# 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:dl 2
doc1:dl 1
doc2:dl 1
doc2:dl 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 dl (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.

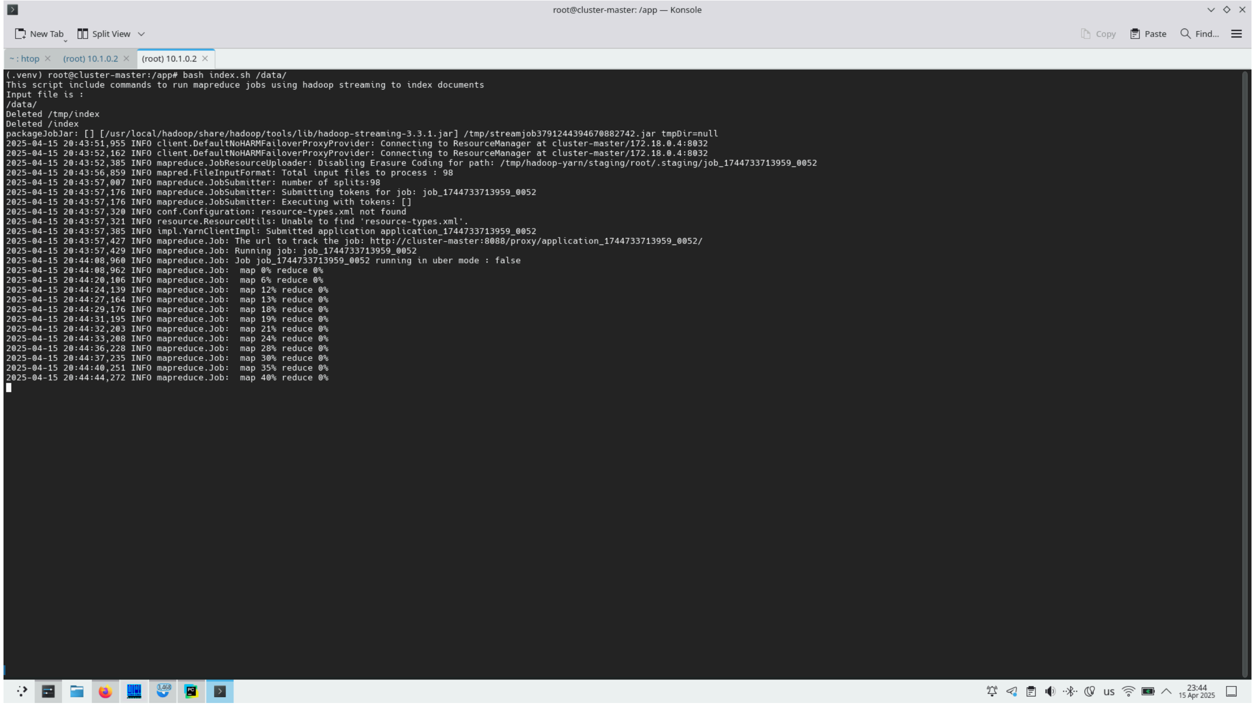
```
agg = dl.agg(avg("len"), count("*")).first()
avgdl, n = agg
```

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

```
tf = (col("freq") * (k+1)) / (col("freq") + k * (1-b + b*n/avgdl))
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"))
```

