

Big Data Assignment 2

Miasoedov Artemii

Methodology

Most search engines consist of 2 parts:

1. Indexer - collects information about the known documents for fast retrieval in future
2. Ranker - uses the prepared index to find documents relevant to the query

Indexer

To efficiently handle large amounts of data we use Hadoop's mapreduce pipeline. I decided to implement a pipeline consisting of 2 steps. First step handles each document separately, calculating term frequency inside each document. Imagine having 2 documents:

```
doc1: hello hello world
doc2: hello friend
```

Mapper receives the text from documents and splits in individual terms, producing output in the following format:

```
friend doc2 1
hello doc1 1
hello doc1 1
hello doc2 1
world doc1 1
```

Reducer calculates the number of term occurrences inside the document and inserts this data in Cassandra database.

```
friend doc2 1
hello doc1 2
hello doc2 1
world doc1 1
```

```
INSERT INTO tdf (term, doc, freq)
VALUES ('friend', 'doc2', 1),
       ('hello', 'doc1', 2),
       ('hello', 'doc2', 1),
       ('world', 'doc1', 1);
```

However, to build TF-IDF ranking we also need to know the length of each document and the total number of occurrences of each token. The data from the previous step might be used to

solve both of the tasks with a single reducer (counting length of each document and overall frequency of term), so mapper should produce data in a format similar to this one:

```
doc1 2
doc1 1
doc2 1
doc2 1
friend 1
hello 2
hello 1
world 1
```

Reducer will sum the values for the same key, thus finding document lengths and token frequencies:

```
doc1 3
doc2 2
friend 1
hello 3
world 1
```

However, we can't distinguish document names and terms for sure, so mapper should add some suffixes, specifying what kind of data it is:

```
doc1:d1 2
doc1:d1 1
doc2:d1 1
doc2:d1 1
friend:tf 1
hello:tf 2
hello:tf 1
world:tf 1
```

Reducer will insert obtained lengths and frequencies in the database:

```
INSERT INTO d1 (doc, len)
VALUES ('doc1', 3),
      ('doc2', 3);

INSERT INTO tf (term, docs, total)
VALUES ('friend', 1, 1),
      ('hello', 2, 3);
```

Now, we have all the required data to do ranking.

Ranking

To find the most relevant documents for a user query, we use a BM25 ranking algorithm implemented with PySpark and Cassandra. The process begins by tokenizing the input query, which breaks it into individual terms.

These query terms are then matched against the term frequency (tf), document frequency (tdf), and document length (dl) data previously stored in Cassandra. Each table is loaded into Spark as a DataFrame, and filters are applied to only consider terms that appear in the current query.

```
tf = spark.read \
    .format(cassandra) \
    .options(table="tf", keyspace=keyspace) \
    .load() \
    .filter(col("term").isin(query)) \
    .alias("tf")
```

The loaded DataFrames are then joined on common fields. The goal is to bring together all relevant term-document pairs with their corresponding frequency and document statistics.

```
index = dl \
    .join(tdf, "doc") \
    .join(tf, "term")
```

To compute BM25 scores, we first calculate the average document length and the total number of documents. In a more efficient scenario, this information should be stored, but for demonstration it is easier to be calculated.

```
agg = dl.agg(avg("len"), count("*")).first()
avgd1, n = agg
```

Finally, obtained metrics are used to calculate BM-25 score:

```
tf = (col("freq") * (k+1)) / (col("freq") + k * (1-b + b*n/avgd1))
idf = log(1 + (n - col("docs") + 0.5) / (col("docs") + 0.5))
index = index \
    .select(col("term"), col("doc"), (tf * idf).alias("score")) \
    .groupby("doc") \
    .agg(sum("score").alias("score")) \
    .orderBy(desc("score"))
```

