJ IDEA community with smart-tomcat configuration.

## 2. Design Overview



### a. Crawler

The crawler is an essential part of the project. The crawler is used for crawling the website to get the content of a website. It will record all the information under <https://www.cse.ust.hk/~kwtleung/COMP4321/testpage.htm> and spread through the child links inside the web page. crawledPage.ser is the object of generated *HashMap<String, WebPage>* saved outside which retrieves all information regarding the webpage. The String URL is the unique identifier of the web page.

### b. Indexer

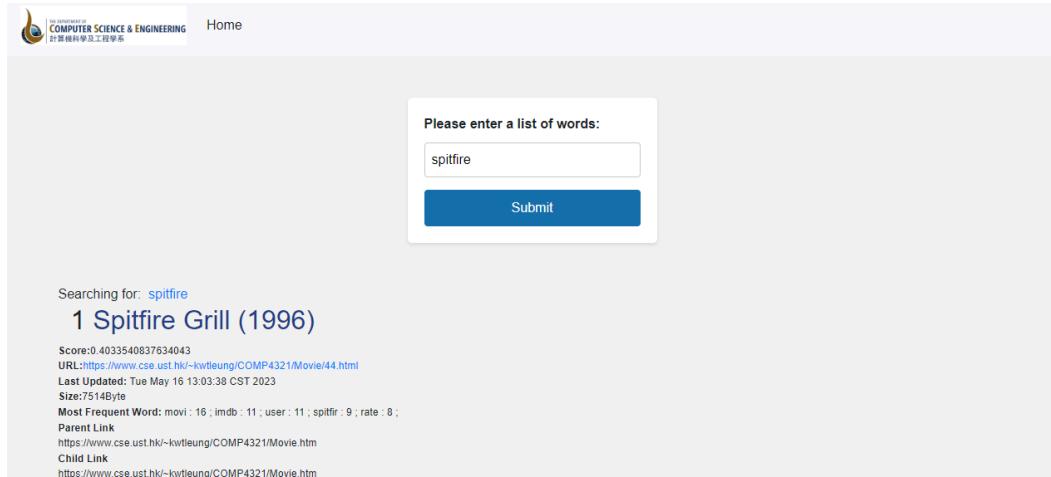
The indexer is a critical component of the document processing pipeline, responsible for performing tasks such as stemming and word processing. By reducing words to their base form through stemming, the indexer enables efficient search and retrieval of information. Users can leverage the stemmed words to conduct targeted searches, gaining access to a broader range of relevant

documents. After extracting the keyword and stemming, the forward index is stored at ForwardIndexTree.db

#### c. Query Handler

The query handler supports input with or without double quote words, where double quote words indicate the phrase search. When there is no double quote, the query will find all the documents that contain at least one word from the query. When there is a double quote, the query will find all the documents that contain the exact phrase in the double quote. Then, the function will determine the cosine similarity between the page and the query vector. And return a score. The score will be modified by the PageRank score, and whether the term appears on the title of the page.

#### d. Web Interface



Please enter a list of words:

Submit

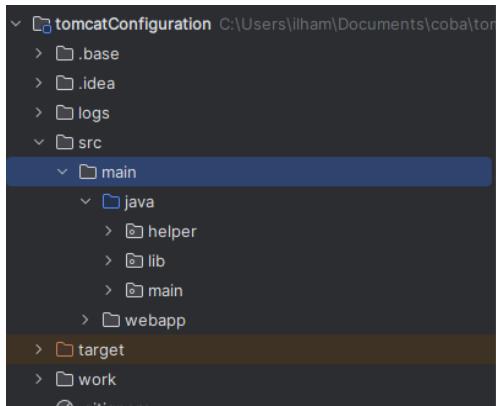
Searching for: [spitfire](#)