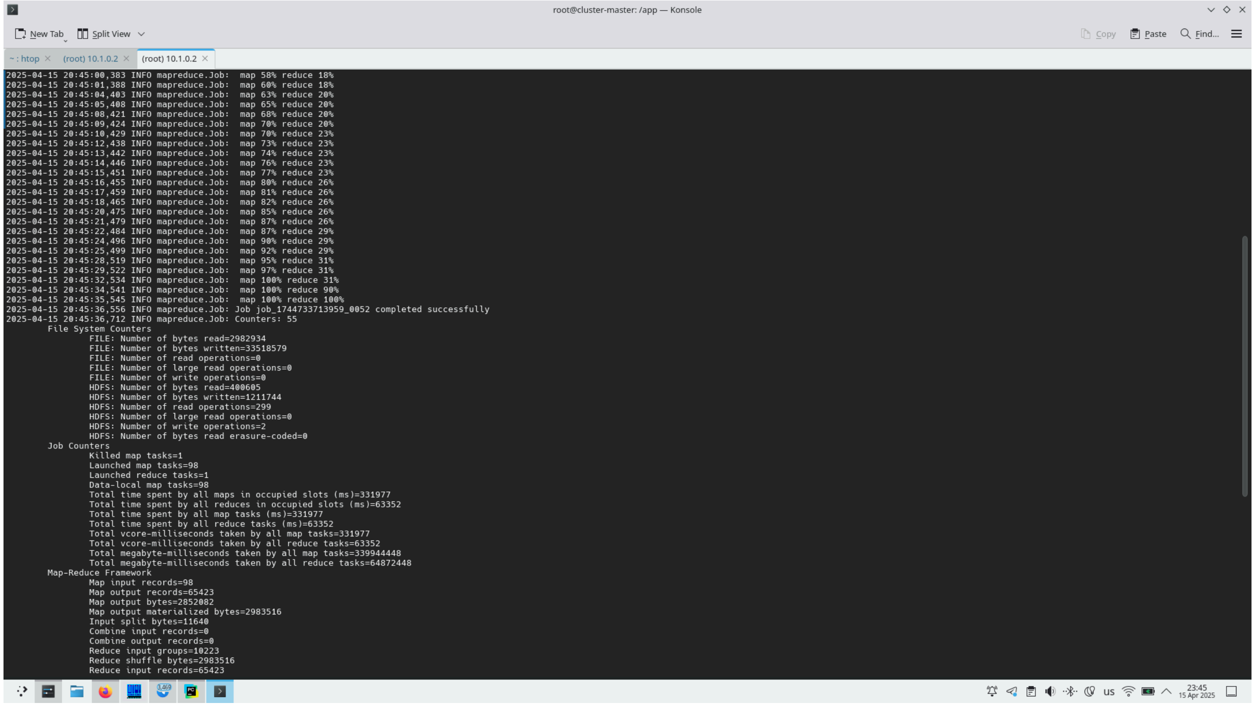
# Demonstration

## Indexing



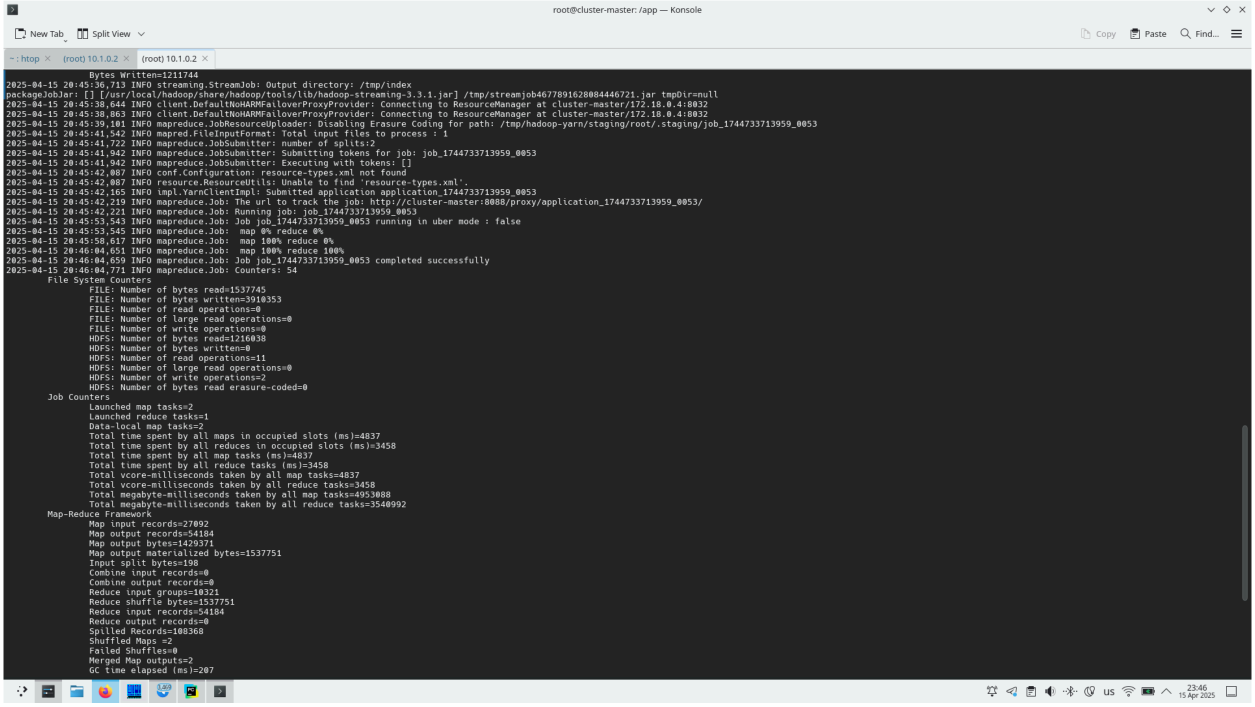
```
(.venv) root@cluster-master:/app# bash index.sh /data/
This script include commands to run mapreduce jobs using hadoop streaming to index documents
Input file is :
/data/
Deleted /tmp/index
Deleted /index
packageJobJar: [] [/usr/local/hadoop/share/hadoop/tools/lib/hadoop-streaming-3.3.1.jar] /tmp/streamjob3791244394670882742.jar tmpDir=null
2025-04-15 20:43:51,955 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at cluster-master/172.18.0.4:8032
2025-04-15 20:43:52,162 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at cluster-master/172.18.0.4:8032
2025-04-15 20:43:52,385 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 20:43:57,007 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 20:43:57,320 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 20:43:57,385 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 20:44:08,960 INFO mapreduce.Job: Job job_1744733713959_0052 running in uber mode : false
2025-04-15 20:44:08,962 INFO mapreduce.Job:  map 0% reduce 0%
2025-04-15 20:44:20,106 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 20:44:31,195 INFO mapreduce.Job:  map 19% 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 30% 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 40% reduce 0%
```

Starting processing files.



```
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 60% reduce 18%
2025-04-15 20:45:04,403 INFO mapreduce.Job:  map 63% reduce 20%
2025-04-15 20:45:05,408 INFO mapreduce.Job:  map 65% 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 70% reduce 20%
2025-04-15 20:45:10,429 INFO mapreduce.Job:  map 70% 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 90% 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 90%
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=33518579
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=400605
                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-milliseconds taken by all map tasks=331977
                Total vcore-milliseconds taken by all reduce tasks=63352
                Total megabyte-milliseconds taken by all map tasks=339944448
                Total megabyte-milliseconds taken by all reduce tasks=64872448
        Map-Reduce Framework
                Map input records=98
                Map output records=65423
                Map output bytes=2852082
                Map output materialized bytes=2983516
                Input split bytes=11640
                Combine input records=0
                Combine output records=0
                Reduce input groups=10223
                Reduce shuffle bytes=2983516
                Reduce input records=65423
```