Demonstration

Indexing

```
root@cluster-master: /app -- Console

New Tab Split View
Copy Paste Find...

root@cluster-master: /app -- Console

root@cluster-master:~# ./index.sh
This script include commands to run mapreduce jobs using hadoop streaming to index documents
Input file is :
/data/
Deleted /tmp/index
Deleted /index
packageJobJar: [/usr/local/hadoop/share/hadoop/tools/lib/hadoop-streaming-3.3.1.jar] /tmp/streamjob3791244394678882742.jar tmpDir=null
2025-04-15 20:43:51,955 INFO client.DefaultHadoopFailoverProxyProvider: Connecting to ResourceManager at cluster-master/172.18.0.4:8032
2025-04-15 20:43:52,162 INFO client.DefaultHadoopFailoverProxyProvider: Connecting to ResourceManager at cluster-master/172.18.0.4:8032
2025-04-15 20:43:52,395 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/root/.staging/job_1744733713959_0052
2025-04-15 20:43:56,859 INFO mapred.FileInputFormat: Total input files to process : 98
2025-04-15 20:43:57,887 INFO mapreduce.JobSubmitter: number of splits:98
2025-04-15 20:43:57,176 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1744733713959_0052
2025-04-15 20:43:57,176 INFO mapreduce.JobSubmitter: Executing with tokens: []
2025-04-15 20:43:57,328 INFO conf.Configuration: resource-types.xml not found
2025-04-15 20:43:57,321 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2025-04-15 20:43:57,395 INFO impl.YarnClientImpl: Submitted application application_1744733713959_0052
2025-04-15 20:43:57,427 INFO mapreduce.Job: The url to track the job: http://cluster-master:8088/proxy/application_1744733713959_0052/
2025-04-15 20:43:57,429 INFO mapreduce.Job: Running job: job_1744733713959_0052
2025-04-15 20:44:00,908 INFO mapreduce.Job: Job job_1744733713959_0052 running in uber mode : false
2025-04-15 20:44:00,962 INFO mapreduce.Job: map 0% reduce 0%
2025-04-15 20:44:19,186 INFO mapreduce.Job: map 6% reduce 0%
2025-04-15 20:44:24,139 INFO mapreduce.Job: map 12% reduce 0%
2025-04-15 20:44:27,164 INFO mapreduce.Job: map 13% reduce 0%
2025-04-15 20:44:29,176 INFO mapreduce.Job: map 18% reduce 0%
2025-04-15 20:44:32,203 INFO mapreduce.Job: map 21% reduce 0%
2025-04-15 20:44:33,208 INFO mapreduce.Job: map 24% reduce 0%
2025-04-15 20:44:36,228 INFO mapreduce.Job: map 28% reduce 0%
2025-04-15 20:44:37,235 INFO mapreduce.Job: map 38% reduce 0%
2025-04-15 20:44:40,251 INFO mapreduce.Job: map 35% reduce 0%
2025-04-15 20:44:44,272 INFO mapreduce.Job: map 48% reduce 0%
```

Starting processing files.

```
root@cluster-master: /app -- Console

New Tab Split View
Copy Paste Find...

root@cluster-master: /app -- Console

root@cluster-master:~# ./index.sh
2025-04-15 20:45:00,383 INFO mapreduce.Job: map 58% reduce 18%
2025-04-15 20:45:01,388 INFO mapreduce.Job: map 68% reduce 18%
2025-04-15 20:45:04,483 INFO mapreduce.Job: map 63% reduce 20%
2025-04-15 20:45:05,488 INFO mapreduce.Job: map 63% reduce 20%
2025-04-15 20:45:08,421 INFO mapreduce.Job: map 68% reduce 20%
2025-04-15 20:45:09,424 INFO mapreduce.Job: map 78% reduce 20%
2025-04-15 20:45:10,429 INFO mapreduce.Job: map 78% reduce 23%
2025-04-15 20:45:12,438 INFO mapreduce.Job: map 73% reduce 23%
2025-04-15 20:45:13,442 INFO mapreduce.Job: map 74% reduce 23%
2025-04-15 20:45:14,446 INFO mapreduce.Job: map 76% reduce 23%
2025-04-15 20:45:15,451 INFO mapreduce.Job: map 77% reduce 23%
2025-04-15 20:45:16,455 INFO mapreduce.Job: map 80% reduce 26%
2025-04-15 20:45:17,459 INFO mapreduce.Job: map 81% reduce 26%
2025-04-15 20:45:18,465 INFO mapreduce.Job: map 82% reduce 26%
2025-04-15 20:45:20,475 INFO mapreduce.Job: map 85% reduce 26%
2025-04-15 20:45:21,479 INFO mapreduce.Job: map 87% reduce 26%
2025-04-15 20:45:22,484 INFO mapreduce.Job: map 87% reduce 29%
2025-04-15 20:45:24,496 INFO mapreduce.Job: map 98% reduce 29%
2025-04-15 20:45:25,499 INFO mapreduce.Job: map 92% reduce 29%
2025-04-15 20:45:28,519 INFO mapreduce.Job: map 95% reduce 31%
2025-04-15 20:45:29,522 INFO mapreduce.Job: map 97% reduce 31%
2025-04-15 20:45:32,534 INFO mapreduce.Job: map 100% reduce 31%
2025-04-15 20:45:34,541 INFO mapreduce.Job: map 100% reduce 98%
2025-04-15 20:45:35,545 INFO mapreduce.Job: map 100% reduce 100%
2025-04-15 20:45:36,556 INFO mapreduce.Job: Job job_1744733713959_0052 completed successfully
2025-04-15 20:45:36,712 INFO mapreduce.Job: Counters: 55

File System Counters
  FILE: Number of bytes read=2982934
  FILE: Number of bytes written=32518579
  FILE: Number of read operations=0
  FILE: Number of write operations=0
  HDFS: Number of bytes read=400695
  HDFS: Number of bytes written=1211744
  HDFS: Number of read operations=299
  HDFS: Number of large read operations=0
  HDFS: Number of write operations=2
  HDFS: Number of bytes read erasure-coded=0

Job Counters
  Killed map tasks=1
  Launched map tasks=98
  Launched reduce tasks=1
  Data-local map tasks=98
  Total time spent by all maps in occupied slots (ms)=331977
  Total time spent by all reduces in occupied slots (ms)=63352
  Total time spent by all map tasks (ms)=331977
  Total time spent by all reduce tasks (ms)=63352
  Total vcore-millisecods taken by all map tasks=331977
  Total vcore-millisecods taken by all reduce tasks=63352
  Total megabyte-millisecods taken by all map tasks=33984448
  Total megabyte-millisecods taken by all reduce tasks=64872448

Map-Reduce Framework
  Map input records=98
  Map output records=85423
  Map output bytes=2872802
  Map output materialized bytes=2983516
  Input split bytes=1648
  Combine input records=0
  Combine output records=0
  Reduce input groups=1823
  Reduce shuffle bytes=2983516
  Reduce input records=85423
```

Successfully completed the first step.