**1 Spitfire Grill (1996)**

Score: 0.4033540837634043  
URL: <https://www.cse.ust.hk/~kwleung/COMP4321/Movie/44.html>  
Last Updated: Tue May 16 13:03:38 CST 2023  
Size: 7514Byte  
Most Frequent Word: movi : 16 ; imdb : 11 ; user : 11 ; spitfir : 9 ; rate : 8 ;  
Parent Link: <https://www.cse.ust.hk/~kwleung/COMP4321/Movie.htm>  
Child Link: <https://www.cse.ust.hk/~kwleung/COMP4321/Movie.htm>

The user's query is transmitted through the web interface, which initiates an HTTP request to the JSP engine running on the server. The JSP engine processes and compiles the JSP (JavaServer Pages) code into a Servlet class. This generated Servlet class is then executed by the Servlet engine. As the Servlet class runs, it dynamically generates HTML-formatted output. The resulting HTML content is sent back to the web server as an HTTP response. Finally, the web browser receives the HTTP response and renders the content on the web interface using the received HTML.

### 3. File Structures



#### tomcatConfiguration

<b>.base</b>	The base configuration for "smart-tomcat"
<b>.idea</b>	The intelliJ configuration for IntelliJ IDEA
<b>logs</b>	This folder stores log files generated by your Tomcat server during runtime.
<b>src</b>	The directory of the program sources
<b>main</b>	for easiness of use
<b>java</b>	The directory of crawler, indexer, search engine
<b>webapps</b>	The directory of web-interface
<b>target</b>	to store compiled output (e.g., compiled Java classes or JAR files)
<b>works</b>	for temporary working files related to JSP compilation

The file directory is divided into two different parts. The first one is to */src/main/java* where it stores all the backend of the crawler, indexer, and search engine. The second is */src/main/webapp* where it stores all the User interfaces for the input queries. Since there are 2 ways to initiate the crawling, .bash is utilized of the purpose of running the backend on the jsp file (search.jsp and index.jsp) while */src/main/java* will store the output from compiling and running directly the main.java under */src/main/java* directory,

## **4. Algorithms Used**

### **a. Crawler**

#### **i. Breadth-First-Search**

Breadth-First-Search (BFS) is a graph traversal algorithm that explores all the vertices of a graph in breadth-first order. This property of BFS makes it suitable for implementing web crawlers because it ensures that the crawler systematically explores webpages in a breadth-first manner, starting from a given seed URL.

### **b. Retrieval**

We utilized many prevalence algorithms in order to retrieve relevant webpage to query input. It combined into one which resulted in PageRank.

#### **i. TF-IDF**

The TF-IDF score is used while calculating the weight of the parameters, It is helpful for getting a correct vector for calculation in the later stage and correctly weight the terms in the document.

#### **ii. Cosine Similarity**

Cosine similarity was used after calculating the vector of both input query and the tf-idf score. It is helpful for deciding which page should be returned, to make our search engine

#### **iii. Page Rank**

After calculating the cosine similarity, a weighted score calculated by PageRank is added to the score, this allow our search engine consider more

#### **iv. Title Favour-Weighted**

After calculating the above score, if a page **title have the query term**, a bonus of (75% \* Number of distinct words in title appear on the query). This helps highly prioritize the page with title matches.

### **c. Stopstem Techniques**

#### **i. Stopword Removal**

Stopwords are commonly used words in a language that do not carry significant meaning and are often removed from text during preprocessing. We utilized this technology in order to remove all unnecessary words that are not related to the information.

#### **ii. Stemming**

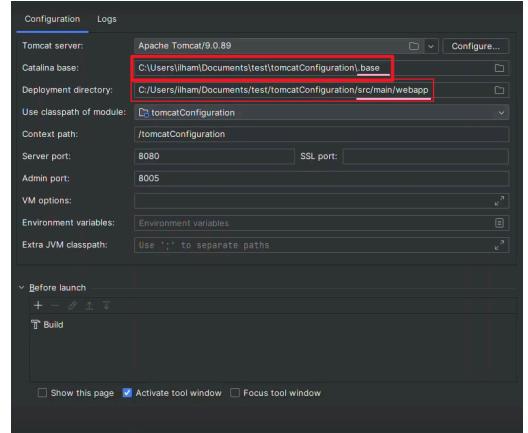
The purpose of stemming is to group together words that have the same meaning but different inflections or variations. For example, stemming would reduce the words "running," "runs," and "ran" to their common stem "run". In this case, it helps us to refer to which webpage is more significant for the query input.