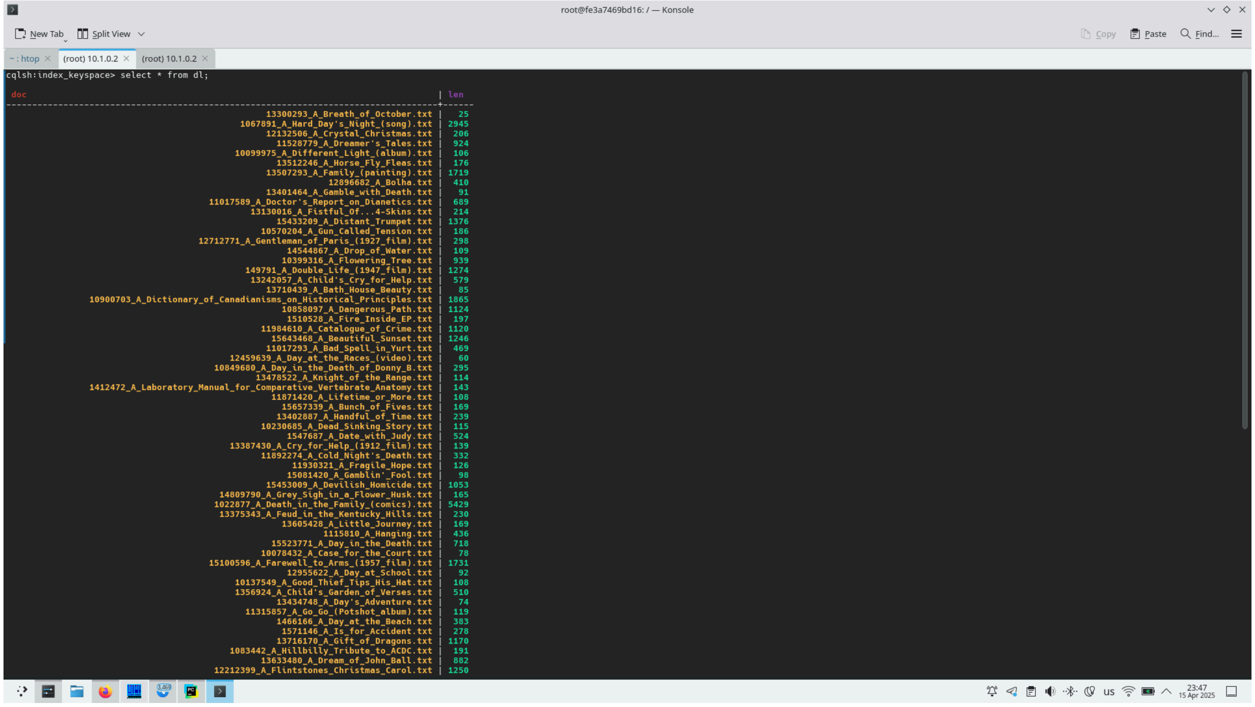
Successfully completed the first step.



```
                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/streamjob4677891628084446721.jar tmpDir=null
2025-04-15 20:45:38,644 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at cluster-master/172.18.0.4:8032
2025-04-15 20:45:38,863 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at cluster-master/172.18.0.4:8032
2025-04-15 20:45:39,101 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 20:45:42,087 INFO conf.Configuration: resource-types.xml not found
2025-04-15 20:45:42,087 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 20:45:58,617 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=1216038
                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=4953088
                Total megabyte-milliseconds taken by all reduce tasks=3540992
        Map-Reduce Framework
                Map input records=27092
                Map output records=54184
                Map output bytes=1429371
                Map output materialized bytes=1537751
                Input split bytes=198
                Combine input records=0
                Combine output records=0
                Reduce input groups=10321
                Reduce shuffle bytes=1537751
                Reduce input records=54184
                Reduce output records=0
                Spilled Records=108368
                Shuffled Maps =2
                Failed Shuffles=0
                Merged Map outputs=2
                GC time elapsed (ms)=207
```

Successfully completed the second step.

