Assignment 3 M2: Search Engine Construction - Developer

Group Members:

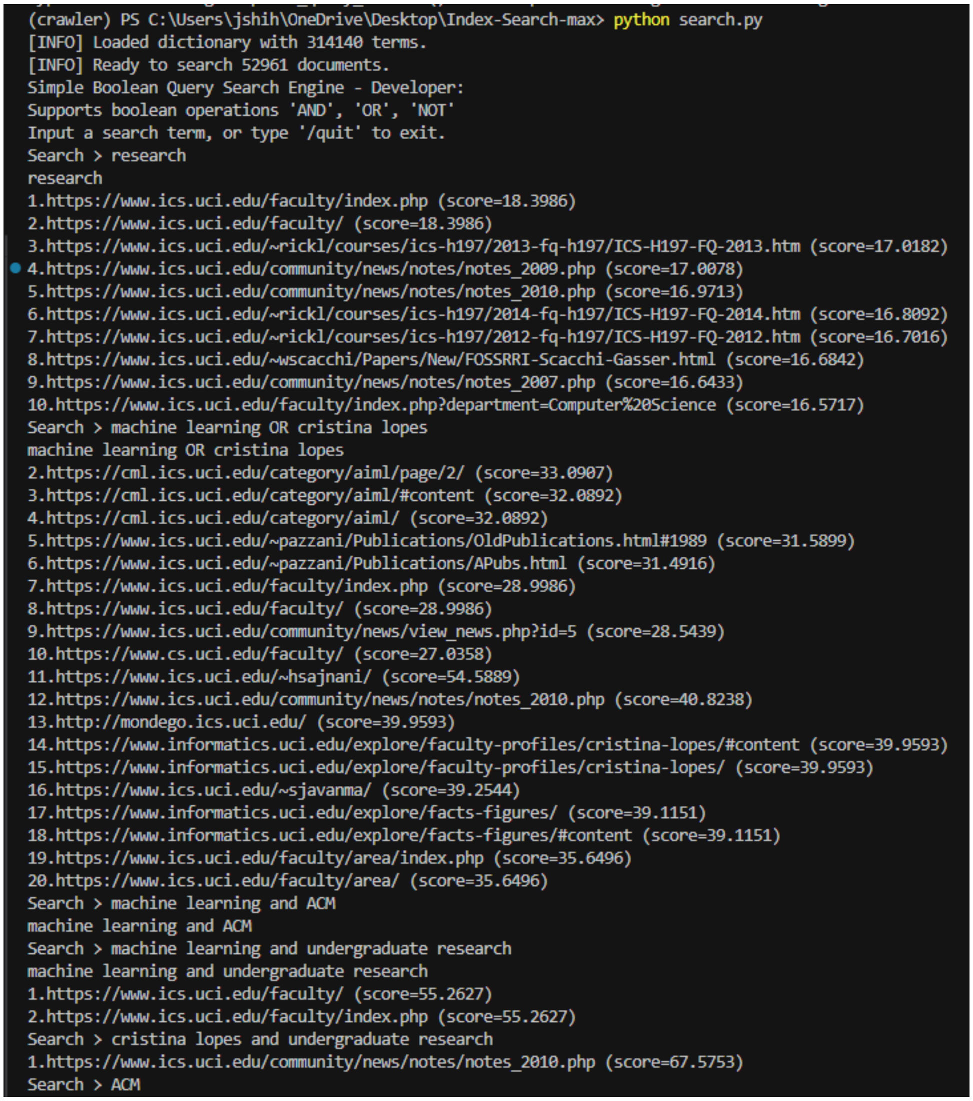
Julie Bui (ID: 59422563)
LoLa Alexis Kim (ID: 52727368)
Maxwell Shih (ID: 88195254)

