# Assignment 3 M3: Search Engine Optimization - Developer

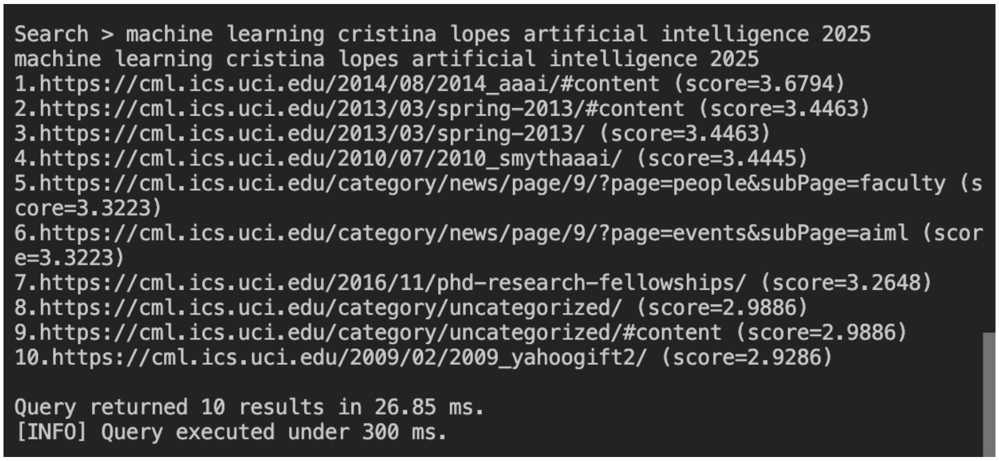
**Group Members:**
Julie Bui (ID: 59422563)
LoLa Alexis Kim (ID: 52727368)
Maxwell Shih (ID: 88195254)

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

## Test Queries:

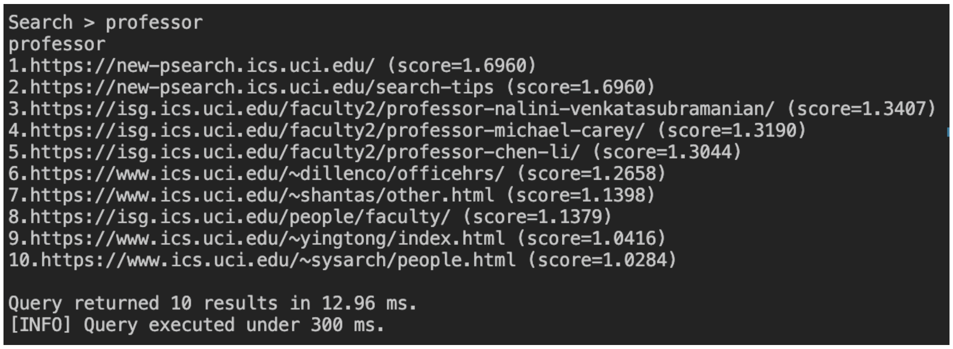
## Good Performance Queries (brief explanation on why):

### Query #1: "Machine learning cristina lopes artificial intelligence 2025"

- Performance: For most queries the results always return within 300 ms.
- UPDATE: we optimized simhash to make sure our queries are even faster, well under 200 ms.
- 

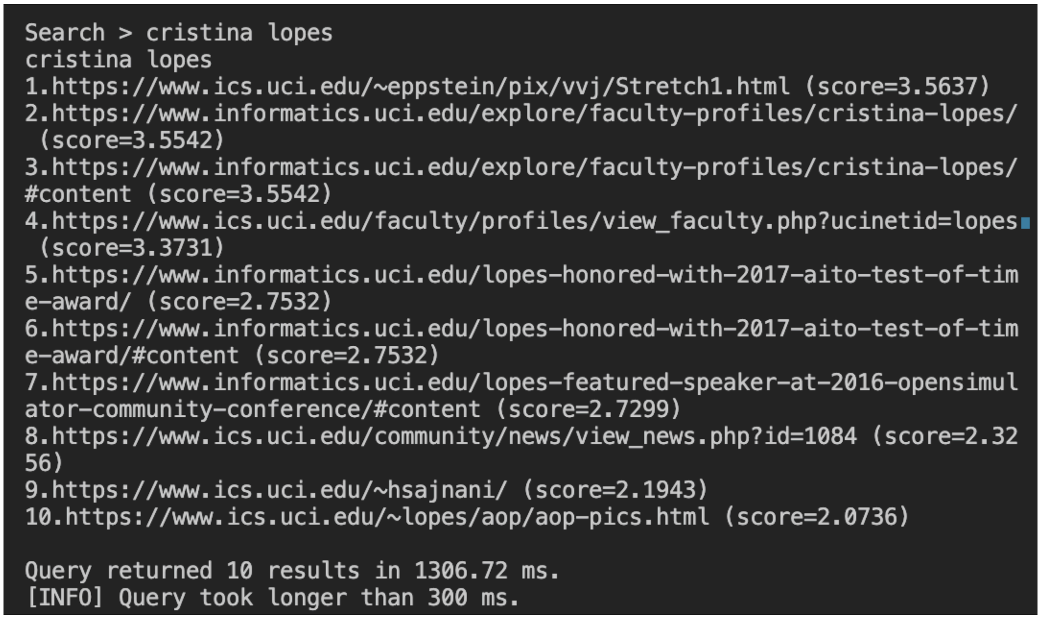
```
Search > machine learning cristina lopes artificial intelligence 2025
machine learning cristina lopes artificial intelligence 2025
1.https://cml.ics.uci.edu/2014/08/2014_aaai/#content (score=3.6794)
2.https://cml.ics.uci.edu/2013/03/spring-2013/#content (score=3.4463)
3.https://cml.ics.uci.edu/2013/03/spring-2013/ (score=3.4463)
4.https://cml.ics.uci.edu/2010/07/2010_smythaaai/ (score=3.4445)
5.https://cml.ics.uci.edu/category/news/page/9/?page=people&subPage=faculty (s
core=3.3223)
6.https://cml.ics.uci.edu/category/news/page/9/?page=events&subPage=aiml (scor
e=3.3223)
7.https://cml.ics.uci.edu/2016/11/phd-research-fellowships/ (score=3.2648)
8.https://cml.ics.uci.edu/category/uncategorized/ (score=2.9886)
9.https://cml.ics.uci.edu/category/uncategorized/#content (score=2.9886)
10.https://cml.ics.uci.edu/2009/02/2009_yahoogift2/ (score=2.9286)

Query returned 10 results in 26.85 ms.
[INFO] Query executed under 300 ms.
```

### Query #2: "faculty" will also give results for "staff" or "professors", but prioritizes the exact term rather than synonyms.

- 

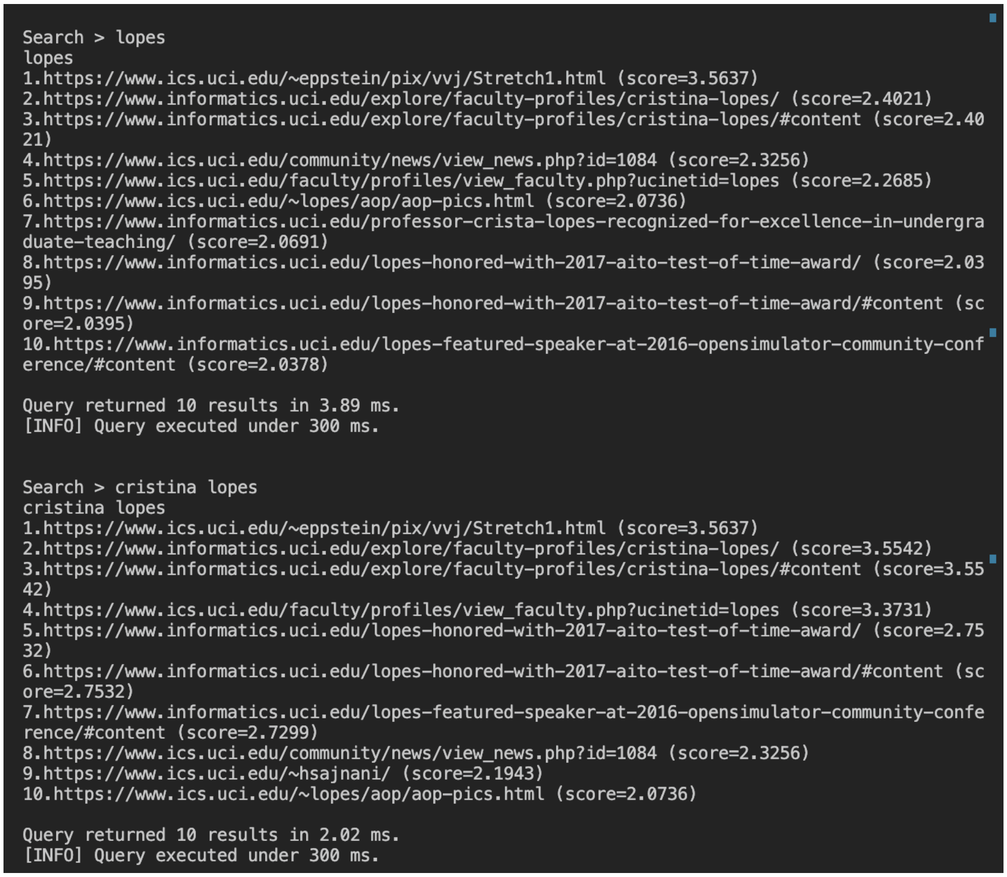
```
Search > professor
professor
1.https://new-psearch.ics.uci.edu/ (score=1.6960)
2.https://new-psearch.ics.uci.edu/search-tips (score=1.6960)
3.https://isg.ics.uci.edu/faculty2/professor-nalini-venkatasubramanian/ (score=1.3407)
4.https://isg.ics.uci.edu/faculty2/professor-michael-carey/ (score=1.3190)
5.https://isg.ics.uci.edu/faculty2/professor-chen-li/ (score=1.3044)
6.https://www.ics.uci.edu/~dillenco/officehrs/ (score=1.2658)
7.https://www.ics.uci.edu/~shantas/other.html (score=1.1398)
8.https://isg.ics.uci.edu/people/faculty/ (score=1.1379)
9.https://www.ics.uci.edu/~yingtong/index.html (score=1.0416)
10.https://www.ics.uci.edu/~sysarch/people.html (score=1.0284)

Query returned 10 results in 12.96 ms.
[INFO] Query executed under 300 ms.
```

**Query #3: “Cristina Lopes”**

- At first it takes longer to query the search, but for subsequent queries it is much faster using OS caching.
- 

```
Search > cristina lopes
cristina lopes
1.https://www.ics.uci.edu/~eppstein/pix/vvj/Stretch1.html (score=3.5637)
2.https://www.informatics.uci.edu/explore/faculty-profiles/cristina-lopes/
 (score=3.5542)
3.https://www.informatics.uci.edu/explore/faculty-profiles/cristina-lopes/
#content (score=3.5542)
4.https://www.ics.uci.edu/faculty/profiles/view_faculty.php?ucinetid=lopes
 (score=3.3731)
5.https://www.informatics.uci.edu/lopes-honored-with-2017-aito-test-of-tim
e-award/ (score=2.7532)
6.https://www.informatics.uci.edu/lopes-honored-with-2017-aito-test-of-tim
e-award/#content (score=2.7532)
7.https://www.informatics.uci.edu/lopes-featured-speaker-at-2016-opensimul
ator-community-conference/#content (score=2.7299)
8.https://www.ics.uci.edu/community/news/view_news.php?id=1084 (score=2.32
56)
9.https://www.ics.uci.edu/~hsajnani/ (score=2.1943)
10.https://www.ics.uci.edu/~lopes/aop/aop-pics.html (score=2.0736)

Query returned 10 results in 1306.72 ms.
[INFO] Query took longer than 300 ms.
```