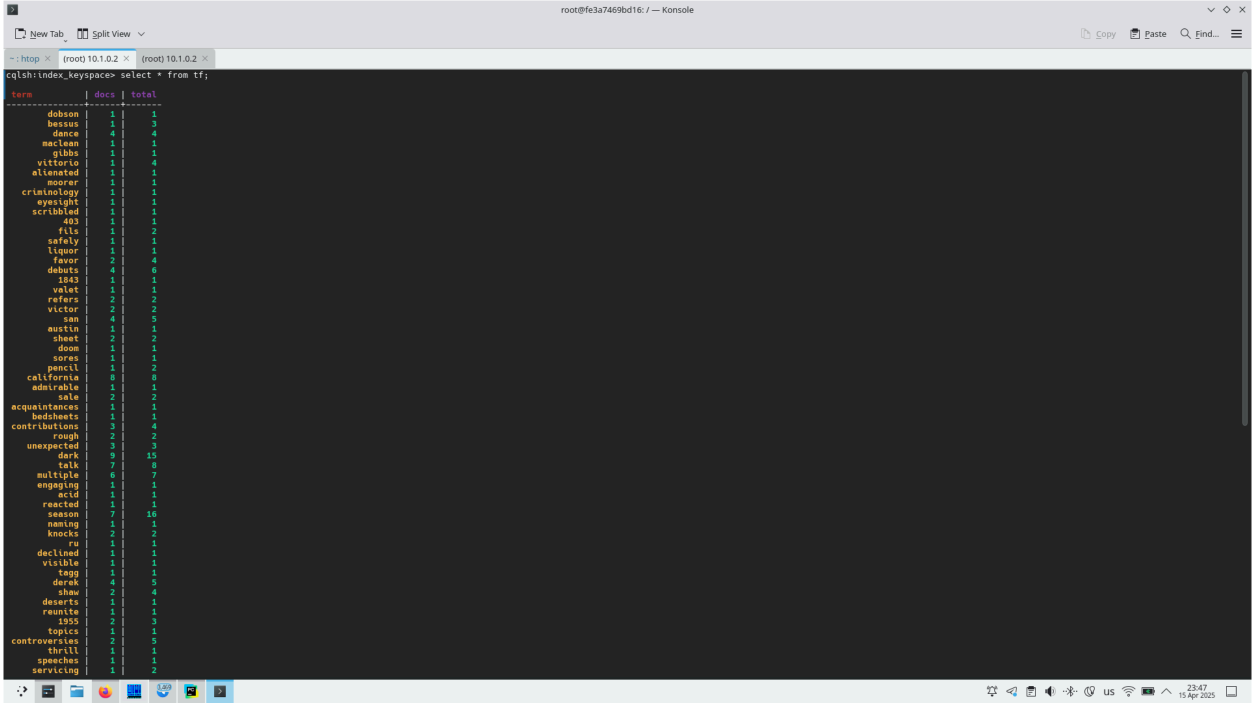
# Cassandra



```
cqlsh:index_keyspace> select * from dl;

 doc                                                                              | len
-----------------------------------------------------------------------------------+------
                                                  13300293_A_Breath_of_October.txt |   25
                                             1067891_A_Hard_Day's_Night_(song).txt | 2945
                                                  12132506_A_Crystal_Christmas.txt |  206
                                                    11528779_A_Dreamer's_Tales.txt |  924
                                            10099975_A_Different_Light_(album).txt |  106
                                                    13512246_A_Horse_Fly_Fleas.txt |  176
                                                  13507293_A_Family_(painting).txt | 1719
                                                              12896682_A_Bolha.txt |  410
                                                  13401464_A_Gamble_with_Death.txt |   91
                                       11017589_A_Doctor's_Report_on_Dianetics.txt |  689
                                               13130016_A_Fistful_Of...4-Skins.txt |  214
                                                    15433209_A_Distant_Trumpet.txt | 1376
                                                 10570204_A_Gun_Called_Tension.txt |  186
                                     12712771_A_Gentleman_of_Paris_(1927_film).txt |  298
                                                      14544867_A_Drop_of_Water.txt |  109
                                                     10399316_A_Flowering_Tree.txt |  939
                                              149791_A_Double_Life_(1947_film).txt | 1274
                                               13242057_A_Child's_Cry_for_Help.txt |  579
                                                  13710439_A_Bath_House_Beauty.txt |   85
                10900703_A_Dictionary_of_Canadianisms_on_Historical_Principles.txt | 1865
                                                     10858097_A_Dangerous_Path.txt | 1124
                                                      1510528_A_Fire_Inside_EP.txt |  197
                                                 11984610_A_Catalogue_of_Crime.txt | 1120
                                                   15643468_A_Beautiful_Sunset.txt | 1246
                                                  11017293_A_Bad_Spell_in_Yurt.txt |  469
                                           12459639_A_Day_at_the_Races_(video).txt |   60
                                        10849680_A_Day_in_the_Death_of_Donny_B.txt |  295
                                                13478522_A_Knight_of_the_Range.txt |  114
                1412472_A_Laboratory_Manual_for_Comparative_Vertebrate_Anatomy.txt |  143
                                                   11871420_A_Lifetime_or_More.txt |  108
                                                     15657339_A_Bunch_of_Fives.txt |  169
                                                    13402887_A_Handful_of_Time.txt |  239
                                                 10230685_A_Dead_Sinking_Story.txt |  115
                                                      1547687_A_Date_with_Judy.txt |  524
                                           13387430_A_Cry_for_Help_(1912_film).txt |  139
                                                 11892274_A_Cold_Night's_Death.txt |  332
                                                       11930321_A_Fragile_Hope.txt |  126
                                                      15081420_A_Gamblin'_Fool.txt |   98
                                                  15453009_A_Devilish_Homicide.txt | 1053
                                         14809790_A_Grey_Sigh_in_a_Flower_Husk.txt |  165
                                        1022877_A_Death_in_the_Family_(comics).txt | 5429
                                         13375343_A_Feud_in_the_Kentucky_Hills.txt |  230
                                                     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_Farewell_to_Arms_(1957_film).txt | 1731
                                                      12955622_A_Day_at_School.txt |   92
                                            10137549_A_Good_Thief_Tips_His_Hat.txt |  108
                                            1356924_A_Child's_Garden_of_Verses.txt |  510
                                                    13434748_A_Day's_Adventure.txt |   74
                                              11315857_A_Go_Go_(Potshot_album).txt |  119
                                                    1466166_A_Day_at_the_Beach.txt |  383
                                                     1571146_A_Is_for_Accident.txt |  278
                                                    13716170_A_Gift_of_Dragons.txt | 1170
                                           1083442_A_Hillbilly_Tribute_to_ACDC.txt |  191
                                                 13633480_A_Dream_of_John_Ball.txt |  882
                                        12212399_A_Flintstones_Christmas_Carol.txt | 1250
```