## **5. Installation Procedures**

- New version of java 20 and JDK20 is used.
  - Put stopwords.txt at root directories  
*(./stopwords.txt)* and helper folder  
*(./src/main/java/helper/stopwords.txt)*

**If you want to use the User Interface:**

1. Open the tomcatConfiguration directory
  2. Use Smart Tomcat configuration
  3. Make sure that the catalina base directory and deployment directory is the same.
  4. Run the smart tomcat.
  5. To crawl for the use of user interface, please open <http://localhost:8080/tomcatConfiguration/> after smart tomcat is working Do not terminate program after crawling.
  6. To search a phrase, go to <http://localhost:8080/tomcatConfiguration/search.jsp>
  7. Please make sure that the crawling page is done before starting to search. You can see the progress under mytomcat terminal



8. If page modification is done, restart 5 to 7 again, to do the re-crawling..

**If you want to crawl and search directly**

1. Open the tomcatConfiguration directory
  2. Run `src/main/java/main/Main.java`
  3. Modify testClass or create testClass object and call `se_search(String[])` where an array of string is the parameter of the queries. It will return the unique identifier of URL which can be used by calling a WebPage object.  
Please see the index.jsp under

## 6. Highlight of features (Bonus Points)

### 1. Similar Page (Bonus 1)

To find a similar page based on the five most frequent words within a webpage, we analyze the content and identify the words that occur most frequently. These frequent words serve as a reference for new input queries. By comparing the word frequencies of different pages, we can identify those that share similar characteristics. This approach helps in finding pages that are likely to have related or relevant content based on the commonality of their most frequently used words.

### 2. User Interface (Bonus 3)

It is important to create a user-friendly interface for the search engine. A well-designed interface helps users easily understand how to utilize the search engine without any inconvenience. By providing clear instructions and intuitive features, users can quickly grasp how to conduct searches effectively.

### 3. PageRank Scoring (Bonus 6)

PageRank is also considered in our implementation. The formula is the same as the Lecture Notes, “ $(1-d)+d(\text{sum of}(PR(T_i)/C(T_i)))$ ”. 2 iterations of the PageRank have been done As there exists pages which have many parent nodes (Movie Index Page), to prevent getting dominated by those pages, a low damping factor (0.015) is used. A weight small have been multiplied, due to the fact that cosine similarity will not be greater than 1, and PageRank score can go up infinitely, especially when a huge amount of parent page, a heavy penalty(multiplied by 1.5%) is applied to prevent dominate by the PageRank score.

### 4. Optimizing (Bonus 7)

We attempted to reduce the access time and total search time **bonus feature 7** by optimizing our code. This reduces the time needed for restraining and storing data from the database. And this increases the time behavior of our search engine.

### 5. Other features

We also support clicking the link of the page, the parent page, and the child page, like the online search engine, it can help users to access the link easier.