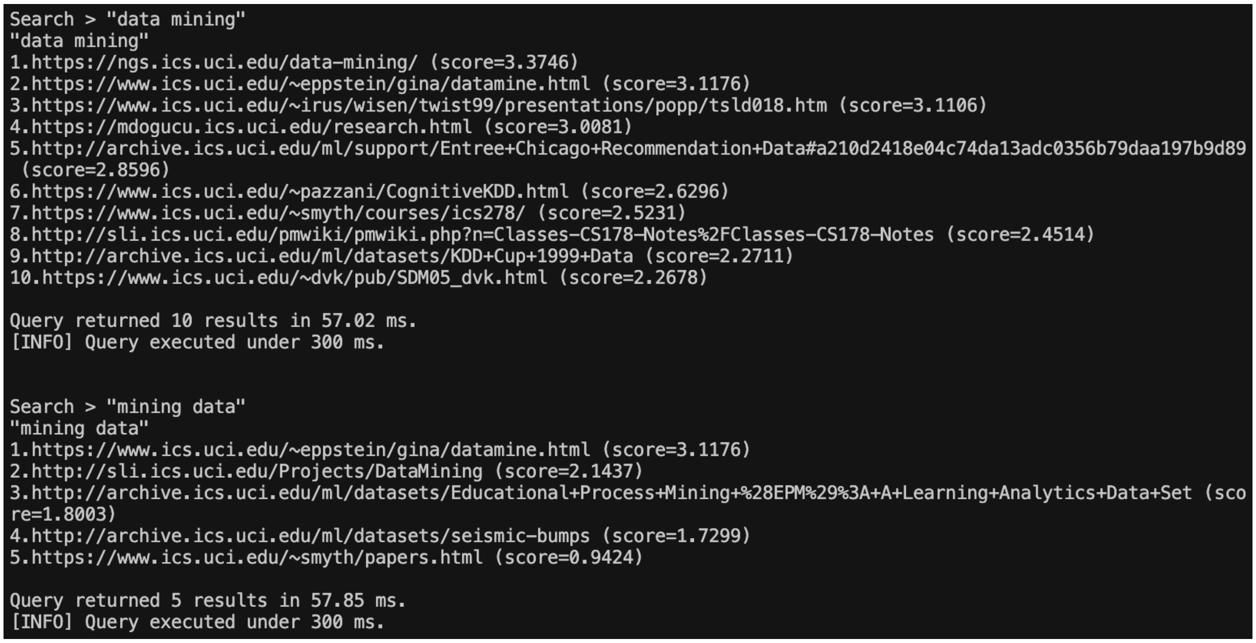
Subsequent Queries:

- 

```
Search > lopes
lopes
1.https://www.ics.uci.edu/~eppstein/pix/vvj/Stretch1.html (score=3.5637)
2.https://www.informatics.uci.edu/explore/faculty-profiles/cristina-lopes/ (score=2.4021)
3.https://www.informatics.uci.edu/explore/faculty-profiles/cristina-lopes/#content (score=2.40
21)
4.https://www.ics.uci.edu/community/news/view_news.php?id=1084 (score=2.3256)
5.https://www.ics.uci.edu/faculty/profiles/view_faculty.php?ucinetid=lopes (score=2.2685)
6.https://www.ics.uci.edu/~lopes/aop/aop-pics.html (score=2.0736)
7.https://www.informatics.uci.edu/professor-crista-lopes-recognized-for-excellence-in-undergra
duate-teaching/ (score=2.0691)
8.https://www.informatics.uci.edu/lopes-honored-with-2017-aito-test-of-time-award/ (score=2.03
95)
9.https://www.informatics.uci.edu/lopes-honored-with-2017-aito-test-of-time-award/#content (sc
ore=2.0395)
10.https://www.informatics.uci.edu/lopes-featured-speaker-at-2016-opensimulator-community-conf
erence/#content (score=2.0378)

Query returned 10 results in 3.89 ms.
[INFO] Query executed under 300 ms.


Search > cristina lopes
cristina lopes
1.https://www.ics.uci.edu/~eppstein/pix/vvj/Stretch1.html (score=3.5637)
2.https://www.informatics.uci.edu/explore/faculty-profiles/cristina-lopes/ (score=3.5542)
3.https://www.informatics.uci.edu/explore/faculty-profiles/cristina-lopes/#content (score=3.55
42)
4.https://www.ics.uci.edu/faculty/profiles/view_faculty.php?ucinetid=lopes (score=3.3731)
5.https://www.informatics.uci.edu/lopes-honored-with-2017-aito-test-of-time-award/ (score=2.75
32)
6.https://www.informatics.uci.edu/lopes-honored-with-2017-aito-test-of-time-award/#content (sc
ore=2.7532)
7.https://www.informatics.uci.edu/lopes-featured-speaker-at-2016-opensimulator-community-confe
rence/#content (score=2.7299)
8.https://www.ics.uci.edu/community/news/view_news.php?id=1084 (score=2.3256)
9.https://www.ics.uci.edu/~hsajnani/ (score=2.1943)
10.https://www.ics.uci.edu/~lopes/aop/aop-pics.html (score=2.0736)

Query returned 10 results in 2.02 ms.
[INFO] Query executed under 300 ms.
```

**Query #4: “data mining” and “mining data”**

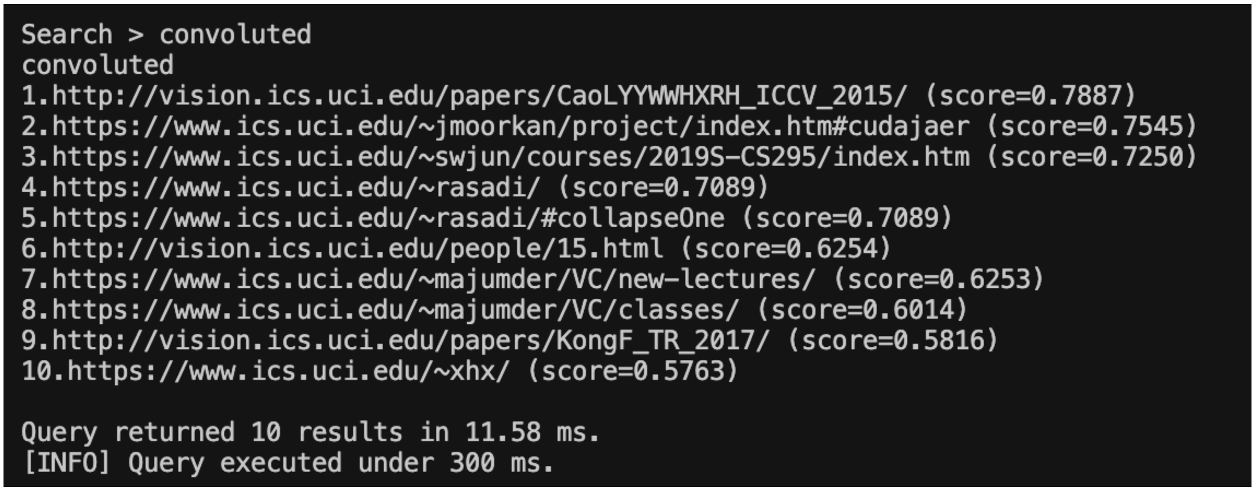
- Although similar, the queries “data mining” and “mining data” produce different results, meaning the search engine considers the order of terms and short term structure, affecting how documents are matched and ranked



```
Search > "data mining"
"data mining"
1.https://ngs.ics.uci.edu/data-mining/ (score=3.3746)
2.https://www.ics.uci.edu/~eppstein/gina/datamine.html (score=3.1176)
3.https://www.ics.uci.edu/~irus/wisen/twist99/presentations/popp/tsld018.htm (score=3.1106)
4.https://mdogucu.ics.uci.edu/research.html (score=3.0081)
5.http://archive.ics.uci.edu/ml/support/Entree+Chicago+Recommendation+Data#a210d2418e04c74da13adc0356b79daa197b9d89
 (score=2.8596)
6.https://www.ics.uci.edu/~pazzani/CognitiveKDD.html (score=2.6296)
7.https://www.ics.uci.edu/~smyth/courses/ics278/ (score=2.5231)
8.http://sli.ics.uci.edu/pmwiki/pmwiki.php?n=Classes-CS178-Notes%2FClasses-CS178-Notes (score=2.4514)
9.http://archive.ics.uci.edu/ml/datasets/KDD+Cup+1999+Data (score=2.2711)
10.https://www.ics.uci.edu/~dvk/pub/SDM05_dvk.html (score=2.2678)

Query returned 10 results in 57.02 ms.
[INFO] Query executed under 300 ms.


Search > "mining data"
"mining data"
1.https://www.ics.uci.edu/~eppstein/gina/datamine.html (score=3.1176)
2.http://sli.ics.uci.edu/Projects/DataMining (score=2.1437)
3.http://archive.ics.uci.edu/ml/datasets/Educational+Process+Mining+%28EPM%29%3A+A+Learning+Analytics+Data+Set (sco
re=1.8003)
4.http://archive.ics.uci.edu/ml/datasets/seismic-bumps (score=1.7299)
5.https://www.ics.uci.edu/~smyth/papers.html (score=0.9424)

Query returned 5 results in 57.85 ms.
[INFO] Query executed under 300 ms.
```

**Query #5: “convoluted”**

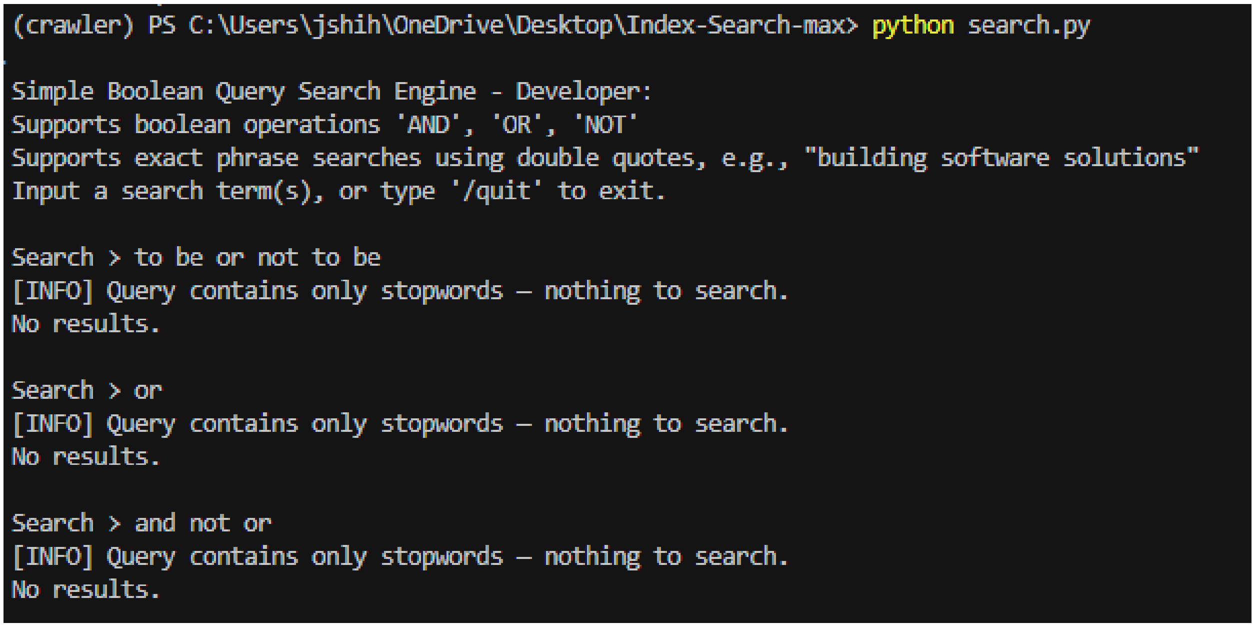
- Although none of the returned documents contained the exact term “convoluted,” the search engine successfully retrieved documents containing related stems such as “convolution” and “convolutional.” This indicates that the stemming process is functioning correctly by mapping the query term to related terms.
- Since these documents do not include the exact words, their relevance scores are naturally lower, but the results still demonstrate that the system retrieves related content when an exact match is unavailable, satisfying the user’s intent.



```
Search > convoluted
convoluted
1.http://vision.ics.uci.edu/papers/CaoLYYWWHXRH_ICCV_2015/ (score=0.7887)
2.https://www.ics.uci.edu/~jmoorkan/project/index.htm#cudajaer (score=0.7545)
3.https://www.ics.uci.edu/~swjun/courses/2019S-CS295/index.htm (score=0.7250)
4.https://www.ics.uci.edu/~rasadi/ (score=0.7089)
5.https://www.ics.uci.edu/~rasadi/#collapseOne (score=0.7089)
6.http://vision.ics.uci.edu/people/15.html (score=0.6254)
7.https://www.ics.uci.edu/~majumder/VC/new-lectures/ (score=0.6253)
8.https://www.ics.uci.edu/~majumder/VC/classes/ (score=0.6014)
9.http://vision.ics.uci.edu/papers/KongF_TR_2017/ (score=0.5816)
10.https://www.ics.uci.edu/~xhx/ (score=0.5763)

Query returned 10 results in 11.58 ms.
[INFO] Query executed under 300 ms.
```

**Query #6: “to be or not to be” Stopword only queries**

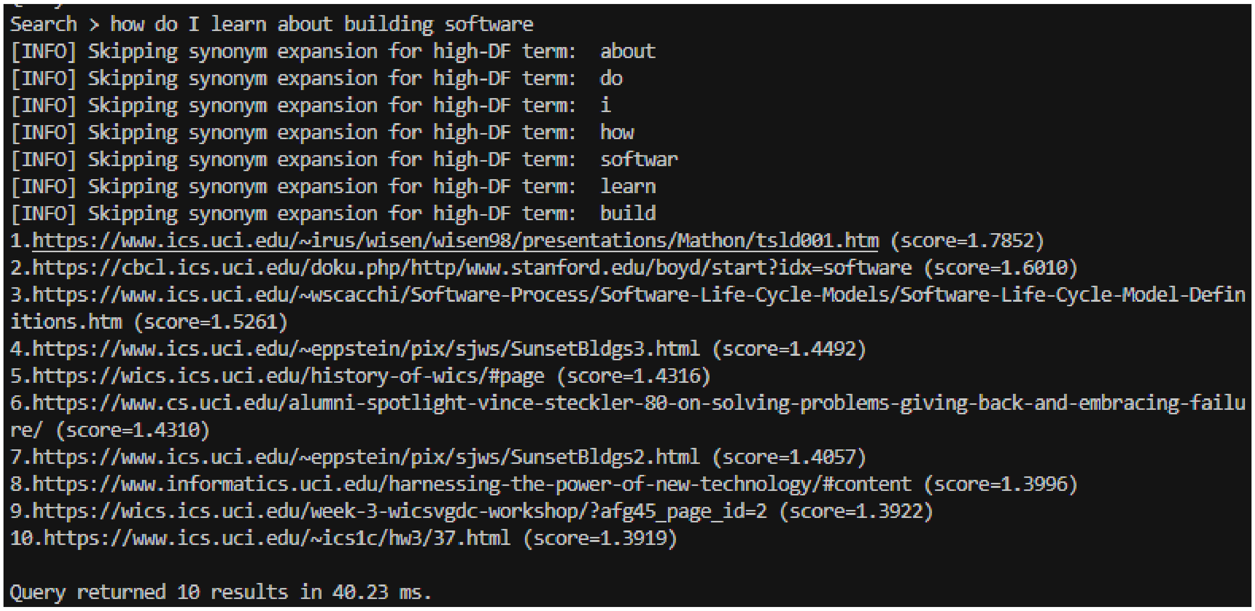
- We handled stopword only queries to not return anything because otherwise, it would just return any webpages that contain those words. We chose to omit any results



```
(crawler) PS C:\Users\jshih\OneDrive\Desktop\Index-Search-max> python search.py

Simple Boolean Query Search Engine - Developer:
Supports boolean operations 'AND', 'OR', 'NOT'
Supports exact phrase searches using double quotes, e.g., "building software solutions"
Input a search term(s), or type '/quit' to exit.

Search > to be or not to be
[INFO] Query contains only stopwords — nothing to search.
No results.

Search > or
[INFO] Query contains only stopwords — nothing to search.
No results.

Search > and not or
[INFO] Query contains only stopwords — nothing to search.
No results.
```