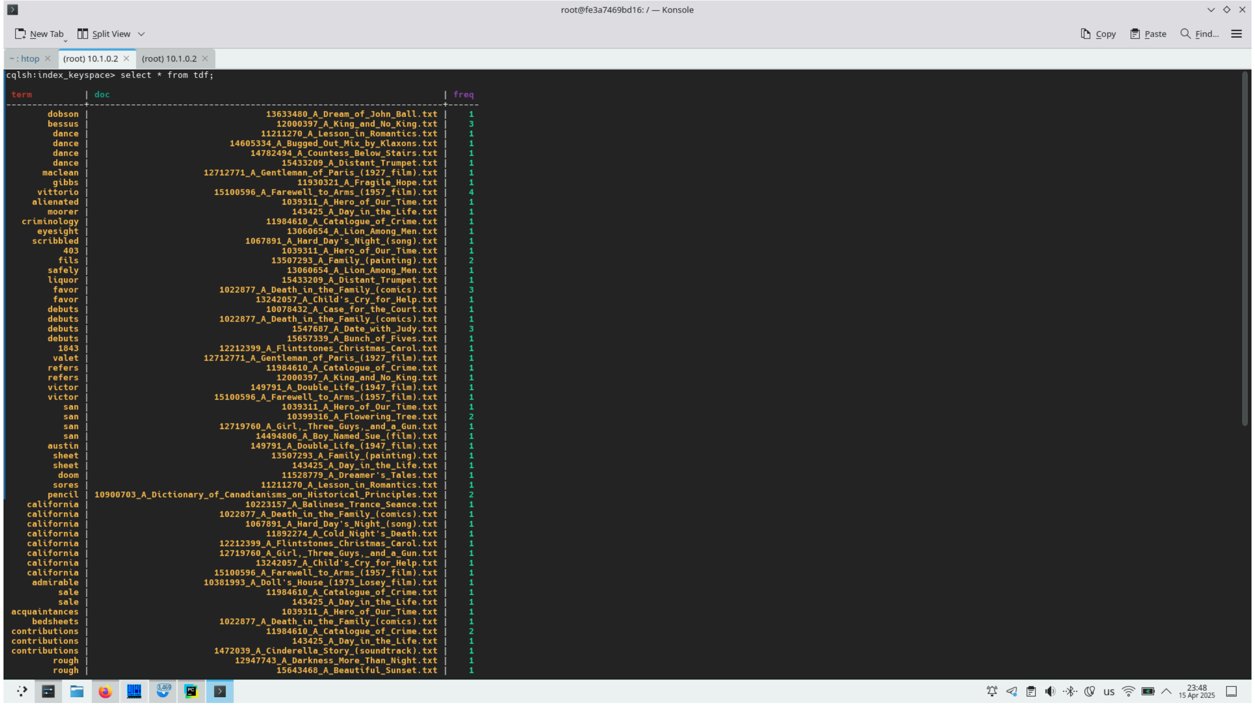
Document lengths.



```
cqlsh:index_keyspace> select * from tf;

 term          | docs | total
---------------+------+-------
         dobson |    1 |     1
         bessus |    1 |     3
          dance |    4 |     4
        maclean |    1 |     1
          gibbs |    1 |     1
       vittorio |    1 |     4
      alienated |    1 |     1
         moorer |    1 |     1
    criminology |    1 |     1
       eyesight |    1 |     1
      scribbled |    1 |     1
            403 |    1 |     1
           fils |    1 |     2
         safely |    1 |     1
         liquor |    1 |     1
          favor |    2 |     4
         debuts |    4 |     6
           1843 |    1 |     1
          valet |    1 |     1
         refers |    2 |     2
         victor |    2 |     2
            san |    4 |     5
         austin |    1 |     1
          sheet |    2 |     2
           doom |    1 |     1
          sores |    1 |     1
         pencil |    1 |     2
     california |    8 |     8
      admirable |    1 |     1
           sale |    2 |     2
  acquaintances |    1 |     1
      bedsheets |    1 |     1
  contributions |    3 |     4
          rough |    2 |     2
     unexpected |    3 |     3
           dark |    9 |    15
           talk |    7 |     8
       multiple |    6 |     7
       engaging |    1 |     1
           acid |    1 |     1
        reacted |    1 |     1
         season |    7 |    16
         naming |    1 |     1
         knocks |    2 |     2
             ru |    1 |     1
       declined |    1 |     1
        visible |    1 |     1
           tagg |    1 |     1
          derek |    4 |     5
           shaw |    2 |     4
        deserts |    1 |     1
        reunite |    1 |     1
           1955 |    2 |     3
         topics |    1 |     1
  controversies |    2 |     5
         thrill |    1 |     1
       speeches |    1 |     1
      servicing |    1 |     2
```

Term frequencies.