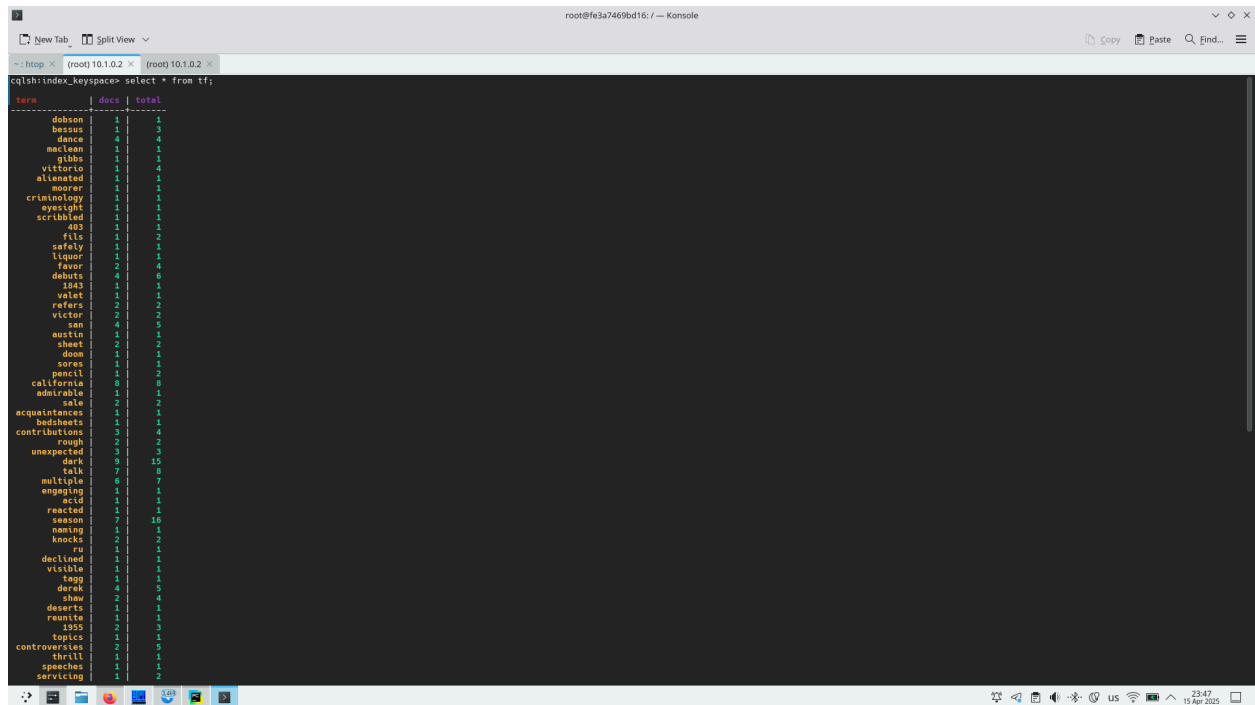
```
root@cluster-master: /app -- Console
New Tab Split View
~: http x (root) 10.1.0.2 x (root) 10.1.0.2 x
Bytes Written=1211744
2025-04-15 20:45:36,713 INFO streaming.StreamJob: Output directory: /tmp/index
packageJobJar: [ [usr/local/hadoop/share/hadoop/tools/lib/hadoop-streaming-3.3.1.jar] /tmp/streamjob6467789162088446721.jar tmpDir=null
2025-04-15 20:45:38,644 INFO Client.DefaultHadoopFailoverProxyProvider: Connecting to ResourceManager at cluster-master/172.18.0.4:8032
2025-04-15 20:45:38,663 INFO Client.DefaultHadoopFailoverProxyProvider: Connecting to ResourceManager at cluster-master/172.18.0.4:8032
2025-04-15 20:45:39,181 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/root/.staging/job_1744733713959_0053
2025-04-15 20:45:41,542 INFO mapred.FileInputFormat: Total input files to process : 1
2025-04-15 20:45:41,722 INFO mapreduce.JobSubmitter: number of splits=2
2025-04-15 20:45:41,942 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1744733713959_0053
2025-04-15 20:45:41,942 INFO mapreduce.JobSubmitter: Executing with tokens: []
2025-04-15 20:45:42,887 INFO conf.Configuration: resource-types.xml not found
2025-04-15 20:45:42,887 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2025-04-15 20:45:42,165 INFO impl.YarnClientImpl: Submitted application application_1744733713959_0053
2025-04-15 20:45:42,219 INFO mapreduce.Job: The url to track the job: http://cluster-master:8088/proxy/application_1744733713959_0053/
2025-04-15 20:45:42,221 INFO mapreduce.Job: Running job: job_1744733713959_0053
2025-04-15 20:45:53,543 INFO mapreduce.Job: Job job_1744733713959_0053 running in uber mode : false
2025-04-15 20:45:53,545 INFO mapreduce.Job: map 0% reduce 0%
2025-04-15 20:45:56,117 INFO mapreduce.Job: map 100% reduce 0%
2025-04-15 20:46:04,651 INFO mapreduce.Job: map 100% reduce 100%
2025-04-15 20:46:04,659 INFO mapreduce.Job: Job job_1744733713959_0053 completed successfully
2025-04-15 20:46:04,771 INFO mapreduce.Job: Counters: 54
File System Counters
  FILE: Number of bytes read=1537745
  FILE: Number of bytes written=3910353
  FILE: Number of read operations=0
  FILE: Number of large read operations=0
  FILE: Number of write operations=0
  HDFS: Number of bytes read=1216838
  HDFS: Number of bytes written=0
  HDFS: Number of read operations=11
  HDFS: Number of large read operations=0
  HDFS: Number of write operations=2
  HDFS: Number of bytes read erasure-coded=0
Job Counters
  Launched map tasks=2
  Launched reduce tasks=1
  Data-local map tasks=2
  Total time spent by all maps in occupied slots (ms)=4837
  Total time spent by all reduces in occupied slots (ms)=3458
  Total time spent by all map tasks (ms)=4837
  Total time spent by all reduce tasks (ms)=3458
  Total vcore-milliseconds taken by all map tasks=4837
  Total vcore-milliseconds taken by all reduce tasks=3458
  Total megabyte-milliseconds taken by all map tasks=4953888
  Total megabyte-milliseconds taken by all reduce tasks=3548992
Map-Reduce Framework
  Map input records=27892
  Map output records=54184
  Map output bytes=5429371
  Map output materialized bytes=1537751
  Input split bytes=198
  Combine input records=0
  Combine output records=0
  Reduce input groups=18321
  Reduce shuffle bytes=1537751
  Reduce input records=54184
  Reduce output records=1
  Spilled Records=188368
  Shuffled Maps=2
  Failed Shuffles=0
  Merged Map outputs=2
  GC time elapsed (ms)=287
```