**Query #7: “How do I learn about building software”**

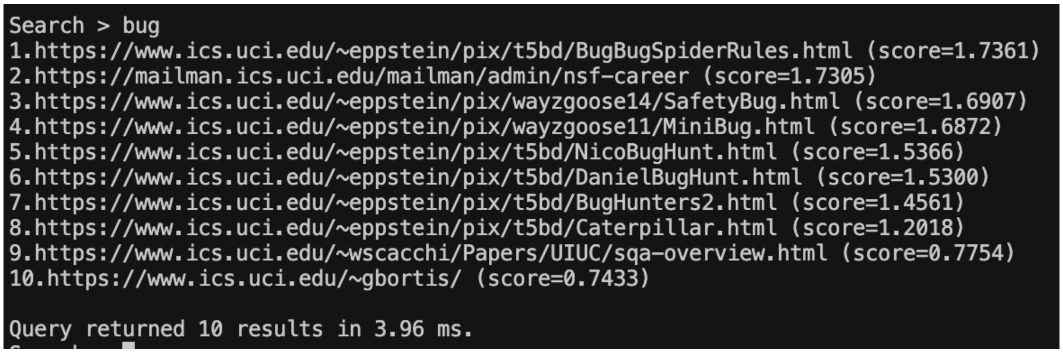
- We added less weight for stopwords, allowing for the query to still return quality results despite majority containing stopwords



```
Search > how do I learn about building software
[INFO] Skipping synonym expansion for high-DF term:  about
[INFO] Skipping synonym expansion for high-DF term:  do
[INFO] Skipping synonym expansion for high-DF term:  i
[INFO] Skipping synonym expansion for high-DF term:  how
[INFO] Skipping synonym expansion for high-DF term:  softwar
[INFO] Skipping synonym expansion for high-DF term:  learn
[INFO] Skipping synonym expansion for high-DF term:  build
1.https://www.ics.uci.edu/~irus/wisen/wisen98/presentations/Mathon/tsld001.htm (score=1.7852)
2.https://cbcl.ics.uci.edu/doku.php/http/www.stanford.edu/boyd/start?idx=software (score=1.6010)
3.https://www.ics.uci.edu/~wscacchi/Software-Process/Software-Life-Cycle-Models/Software-Life-Cycle-Model-Defin
itions.htm (score=1.5261)
4.https://www.ics.uci.edu/~eppstein/pix/sjws/SunsetBldgs3.html (score=1.4492)
5.https://wics.ics.uci.edu/history-of-wics/#page (score=1.4316)
6.https://www.cs.uci.edu/alumni-spotlight-vince-steckler-80-on-solving-problems-giving-back-and-embracing-failu
re/ (score=1.4310)
7.https://www.ics.uci.edu/~eppstein/pix/sjws/SunsetBldgs2.html (score=1.4057)
8.https://www.informatics.uci.edu/harnessing-the-power-of-new-technology/#content (score=1.3996)
9.https://wics.ics.uci.edu/week-3-wicsvgdc-workshop/?afg45_page_id=2 (score=1.3922)
10.https://www.ics.uci.edu/~ics1c/hw3/37.html (score=1.3919)

Query returned 10 results in 40.23 ms.
```

### Query #8: “bug”

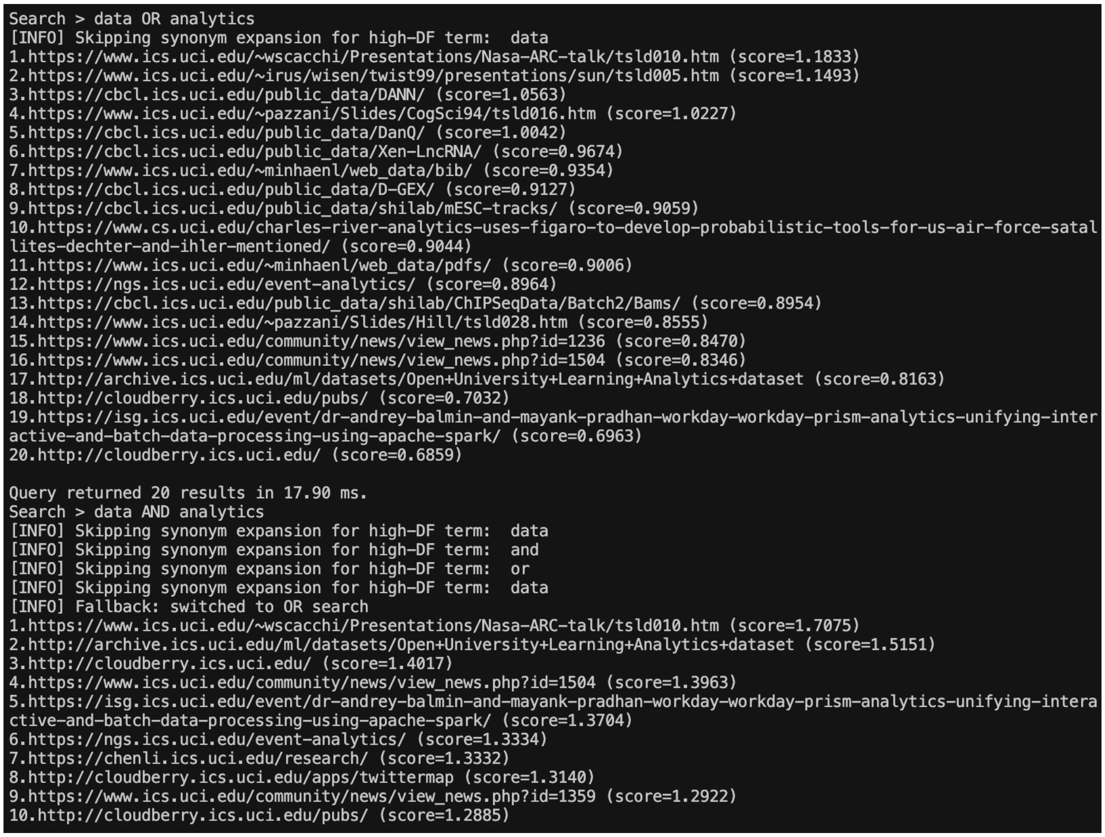
- Bug can mean an insect, a software defect, or a glitch. Because the query is ambiguous, the search engine provides results for all possible interpretations, effectively supporting polysemy.



```
Search > bug
1.https://www.ics.uci.edu/~eppstein/pix/t5bd/BugBugSpiderRules.html (score=1.7361)
2.https://mailman.ics.uci.edu/mailman/admin/nsf-career (score=1.7305)
3.https://www.ics.uci.edu/~eppstein/pix/wayzgoose14/SafetyBug.html (score=1.6907)
4.https://www.ics.uci.edu/~eppstein/pix/wayzgoose11/MiniBug.html (score=1.6872)
5.https://www.ics.uci.edu/~eppstein/pix/t5bd/NicoBugHunt.html (score=1.5366)
6.https://www.ics.uci.edu/~eppstein/pix/t5bd/DanielBugHunt.html (score=1.5300)
7.https://www.ics.uci.edu/~eppstein/pix/t5bd/BugHunters2.html (score=1.4561)
8.https://www.ics.uci.edu/~eppstein/pix/t5bd/Caterpillar.html (score=1.2018)
9.https://www.ics.uci.edu/~wscacchi/Papers/UIUC/sqa-overview.html (score=0.7754)
10.https://www.ics.uci.edu/~gbortis/ (score=0.7433)

Query returned 10 results in 3.96 ms.
```

### Query #9: Boolean queries → “data OR analytics” and “data AND analytics”

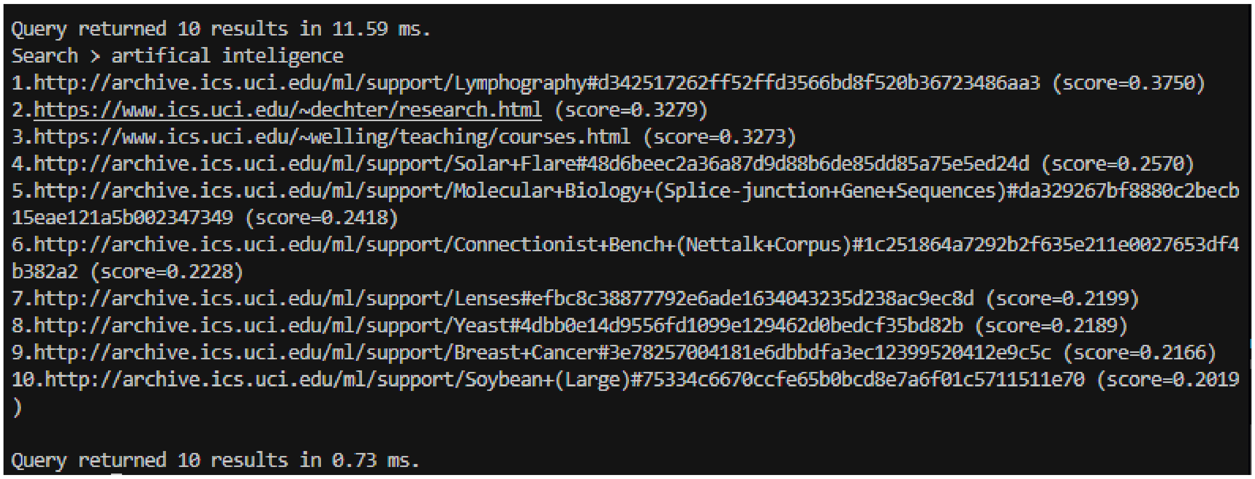
- Our search engine effectively supports boolean queries and provides accurate results that respect AND, OR, and NOT operators.
- Once AND queries are exhausted, fallback will occur, filling the remaining top-k results with an OR query to ensure that relevant documents are still retrieved even if they don’t match all terms.



```
Search > data OR analytics
[INFO] Skipping synonym expansion for high-DF term:  data
1.https://www.ics.uci.edu/~wscacchi/Presentations/Nasa-ARC-talk/tsld010.htm (score=1.1833)
2.https://www.ics.uci.edu/~irus/wisen/twist99/presentations/sun/tsld005.htm (score=1.1493)
3.https://cbcl.ics.uci.edu/public_data/DANN/ (score=1.0563)
4.https://www.ics.uci.edu/~pazzani/Slides/CogSci94/tsld016.htm (score=1.0227)
5.https://cbcl.ics.uci.edu/public_data/DanQ/ (score=1.0042)
6.https://cbcl.ics.uci.edu/public_data/Xen-LncRNA/ (score=0.9674)
7.https://www.ics.uci.edu/~minhaenl/web_data/bib/ (score=0.9354)
8.https://cbcl.ics.uci.edu/public_data/D-GEX/ (score=0.9127)
9.https://cbcl.ics.uci.edu/public_data/shilab/mESC-tracks/ (score=0.9059)
10.https://www.cs.uci.edu/charles-river-analytics-uses-figaro-to-develop-probabilistic-tools-for-us-air-force-satal
lites-dechter-and-ihler-mentioned/ (score=0.9044)
11.https://www.ics.uci.edu/~minhaenl/web_data/pdfs/ (score=0.9006)
12.https://ngs.ics.uci.edu/event-analytics/ (score=0.8964)
13.https://cbcl.ics.uci.edu/public_data/shilab/ChIPSeqData/Batch2/Bams/ (score=0.8954)
14.https://www.ics.uci.edu/~pazzani/Slides/Hill/tsld028.htm (score=0.8555)
15.https://www.ics.uci.edu/community/news/view_news.php?id=1236 (score=0.8470)
16.https://www.ics.uci.edu/community/news/view_news.php?id=1504 (score=0.8346)
17.http://archive.ics.uci.edu/ml/datasets/Open+University+Learning+Analytics+dataset (score=0.8163)
18.http://cloudberry.ics.uci.edu/pubs/ (score=0.7032)
19.https://isg.ics.uci.edu/event/dr-andrey-balmin-and-mayank-pradhan-workday-workday-prism-analytics-unifying-inter
active-and-batch-data-processing-using-apache-spark/ (score=0.6963)
20.http://cloudberry.ics.uci.edu/ (score=0.6859)

Query returned 20 results in 17.90 ms.
Search > data AND analytics
[INFO] Skipping synonym expansion for high-DF term:  data
[INFO] Skipping synonym expansion for high-DF term:  and
[INFO] Skipping synonym expansion for high-DF term:  or
[INFO] Skipping synonym expansion for high-DF term:  data
[INFO] Fallback: switched to OR search
1.https://www.ics.uci.edu/~wscacchi/Presentations/Nasa-ARC-talk/tsld010.htm (score=1.7075)
2.http://archive.ics.uci.edu/ml/datasets/Open+University+Learning+Analytics+dataset (score=1.5151)
3.http://cloudberry.ics.uci.edu/ (score=1.4017)
4.https://www.ics.uci.edu/community/news/view_news.php?id=1504 (score=1.3963)
5.https://isg.ics.uci.edu/event/dr-andrey-balmin-and-mayank-pradhan-workday-workday-prism-analytics-unifying-intera
ctive-and-batch-data-processing-using-apache-spark/ (score=1.3704)
6.https://ngs.ics.uci.edu/event-analytics/ (score=1.3334)
7.https://chenli.ics.uci.edu/research/ (score=1.3332)
8.http://cloudberry.ics.uci.edu/apps/twittermap (score=1.3140)
9.https://www.ics.uci.edu/community/news/view_news.php?id=1359 (score=1.2922)
10.http://cloudberry.ics.uci.edu/pubs/ (score=1.2885)
```

**Query #10: “Artifical inteligence”**

- Even for slightly misspelled queries, the search engine is still able to find relevant queries using synonym expansion, at the expense of getting lower TF-IDF ranking scores.