Github link: https://github.com/Skipper321/Index-Search

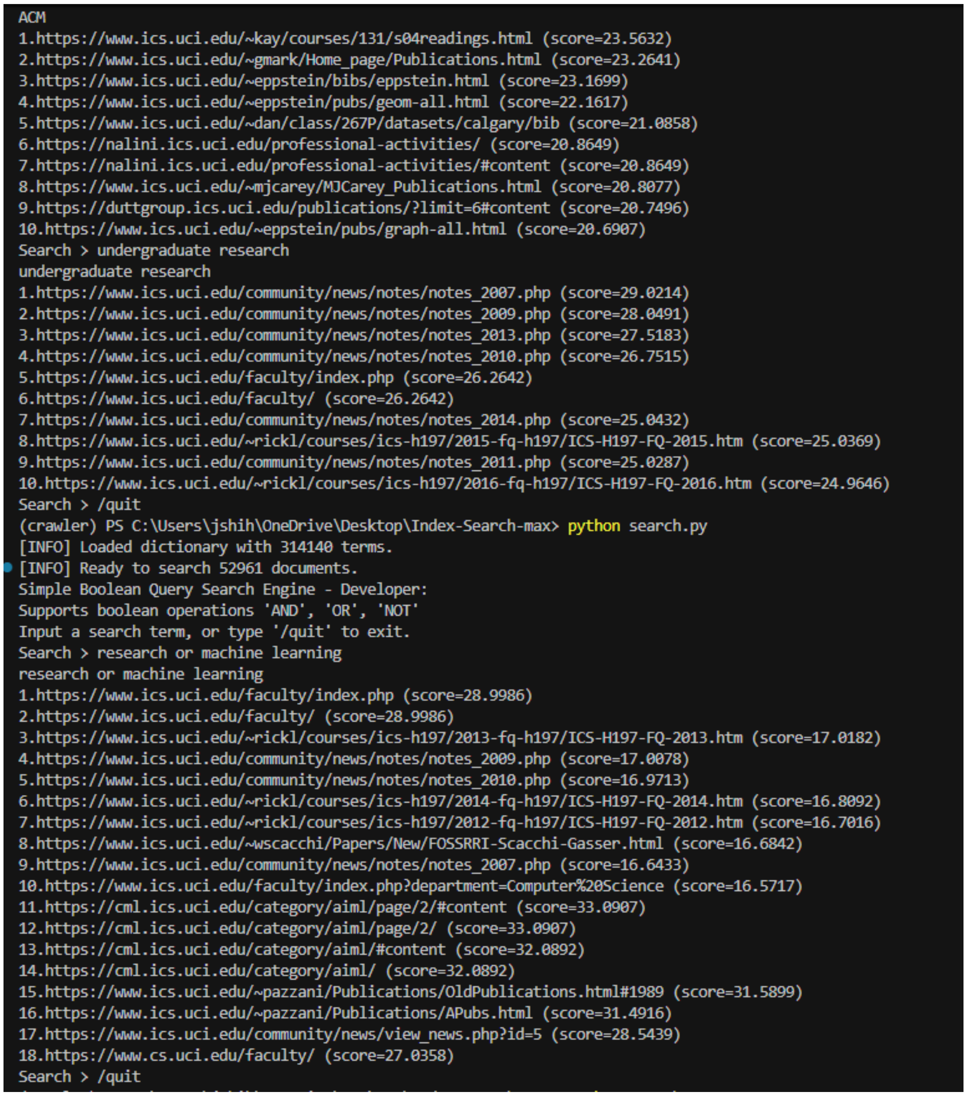
Performance:

Here is our search engine working for our developer set of webpages that have been indexed using TF-IDF.