Successfully completed the second step.

Cassandra

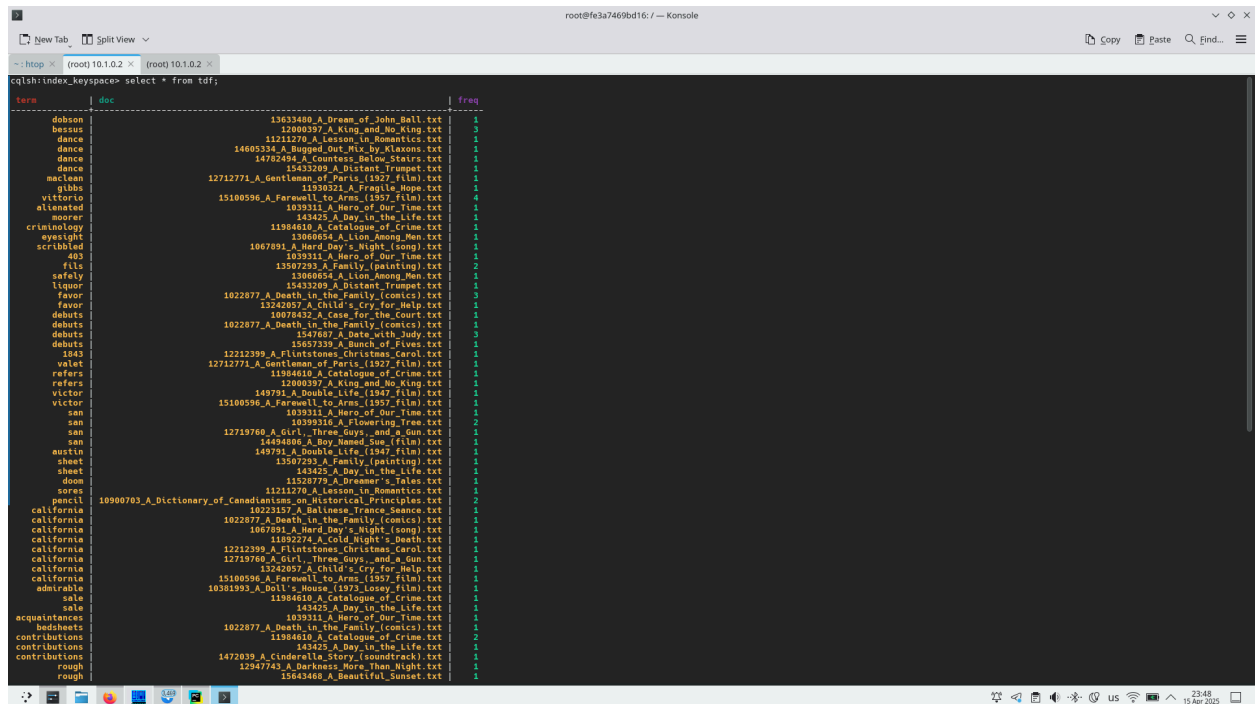
```
root@fe3a7469bd16: / -- Console
New Tab Split View
~: http x (root) 10.1.0.2 x (root) 10.1.0.2 x
cqlsh:index_keyspace> select * from dl;
+-----+-----+-----+
| doc   |      | len |
+-----+-----+-----+
| 13380293_A_Breath_of_October.txt | 28 |
| 1067891_A_Hard_Day's_Night_(song).txt | 2945 |
| 1213258_A_Crystal_Christmas.txt | 206 |
| 11520779_A_Dreamer's_Tales.txt | 924 |
| 10099975_A_Different_Light_(album).txt | 106 |
| 1211248_A_Herrie_Fly_Flores.txt | 176 |
| 13507293_A_Family_(painting).txt | 1719 |
| 12894682_A_Bolha.txt | 410 |
| 13401464_A_Gambol_with_Death.txt | 51 |
| 11017589_A_Doctor's_Report_on_Dianetics.txt | 689 |
| 15130016_A_Faithful_Of..._4_Skate.txt | 214 |
| 15433209_A_Distant_Trumpet.txt | 1376 |
| 10570204_A_Son_Called_Tension.txt | 186 |
| 12712771_A_Gentleman_of_Paris_(1927_film).txt | 288 |
| 14544487_A_Drop_of_Water.txt | 109 |
| 18395011_A_Flowering_Tree.txt | 959 |
| 149791_A_Domino_Life_(1947_film).txt | 1274 |
| 13242057_A_Child's_Cry_for_Help.txt | 579 |
| 12714432_A_Goth_House_Beneath.txt | 65 |
| 10900703_A_Dictionary_of_Canadianisms_on_Historical_Principles.txt | 1065 |
| 10056097_A_Dangerous_Path.txt | 1124 |
| 15105252_A_Fire_Inside_IP.txt | 197 |
| 11984810_A_Catalogue_of_Crime.txt | 1120 |
| 15643684_A_Beautiful_Sunset.txt | 1246 |
| 11017293_A_Bad_Spell_in_Yurt.txt | 469 |
| 12459639_A_Day_at_the_Races_(video).txt | 60 |
| 10849608_A_Day_in_the_Death_of_Danny_B.txt | 285 |
| 13478522_A_Knight_of_the_Range.txt | 114 |
| 1412472_A_Laboratory_Manual_for_Comparative_Herpetbrate_Anatomy.txt | 143 |
| 11671420_A_Lifetime_of_More.txt | 108 |
| 15057339_A_Bunch_of_Fives.txt | 169 |
| 13402887_A_Bandful_of_Time.txt | 239 |
| 10230685_A_Dead_Sinking_Story.txt | 115 |
| 1547663_A_Date_with_Judy.txt | 524 |
| 13307430_A_Cry_for_Help_(1912_film).txt | 139 |
| 11892274_A_Cold_Night's_Death.txt | 332 |
| 1133231_A_Fragile_Hope.txt | 126 |
| 15081420_A_Gamblin'_Fool.txt | 98 |
| 10453695_A_Devilish_Homicide.txt | 1055 |