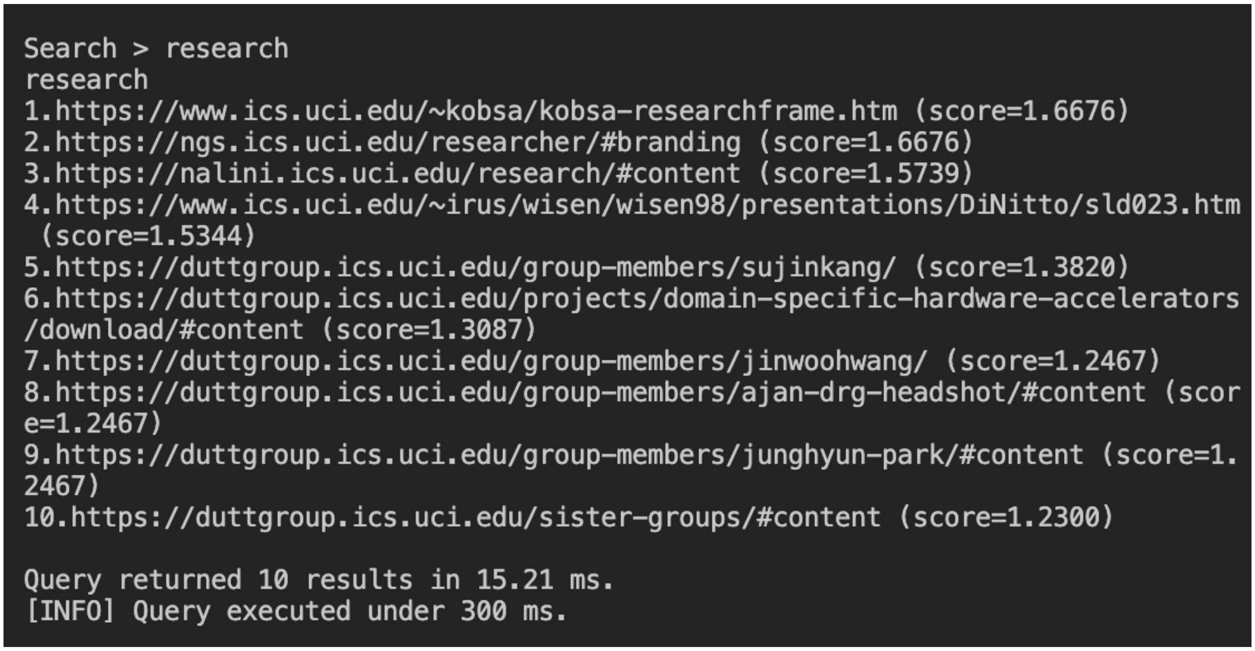


```
Query returned 10 results in 11.59 ms.
Search > artifical inteligence
1.http://archive.ics.uci.edu/ml/support/Lymphography#d342517262ff52ffd3566bd8f520b36723486aa3 (score=0.3750)
2.https://www.ics.uci.edu/~dechter/research.html (score=0.3279)
3.https://www.ics.uci.edu/~welling/teaching/courses.html (score=0.3273)
4.http://archive.ics.uci.edu/ml/support/Solar+Flare#48d6beec2a36a87d9d88b6de85dd85a75e5ed24d (score=0.2570)
5.http://archive.ics.uci.edu/ml/support/Molecular+Biology+(Splice-junction+Gene+Sequences)#da329267bf8880c2becb
15eae121a5b002347349 (score=0.2418)
6.http://archive.ics.uci.edu/ml/support/Connectionist+Bench+(Nettalk+Corpus)#1c251864a7292b2f635e211e0027653df4
b382a2 (score=0.2228)
7.http://archive.ics.uci.edu/ml/support/Lenses#efbc8c38877792e6ade1634043235d238ac9ec8d (score=0.2199)
8.http://archive.ics.uci.edu/ml/support/Yeast#4dbb0e14d9556fd1099e129462d0bedcf35bd82b (score=0.2189)
9.http://archive.ics.uci.edu/ml/support/Breast+Cancer#3e78257004181e6dbbdfa3ec12399520412e9c5c (score=0.2166)
10.http://archive.ics.uci.edu/ml/support/Soybean+(Large)#75334c6670ccfe65b0bcd8e7a6f01c5711511e70 (score=0.2019
)

Query returned 10 results in 0.73 ms.
```

## Bad Performance Queries (brief explanation & fix):

**Query #1: “research”**

- Research is such a commonly used term within the ICS space that it is difficult to rank documents, we used TF-IDF to better help rank this using a ranking system.

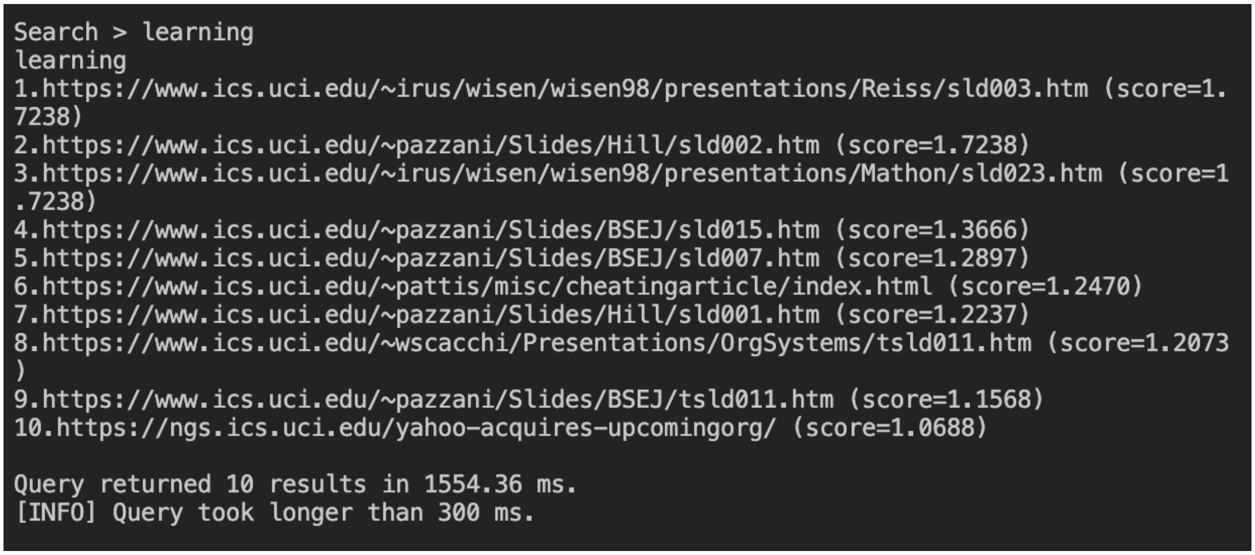


```
Search > research
research
1.https://www.ics.uci.edu/~kobsa/kobsa-researchframe.htm (score=1.6676)
2.https://ngs.ics.uci.edu/researcher/#branding (score=1.6676)
3.https://nalini.ics.uci.edu/research/#content (score=1.5739)
4.https://www.ics.uci.edu/~irus/wisen/wisen98/presentations/DiNitto/sld023.htm
 (score=1.5344)
5.https://duttgroup.ics.uci.edu/group-members/sujinkang/ (score=1.3820)
6.https://duttgroup.ics.uci.edu/projects/domain-specific-hardware-accelerators
/download/#content (score=1.3087)
7.https://duttgroup.ics.uci.edu/group-members/jinwoohwang/ (score=1.2467)
8.https://duttgroup.ics.uci.edu/group-members/ajan-drg-headshot/#content (scor
e=1.2467)
9.https://duttgroup.ics.uci.edu/group-members/junghyun-park/#content (score=1.
2467)
10.https://duttgroup.ics.uci.edu/sister-groups/#content (score=1.2300)

Query returned 10 results in 15.21 ms.
[INFO] Query executed under 300 ms.
```

-

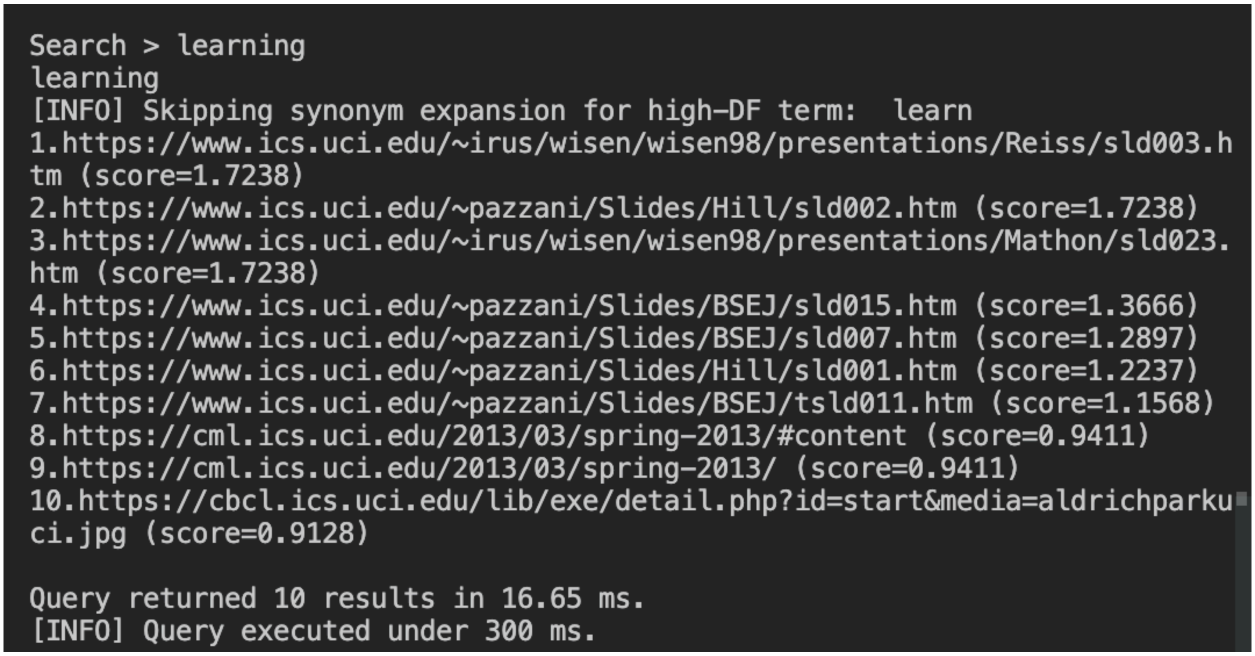
Synonym matching:

- Example query: “learning” can dramatically reduce the speed of the engine.



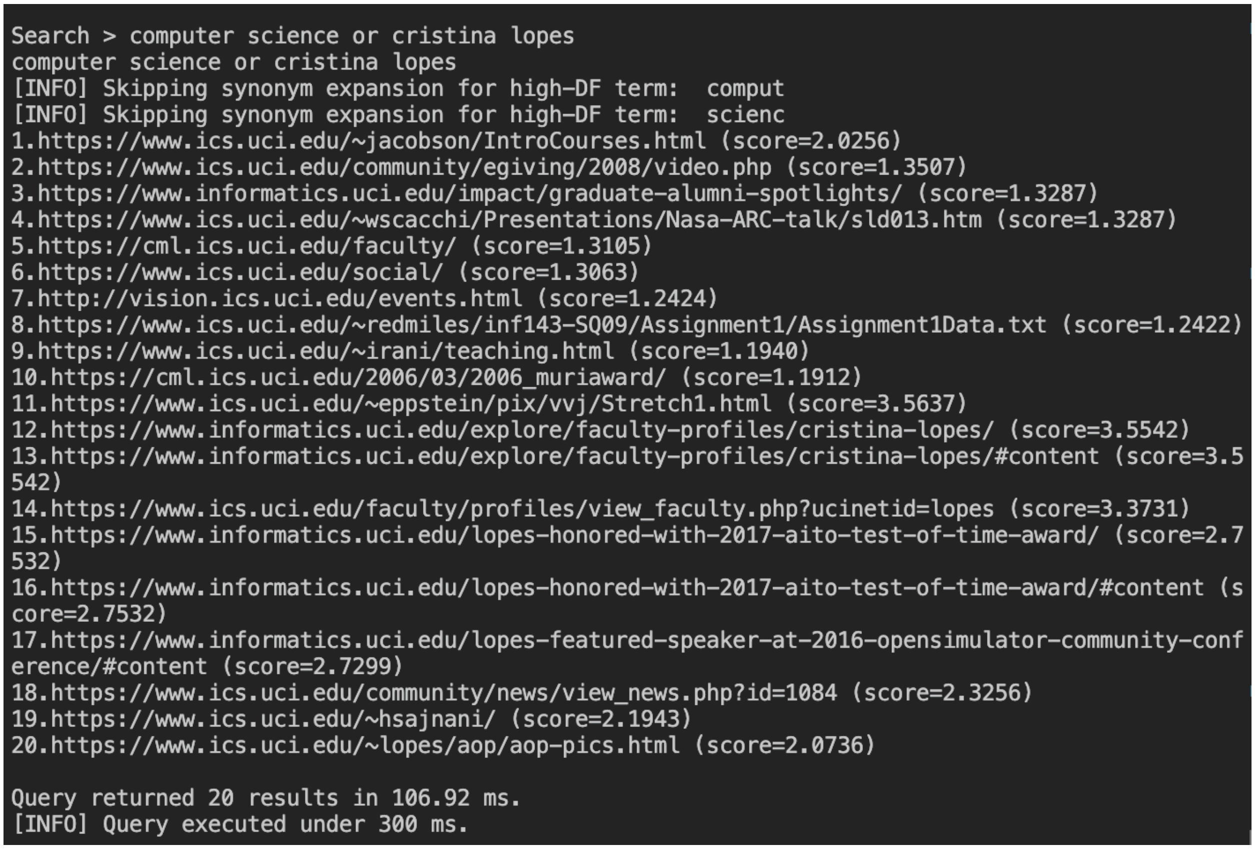
```
Search > learning
learning
1.https://www.ics.uci.edu/~irus/wisen/wisen98/presentations/Reiss/sld003.htm (score=1.
7238)
2.https://www.ics.uci.edu/~pazzani/Slides/Hill/sld002.htm (score=1.7238)
3.https://www.ics.uci.edu/~irus/wisen/wisen98/presentations/Mathon/sld023.htm (score=1
.7238)
4.https://www.ics.uci.edu/~pazzani/Slides/BSEJ/sld015.htm (score=1.3666)
5.https://www.ics.uci.edu/~pazzani/Slides/BSEJ/sld007.htm (score=1.2897)
6.https://www.ics.uci.edu/~pattis/misc/cheatingarticle/index.html (score=1.2470)
7.https://www.ics.uci.edu/~pazzani/Slides/Hill/sld001.htm (score=1.2237)
8.https://www.ics.uci.edu/~wscacchi/Presentations/OrgSystems/tsld011.htm (score=1.2073
)
9.https://www.ics.uci.edu/~pazzani/Slides/BSEJ/tsld011.htm (score=1.1568)
10.https://ngs.ics.uci.edu/yahoo-acquires-upcomingorg/ (score=1.0688)

Query returned 10 results in 1554.36 ms.
[INFO] Query took longer than 300 ms.
```