## 7. Functions Implemented Testing

**Main.main()** is used to start crawling procedures.

```
1 package main;
2
3
4 > import ...
5
6
7 public class Main {
8
9     /**
10      * @param url The url for crawling
11      * @param numPage The number of page need to crawl, and store in JDBM
12      * @return a Hashmap, key is the URL, Content is the URL Object
13
14     */
15
16     @Public static HashMap<String,WebPage> crawling(String url,int numPage){
17
18
19 }
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
79
80
81
82
83
84
85
86
87
88
89
89
90
91
92
93
94
95
96
97
98
99
99
100
101
102
103
104
105
106
107
108
109
109
110
111
112
113
114
115
116
117
118
119
119
120
121
122
123
124
125
126
127
128
129
129
130
131
132
133
134
135
136
137
138
139
139
140
141
142
143
144
145
146
147
148
149
149
150
151
152
153
154
155
156
157
158
159
159
160
161
162
163
164
165
166
167
168
169
169
170
171
172
173
174
175
176
177
178
179
179
180
181
182
183
184
185
186
187
188
189
189
190
191
192
193
194
195
196
197
198
199
199
200
201
202
203
204
205
206
207
208
209
209
210
211
212
213
214
215
216
217
218
219
219
220
221
222
223
224
225
226
227
227
228
229
229
230
231
232
233
234
235
236
237
237
238
239
239
240
241
242
243
244
245
245
246
247
247
248
249
249
250
251
252
253
254
255
255
256
257
257
258
259
259
260
261
262
263
264
265
265
266
267
267
268
269
269
270
271
272
273
274
275
275
276
277
277
278
279
279
280
281
282
283
284
285
285
286
287
287
288
289
289
290
291
292
293
294
295
295
296
297
297
298
299
299
300
301
302
303
304
305
305
306
307
307
308
309
309
310
311
311
312
313
313
314
315
315
316
317
317
318
319
319
320
321
321
322
323
323
324
325
325
326
327
327
328
329
329
330
331
331
332
333
333
334
335
335
336
337
337
338
339
339
340
341
341
342
343
343
344
345
345
346
347
347
348
349
349
350
351
351
352
353
353
354
355
355
356
357
357
358
359
359
360
361
361
362
363
363
364
365
365
366
367
367
368
369
369
370
371
371
372
373
373
374
375
375
376
377
377
378
379
379
380
381
381
382
383
383
384
385
385
386
387
387
388
389
389
390
391
391
392
393
393
394
395
395
396
397
397
398
399
399
400
401
401
402
403
403
404
405
405
406
407
407
408
409
409
4010
4011
4012
4013
4014
4015
4016
4017
4018
4019
4020
4021
4022
4023
4024
4025
4026
4027
4028
4029
4029
4030
4031
4032
4033
4034
4035
4036
4037
4038
4039
4039
4040
4041
4042
4043
4044
4045
4046
4047
4048
4049
4049
4050
4051
4052
4053
4054
4055
4056
4057
4058
4059
4059
4060
4061
4062
4063
4064
4065
4066
4067
4068
4069
4069
4070
4071
4072
4073
4074
4075
4076
4077
4078
4079
4079
4080
4081
4082
4083
4084
4085
4086
4087
4088
4089
4089
4090
4091
4092
4093
4094
4095
4096
4097
4098
4099
4099
40100
40101
40102
40103
40104
40105
40106
40107
40108
40109
40109
40110
40111
40112
40113
40114
40115
40115
40116
40117
40117
40118
40119
40119
40120
40121
40121
40122
40123
40123
40124
40125
40125
40126
40127
40127
40128
40129
40129
40130
40131
40131
40132
40133
40133
40134
40135
40135
40136
40137
40137
40138
40139
40139
40140
40141
40141
40142
40143
40143
40144
40145
40145
40146
40147
40147
40148
40149
40149
40150
40151
40151
40152
40153
40153
40154
40155
40155
40156
40157
40157
40158
40159
40159
40160
40161
40161
40162
40163
40163
40164
40165
40165
40166
40167
40167
40168
40169
40169
40170
40171
40171
40172
40173
40173
40174
40175
40175
40176
40177
40177
40178
40179
40179
40180
40181
40181
40182
40183
40183
40184
40185
40185
40186
40187
40187
40188
40189
40189
40190
40191
40191
40192
40193
40193
40194
40195
40195
40196
40197
40197
40198
40199
40199
401200
401201
401202
401203
401204
401205
401206
401207
401208
401208
401209
401210
401211
401212
401213
401213
401214
401215
401215
401216
401217
401217
401218
401219
401219
401220
401221
401221
401222
401223
401223
401224
401225
401225
401226
401227
401227
401228
401229
401229
401230
401231
401231
401232
401233
401233
401234
401235
401235
401236
401237
401237
401238
401239
401239
401240
401241
401241
401242
401243
401243
401244
401245
401245
401246
401247
401247
401248
401249
401249
401250
401251
401251
401252
401253
401253
401254
401255
401255
401256
401257
401257
401258
401259
401259
401260
401261
401261
401262
401263
401263
401264
401265
401265