| 14809790_A_Gray_Sigb_in_a_Flower_Husk.txt | 145 |
| 1022877_A_Death_in_the_Family_(comics).txt | 5429 |
| 13375345_A_Few_in_the_Sentucky_Hills.txt | 228 |
| 13605428_A_Little_Journey.txt | 169 |
| 1115810_A_Hanging.txt | 436 |
| 15523771_A_Day_in_the_Death.txt | 718 |
| 10078432_A_Case_for_the_Court.txt | 78 |
| 15100596_A_Forewell_to_Arms_(1957_film).txt | 1731 |
| 12955622_A_Day_at_School.txt | 92 |
| 10137548_A_Good_Thief_Tips_His_Hat.txt | 180 |
| 1356264_A_Child's_Garden_of_Verues.txt | 518 |
| 13434748_A_Day's_Adventure.txt | 74 |
| 11315057_A_Go_Go_(Portrait_Album).txt | 119 |
| 1466186_A_Day_at_the_Beach.txt | 383 |
| 1571146_A_Is_for_Accident.txt | 278 |
| 13716170_A_Sift_of_Dragons.txt | 1170 |
| 1003442_A_Hillbilly_Tribute_to_ACDU.txt | 191 |
| 13633468_A_Dream_of_John_Bell.txt | 882 |
| 12212399_A_Filutstone_Christmas_Carol.txt | 1258
```

Document lengths.

A terminal window titled 'root@fe3a7469bd16: / - Konsole' with a split view. The left pane shows a command prompt 'colsh:index_keyspace> select * from tf;' and a table of document lengths. The right pane is empty. The table has four columns: 'term', 'doc', 'count', and 'total'. The data is sorted by 'total' in descending order.

term	doc	count	total
dobson	1	1	1
bessus	1	1	1
dance	4	4	4
maclean	1	1	1
gibbs	1	1	1
vittorio	1	1	1
alienated	1	1	1
moorer	1	1	1
criminology	1	1	1
eyesight	1	1	1
scribbled	1	1	1
403	1	1	1
files	1	1	1
safely	1	1	1
liquor	1	1	1
favor	2	2	4
debuts	4	4	4
1843	1	1	1
valet	1	1	1
refers	2	2	2
victor	2	2	2
san	4	4	4
austin	1	1	1
sheet	2	2	2
doom	1	1	1
sorens	1	1	1
pencil	1	1	1
california	8	8	8
admirable	1	1	1
sale	2	2	2
acquaintances	1	1	1
bedsheets	1	1	1
contributions	3	3	3
rough	2	2	2
unexpected	1	1	1
dark	9	9	15
talk	6	6	7
multiple	1	1	1
engaging	1	1	1
acid	1	1	1
reacted	1	1	1
season	7	7	16
naming	1	1	1
knocks	2	2	2
ru	1	1	1
declined	1	1	1
visible	1	1	1
tagg	1	1	1
derek	4	4	5
show	2	2	2
deserts	1	1	1
reunite	1	1	1
1855	1	1	1
topics	1	1	1
controversies	2	2	1
thrill	1	1	1
speeches	1	1	1
serving	1	1	2

Term frequencies.