-
- Fix: Disabled synonym expansion for high-DF terms (if DF generates more than 2000 synonyms, skip)

- 

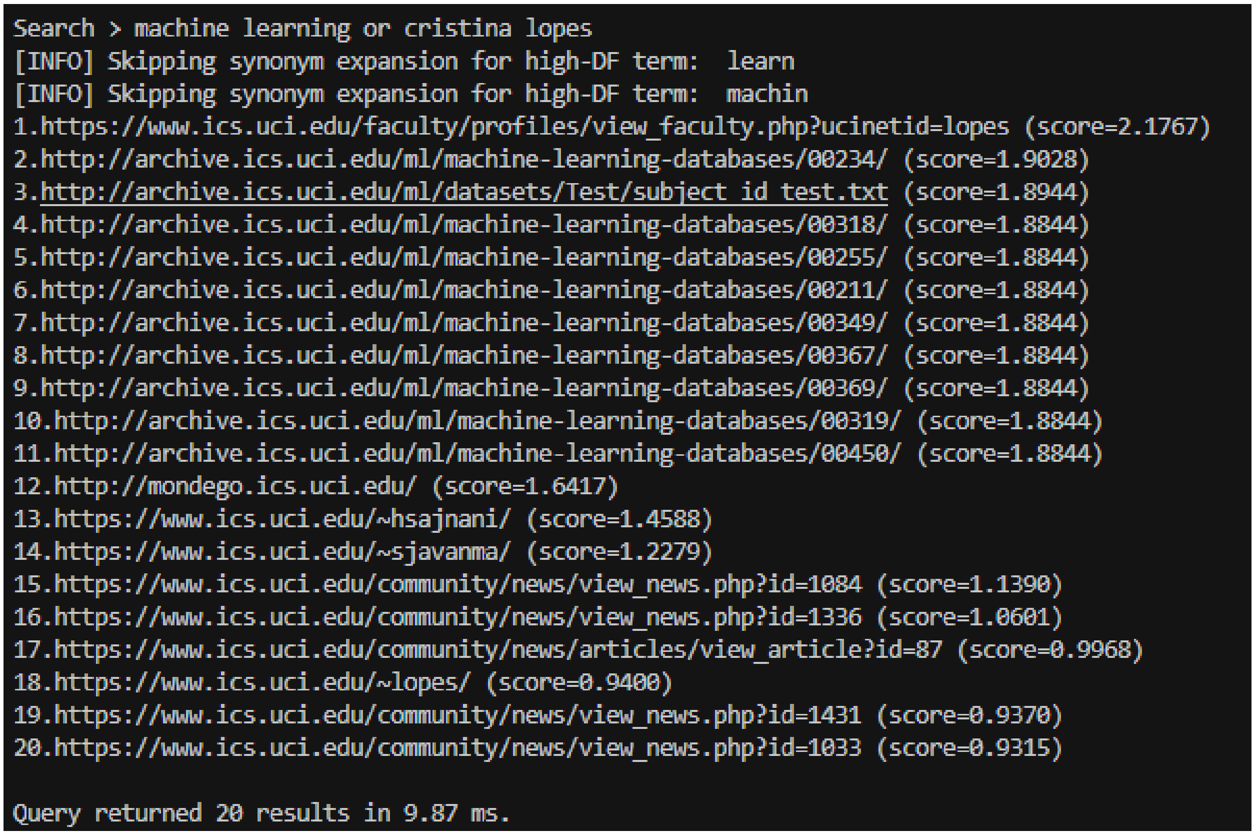
```
Search > learning
learning
[INFO] Skipping synonym expansion for high-DF term:  learn
1.https://www.ics.uci.edu/~irus/wisen/wisen98/presentations/Reiss/sld003.h
tm (score=1.7238)
2.https://www.ics.uci.edu/~pazzani/Slides/Hill/sld002.htm (score=1.7238)
3.https://www.ics.uci.edu/~irus/wisen/wisen98/presentations/Mathon/sld023.
htm (score=1.7238)
4.https://www.ics.uci.edu/~pazzani/Slides/BSEJ/sld015.htm (score=1.3666)
5.https://www.ics.uci.edu/~pazzani/Slides/BSEJ/sld007.htm (score=1.2897)
6.https://www.ics.uci.edu/~pazzani/Slides/Hill/sld001.htm (score=1.2237)
7.https://www.ics.uci.edu/~pazzani/Slides/BSEJ/tsld011.htm (score=1.1568)
8.https://cml.ics.uci.edu/2013/03/spring-2013/#content (score=0.9411)
9.https://cml.ics.uci.edu/2013/03/spring-2013/ (score=0.9411)
10.https://cbcl.ics.uci.edu/lib/exe/detail.php?id=start&media=aldrichparku
ci.jpg (score=0.9128)

Query returned 10 results in 16.65 ms.
[INFO] Query executed under 300 ms.
```

**Query #2: “computer science or cristina lopes”**

- May have a bug with the ranking order of TF-IDF when searching for multiple terms (see result #12)
- Also gives too many results
- 

```
Search > computer science or cristina lopes
computer science or cristina lopes
[INFO] Skipping synonym expansion for high-DF term:  comput
[INFO] Skipping synonym expansion for high-DF term:  scienc
1.https://www.ics.uci.edu/~jacobson/IntroCourses.html (score=2.0256)
2.https://www.ics.uci.edu/community/egiving/2008/video.php (score=1.3507)
3.https://www.informatics.uci.edu/impact/graduate-alumni-spotlights/ (score=1.3287)
4.https://www.ics.uci.edu/~wscacchi/Presentations/Nasa-ARC-talk/sld013.htm (score=1.3287)
5.https://cml.ics.uci.edu/faculty/ (score=1.3105)
6.https://www.ics.uci.edu/social/ (score=1.3063)
7.http://vision.ics.uci.edu/events.html (score=1.2424)
8.https://www.ics.uci.edu/~redmiles/inf143-SQ09/Assignment1/Assignment1Data.txt (score=1.2422)
9.https://www.ics.uci.edu/~irani/teaching.html (score=1.1940)
10.https://cml.ics.uci.edu/2006/03/2006_muriaward/ (score=1.1912)
11.https://www.ics.uci.edu/~eppstein/pix/vvj/Stretch1.html (score=3.5637)
12.https://www.informatics.uci.edu/explore/faculty-profiles/cristina-lopes/ (score=3.5542)
13.https://www.informatics.uci.edu/explore/faculty-profiles/cristina-lopes/#content (score=3.5
542)
14.https://www.ics.uci.edu/faculty/profiles/view_faculty.php?ucinetid=lopes (score=3.3731)
15.https://www.informatics.uci.edu/lopes-honored-with-2017-aito-test-of-time-award/ (score=2.7
532)
16.https://www.informatics.uci.edu/lopes-honored-with-2017-aito-test-of-time-award/#content (s
core=2.7532)
17.https://www.informatics.uci.edu/lopes-featured-speaker-at-2016-opensimulator-community-conf
erence/#content (score=2.7299)
18.https://www.ics.uci.edu/community/news/view_news.php?id=1084 (score=2.3256)
19.https://www.ics.uci.edu/~hsajnani/ (score=2.1943)
20.https://www.ics.uci.edu/~lopes/aop/aop-pics.html (score=2.0736)

Query returned 20 results in 106.92 ms.
[INFO] Query executed under 300 ms.
```

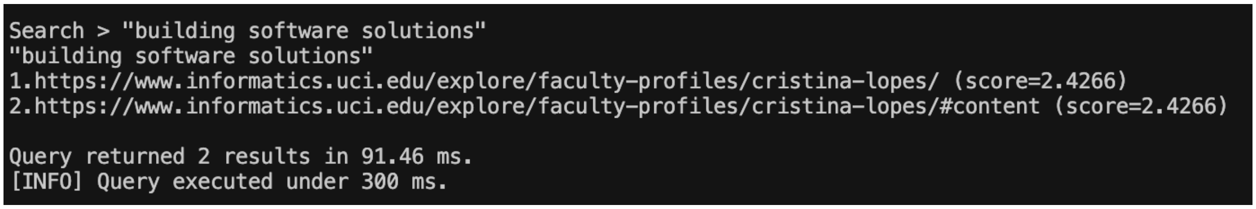
- Fix: Adjusted search results by adding a sorting algorithm



```
Search > machine learning or cristina lopes
[INFO] Skipping synonym expansion for high-DF term:  learn
[INFO] Skipping synonym expansion for high-DF term:  machin
1.https://www.ics.uci.edu/faculty/profiles/view_faculty.php?ucinetid=lopes (score=2.1767)
2.http://archive.ics.uci.edu/ml/machine-learning-databases/00234/ (score=1.9028)
3.http://archive.ics.uci.edu/ml/datasets/Test/subject_id_test.txt (score=1.8944)
4.http://archive.ics.uci.edu/ml/machine-learning-databases/00318/ (score=1.8844)
5.http://archive.ics.uci.edu/ml/machine-learning-databases/00255/ (score=1.8844)
6.http://archive.ics.uci.edu/ml/machine-learning-databases/00211/ (score=1.8844)
7.http://archive.ics.uci.edu/ml/machine-learning-databases/00349/ (score=1.8844)
8.http://archive.ics.uci.edu/ml/machine-learning-databases/00367/ (score=1.8844)
9.http://archive.ics.uci.edu/ml/machine-learning-databases/00369/ (score=1.8844)
10.http://archive.ics.uci.edu/ml/machine-learning-databases/00319/ (score=1.8844)
11.http://archive.ics.uci.edu/ml/machine-learning-databases/00450/ (score=1.8844)
12.http://mondego.ics.uci.edu/ (score=1.6417)
13.https://www.ics.uci.edu/~hsajnani/ (score=1.4588)
14.https://www.ics.uci.edu/~sjavanma/ (score=1.2279)
15.https://www.ics.uci.edu/community/news/view_news.php?id=1084 (score=1.1390)
16.https://www.ics.uci.edu/community/news/view_news.php?id=1336 (score=1.0601)
17.https://www.ics.uci.edu/community/news/articles/view_article?id=87 (score=0.9968)
18.https://www.ics.uci.edu/~lopes/ (score=0.9400)
19.https://www.ics.uci.edu/community/news/view_news.php?id=1431 (score=0.9370)
20.https://www.ics.uci.edu/community/news/view_news.php?id=1033 (score=0.9315)

Query returned 20 results in 9.87 ms.
```

**Query #3: “building software solutions”**

- Earlier in the project, we added support for exact phrase queries containing more than 2 terms. However, the search engine returned only documents containing the exact phrase , which often resulted in fewer than k results and no similar or partially matching alternatives.



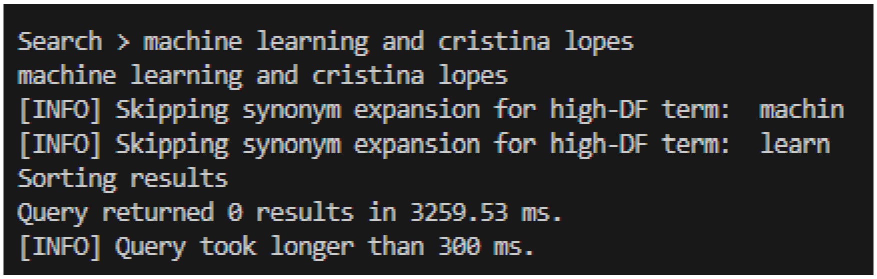
```
Search > "building software solutions"
"building software solutions"
1.https://www.informatics.uci.edu/explore/faculty-profiles/cristina-lopes/ (score=2.4266)
2.https://www.informatics.uci.edu/explore/faculty-profiles/cristina-lopes/#content (score=2.4266)

Query returned 2 results in 91.46 ms.
[INFO] Query executed under 300 ms.
```

- Fix: If an exact phrase match returns fewer than k results, the system will apply query relaxation and fallback using boolean AND sub queries. For this example, the phrase would be split into 2 AND queries:
    1. building AND software
    2. software AND solutions)

  These relaxed sub queries allow the engine to retrieve documents that contain related term combinations if the full phrase does not appear to fill the top k results.