401266
401267
401267
401268
401269
401269
401270
401271
401271
401272
401273
401273
401274
401275
401275
401276
401277
401277
401278
401279
401279
401280
401281
401281
401282
401283
401283
401284
401285
401285
401286
401287
401287
401288
401289
401289
401290
401291
401291
401292
401293
401293
401294
401295
401295
401296
401297
401297
401298
401299
401299
401300
401301
401301
401302
401303
401303
401304
401305
401305
401306
401307
401307
401308
401309
401309
401310
401311
401311
401312
401313
401313
401314
401315
401315
401316
401317
401317
401318
401319
401319
401320
401321
401321
401322
401323
401323
401324
401325
401325
401326
401327
401327
401328
401329
401329
401330
401331
401331
401332
401333
401333
401334
401335
401335
401336
401337
401337
401338
401339
401339
401340
401341
401341
401342
401343
401343
401344
401345
401345
401346
401347
401347
401348
401349
401349
401350
401351
401351
401352
401353
401353
401354
401355
401355
401356
401357
401357
401358
401359
401359
401360
401361
401361
401362
401363
401363
401364
401365
401365
401366
401367
401367
401368
401369
401369
401370
401371
401371
401372
401373
401373
401374
401375
401375
401376
401377
401377
401378
401379
401379
401380
401381
401381
401382
401383
401383
401384
401385
401385
401386
401387
401387
401388
401389
401389
401390
401391
401391
401392
401393
401393
401394
401395
401395
401396
401397
401397
401398
401399
401399
401400
401401
401401
401402
401403
401403
401404
401405
401405
401406
401407
401407
401408
401409
401409
401410
401411
401411
401412
401413
401413
401414
401415
401415
401416
401417
401417
401418
401419
401419
401420
401421
401421
401422
401423
401423
401424
401425
401425
401426
401427
401427
401428
401429
401429
401430
401431
401431
401432
401433
401433
401434
401435
401435
401436
401437
401437
401438
401439
401439
401440
401441
401441
401442
401443
401443
401444
401445
401445
401446
401447
401447
401448
401449
401449
401450
401451
401451
401452
401453
401453
401454
401455
401455
401456
401457
401457
401458
401459
401459
401460
401461
401461
401462
401463
401463
401464
401465
401465
401466
401467
401467
401468
401469
401469
401470
401471
401471
401472
401473
401473
401474
401475
401475
401476
401477
401477
401478
401479
401479
401480
401481
401481
401482
401483
401483
401484
401485
401485
401486
401487
401487
401488
401489
401489
401490
401491
401491
401492
401493
401493
401494
401495
401495
401496
401497
401497
401498
401499
401499
401500
401501
401501
401502
401503
401503
401504
401505
401505
401506
401507
401507
401508
401509
401509
401510
401511
401511
401512
401513
401513
401514
401515
401515
401516
401517
401517
401518
401519
401519
401520
401521
401521
401522
401523
401523
401524
401525
401525
401526
401527
401527
401528
401529
401529
401530
401531
401531
401532
401533
401533
401534
401535
401535
401536
401537
401537
401538
401539
401539
401540
401541
401541
401542
401543
401543
401544
401545
401545
401546
401547
401547
401548
401549
401549
401550
401551
401551
401552
401553
401553
401554
401555
401555
401556
401557
401557
401558
401559
401559
401560
401561
401561
401562
401563
401563
401564
401565
401565
401566
401567
401567
401568
401569
401569
401570
401571
401571
401572
401573
401573
401574
401575
401575
401576
401577
401577
401578
401579
401579
401580
401581
401581
401582
401583
401583
401584
401585
401585
401586
401587
401587
401588
401589
401589
401590
401591
401591
401592
401593
401593
401594
401595
401595
401596
401597
401597
401598
401599
401599
401600
401601
401601
401602
401603
401603
401604
401605
401605
401606
401607
401607
401608
401609
401609
401610
401611
401611
401612
401613
401613
401614
401615
401615
401616
401617
401617
401618
401619
401619
401620
401621
401621
401622
401623
401623
401624
401625
401625
401626
401627
401627
401628
401629
401629
401630
401631
401631
401632
401633
401633
401634
401635
401635
401636
401637
401637
401638
401639
401639
401640
401641
401641
401642
401643
401643
401644
401645
401645
401646
401647
401647
401648
401649
401649
401650
401651
401651
401652
401653
401653
401654
401655
401655
401656
401657
401657
401658
401659
401659
401660
401661
401661
401662
401663
401663
401664
401665