A terminal window titled 'root@fe3a7469bd16: / - Konsole' with a split view. The left pane shows a command prompt 'colsh:index_keyspace> select * from tdf;' and a table of term frequencies. The right pane is empty. The table has three columns: 'term', 'doc', and 'freq'. The data is sorted by 'freq' in descending order.

term	doc	freq
dobson	13633488 A Dream of John Bull.txt	1
bessus	12000397 A King and No King.txt	1
dance	11212779 A Lesson in Romantics.txt	1
dance	14605334 A Ragged Out Mix by Klaxons.txt	1
dance	14782494 A Countess Below Stairs.txt	1
dance	15433289 A Distant Trumpet.txt	1
maclean	12712771 A Gentleman of Paris (1927 film).txt	1
gibbs	18398321 A Fragile Hope.txt	1
vittorio	15100596 A Farewell to Arms (1957 film).txt	4
alienated	1039311 A Hero of Our Time.txt	1
moorer	143425 A Day in the Life.txt	1
criminology	11984610 A Catalogue of Crime.txt	1
eyesight	13066654 A Lion Among Men.txt	1
scribbled	1067891 A Hard Day's Night (song).txt	1
403	1039311 A Hero of Our Time.txt	1
files	13972293 A Family (painting).txt	2
safely	13066654 A Lion Among Men.txt	1
liquor	15433289 A Distant Trumpet.txt	1
favor	1022877 A Death in the Family (comics).txt	3
favor	13242057 A Child's Cry for Help.txt	1
debuts	18078432 A Case for the Court.txt	1
debuts	1022877 A Death in the Family (comics).txt	1
debuts	1547857 A Date with Judy.txt	3
debuts	15627330 A Bunch of Flies.txt	1
1843	12212399 A Flintstones Christmas Carol.txt	1
valet	12712771 A Gentleman of Paris (1927 film).txt	1
refers	11984610 A Catalogue of Crime.txt	1
refers	12000397 A King and No King.txt	1
victor	149791 A Double Life (1947 film).txt	1
victor	15100596 A Farewell to Arms (1957 film).txt	1
san	1039311 A Hero of Our Time.txt	1
san	1039311 A Hero of Our Time.txt	2
austin	12719760 A Girl, Three Guys, and a Gun.txt	1
sheet	1448488 A Day Ahead, Inc.(film).txt	1
sheet	149791 A Double Life (1947 film).txt	1
sheet	13507253 A Family (painting).txt	1
doom	143425 A Day in the Life.txt	1
sorens	11528779 A Dreamer's Tales.txt	1
pencil	10980703 A Dictionary of Canadianisms on Historical Principles.txt	2
california	10223157 A Saline Seance.txt	1
california	1022877 A Death in the Family (comics).txt	1
california	1067891 A Hard Day's Night (song).txt	1
california	11022779 A Cold Right's Death.txt	1
california	12212399 A Flintstones Christmas Carol.txt	1
california	12719760 A Girl, Three Guys, and a Gun.txt	1
california	13242057 A Child's Cry for Help.txt	1
california	15100596 A Farewell to Arms (1957 film).txt	1
admirable	10301999 A Doll House (1975 play).txt	1
sale	11984610 A Catalogue of Crime.txt	1
sale	143425 A Day in the Life.txt	1
acquaintances	1039311 A Hero of Our Time.txt	1
bedsheets	1022877 A Death in the Family (comics).txt	1
contributions	11984610 A Catalogue of Crime.txt	2
contributions	143425 A Day in the Life.txt	1
contributions	1472039 A Underella Story (soundtrack).txt	1
rough	12947742 A Dreamer's Tale, Night.txt	1
rough	15643668 A Beautiful Sunset.txt	1

Term frequencies per each document.