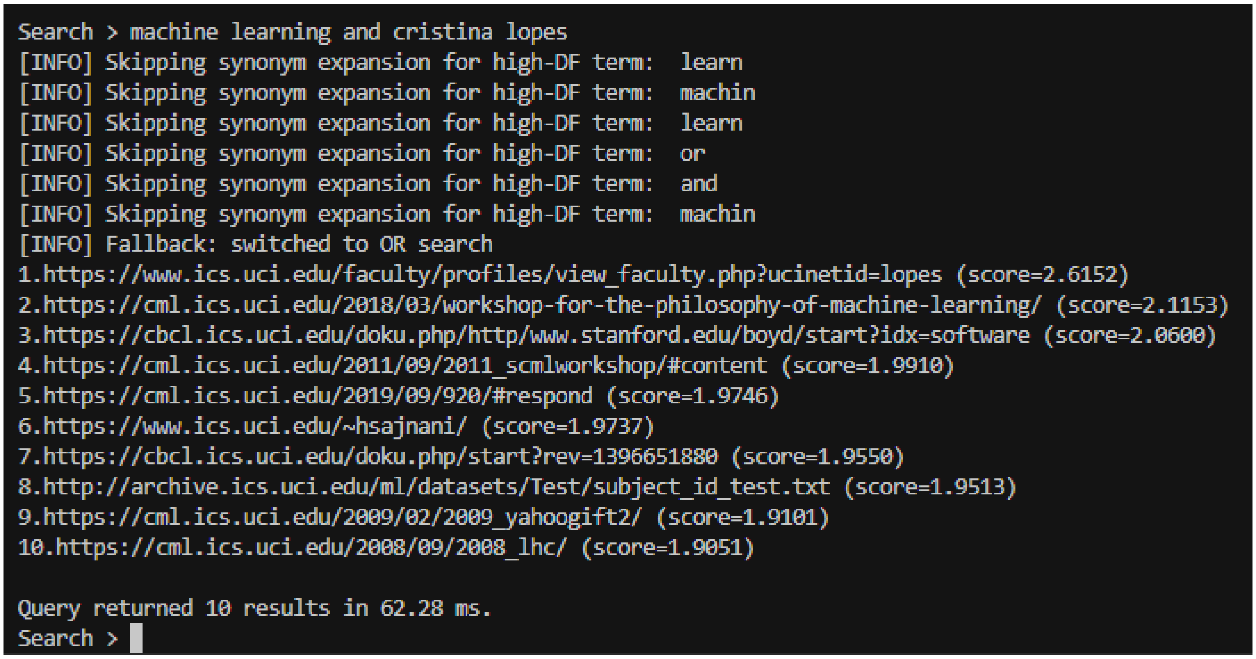
**Query #4: “Machine learning and cristina lopes” Fallback search for 0 search results**

- Even with 0 search results, the query will try for “or” if the query fails. The results with higher TF-IDF will show up first
- Original:



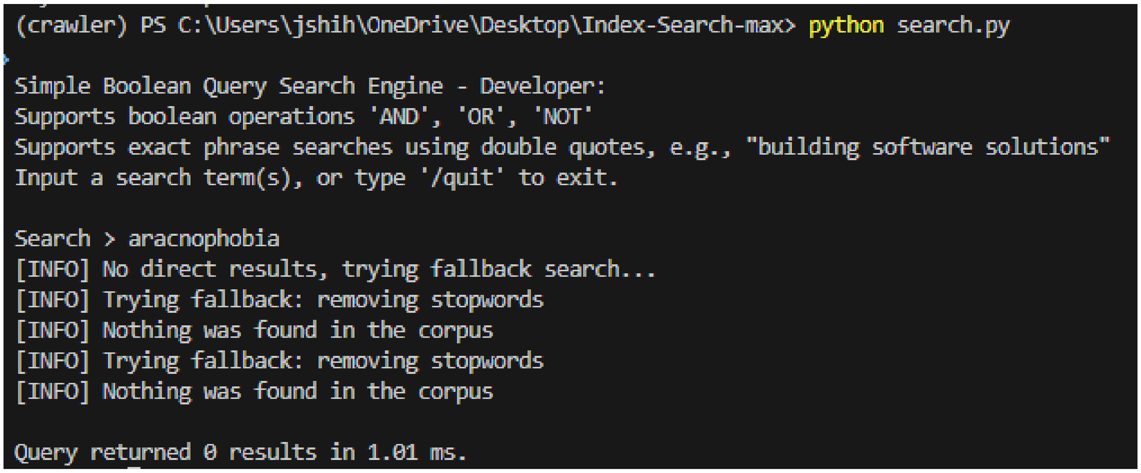
```
Search > machine learning and cristina lopes
machine learning and cristina lopes
[INFO] Skipping synonym expansion for high-DF term:  machin
[INFO] Skipping synonym expansion for high-DF term:  learn
Sorting results
Query returned 0 results in 3259.53 ms.
[INFO] Query took longer than 300 ms.
```

- Fix:



```
Search > machine learning and cristina lopes
[INFO] Skipping synonym expansion for high-DF term:  learn
[INFO] Skipping synonym expansion for high-DF term:  machin
[INFO] Skipping synonym expansion for high-DF term:  learn
[INFO] Skipping synonym expansion for high-DF term:  or
[INFO] Skipping synonym expansion for high-DF term:  and
[INFO] Skipping synonym expansion for high-DF term:  machin
[INFO] Fallback: switched to OR search
1.https://www.ics.uci.edu/faculty/profiles/view_faculty.php?ucinetid=lopes (score=2.6152)
2.https://cml.ics.uci.edu/2018/03/workshop-for-the-philosophy-of-machine-learning/ (score=2.1153)
3.https://cbcl.ics.uci.edu/doku.php/http/www.stanford.edu/boyd/start?idx=software (score=2.0600)
4.https://cml.ics.uci.edu/2011/09/2011_scmlworkshop/#content (score=1.9910)
5.https://cml.ics.uci.edu/2019/09/920/#respond (score=1.9746)
6.https://www.ics.uci.edu/~hsajnani/ (score=1.9737)
7.https://cbcl.ics.uci.edu/doku.php/start?rev=1396651880 (score=1.9550)
8.http://archive.ics.uci.edu/ml/datasets/Test/subject_id_test.txt (score=1.9513)
9.https://cml.ics.uci.edu/2009/02/2009_yahoogift2/ (score=1.9101)
10.https://cml.ics.uci.edu/2008/09/2008_lhc/ (score=1.9051)

Query returned 10 results in 62.28 ms.
Search >
```

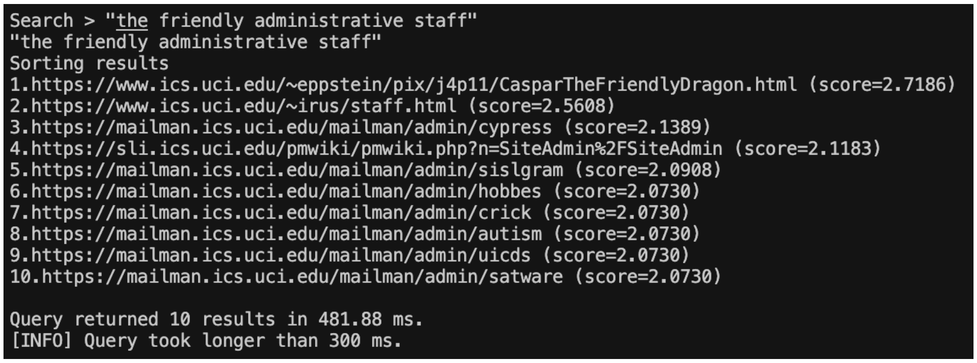
**Query #5: “Aracnophobia” Fallback search infinite loop, fixed.**



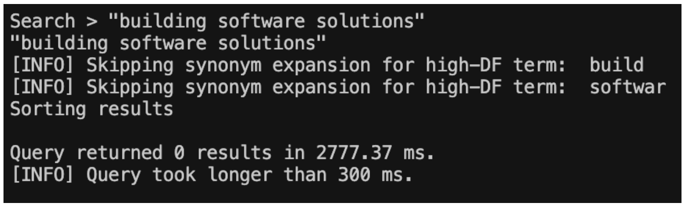
```
(crawler) PS C:\Users\jshih\OneDrive\Desktop\Index-Search-max> python search.py

Simple Boolean Query Search Engine - Developer:
Supports boolean operations 'AND', 'OR', 'NOT'
Supports exact phrase searches using double quotes, e.g., "building software solutions"
Input a search term(s), or type '/quit' to exit.

Search > aracnophobia
[INFO] No direct results, trying fallback search...
[INFO] Trying fallback: removing stopwords
[INFO] Nothing was found in the corpus
[INFO] Trying fallback: removing stopwords
[INFO] Nothing was found in the corpus

Query returned 0 results in 1.01 ms.
```

**Query #6: “the friendly administrative staff” and “building software solutions”**

- After implementing synonym expansion in our search engine, multi-term queries took an exceedingly large amount of time to run, up to 3000 ms.



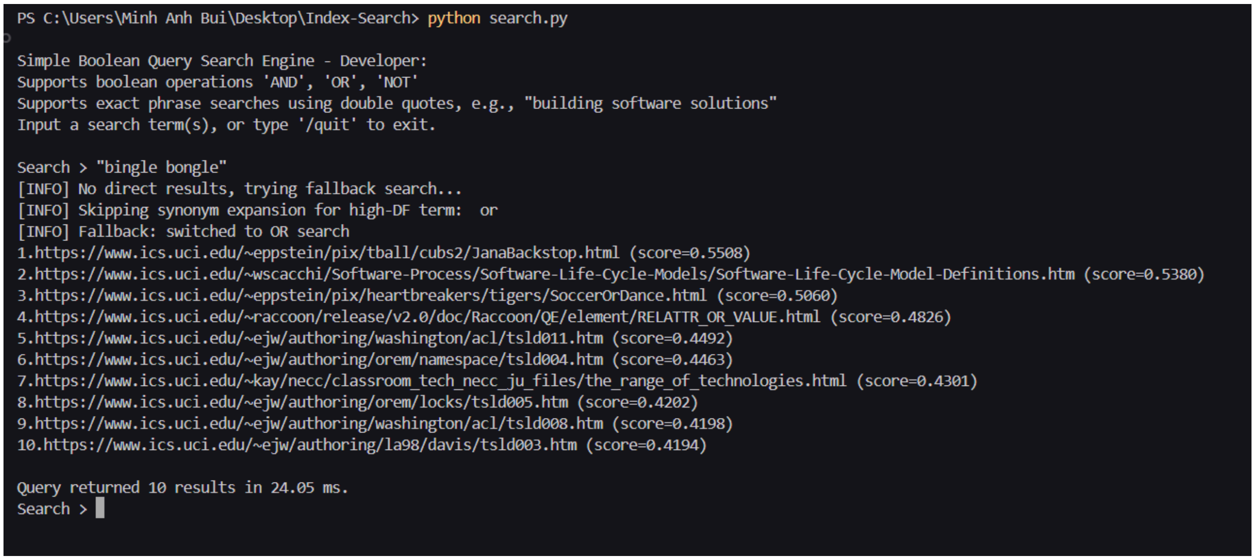
```
Search > "the friendly administrative staff"
"the friendly administrative staff"
Sorting results
1.https://www.ics.uci.edu/~eppstein/pix/j4p11/CasparTheFriendlyDragon.html (score=2.7186)
2.https://www.ics.uci.edu/~irus/staff.html (score=2.5608)
3.https://mailman.ics.uci.edu/mailman/admin/cypress (score=2.1389)
4.https://sli.ics.uci.edu/pmwiki/pmwiki.php?n=SiteAdmin%2FSiteAdmin (score=2.1183)
5.https://mailman.ics.uci.edu/mailman/admin/sislgram (score=2.0908)
6.https://mailman.ics.uci.edu/mailman/admin/hobbes (score=2.0730)
7.https://mailman.ics.uci.edu/mailman/admin/crick (score=2.0730)
8.https://mailman.ics.uci.edu/mailman/admin/autism (score=2.0730)
9.https://mailman.ics.uci.edu/mailman/admin/uicds (score=2.0730)
10.https://mailman.ics.uci.edu/mailman/admin/satware (score=2.0730)

Query returned 10 results in 481.88 ms.
[INFO] Query took longer than 300 ms.
```



```
Search > "building software solutions"
"building software solutions"
[INFO] Skipping synonym expansion for high-DF term:  build
[INFO] Skipping synonym expansion for high-DF term:  softwar
Sorting results

Query returned 0 results in 2777.37 ms.
[INFO] Query took longer than 300 ms.
```

- Fix: To reduce query time (especially with synonym expansion) to under 300 ms, we precomputed synonyms and stored them in a separate file, synonyms.json. This file holds up to three synonyms for each indexed term, enabling fast dictionary lookups during search instead of recomputing synonyms on every query.