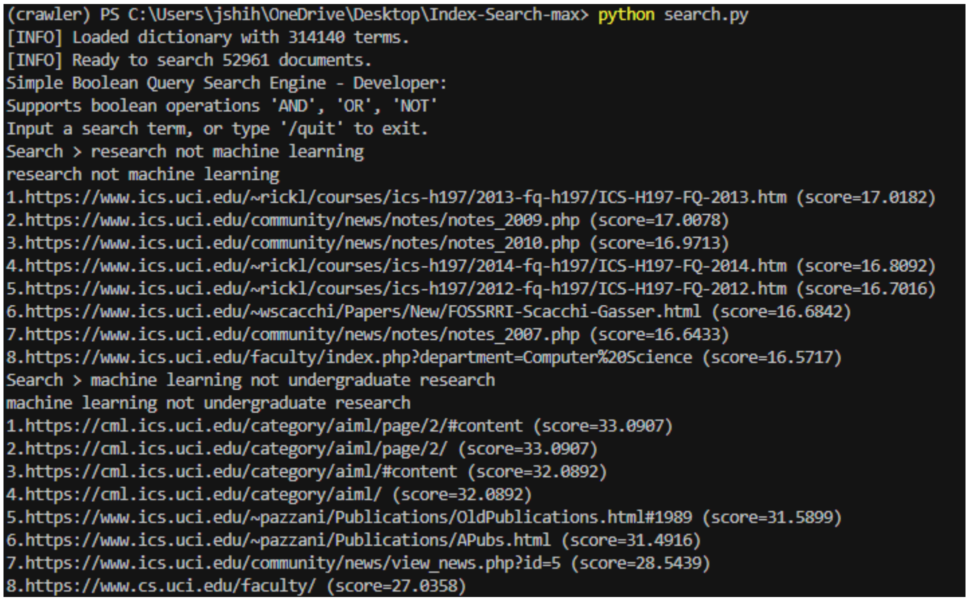
Testing individual queries, boolean operations, and queries other than sample provided:



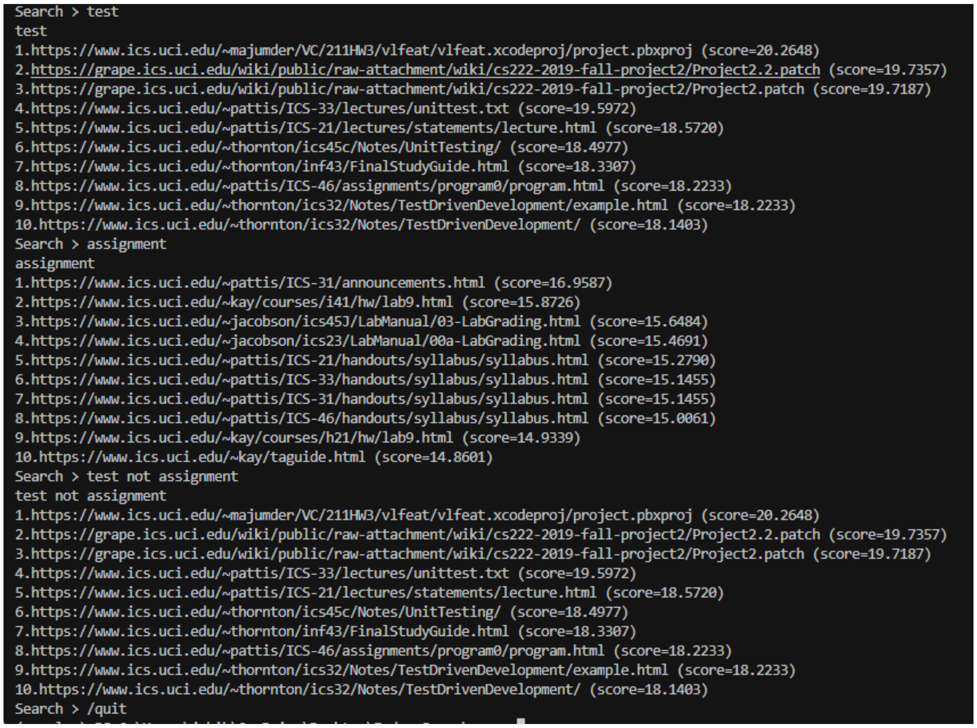
```
(crawler) PS C:\Users\jshih\OneDrive\Desktop\Index-Search-max> python search.py
[INFO] Loaded dictionary with 314140 terms.
[INFO] Ready to search 52961 documents.
Simple Boolean Query Search Engine - Developer:
Supports boolean operations 'AND', 'OR', 'NOT'
Input a search term, or type '/quit' to exit.
Search > research
research
1.https://www.ics.uci.edu/faculty/index.php (score=18.3986)
2.https://www.ics.uci.edu/faculty/ (score=18.3986)
3.https://www.ics.uci.edu/~rickl/courses/ics-h197/2013-fq-h197/ICS-H197-FQ-2013.htm (score=17.0182)
4.https://www.ics.uci.edu/community/news/notes/notes_2009.php (score=17.0078)
5.https://www.ics.uci.edu/community/news/notes/notes_2010.php (score=16.9713)
6.https://www.ics.uci.edu/~rickl/courses/ics-h197/2014-fq-h197/ICS-H197-FQ-2014.htm (score=16.8092)
7.https://www.ics.uci.edu/~rickl/courses/ics-h197/2012-fq-h197/ICS-H197-FQ-2012.htm (score=16.7016)
8.https://www.ics.uci.edu/~wscacchi/Papers/New/FOSSRRI-Scacchi-Gasser.html (score=16.6842)
9.https://www.ics.uci.edu/community/news/notes/notes_2007.php (score=16.6433)
10.https://www.ics.uci.edu/faculty/index.php?department=Computer%20Science (score=16.5717)
Search > machine learning OR cristina lopes
machine learning OR cristina lopes
2.https://cml.ics.uci.edu/category/aiml/page/2/ (score=33.0907)
3.https://cml.ics.uci.edu/category/aiml/#content (score=32.0892)
4.https://cml.ics.uci.edu/category/aiml/ (score=32.0892)
5.https://www.ics.uci.edu/~pazzani/Publications/OldPublications.html#1989 (score=31.5899)
6.https://www.ics.uci.edu/~pazzani/Publications/APubs.html (score=31.4916)
7.https://www.ics.uci.edu/faculty/index.php (score=28.9986)
8.https://www.ics.uci.edu/faculty/ (score=28.9986)
9.https://www.ics.uci.edu/community/news/view_news.php?id=5 (score=28.5439)
10.https://www.cs.uci.edu/faculty/ (score=27.0358)
11.https://www.ics.uci.edu/~hsajnani/ (score=54.5889)
12.https://www.ics.uci.edu/community/news/notes/notes_2010.php (score=40.8238)
13.http://mondego.ics.uci.edu/ (score=39.9593)
14.https://www.informatics.uci.edu/explore/faculty-profiles/cristina-lopes/#content (score=39.9593)
15.https://www.informatics.uci.edu/explore/faculty-profiles/cristina-lopes/ (score=39.9593)
16.https://www.ics.uci.edu/~sjavanma/ (score=39.2544)
17.https://www.informatics.uci.edu/explore/facts-figures/ (score=39.1151)
18.https://www.informatics.uci.edu/explore/facts-figures/#content (score=39.1151)
19.https://www.ics.uci.edu/faculty/area/index.php (score=35.6496)
20.https://www.ics.uci.edu/faculty/area/ (score=35.6496)
Search > machine learning and ACM
machine learning and ACM
Search > machine learning and undergraduate research
machine learning and undergraduate research
1.https://www.ics.uci.edu/faculty/ (score=55.2627)
2.https://www.ics.uci.edu/faculty/index.php (score=55.2627)
Search > cristina lopes and undergraduate research
1.https://www.ics.uci.edu/community/news/notes/notes_2010.php (score=67.5753)
Search > ACM
```



```
ACM
1.https://www.ics.uci.edu/~kay/courses/131/s04readings.html (score=23.5632)
2.https://www.ics.uci.edu/~gmark/Home_page/Publications.html (score=23.2641)
3.https://www.ics.uci.edu/~eppstein/bibs/eppstein.html (score=23.1699)
4.https://www.ics.uci.edu/~eppstein/pubs/geom-all.html (score=22.1617)
5.https://www.ics.uci.edu/~dan/class/267P/datasets/calgary/bib (score=21.0858)
6.https://nalini.ics.uci.edu/professional-activities/ (score=20.8649)
7.https://nalini.ics.uci.edu/professional-activities/#content (score=20.8649)
8.https://www.ics.uci.edu/~mjcarey/MJCarey_Publications.html (score=20.8077)
9.https://duttgroup.ics.uci.edu/publications/?limit=6#content (score=20.7496)