Queries

```
root@cluster-master: /app — Console

[venv] root@cluster-master:/app# python query.py 'Crunchyroll anime series'
:: loading settings :: url = jar:file:/usr/local/spark/jars/ivy-2.5.1-jar!org/apache/ivy/core/settings/ivysettings.xml
ivy Default Cache set to: /root/.ivy2/cache
The jars for the packages stored in: /root/.ivy2/jars
com.datastax.spark#spark-cassandra-connector_2.12 added as a dependency
:: resolving dependencies :: org.apache.spark#spark-submit-parent-7b0fcb0-b237-49e4-a02c-db971a0dc243;1.0
  confs: [default]
  found com.datastax.spark#spark-cassandra-connector_2.12;3.5.1 in central
  found com.datastax.spark#spark-cassandra-connector-driver_2.12;3.5.1 in central
  found org.scala-lang.modules#scala-collection-compat_2.12;2.11.0 in central
  found org.apache.cassandra#java-driver-core-shaded;4.18.1 in central
  found com.datastax.oss#native-protocol;1.5.1 in central
  found com.datastax.oss#java-driver-shaded-guava;25.1-jre-graal-sub-1 in central
  found com.typesafe#config;1.4.1 in central
  found org.slf4j#slf4j-api;1.7.26 in central
  found io.dropwizard.metrics#metrics-core;4.1.18 in central
  found org.hdrhistogram#hdrhistogram;2.1.12 in central
  found org.reactivestreams#reactive-streams;1.0.3 in central
  found org.apache.cassandra#java-driver-mapper-runtime;4.18.1 in central
  found org.apache.cassandra#java-driver-query-builder;4.18.1 in central
  found org.apache.commons#commons-lang3;3.10 in central
  found com.thoughtworks.paranamer#paranamer;2.8 in central
  found org.scala-lang#scala-reflect;2.12.19 in central
:: resolution report :: resolve 381ms :: artifacts dl 16ms
  :: modules in use:
    com.datastax.oss#java-driver-shaded-guava;25.1-jre-graal-sub-1 from central in [default]
    com.datastax.oss#native-protocol;1.5.1 from central in [default]
    com.datastax.spark#spark-cassandra-connector-driver_2.12;3.5.1 from central in [default]
    com.datastax.spark#spark-cassandra-connector_2.12;3.5.1 from central in [default]
    com.thoughtworks.paranamer#paranamer;2.8 from central in [default]
    com.typesafe#config;1.4.1 from central in [default]
    io.dropwizard.metrics#metrics-core;4.1.18 from central in [default]
    org.apache.cassandra#java-driver-core-shaded;4.18.1 from central in [default]
    org.apache.cassandra#java-driver-mapper-runtime;4.18.1 from central in [default]
    org.apache.cassandra#java-driver-query-builder;4.18.1 from central in [default]
    org.apache.commons#commons-lang3;3.10 from central in [default]
    org.hdrhistogram#hdrhistogram;2.1.12 from central in [default]
    org.reactivestreams#reactive-streams;1.0.3 from central in [default]
    org.scala-lang#scala-reflect;2.12.19 from central in [default]
    org.scala-lang.modules#scala-collection-compat_2.12;2.11.0 from central in [default]
    org.slf4j#slf4j-api;1.7.26 from central in [default]
  -----
  | conf |      | modules |      | artifacts | | | |
|---|---|---|---|---|---|---|---|
  | default | 16 | 0 | 0 | 0 | 0 | 16 | 0 |
  -----
:: retrieving :: org.apache.spark#spark-submit-parent-7b0fcb0-b237-49e4-a02c-db971a0dc243
  confs: [default]
  0 artifacts copied, 16 already retrieved (0kB/12ms)
25/04/15 20:50:51 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
-----
| doc |      | score |
|-----|-----|-----|
|15170076_A_Gentle...|10.156317063274294|
|1547687_A_Date_w...|3.586387207794337|
|1822877_A_Death...|3.537314114248527|
|15657339_A_Bunch...|3.5818589413548897|
|15523771_A_Day_in...|3.338387411577756|
|15643468_A_Beaut...|3.338387411577756|
```