401665
401666
401667
401667
401668
401669
401669
401670
401671
401671
401672
401673
401673
401674
401675
401675
401676
401677
401677
401678
401679
401679
401680
401681
401681
401682
401683
401683
401684
401685
401685
401686
401687
401687
401688
401689
401689
401690
401691
401691
401692
401693
401693
401694
401695
401695
401696
401697
401697
401698
401699
401699
401700
401701
401701
401702
401703
401703
401704
401705
401705
401706
401707
401707
401708
401709
401709
401710
401711
401711
401712
401713
401713
401714
401715
401715
401716
401717
401717
401718
401719
401719
401720
401721
401721
401722
401723
401723
401724
401725
401725
401726
401727
401727
401728
401729
401729
401730
401731
401731
401732
401733
401733
401734
401735
401735
401736
401737
401737
401738
401739
401739
401740
401741
401741
401742
401743
401743
401744
401745
401745
401746
401747
401747
401748
401749
401749
401750
401751
401751
401752
401753
401753
401754
401755
401755
401756
401757
401757
401758
401759
401759
401760
401761
401761
401762
401763
401763
401764
401765
401765
401766
401767
401767
401768
401769
401769
401770
401771
401771
401772
401773
401773
401774
401775
401775
401776
401777
401777
401778
401779
401779
401780
401781
401781
401782
401783
401783
401784
401785
401785
401786
401787
401787
401788
401789
401789
401790
401791
401791
401792
401793
401793
401794
401795
401795
401796
401797
401797
401798
401799
401799
401800
401801
401801
401802
401803
401803
401804
401805
401805
401806
401807
401807
401808
401809
401809
401810
401811
401811
401812
401813
401813
401814
401815
401815
401816
401817
401817
401818
401819
401819
401820
401821
401821
401822
401823
401823
401824
401825
401825
401826
401827
401827
401828
401829
401829
401830
401831
401831
401832
401833
401833
401834
401835
401835
401836
401837
401837
401838
401839
401839
401840
401841
401841
401842
401843
401843
401844
401845
401845
401846
401847
401847
401848
401849
401849
401850
401851
401851
401852
401853
401853
401854
401855
401855
401856
401857
401857
401858
401859
401859
401860
401861
401861
401862
401863
401863
401864
401865
401865
401866
401867
401867
401868
401869
401869
401870
401871
401871
401872
401873
401873
401874
401875
401875
401876
401877
401877
401878
401879
401879
401880
401881
401881
401882
401883
401883
401884
401885
401885
401886
401887
401887
401888
401889
401889
401890
401891
401891
401892
401893
401893
401894
401895
401895
401896
401897
401897
401898
401899
401899
401900
401901
401901
401902
401903
401903
401904
401905
401905
401906
401907
401907
401908
401909
401909
401910
401911
401911
401912
401913
401913
401914
401915
401915
401916
401917
401917
401918
401919
401919
401920
401921
401921
401922
401923
401923
401924
401925
401925
401926
401927
401927
401928
401929
401929
401930
401931
401931
401932
401933
401933
401934
401935
401935
401936
401937
401937
401938
401939
401939
401940
401941
401941
401942
401943
401943
401944
401945
401945
401946
401947
401947
401948
401949
401949
401950
401951
401951
401952
401953
401953
401954
401955
401955
401956
401957
401957
401958
401959
401959
401960
401961
401961
401962
401963
401963
401964
401965
401965
401966
401967
401967
401968
401969
401969
401970
401971
401971
401972
401973
401973
401974
401975
401975
401976
401977
401977
401978
401979
401979
401980
401981
401981
401982
401983
401983
401984
401985
401985
401986
401987
401987
401988
401989
401989
401990
401991
401991
401992
401993
401993
401994
401995
401995
401996
401997
401997
401998
401999
401999
402000
402001
402001
402002
402003
402003
402004
402005
402005
402006
402007
402007
402008
402009
402009
402010
402011
402011
402012
402013
402013
402014
402015
402015
402016
402017
402017
402018
402019
402019
402020
402021
402021
402022
402023
402023
402024
402025
402025
402026
402027
402027
402028
402029
402029
402030
402031
402031
402032
402033
402033
402034
402035
402035
402036
402037
402037
402038
402039
402039
402040
402041
402041
402042
402043
402043
402044
402045
402045
402046
402047
402047
402048
402049
402049
402050
402051
402051
402052
402053
402053
402054
402055
402055
402056
402057
402057
402058
402059
402059
402060
402061
402061
402062
402063
4
```