10.https://www.ics.uci.edu/~eppstein/pubs/graph-all.html (score=20.6907)
Search > undergraduate research
undergraduate research
1.https://www.ics.uci.edu/community/news/notes/notes_2007.php (score=29.0214)
2.https://www.ics.uci.edu/community/news/notes/notes_2009.php (score=28.0491)
3.https://www.ics.uci.edu/community/news/notes/notes_2013.php (score=27.5183)
4.https://www.ics.uci.edu/community/news/notes/notes_2010.php (score=26.7515)
5.https://www.ics.uci.edu/faculty/index.php (score=26.2642)
6.https://www.ics.uci.edu/faculty/ (score=26.2642)
7.https://www.ics.uci.edu/community/news/notes/notes_2014.php (score=25.0432)
8.https://www.ics.uci.edu/~rickl/courses/ics-h197/2015-fq-h197/ICS-H197-FQ-2015.htm (score=25.0369)
9.https://www.ics.uci.edu/community/news/notes/notes_2011.php (score=25.0287)
10.https://www.ics.uci.edu/~rickl/courses/ics-h197/2016-fq-h197/ICS-H197-FQ-2016.htm (score=24.9646)
Search > /quit
(crawler) PS C:\Users\jshih\OneDrive\Desktop\Index-Search-max> python search.py
[INFO] Loaded dictionary with 314140 terms.
[INFO] Ready to search 52961 documents.
Simple Boolean Query Search Engine - Developer:
Supports boolean operations 'AND', 'OR', 'NOT'
Input a search term, or type '/quit' to exit.
Search > research or machine learning
research or machine learning
1.https://www.ics.uci.edu/faculty/index.php (score=28.9986)
2.https://www.ics.uci.edu/faculty/ (score=28.9986)
3.https://www.ics.uci.edu/~rickl/courses/ics-h197/2013-fq-h197/ICS-H197-FQ-2013.htm (score=17.0182)
4.https://www.ics.uci.edu/community/news/notes/notes_2009.php (score=17.0078)
5.https://www.ics.uci.edu/community/news/notes/notes_2010.php (score=16.9713)
6.https://www.ics.uci.edu/~rickl/courses/ics-h197/2014-fq-h197/ICS-H197-FQ-2014.htm (score=16.8092)
7.https://www.ics.uci.edu/~rickl/courses/ics-h197/2012-fq-h197/ICS-H197-FQ-2012.htm (score=16.7016)
8.https://www.ics.uci.edu/~wscacchi/Papers/New/FOSSRRI-Scacchi-Gasser.html (score=16.6842)
9.https://www.ics.uci.edu/community/news/notes/notes_2007.php (score=16.6433)
10.https://www.ics.uci.edu/faculty/index.php?department=Computer%20Science (score=16.5717)
11.https://cml.ics.uci.edu/category/aiml/page/2/#content (score=33.0907)
12.https://cml.ics.uci.edu/category/aiml/page/2/ (score=33.0907)
13.https://cml.ics.uci.edu/category/aiml/#content (score=32.0892)
14.https://cml.ics.uci.edu/category/aiml/ (score=32.0892)
15.https://www.ics.uci.edu/~pazzani/Publications/OldPublications.html#1989 (score=31.5899)
16.https://www.ics.uci.edu/~pazzani/Publications/APubs.html (score=31.4916)
17.https://www.ics.uci.edu/community/news/view_news.php?id=5 (score=28.5439)
18.https://www.cs.uci.edu/faculty/ (score=27.0358)
Search > /quit
```



```
(crawler) PS C:\Users\jshih\OneDrive\Desktop\Index-Search-max> python search.py
[INFO] Loaded dictionary with 314140 terms.
[INFO] Ready to search 52961 documents.
Simple Boolean Query Search Engine - Developer:
Supports boolean operations 'AND', 'OR', 'NOT'
Input a search term, or type '/quit' to exit.
Search > research not machine learning
research not machine learning
1.https://www.ics.uci.edu/~rickl/courses/ics-h197/2013-fq-h197/ICS-H197-FQ-2013.htm (score=17.0182)
2.https://www.ics.uci.edu/community/news/notes/notes_2009.php (score=17.0078)
3.https://www.ics.uci.edu/community/news/notes/notes_2010.php (score=16.9713)
4.https://www.ics.uci.edu/~rickl/courses/ics-h197/2014-fq-h197/ICS-H197-FQ-2014.htm (score=16.8092)
5.https://www.ics.uci.edu/~rickl/courses/ics-h197/2012-fq-h197/ICS-H197-FQ-2012.htm (score=16.7016)
6.https://www.ics.uci.edu/~wscacchi/Papers/New/FOSSRRI-Scacchi-Gasser.html (score=16.6842)