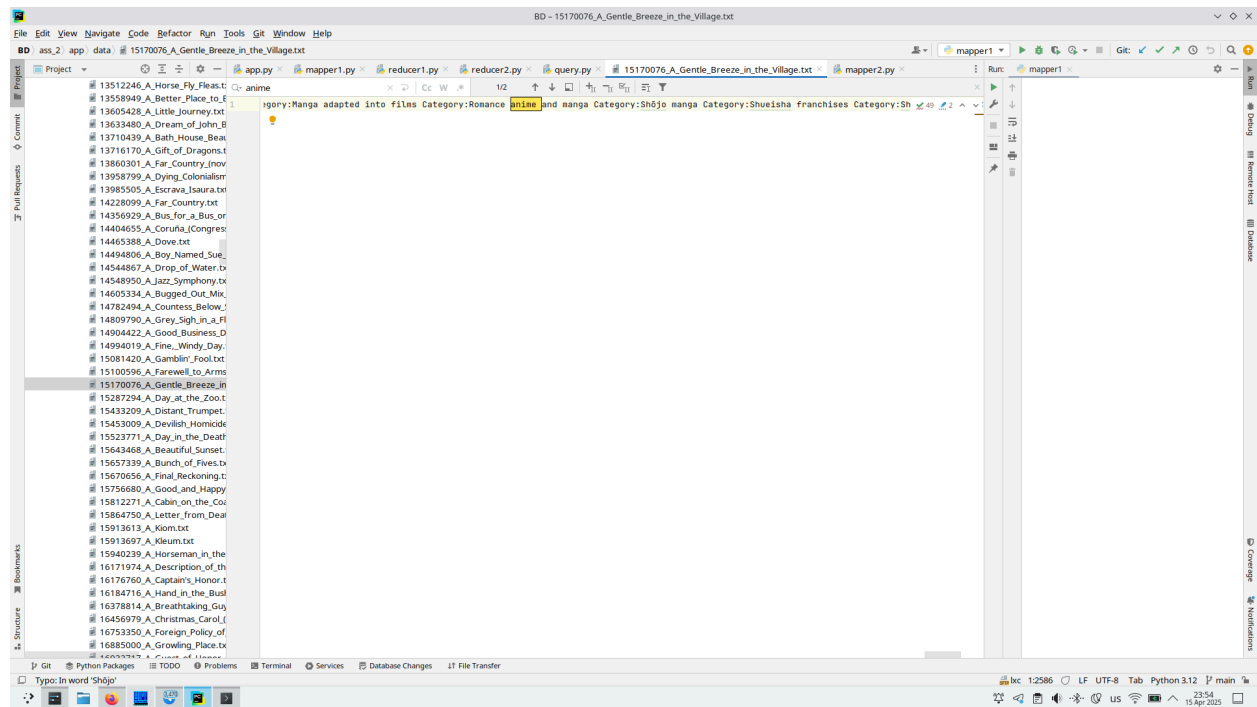
Query: “Crunchyroll anime series”

```
root@cluster-master: /app — Console

[venv] root@cluster-master:/app# python query.py 'anime series'
:: loading settings :: url = jar:file:/usr/local/spark/jars/ivy-2.5.1-jar!org/apache/ivy/core/settings/ivysettings.xml
ivy Default Cache set to: /root/.ivy2/cache
The jars for the packages stored in: /root/.ivy2/jars
com.datastax.spark#spark-cassandra-connector_2.12 added as a dependency
:: resolving dependencies :: org.apache.spark#spark-submit-parent-58dfc877-2763-436d-8132-ca1bc31b8362;1.0
  confs: [default]
  found com.datastax.spark#spark-cassandra-connector_2.12;3.5.1 in central
  found com.datastax.spark#spark-cassandra-connector-driver_2.12;3.5.1 in central
  found org.scala-lang.modules#scala-collection-compat_2.12;2.11.0 in central
  found org.apache.cassandra#java-driver-core-shaded;4.18.1 in central
  found com.datastax.oss#native-protocol;1.5.1 in central
  found com.datastax.oss#java-driver-shaded-guava;25.1-jre-graal-sub-1 in central
  found com.typesafe#config;1.4.1 in central
  found org.slf4j#slf4j-api;1.7.26 in central
  found io.dropwizard.metrics#metrics-core;4.1.18 in central
  found org.hdrhistogram#hdrhistogram;2.1.12 in central
  found org.reactivestreams#reactive-streams;1.0.3 in central
  found org.apache.cassandra#java-driver-mapper-runtime;4.18.1 in central
  found org.apache.cassandra#java-driver-query-builder;4.18.1 in central
  found org.apache.commons#commons-lang3;3.10 in central
  found com.thoughtworks.paranamer#paranamer;2.8 in central
  found org.scala-lang#scala-reflect;2.12.19 in central
:: resolution report :: resolve 374ms :: artifacts dl 18ms
  :: modules in use:
    com.datastax.oss#java-driver-shaded-guava;25.1-jre-graal-sub-1 from central in [default]
    com.datastax.oss#native-protocol;1.5.1 from central in [default]
    com.datastax.spark#spark-cassandra-connector-driver_2.12;3.5.1 from central in [default]
    com.datastax.spark#spark-cassandra-connector_2.12;3.5.1 from central in [default]
    com.thoughtworks.paranamer#paranamer;2.8 from central in [default]
    com.typesafe#config;1.4.1 from central in [default]
    io.dropwizard.metrics#metrics-core;4.1.18 from central in [default]
    org.apache.cassandra#java-driver-core-shaded;4.18.1 from central in [default]
    org.apache.cassandra#java-driver-mapper-runtime;4.18.1 from central in [default]
    org.apache.cassandra#java-driver-query-builder;4.18.1 from central in [default]
    org.apache.commons#commons-lang3;3.10 from central in [default]
    org.hdrhistogram#hdrhistogram;2.1.12 from central in [default]
    org.reactivestreams#reactive-streams;1.0.3 from central in [default]
    org.scala-lang#scala-reflect;2.12.19 from central in [default]
    org.scala-lang.modules#scala-collection-compat_2.12;2.11.0 from central in [default]
    org.slf4j#slf4j-api;1.7.26 from central in [default]
  -----
  | conf |      | modules |      | artifacts | | | |
|---|---|---|---|---|---|---|---|
  | default | 16 | 0 | 0 | 0 | 0 | 16 | 0 |
  -----
:: retrieving :: org.apache.spark#spark-submit-parent-58dfc877-2763-436d-8132-ca1bc31b8362
  confs: [default]
  0 artifacts copied, 16 already retrieved (0kB/9ms)
25/04/15 20:53:16 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
-----
| doc |      | score |
|-----|-----|-----|
|15170076_A_Gentle...|10.156317063274294|
|1547687_A_Date_w...|3.586387207794337|
|1822877_A_Death...|3.537314114248527|
|15657339_A_Bunch...|3.5818589413548897|
|15523771_A_Day_in...|3.338387411577756|
|15643468_A_Beaut...|3.338387411577756|
```

Query: “anime series”

The first document “15170076_A_Gentle...” has anime:



As it can be seen from the screenshots, similar queries lead to similar results and document really contains keywords.