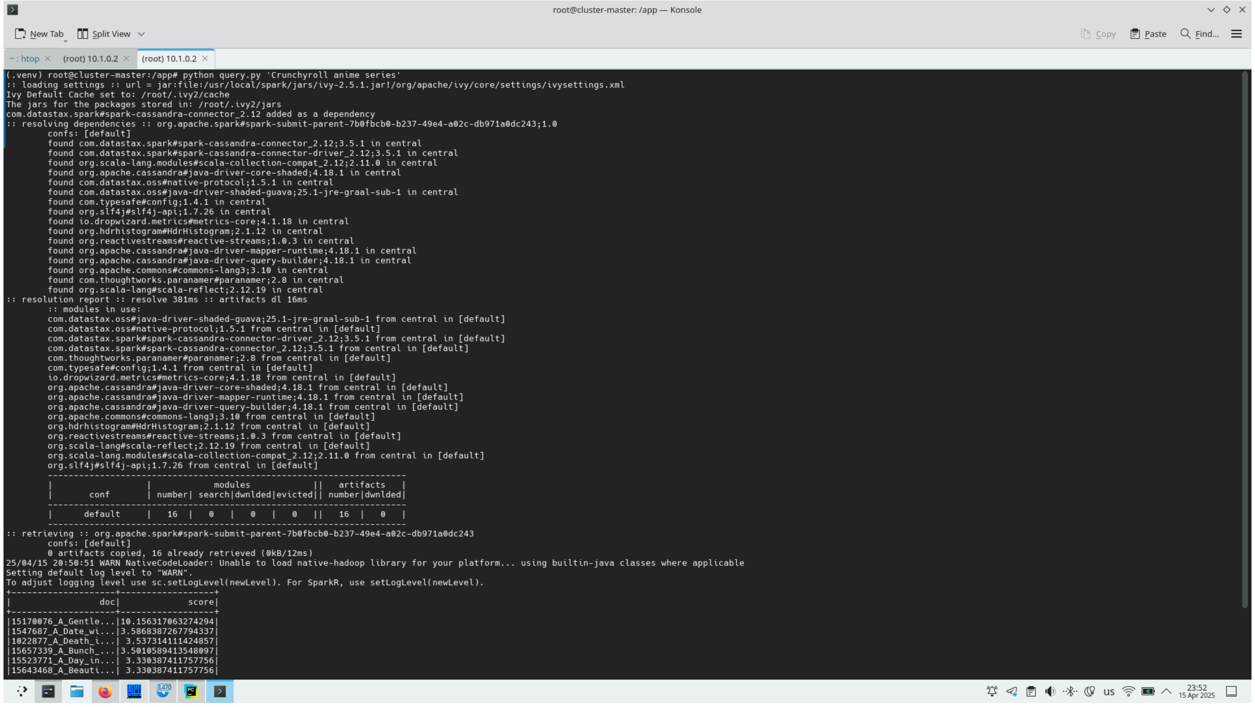


```
cqlsh:index_keyspace> select * from tdf;

 term          | doc                                                                | freq
---------------+--------------------------------------------------------------------+------
         dobson |                                  13633480_A_Dream_of_John_Ball.txt |    1
         bessus |                                    12000397_A_King_and_No_King.txt |    3
          dance |                                 11211270_A_Lesson_in_Romantics.txt |    1
          dance |                           14605334_A_Bugged_Out_Mix_by_Klaxons.txt |    1
          dance |                               14782494_A_Countess_Below_Stairs.txt |    1
          dance |                                     15433209_A_Distant_Trumpet.txt |    1
        maclean |                      12712771_A_Gentleman_of_Paris_(1927_film).txt |    1
          gibbs |                                        11930321_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 |                                      13060654_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
           fils |                                   13507293_A_Family_(painting).txt |    2
         safely |                                      13060654_A_Lion_Among_Men.txt |    1
         liquor |                                     15433209_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 |                                  10078432_A_Case_for_the_Court.txt |    1
         debuts |                         1022877_A_Death_in_the_Family_(comics).txt |    1
         debuts |                                       1547687_A_Date_with_Judy.txt |    3
         debuts |                                      15657339_A_Bunch_of_Fives.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 |                                      10399316_A_Flowering_Tree.txt |    2
            san |                         12719760_A_Girl,_Three_Guys,_and_a_Gun.txt |    1
            san |                                14494806_A_Boy_Named_Sue_(film).txt |    1
         austin |                               149791_A_Double_Life_(1947_film).txt |    1
          sheet |                                   13507293_A_Family_(painting).txt |    1
          sheet |                                       143425_A_Day_in_the_Life.txt |    1
           doom |                                     11528779_A_Dreamer's_Tales.txt |    1
          sores |                                 11211270_A_Lesson_in_Romantics.txt |    1
         pencil | 10900703_A_Dictionary_of_Canadianisms_on_Historical_Principles.txt |    2
     california |                              10223157_A_Balinese_Trance_Seance.txt |    1
     california |                         1022877_A_Death_in_the_Family_(comics).txt |    1
     california |                              1067891_A_Hard_Day's_Night_(song).txt |    1
     california |                                  11892274_A_Cold_Night'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 |                      10381993_A_Doll's_House_(1973_Losey_film).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_Cinderella_Story_(soundtrack).txt |    1
          rough |                            12947743_A_Darkness_More_Than_Night.txt |    1
          rough |                                    15643468_A_Beautiful_Sunset.txt |    1
```

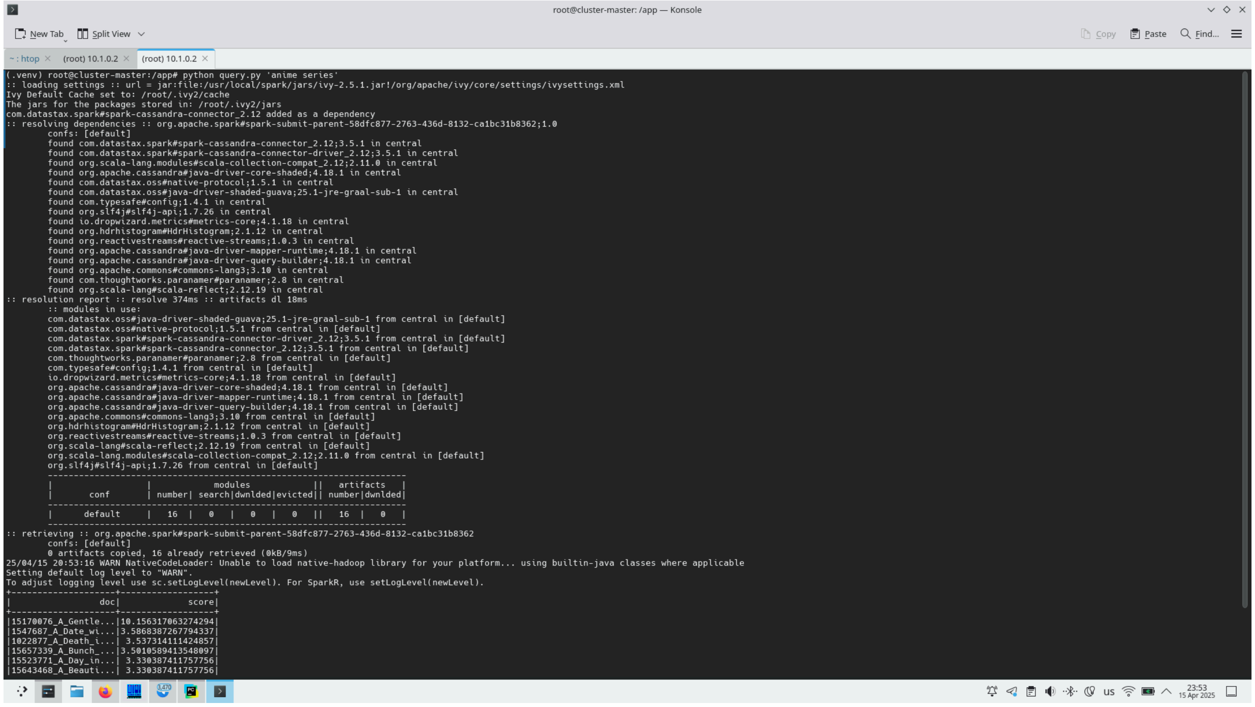
Term frequencies per each document.