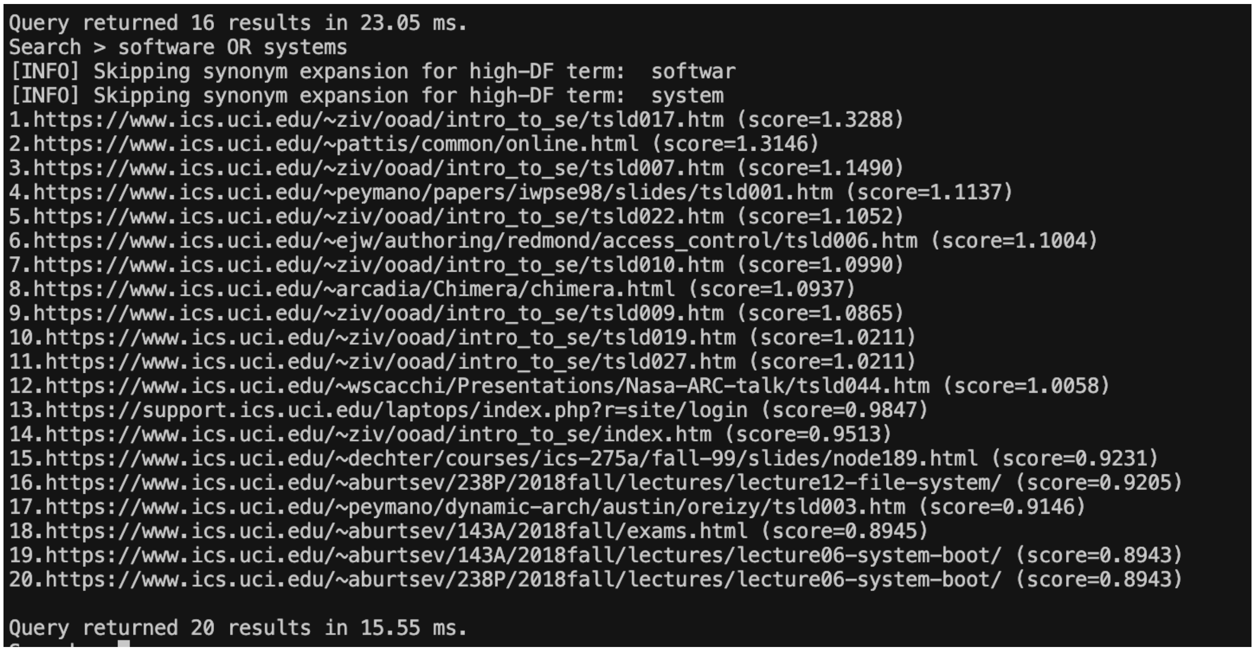
**Query #7: “bingle bongle” → Fallback_search, when primary search returns 0 results, will try weaker searches to get something useful**



```
PS C:\Users\Minh Anh Bui\Desktop\Index-Search> python search.py

Simple Boolean Query Search Engine - Developer:
Supports boolean operations 'AND', 'OR', 'NOT'
Supports exact phrase searches using double quotes, e.g., "building software solutions"
Input a search term(s), or type '/quit' to exit.

Search > "bingle bongle"
[INFO] No direct results, trying fallback search...
[INFO] Skipping synonym expansion for high-DF term:  or
[INFO] Fallback: switched to OR search
1.https://www.ics.uci.edu/~eppstein/pix/tball/cubs2/JanaBackstop.html (score=0.5508)
2.https://www.ics.uci.edu/~wscacchi/Software-Process/Software-Life-Cycle-Models/Software-Life-Cycle-Model-Definitions.htm (score=0.5380)
3.https://www.ics.uci.edu/~eppstein/pix/heartbreakers/tigers/SoccerOrDance.html (score=0.5060)
4.https://www.ics.uci.edu/~raccoon/release/v2.0/doc/Raccoon/QE/element/RELATTR_OR_VALUE.html (score=0.4826)
5.https://www.ics.uci.edu/~ejw/authoring/washington/acl/tsld011.htm (score=0.4492)
6.https://www.ics.uci.edu/~ejw/authoring/orem/namespace/tsld004.htm (score=0.4463)
7.https://www.ics.uci.edu/~kay/necc/classroom_tech_necc_ju_files/the_range_of_technologies.html (score=0.4301)
8.https://www.ics.uci.edu/~ejw/authoring/orem/locks/tsld005.htm (score=0.4202)
9.https://www.ics.uci.edu/~ejw/authoring/washington/acl/tsld008.htm (score=0.4198)
10.https://www.ics.uci.edu/~ejw/authoring/la98/davis/tsld003.htm (score=0.4194)

Query returned 10 results in 24.05 ms.
Search > 
```

### Query #8: “software OR systems”

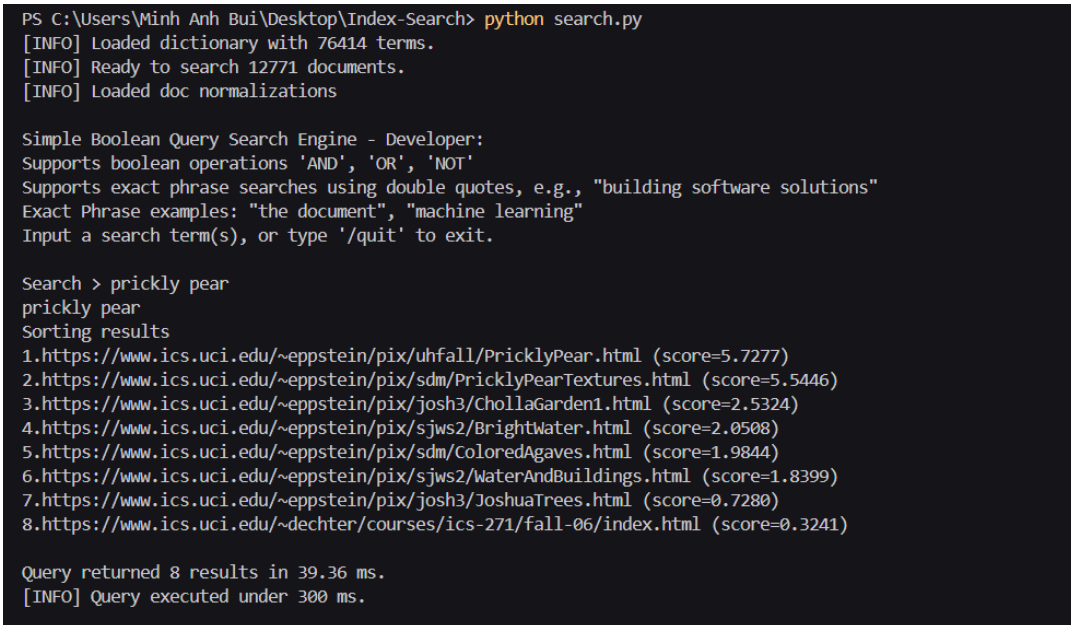
- Boolean OR queries currently exceed the intended top-10 limit, and we need to ensure they return only the highest-scoring 10 documents.



```
Query returned 16 results in 23.05 ms.
Search > software OR systems
[INFO] Skipping synonym expansion for high-DF term:  softwar
[INFO] Skipping synonym expansion for high-DF term:  system
1.https://www.ics.uci.edu/~ziv/ooad/intro_to_se/tsld017.htm (score=1.3288)
2.https://www.ics.uci.edu/~pattis/common/online.html (score=1.3146)
3.https://www.ics.uci.edu/~ziv/ooad/intro_to_se/tsld007.htm (score=1.1490)
4.https://www.ics.uci.edu/~peymano/papers/iwpse98/slides/tsld001.htm (score=1.1137)
5.https://www.ics.uci.edu/~ziv/ooad/intro_to_se/tsld022.htm (score=1.1052)
6.https://www.ics.uci.edu/~ejw/authoring/redmond/access_control/tsld006.htm (score=1.1004)
7.https://www.ics.uci.edu/~ziv/ooad/intro_to_se/tsld010.htm (score=1.0990)
8.https://www.ics.uci.edu/~arcadia/Chimera/chimera.html (score=1.0937)
9.https://www.ics.uci.edu/~ziv/ooad/intro_to_se/tsld009.htm (score=1.0865)
10.https://www.ics.uci.edu/~ziv/ooad/intro_to_se/tsld019.htm (score=1.0211)
11.https://www.ics.uci.edu/~ziv/ooad/intro_to_se/tsld027.htm (score=1.0211)
12.https://www.ics.uci.edu/~wscacchi/Presentations/Nasa-ARC-talk/tsld044.htm (score=1.0058)
13.https://support.ics.uci.edu/laptops/index.php?r=site/login (score=0.9847)
14.https://www.ics.uci.edu/~ziv/ooad/intro_to_se/index.htm (score=0.9513)
15.https://www.ics.uci.edu/~dechter/courses/ics-275a/fall-99/slides/node189.html (score=0.9231)
16.https://www.ics.uci.edu/~aburtsev/238P/2018fall/lectures/lecture12-file-system/ (score=0.9205)
17.https://www.ics.uci.edu/~peymano/dynamic-arch/austin/oreizy/tsld003.htm (score=0.9146)
18.https://www.ics.uci.edu/~aburtsev/143A/2018fall/exams.html (score=0.8945)
19.https://www.ics.uci.edu/~aburtsev/143A/2018fall/lectures/lecture06-system-boot/ (score=0.8943)
20.https://www.ics.uci.edu/~aburtsev/238P/2018fall/lectures/lecture06-system-boot/ (score=0.8943)

Query returned 20 results in 15.55 ms.
```

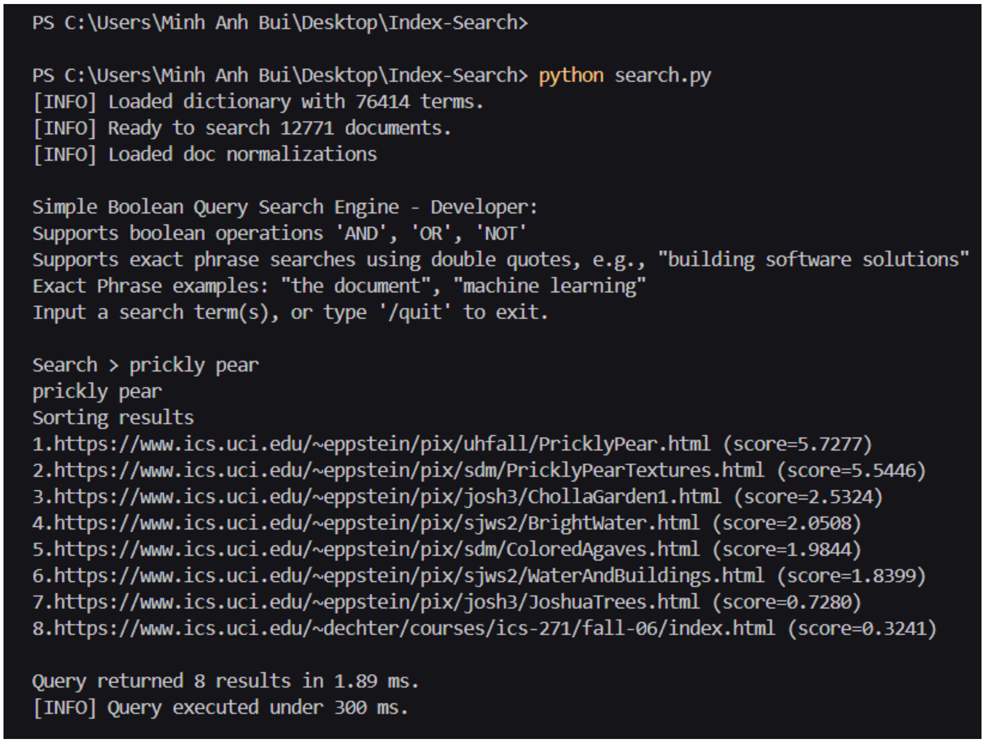
- Fix: Limit boolean OR queries to the first k (10) results only

**Query #9: Testing simhash**
**Without:**



```
PS C:\Users\Minh Anh Bui\Desktop\Index-Search> python search.py
[INFO] Loaded dictionary with 76414 terms.
[INFO] Ready to search 12771 documents.
[INFO] Loaded doc normalizations

Simple Boolean Query Search Engine - Developer:
Supports boolean operations 'AND', 'OR', 'NOT'
Supports exact phrase searches using double quotes, e.g., "building software solutions"
Exact Phrase examples: "the document", "machine learning"
Input a search term(s), or type '/quit' to exit.

Search > prickly pear
prickly pear
Sorting results
1.https://www.ics.uci.edu/~eppstein/pix/uhfall/PricklyPear.html (score=5.7277)
2.https://www.ics.uci.edu/~eppstein/pix/sdm/PricklyPearTextures.html (score=5.5446)
3.https://www.ics.uci.edu/~eppstein/pix/josh3/ChollaGarden1.html (score=2.5324)
4.https://www.ics.uci.edu/~eppstein/pix/sjws2/BrightWater.html (score=2.0508)
5.https://www.ics.uci.edu/~eppstein/pix/sdm/ColoredAgaves.html (score=1.9844)
6.https://www.ics.uci.edu/~eppstein/pix/sjws2/WaterAndBuildings.html (score=1.8399)
7.https://www.ics.uci.edu/~eppstein/pix/josh3/JoshuaTrees.html (score=0.7280)
8.https://www.ics.uci.edu/~dechter/courses/ics-271/fall-06/index.html (score=0.3241)

Query returned 8 results in 39.36 ms.
[INFO] Query executed under 300 ms.
```

**With simhash:**

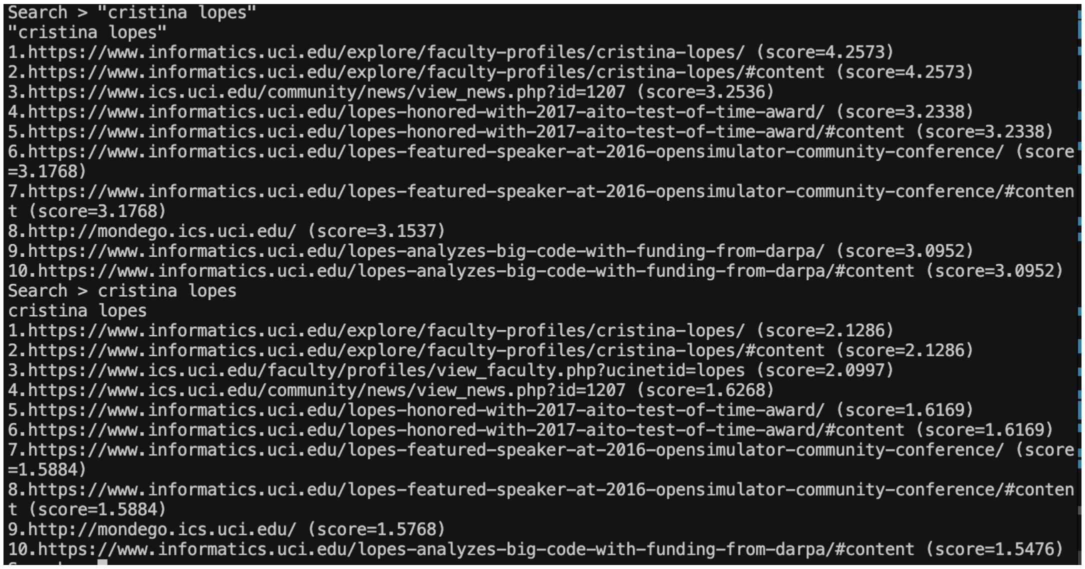


```
PS C:\Users\Minh Anh Bui\Desktop\Index-Search>

PS C:\Users\Minh Anh Bui\Desktop\Index-Search> python search.py
[INFO] Loaded dictionary with 76414 terms.
[INFO] Ready to search 12771 documents.
[INFO] Loaded doc normalizations

Simple Boolean Query Search Engine - Developer:
Supports boolean operations 'AND', 'OR', 'NOT'
Supports exact phrase searches using double quotes, e.g., "building software solutions"
Exact Phrase examples: "the document", "machine learning"
Input a search term(s), or type '/quit' to exit.

Search > prickly pear
prickly pear
Sorting results
1.https://www.ics.uci.edu/~eppstein/pix/uhfall/PricklyPear.html (score=5.7277)
2.https://www.ics.uci.edu/~eppstein/pix/sdm/PricklyPearTextures.html (score=5.5446)
3.https://www.ics.uci.edu/~eppstein/pix/josh3/ChollaGarden1.html (score=2.5324)
4.https://www.ics.uci.edu/~eppstein/pix/sjws2/BrightWater.html (score=2.0508)
5.https://www.ics.uci.edu/~eppstein/pix/sdm/ColoredAgaves.html (score=1.9844)
6.https://www.ics.uci.edu/~eppstein/pix/sjws2/WaterAndBuildings.html (score=1.8399)
7.https://www.ics.uci.edu/~eppstein/pix/josh3/JoshuaTrees.html (score=0.7280)
8.https://www.ics.uci.edu/~dechter/courses/ics-271/fall-06/index.html (score=0.3241)

Query returned 8 results in 1.89 ms.
[INFO] Query executed under 300 ms.
```

