

Program Log

```
1 C:\Python27\python.exe C:/Users/PrestonSSD2/PycharmProjects/WebCrawler/crawler.py
2 Crawling...
3   index.html ... indexed
4     SMU-CSE-LOGO.gif ... image
5     mailto:fmoore@lyle.smu.edu ... external
6     CSE5337_syllabus.pdf ... skipped
7     CSE7337_syllabus.pdf ... skipped
8     schedule.htm ... new
9     https://smu.instructure.com ... external
10    dontgohere/badfile1.html ... new
11    SMU-CSE-LOGO.gif ... image
12    dontgohere/badfile1.html ... restricted
13    schedule.htm ... indexed
14      index.htm ... new
15      index_duplicate.htm ... new
16      does_not_exist.htm ... broken
17      misc/count_letters.txt ... new
18      misc/count_letters_duplicate.txt ... new
19      misc/google-and-or-example.jpg ... image
20      misc/tokenizing_exercise.pdf ... skipped
21      misc/algorithm-1-7.pdf ... skipped
22      misc/jaccard_example.pdf ... skipped
23      misc/mercator-article.pdf ... skipped
24      misc/levenshtein.html ... new
25      http://www.gedpage.com/soundex.html ... external
26      misc/porter_stemmer_example.html ... new
27      misc/word-morphing-puzzle.pdf ... skipped
28      misc/urlexample1.htm ... new
29      this_aint_gonna_work.htm ... broken
30      misc/example-2017-03-02.xlsx ... skipped
31      misc/examl.html ... new
32      misc/permuterminindex-example.jpg ... image
33      misc/excel-examples.html ... new
34      misc/activity-mar28.pdf ... skipped
35      misc/useragent.html ... new
36      misc/sim_example.xlsx ... skipped
37      misc/roadmap-04-06-2017.jpg ... image
38      misc/text/index.php ... new
39      misc/seo_email.txt ... new
40      misc/good_bad_page.htm ... new
41      misc/bayes.html ... new
42      misc/poem-classification.pptx ... skipped
43      misc/bayes-poem.html ... new
44      misc/poem-classification.xlsx ... skipped
45      http://www.wessa.net/rwasp_hierarchicalclustering.wasp ... external
46      signup.php ... new
47      http://www.smu.edu/EnrollmentServices/Registrar/Enrollment/FinalExamSchedule/Spring2017 ...
external
48      signup.php ... indexed
49      mailto:fmoore@lyle.smu.edu ... skipped
50      misc/bayes-poem.html ... indexed
51      misc/bayes.js ... skipped
52      misc/bayes.html ... indexed
53      misc/bayes.js ... skipped
54      misc/good_bad_page.htm ... indexed
55      misc/seo_email.txt ... indexed
56      misc/text/index.php ... indexed
57        misc/text/mockingbird1.html ... new
58        misc/text/mockingbird2.html ... new
59        misc/text/mockingbird3.html ... new
60        misc/text/mockingbird4.html ... new
61        misc/text/mockingbird5.html ... new
62        misc/text/magictext.html ... new
63      misc/text/magictext.html ... indexed
64      misc/text/mockingbird5.html ... indexed
65      misc/text/mockingbird4.html ... indexed
66      misc/text/mockingbird3.html ... indexed
67      misc/text/mockingbird2.html ... indexed
68      misc/text/mockingbird1.html ... indexed
69      misc/useragent.html ... indexed
70      misc/excel-examples.html ... indexed
71        misc/pagel13.xlsx ... skipped
72        misc/excel-icon.jpg ... image
73        misc/pagel13-ex2.xlsx ... skipped
74        misc/excel-icon.jpg ... image
75      misc/examl.html ... indexed
76      misc/urlexample1.htm ... indexed
```

Program Log