7.https://www.ics.uci.edu/community/news/notes/notes_2007.php (score=16.6433)
8.https://www.ics.uci.edu/faculty/index.php?department=Computer%20Science (score=16.5717)
Search > machine learning not undergraduate research
machine learning not undergraduate research
1.https://cml.ics.uci.edu/category/aiml/page/2/#content (score=33.0907)
2.https://cml.ics.uci.edu/category/aiml/page/2/ (score=33.0907)
3.https://cml.ics.uci.edu/category/aiml/#content (score=32.0892)
4.https://cml.ics.uci.edu/category/aiml/ (score=32.0892)
5.https://www.ics.uci.edu/~pazzani/Publications/OldPublications.html#1989 (score=31.5899)
6.https://www.ics.uci.edu/~pazzani/Publications/APubs.html (score=31.4916)
7.https://www.ics.uci.edu/community/news/view_news.php?id=5 (score=28.5439)
8.https://www.cs.uci.edu/faculty/ (score=27.0358)
```



```
Search > test
test
1.https://www.ics.uci.edu/~majumder/VC/211HW3/vlfeat/vlfeat.xcodeproj/project.pbxproj (score=20.2648)
2.https://grape.ics.uci.edu/wiki/public/raw-attachment/wiki/cs222-2019-fall-project2/Project2.2.patch (score=19.7357)
3.https://grape.ics.uci.edu/wiki/public/raw-attachment/wiki/cs222-2019-fall-project2/Project2.patch (score=19.7187)
4.https://www.ics.uci.edu/~pattis/ICS-33/lectures/unittest.txt (score=19.5972)
5.https://www.ics.uci.edu/~pattis/ICS-21/lectures/statements/lecture.html (score=18.5720)
6.https://www.ics.uci.edu/~thornton/ics45c/Notes/UnitTesting/ (score=18.4977)
7.https://www.ics.uci.edu/~thornton/inf43/FinalStudyGuide.html (score=18.3307)
8.https://www.ics.uci.edu/~pattis/ICS-46/assignments/program0/program.html (score=18.2233)
9.https://www.ics.uci.edu/~thornton/ics32/Notes/TestDrivenDevelopment/example.html (score=18.2233)
10.https://www.ics.uci.edu/~thornton/ics32/Notes/TestDrivenDevelopment/ (score=18.1403)
Search > assignment
assignment
1.https://www.ics.uci.edu/~pattis/ICS-31/announcements.html (score=16.9587)
2.https://www.ics.uci.edu/~kay/courses/i41/hw/lab9.html (score=15.8726)
3.https://www.ics.uci.edu/~jacobson/ics45J/LabManual/03-LabGrading.html (score=15.6484)
4.https://www.ics.uci.edu/~jacobson/ics23/LabManual/00a-LabGrading.html (score=15.4691)
5.https://www.ics.uci.edu/~pattis/ICS-21/handouts/syllabus/syllabus.html (score=15.2790)
6.https://www.ics.uci.edu/~pattis/ICS-33/handouts/syllabus/syllabus.html (score=15.1455)
7.https://www.ics.uci.edu/~pattis/ICS-31/handouts/syllabus/syllabus.html (score=15.1455)
8.https://www.ics.uci.edu/~pattis/ICS-46/handouts/syllabus/syllabus.html (score=15.0061)
9.https://www.ics.uci.edu/~kay/courses/h21/hw/lab9.html (score=14.9339)
10.https://www.ics.uci.edu/~kay/taguide.html (score=14.8601)
Search > test not assignment
test not assignment
1.https://www.ics.uci.edu/~majumder/VC/211HW3/vlfeat/vlfeat.xcodeproj/project.pbxproj (score=20.2648)
2.https://grape.ics.uci.edu/wiki/public/raw-attachment/wiki/cs222-2019-fall-project2/Project2.2.patch (score=19.7357)
3.https://grape.ics.uci.edu/wiki/public/raw-attachment/wiki/cs222-2019-fall-project2/Project2.patch (score=19.7187)
4.https://www.ics.uci.edu/~pattis/ICS-33/lectures/unittest.txt (score=19.5972)
5.https://www.ics.uci.edu/~pattis/ICS-21/lectures/statements/lecture.html (score=18.5720)
6.https://www.ics.uci.edu/~thornton/ics45c/Notes/UnitTesting/ (score=18.4977)
7.https://www.ics.uci.edu/~thornton/inf43/FinalStudyGuide.html (score=18.3307)
8.https://www.ics.uci.edu/~pattis/ICS-46/assignments/program0/program.html (score=18.2233)
9.https://www.ics.uci.edu/~thornton/ics32/Notes/TestDrivenDevelopment/example.html (score=18.2233)
10.https://www.ics.uci.edu/~thornton/ics32/Notes/TestDrivenDevelopment/ (score=18.1403)
Search > /quit
```