**Query #10: “cristina lopes” vs. cristina lopes**

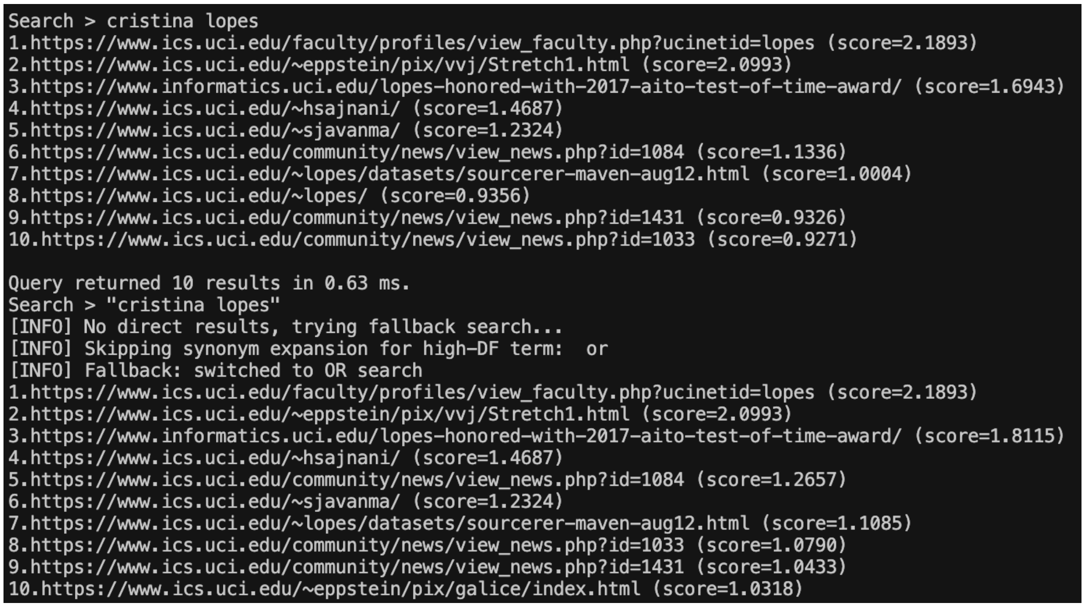
- In our search engine, users can specify phrase queries by enclosing them in quotation marks, distinguishing them from regular Boolean queries. Previously, phrase matches received a score boost because documents containing the full phrase were considered rarer. However, this caused their scores to become disproportionately high compared to other relevant results.

Previous:



```
Search > "cristina lopes"
"cristina lopes"
1.https://www.informatics.uci.edu/explore/faculty-profiles/cristina-lopes/ (score=4.2573)
2.https://www.informatics.uci.edu/explore/faculty-profiles/cristina-lopes/#content (score=4.2573)
3.https://www.ics.uci.edu/community/news/view_news.php?id=1207 (score=3.2536)
4.https://www.informatics.uci.edu/lopes-honored-with-2017-aito-test-of-time-award/ (score=3.2338)
5.https://www.informatics.uci.edu/lopes-honored-with-2017-aito-test-of-time-award/#content (score=3.2338)
6.https://www.informatics.uci.edu/lopes-featured-speaker-at-2016-opensimulator-community-conference/ (score
=3.1768)
7.https://www.informatics.uci.edu/lopes-featured-speaker-at-2016-opensimulator-community-conference/#conten
t (score=3.1768)
8.http://mondego.ics.uci.edu/ (score=3.1537)
9.https://www.informatics.uci.edu/lopes-analyzes-big-code-with-funding-from-darpa/ (score=3.0952)
10.https://www.informatics.uci.edu/lopes-analyzes-big-code-with-funding-from-darpa/#content (score=3.0952)
Search > cristina lopes
cristina lopes
1.https://www.informatics.uci.edu/explore/faculty-profiles/cristina-lopes/ (score=2.1286)
2.https://www.informatics.uci.edu/explore/faculty-profiles/cristina-lopes/#content (score=2.1286)
3.https://www.ics.uci.edu/faculty/profiles/view_faculty.php?ucinetid=lopes (score=2.0997)
4.https://www.ics.uci.edu/community/news/view_news.php?id=1207 (score=1.6268)
5.https://www.informatics.uci.edu/lopes-honored-with-2017-aito-test-of-time-award/ (score=1.6169)
6.https://www.informatics.uci.edu/lopes-honored-with-2017-aito-test-of-time-award/#content (score=1.6169)
7.https://www.informatics.uci.edu/lopes-featured-speaker-at-2016-opensimulator-community-conference/ (score
=1.5884)
8.https://www.informatics.uci.edu/lopes-featured-speaker-at-2016-opensimulator-community-conference/#conten
t (score=1.5884)
9.http://mondego.ics.uci.edu/ (score=1.5768)
10.https://www.informatics.uci.edu/lopes-analyzes-big-code-with-funding-from-darpa/#content (score=1.5476)
```

Current:



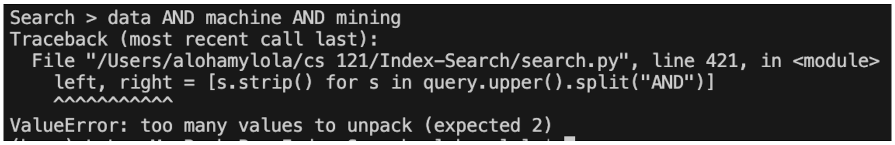
```
Search > cristina lopes
1.https://www.ics.uci.edu/faculty/profiles/view_faculty.php?ucinetid=lopes (score=2.1893)
2.https://www.ics.uci.edu/~eppstein/pix/vvj/Stretch1.html (score=2.0993)
3.https://www.informatics.uci.edu/lopes-honored-with-2017-aito-test-of-time-award/ (score=1.6943)
4.https://www.ics.uci.edu/~hsajnani/ (score=1.4687)
5.https://www.ics.uci.edu/~sjavanma/ (score=1.2324)
6.https://www.ics.uci.edu/community/news/view_news.php?id=1084 (score=1.1336)
7.https://www.ics.uci.edu/~lopes/datasets/sourcerer-maven-aug12.html (score=1.0004)
8.https://www.ics.uci.edu/~lopes/ (score=0.9356)
9.https://www.ics.uci.edu/community/news/view_news.php?id=1431 (score=0.9326)
10.https://www.ics.uci.edu/community/news/view_news.php?id=1033 (score=0.9271)

Query returned 10 results in 0.63 ms.
Search > "cristina lopes"
[INFO] No direct results, trying fallback search...
[INFO] Skipping synonym expansion for high-DF term:  or
[INFO] Fallback: switched to OR search
1.https://www.ics.uci.edu/faculty/profiles/view_faculty.php?ucinetid=lopes (score=2.1893)
2.https://www.ics.uci.edu/~eppstein/pix/vvj/Stretch1.html (score=2.0993)
3.https://www.informatics.uci.edu/lopes-honored-with-2017-aito-test-of-time-award/ (score=1.8115)
4.https://www.ics.uci.edu/~hsajnani/ (score=1.4687)
5.https://www.ics.uci.edu/community/news/view_news.php?id=1084 (score=1.2657)
6.https://www.ics.uci.edu/~sjavanma/ (score=1.2324)
7.https://www.ics.uci.edu/~lopes/datasets/sourcerer-maven-aug12.html (score=1.1085)
8.https://www.ics.uci.edu/community/news/view_news.php?id=1033 (score=1.0790)
9.https://www.ics.uci.edu/community/news/view_news.php?id=1431 (score=1.0433)
10.https://www.ics.uci.edu/~eppstein/pix/galice/index.html (score=1.0318)
```

- Fix: Adjusted phrase query scoring by multiplying the existing TF-IDF score by a fixed factor (2.0). This ensures that phrase matches are ranked higher than individual term matches while remaining on the same scale as regular TF-IDF scores. After this change, the difference between phrase match scores and Boolean query scores is only marginal.

**Query #11: “data AND mining AND machine”**

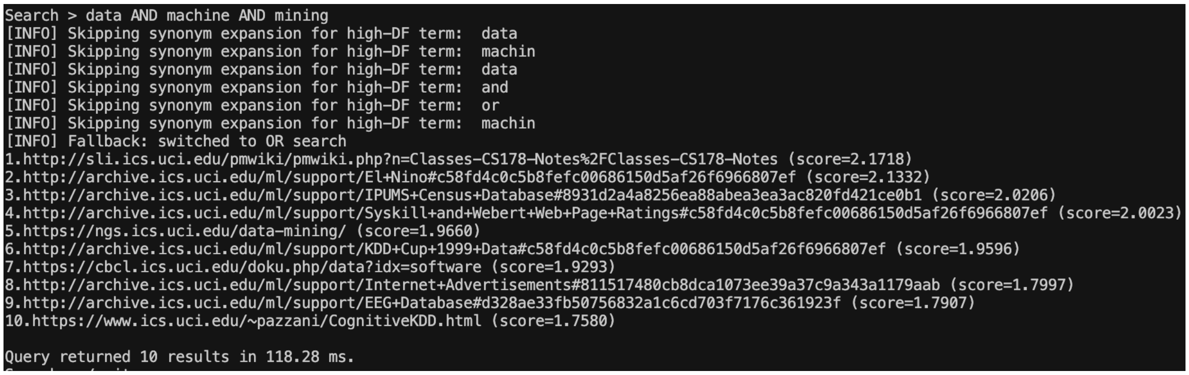
- The previous system could not handle boolean queries containing multiple operators. If a user were to enter a query with more than one boolean operator, the system failed to execute unless the query was enclosed in quotation marks, which the system would interpret as a phrase search.

Previous:



```
Search > data AND machine AND mining
Traceback (most recent call last):
  File "/Users/alohamylola/cs 121/Index-Search/search.py", line 421, in <module>
    left, right = [s.strip() for s in query.upper().split("AND")]
    ^^^^^^^^^^^
ValueError: too many values to unpack (expected 2)
```

Current:



```
Search > data AND machine AND mining
[INFO] Skipping synonym expansion for high-DF term:  data
[INFO] Skipping synonym expansion for high-DF term:  machin
[INFO] Skipping synonym expansion for high-DF term:  data
[INFO] Skipping synonym expansion for high-DF term:  and
[INFO] Skipping synonym expansion for high-DF term:  or
[INFO] Skipping synonym expansion for high-DF term:  machin
[INFO] Fallback: switched to OR search
1.http://sli.ics.uci.edu/pmwiki/pmwiki.php?n=Classes-CS178-Notes%2FClasses-CS178-Notes (score=2.1718)
2.http://archive.ics.uci.edu/ml/support/El+Nino#c58fd4c0c5b8fefc00686150d5af26f6966807ef (score=2.1332)
3.http://archive.ics.uci.edu/ml/support/IPUMS+Census+Database#8931d2a4a8256ea88abea3ea3ac820fd421ce0b1 (score=2.0206)
4.http://archive.ics.uci.edu/ml/support/Syskill+and+Webert+Web+Page+Ratings#c58fd4c0c5b8fefc00686150d5af26f6966807ef (score=2.0023)
5.https://ngs.ics.uci.edu/data-mining/ (score=1.9660)
6.http://archive.ics.uci.edu/ml/support/KDD+Cup+1999+Data#c58fd4c0c5b8fefc00686150d5af26f6966807ef (score=1.9596)
7.https://cbcl.ics.uci.edu/doku.php/data?idx=software (score=1.9293)
8.http://archive.ics.uci.edu/ml/support/Internet+Advertisements#811517480cb8dca1073ee39a37c9a343a1179aab (score=1.7997)
9.http://archive.ics.uci.edu/ml/support/EEG+Database#d328ae33fb50756832a1c6cd703f7176c361923f (score=1.7907)
10.https://www.ics.uci.edu/~pazzani/CognitiveKDD.html (score=1.7580)

Query returned 10 results in 118.28 ms.
```

- Fix: Implemented a multi-operator boolean evaluation system by introducing an enhanced query handler to process multiple operators within a single query.

## Extra credit:

- “Detect and eliminate duplicate pages. (1 point for exact, 2 points for near)” → We implemented simhash (see simhash.py) and skips files that are too similar (threshold = 0.9)
- “Index anchor words for the target pages (1 point)” → See tokenizer.py for sample text