```
77      ... skipped
78      index.htm ... new
79      index.htm ... skipped
80      ... skipped
81      http://lyle.smu.edu/~fmoore ... external
82      index.htm ... duplicate
83      misc/porter_stemmer_example.html ... indexed
84      http://tartarus.org/~martin/PorterStemmer/ ... external
85      http://en.wikipedia.org/wiki/Document_classification ... external
86      http://en.wikipedia.org/wiki/Stop_words ... external
87      http://en.wikipedia.org/wiki/Tf*idf ... external
88      http://lucene.apache.org/core/ ... external
89      http://search.carrot2.org/stable/search ... external
90      http://9ol.es/porter_js_demo.html ... external
91      misc/levenshtein.html ... indexed
92      misc/count_letters_duplicate.txt ... indexed
93      misc/count_letters.txt ... duplicate
94      index_duplicate.htm ... duplicate
95      index.htm ... already visited
96 Crawling finished
97 #####
98 #####
99 ##### Preston and Arturo's Web Crawler #####
100 #####
101 #####
102
103 Please enter a query to search the lyle.smu.edu/~fmoore domain.
104 Search will display top 6 results or all results that query is found on.
105 Type 'quit' to exit the search engine
106 > moore smu
107 Score | url (title) | first 20 words
108 1.491434 | http://lyle.smu.edu/~fmoore/signup.php (SMU CSE 5337/7337 Demo sign up) | smu cse
53377337 demo sign up smu spring 2017 cse 53377337 demo sign up final project grade requir dump your
109 1.477080 | http://lyle.smu.edu/~fmoore/index.html (Freeman Moore - SMU Spring 2017) | freeman moor
smu spring 2017 spring 2017 freeman l moor phd email fmoorelylesmuedu fall 2016 cse 53307330 spring
2017 tuesdaythursdai
110 1.137557 | http://lyle.smu.edu/~fmoore/schedule.htm (SMU CSE 5337/7337 Spring 2017 Schedule) | smu
cse 53377337 spring 2017 schedul 2017 preliminari schedul page maintain latest schedul content and
activ date topic activ jan
111
112 > Bob Ewell where Scout
113 Score | url (title) | first 20 words
114 0.995715 | http://lyle.smu.edu/~fmoore/misc/text/mockingbird5.html (Mockingbird part 5) |
mockingbird part 5 mockingbird part 5 attic do not want jem and scout present tom robinson trial no
seat avail
115 0.645121 | http://lyle.smu.edu/~fmoore/misc/text/mockingbird4.html (Mockingbird part 4) |
mockingbird part 4 mockingbird part 4 stori take place dure three year 1933â35 great depress
fiction tire old town maycomb
116 0.349527 | http://lyle.smu.edu/~fmoore/misc/text/mockingbird3.html (Mockingbird part 3) |
mockingbird part 3 mockingbird part 3 suddenli scout and jem have toler barrag racial slur and insult
becaus attic role
117 0.349527 | http://lyle.smu.edu/~fmoore/misc/text/mockingbird2.html (Mockingbird part 2) |
mockingbird part 2 mockingbird part 2 onli neighbor puzzl them mysteri arthur radlei nicknam boo
never come outsid dill anoth
118 0.349527 | http://lyle.smu.edu/~fmoore/misc/text/mockingbird1.html (Mockingbird novel part 1) |
mockingbird novel part 1 mockingbird part 1 kill mockingbird primarili novel grow up under
extraordinari circumst 1930 southern unit state
119
120 > three year story
121 Score | url (title) | first 20 words
122 1.000000 | http://lyle.smu.edu/~fmoore/misc/text/mockingbird1.html (Mockingbird novel part 1) |
mockingbird novel part 1 mockingbird part 1 kill mockingbird primarili novel grow up under
extraordinari circumst 1930 southern unit state
123 0.994792 | http://lyle.smu.edu/~fmoore/misc/text/mockingbird4.html (Mockingbird part 4) |
mockingbird part 4 mockingbird part 4 stori take place dure three year 1933â35 great depress
fiction tire old town maycomb
124 0.736033 | http://lyle.smu.edu/~fmoore/misc/text/mockingbird2.html (Mockingbird part 2) |
mockingbird part 2 mockingbird part 2 onli neighbor puzzl them mysteri arthur radlei nicknam boo
never come outsid dill anoth
125 0.477393 | http://lyle.smu.edu/~fmoore/misc/text/mockingbird5.html (Mockingbird part 5) |
mockingbird part 5 mockingbird part 5 attic do not want jem and scout present tom robinson trial no
seat avail
126
127 > Atticus to defend Maycomb
128 Score | url (title) | first 20 words
129 0.991178 | http://lyle.smu.edu/~fmoore/misc/text/mockingbird4.html (Mockingbird part 4) |
mockingbird part 4 mockingbird part 4 stori take place dure three year 1933â35 great depress
```

Program Log

```
129 fiction tire old town maycomb
130 0.641846 | http://lyle.smu.edu/~fmoore/misc/text/mockingbird1.html (Mockingbird novel part 1) |
mockingbird novel part 1 mockingbird part 1 kill mockingbird primarili novel grow up under
extraordinari circumst 1930 southern unit state
131 0.489463 | http://lyle.smu.edu/~fmoore/misc/text/mockingbird2.html (Mockingbird part 2) |
mockingbird part 2 mockingbird part 2 onli neighbor puzzl them mysteri arthur radlei nicknam boo
never come outsid dill anoth
132 0.417101 | http://lyle.smu.edu/~fmoore/misc/text/mockingbird3.html (Mockingbird part 3) |
mockingbird part 3 mockingbird part 3 suddenli scout and jem have toler barrag racial slur and insult
becaus atticu role
133 0.417101 | http://lyle.smu.edu/~fmoore/misc/text/mockingbird5.html (Mockingbird part 5) |
mockingbird part 5 mockingbird part 5 atticu doe not want jem and scout present tom robinson trial no
seat avail
134
135 > hocuspocus thisworks
136 Score | url (title) | first 20 words
137 1.207107 | http://lyle.smu.edu/~fmoore/misc/text/magictext.html (This is the magic file) | magic
file magic show up here and titl brown beig tan auburn
138 0.707107 | http://lyle.smu.edu/~fmoore/misc/seo_email.txt (No page title) | origin messag atul
kumar mailtoatul1930outlookcom sent thursdai april 13 2017 1258 am subject first page googl 99 per