# Queries



```
(.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-7b0fbcb0-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]
        ---------------------------------------------------------------------
        |                  |            modules            ||   artifacts   |
        |       conf       | number| search|dwnlded|evicted|| number|dwnlded|
        ---------------------------------------------------------------------
        |      default     |   16  |   0   |   0   |   0   ||   16  |   0   |
        ---------------------------------------------------------------------
:: retrieving :: org.apache.spark#spark-submit-parent-7b0fbcb0-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_wi...|3.5868387267794337|
|1022877_A_Death_i...| 3.537314111424857|
|15657339_A_Bunch_...|3.5010589413548097|
|15523771_A_Day_in...| 3.330387411757756|
|15643468_A_Beauti...| 3.330387411757756|
```

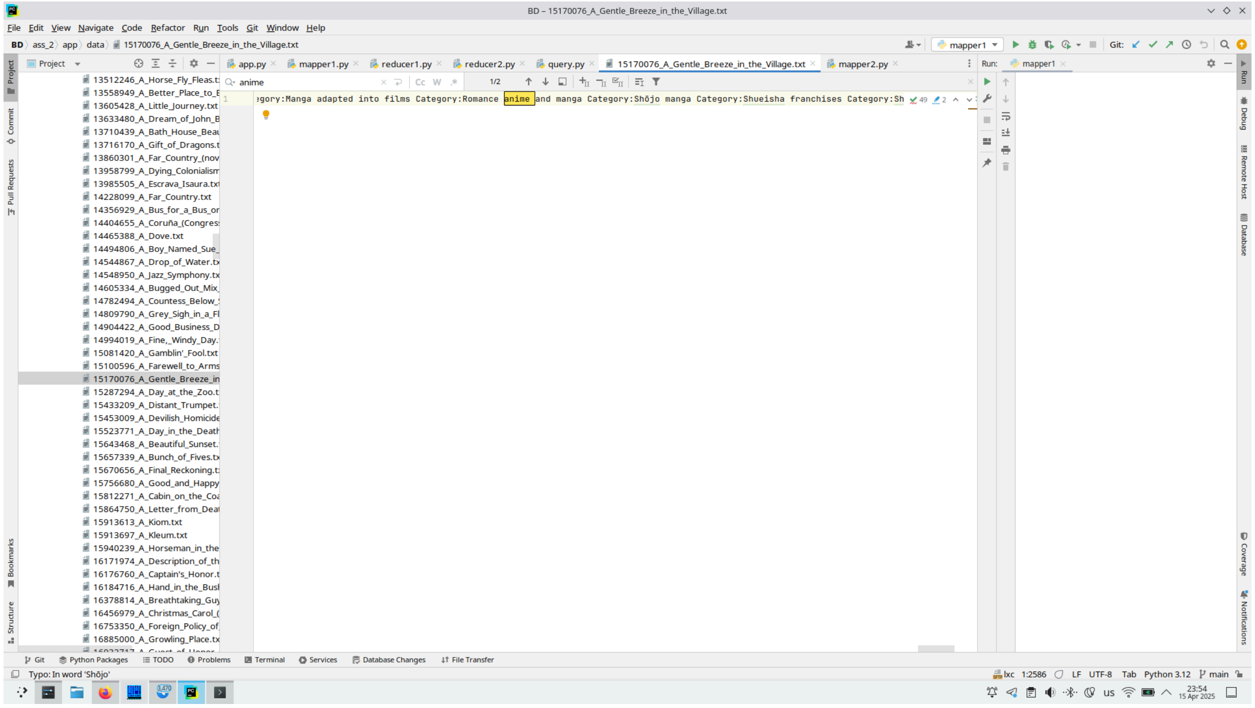
Query: “Crunchyroll anime series”



```
(.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]
        ---------------------------------------------------------------------
        |                  |            modules            ||   artifacts   |
        |       conf       | number| search|dwnlded|evicted|| number|dwnlded|
        ---------------------------------------------------------------------
        |      default     |   16  |   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_wi...|3.5868387267794337|
|1022877_A_Death_i...| 3.537314111424857|
|15657339_A_Bunch_...|3.5010589413548097|
|15523771_A_Day_in...| 3.330387411757756|
|15643468_A_Beauti...| 3.330387411757756|
```