**Index.jsp** : is used to start crawling procedures for the web interface.

Project tomcatConfiguration Version control Unnamed - G O S +

src

main

java

webapp

WEB-INF

index.css

index.jsp

Run Unnamed -

Console Tomcat Localhost Log Tomcat Access Log

```
11 words[] = "test";
12 TestClass tc = new TestTestClass();
13 Main.main(words);
14 MapString.Double result = tc.to_se_search(words);
15 for (Map.Entry<String, Double> r : result.entrySet()) {
16     out.println(r.getKey());
17     out.println(r.getValue());
18 }
```

12-May-2024 19:27:39.483 INFO [main] org.apache.catalina.startup.Catalina.start Server startup in [1101] milliseconds

12-May-2024 19:27:39.483 INFO [main] org.apache.catalina.core.ContainerBase.[Catalina].localhost.[localhost].[/] CATALINA\_STARTED

http://localhost:8080/tomcatConfiguration

150 page have been crawled

150 page have been crawled

Crawling finished

Crawled Page saved successfully at crawledPage.ser.

Score: 0.4557496886997944

Title: Test page

url: http://www.jianshu.com/age.us1.hk/~kevleung/ICMP&21/testpage.htm

Last Update Date: Tue May 15 05:16:51 CST 2023 ,005 byte

Most Frequent word:

test : 2 ; Page : 2 ; new : 1 ; read : 1 ; intern : 1 ;

PowerLink:

**search.jsp()**: The search.jsp is the interface for query input.

 Home

Please enter a list of words:

Submit

## **8. Summary**

The project involves the development of a search engine for web and enterprise data. It consists of components such as the crawler, indexer, query handler, and web interface. The crawler utilizes the Breadth-First Search (BFS) algorithm to systematically explore web pages starting from a seed URL, recording information and spreading to child links. The indexer processes the crawled data, performs tasks like stemming and word processing, and builds an index for efficient search and retrieval. The query handler supports different types of searches, including phrase searches, using algorithms like TF-IDF and cosine similarity. The web interface allows users to input queries, initiates HTTP requests, and displays HTML-formatted search results. The project also incorporates algorithms like PageRank for retrieval and scoring, as well as techniques like stopword removal and stemming for text preprocessing. Bonus features include finding similar pages based on frequent words, a user-friendly interface, and optimization techniques for enhanced performance.

One of the weaknesses in our search engine lies in the implementation of the linear search algorithm. While this algorithm allows users to perform phrase searches by specifying multiple words in a specific order, it relies on a sequential scan of indexed documents. This approach can result in slower search performance, especially when dealing with a large number of documents. To change this, our solution if we re-implement is to have a better vector space which contains a large amount of correlation variable between one page and another. For example, implementing GloVe is one way.

Overall, the project aims to create an effective search engine that provides relevant results by leveraging crawling, indexing, and various algorithms.