month hi
139
140 > the
141 Only disallowed words were searched upon. Please add more words to your query.
142 > spring
143 Score | url (title) | first 20 words
144 1.500000 | http://lyle.smu.edu/~fmoore/index.html (Freeman Moore - SMU Spring 2017) | freeman moor
smu spring 2017 spring 2017 freeman l moor phd email fmoorelylesmu.edu fall 2016 cse 53307330 spring
2017 tuesdaythursdai
145 1.500000 | http://lyle.smu.edu/~fmoore/schedule.htm (SMU CSE 5337/7337 Spring 2017 Schedule) | smu
cse 53377337 spring 2017 schedul 2017 preliminari schedul page maintain latest schedul content and
activ date topic activ jan
146 1.500000 | http://lyle.smu.edu/~fmoore/misc/examl.html (CSE 7337 Spring 2017 distance students
exam 1 location) | cse 7337 spring 2017 distanc student exam 1 locat cse 7337 distanc student exam 1
locat numberpersonlocationrequest date lsurawashi anuja
147 1.000000 | http://lyle.smu.edu/~fmoore/signup.php (SMU CSE 5337/7337 Demo sign up) | smu cse
53377337 demo sign up smu spring 2017 cse 53377337 demo sign up final project grade requir dump your
148
149 > all course information
150 Score | url (title) | first 20 words
151 0.991546 | http://lyle.smu.edu/~fmoore/index.html (Freeman Moore - SMU Spring 2017) | freeman moor
smu spring 2017 spring 2017 freeman l moor phd email fmoorelylesmu.edu fall 2016 cse 53307330 spring
2017 tuesdaythursdai
152 0.792175 | http://lyle.smu.edu/~fmoore/schedule.htm (SMU CSE 5337/7337 Spring 2017 Schedule) | smu
cse 53377337 spring 2017 schedul 2017 preliminari schedul page maintain latest schedul content and
activ date topic activ jan
153 0.610294 | http://lyle.smu.edu/~fmoore/misc/bayes-poem.html (A simple naive bayes classifier
implemented in javascript) | simpl naiv bay classifi implement javascript train section these four
area contain four can poem two frost and two whitman
154 0.505057 | http://lyle.smu.edu/~fmoore/misc/useragent.html (CSE 5337/7337 User-Agent) | cse
53377337 userag userag inform receiv var x userag header sent navigatoruserag
documentgetelementbyiddemoinnerhtml x if you crawler did you defin
155
156 > spring break
157 Score | url (title) | first 20 words
158 1.491795 | http://lyle.smu.edu/~fmoore/schedule.htm (SMU CSE 5337/7337 Spring 2017 Schedule) | smu
cse 53377337 spring 2017 schedul 2017 preliminari schedul page maintain latest schedul content and
activ date topic activ jan
159 1.076320 | http://lyle.smu.edu/~fmoore/index.html (Freeman Moore - SMU Spring 2017) | freeman moor
smu spring 2017 spring 2017 freeman l moor phd email fmoorelylesmu.edu fall 2016 cse 53307330 spring
2017 tuesdaythursdai
160 1.076320 | http://lyle.smu.edu/~fmoore/misc/examl.html (CSE 7337 Spring 2017 distance students
exam 1 location) | cse 7337 spring 2017 distanc student exam 1 locat cse 7337 distanc student exam 1
locat numberpersonlocationrequest date lsurawashi anuja
161 0.817224 | http://lyle.smu.edu/~fmoore/misc/text/mockingbird5.html (Mockingbird part 5) |
mockingbird part 5 mockingbird part 5 atticu doe not want jem and scout present tom robinson trial no
seat avail
162 0.576320 | http://lyle.smu.edu/~fmoore/signup.php (SMU CSE 5337/7337 Demo sign up) | smu cse
53377337 demo sign up smu spring 2017 cse 53377337 demo sign up final project grade requir dump your
163
164 > smu
165 Score | url (title) | first 20 words
166 1.500000 | http://lyle.smu.edu/~fmoore/index.html (Freeman Moore - SMU Spring 2017) | freeman moor
smu spring 2017 spring 2017 freeman l moor phd email fmoorelylesmu.edu fall 2016 cse 53307330 spring
2017 tuesdaythursdai
167 1.500000 | http://lyle.smu.edu/~fmoore/schedule.htm (SMU CSE 5337/7337 Spring 2017 Schedule) | smu
cse 53377337 spring 2017 schedul 2017 preliminari schedul page maintain latest schedul content and
```

Program Log

```
167 activ date topic activ jan
168 1.500000 | http://lyle.smu.edu/~fmoore/signup.php (SMU CSE 5337/7337 Demo sign up) | smu cse
53377337 demo sign up smu spring 2017 cse 53377337 demo sign up final project grade requir dump your
169
170 > vector
171 Score | url (title) | first 20 words
172 1.000000 | http://lyle.smu.edu/~fmoore/schedule.htm (SMU CSE 5337/7337 Spring 2017 Schedule) | smu
cse 53377337 spring 2017 schedul 2017 preliminari schedul page maintain latest schedul content and
activ date topic activ jan
173
174 > quit
175
176 Process finished with exit code 0
177
```