Top 5 Search Results for each sample query:

“machine learning” top 5 urls:
1.https://cml.ics.uci.edu/category/aiml/page/2/#content (score=33.0907)
2.https://cml.ics.uci.edu/category/aiml/page/2/ (score=33.0907)
3.https://cml.ics.uci.edu/category/aiml/#content (score=32.0892)
4.https://cml.ics.uci.edu/category/aiml/ (score=32.0892)
5.https://www.ics.uci.edu/~pazzani/Publications/OldPublications.html#1989 (score=31.5899)

“Cristina lopes” top 5 urls:
1.https://www.ics.uci.edu/~hsajnani/ (score=54.5889)
2.https://www.ics.uci.edu/community/news/notes/notes_2010.php (score=40.8238)
3.http://mondego.ics.uci.edu/ (score=39.9593)
4.https://www.informatics.uci.edu/explore/faculty-profiles/cristina-lopes/#content (score=39.9593)
5.https://www.informatics.uci.edu/explore/faculty-profiles/cristina-lopes/ (score=39.9593)

“ACM” top 5 urls:
1.https://www.ics.uci.edu/~kay/courses/131/s04readings.html (score=23.5632)
2.https://www.ics.uci.edu/~gmark/Home_page/Publications.html (score=23.2641)
3.https://www.ics.uci.edu/~eppstein/bibs/eppstein.html (score=23.1699)
4.https://www.ics.uci.edu/~eppstein/pubs/geom-all.html (score=22.1617)
5.https://www.ics.uci.edu/~dan/class/267P/datasets/calgary/bib (score=21.0858)

“Master of Software Engineering” top 5 urls:
1.https://www.ics.uci.edu/~pattis/quotations.html (score=48.4957)
2.https://www.ics.uci.edu/~neno/vita/vita.html (score=44.9382)
3.https://www.ics.uci.edu/~taylor/Publications.htm (score=43.5117)
4.https://www.ics.uci.edu/community/news/notes/notes_2014.php (score=43.4287)
5.https://mswe.ics.uci.edu/program/curriculum/ (score=43.0896)

Evaluation:

- These reported urls are plausible because if one were to click into these links, the term they were looking for would appear on the page.
- The TF-IDF scores make sense because the first few results have a high frequency of the search term in the webpage. As you go down the search results, the frequency of the search term decreases.