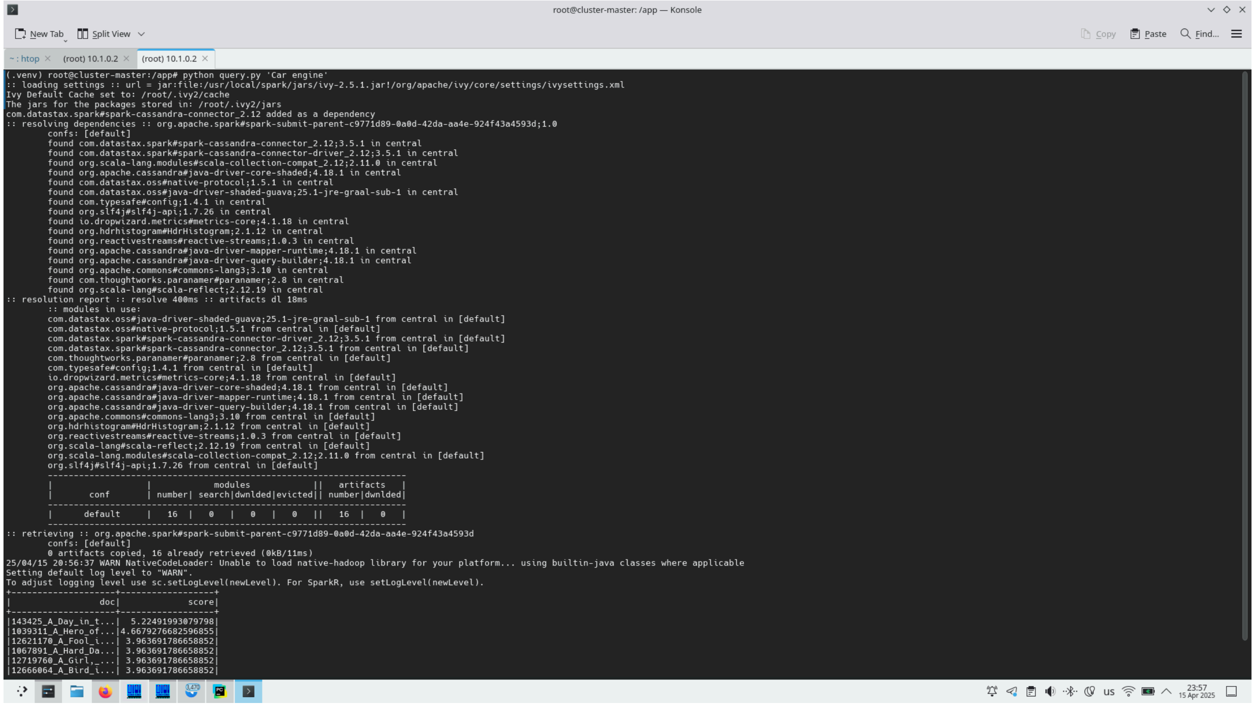
Query: “anime series”

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


BD – 15170076_A_Gentle_Breeze_in_the_Village.txt
File Edit View Navigate Code Refactor Run Tools Git Window Help
BD › ass_2 › app › data › 15170076_A_Gentle_Breeze_in_the_Village.txt
mapper1
Project
app.py
mapper1.py
reducer1.py
reducer2.py
query.py
15170076_A_Gentle_Breeze_in_the_Village.txt
mapper2.py
Run:
mapper1
13512246_A_Horse_Fly_Fleas.t
13558949_A_Better_Place_to_E
13605428_A_Little_Journey.txt
13633480_A_Dream_of_John_B
13710439_A_Bath_House_Beau
13716170_A_Gift_of_Dragons.t
13860301_A_Far_Country_(nov
13958799_A_Dying_Colonialism
13985505_A_Escrava_Isaura.txt
14228099_A_Far_Country.txt
14356929_A_Bus_for_a_Bus_or
14404655_A_Coruña_(Congres
14465388_A_Dove.txt
14494806_A_Boy_Named_Sue_
14544867_A_Drop_of_Water.tx
14548950_A_Jazz_Symphony.tx
14605334_A_Bugged_Out_Mix
14782494_A_Countess_Below_
14809790_A_Grey_Sigh_in_a_Fl
14904422_A_Good_Business_D
14994019_A_Fine,_Windy_Day.
15081420_A_Gamblin'_Fool.txt
15100596_A_Farewell_to_Arms
15170076_A_Gentle_Breeze_in
15287294_A_Day_at_the_Zoo.t
15433209_A_Distant_Trumpet.
15453009_A_Devilish_Homicide
15523771_A_Day_in_the_Death
15643468_A_Beautiful_Sunset.
15657339_A_Bunch_of_Fives.tx
15670656_A_Final_Reckoning.t
15756680_A_Good_and_Happy
15812271_A_Cabin_on_the_Coa
15864750_A_Letter_from_Deat
15913613_A_Kiom.txt
15913697_A_Kleum.txt
15940239_A_Horseman_in_the
16171974_A_Description_of_th
16176760_A_Captain's_Honor.t
16184716_A_Hand_in_the_Busl
16378814_A_Breathtaking_Guy
16456979_A_Christmas_Carol_(
16753350_A_Foreign_Policy_of
16885000_A_Growling_Place.tx
anime
1/2
Git
Python Packages
TODO
Problems
Terminal
Services
Database Changes
File Transfer
Typo: In word 'Shōjo'
lxc 1:2586 LF UTF-8 Tab Python 3.12 main
23:54
15 Apr 2025

```
:gory:Manga adapted into films Category:Romance anime and manga Category:Shōjo manga Category:Shueisha franchises Category:Sh
```

As it can be seen from the screenshots, similar queries lead to similar results and documents really contain keywords, and different query leads to different results.



```
(.venv) root@cluster-master:/app# python query.py 'Car engine'
:: 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-c9771d89-0a0d-42da-aa4e-924f43a4593d;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 400ms :: 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]
	---------------------------------------------------------------------
	|                  |            modules            ||   artifacts   |
	|       conf       | number| search|dwnlded|evicted|| number|dwnlded|
	---------------------------------------------------------------------
	|      default     |   16  |   0   |   0   |   0   ||   16  |   0   |
	---------------------------------------------------------------------
:: retrieving :: org.apache.spark#spark-submit-parent-c9771d89-0a0d-42da-aa4e-924f43a4593d
	confs: [default]
	0 artifacts copied, 16 already retrieved (0kB/11ms)
25/04/15 20:56:37 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|
+--------------------+------------------+
|143425_A_Day_in_t...|  5.22491993079798|
|1039311_A_Hero_of...|4.6679276682596855|
|12621170_A_Fool_i...| 3.963691786658852|
|1067891_A_Hard_Da...| 3.963691786658852|
|12719760_A_Girl,_...| 3.963691786658852|
|12666064_A_Bird_i...| 3.963691786658852|
```