Introduction to Big Data: Assignment 2

Author: Anastasiia Shvets
Group: B22-DS-02

Methodology

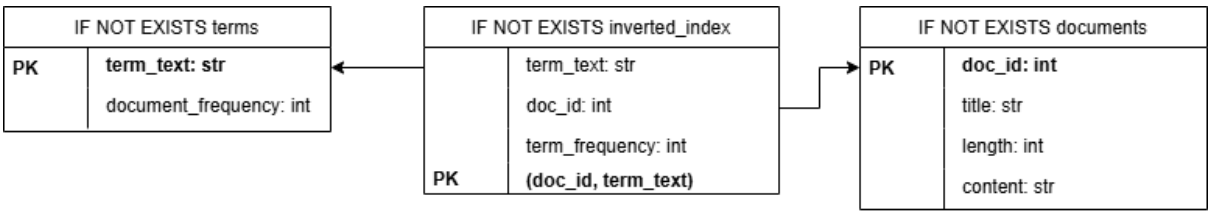
I followed the structure proposed in the github repository. However, I have made some design choices that are not present in the original solution. Here are the most import notes about my solution:

Data preparation

Data preparation part is taken fully from the repository, nothing was changed significantly. However, I did use another parquet file from Kaggle dataset: k.parquet, since I wanted to add some twist.

Cassandra schema

I created the following schema for the solution:



It is simple enough, but fits the queries I will use in the ranker (getting all documents, getting term frequency, getting inverted index for term and document).

Indexer

The indexer solution is presented in the indes.sh file. First it checks if user input is provided. If it is the case, it transforms the input to one csv file in hdfs to be passed to mapreduce. Otherwise it defaults to using csv file created in /index/data during data preparation.
The mapreduce consists of two parts:

1. mapper.py: adds every document to the Cassandra database, and splits the document into (term, doc_id, 1) triples, where 1 is then used to count the number of occurrences. The solution is not too

elegant, since I do not count the number of occurrences here, however, it works anyway.

2. reducer.py: gets the triples from mapper.py and counts statistics for a term in the document. Since the indexer can be run multiple times, the reducer also gets previously collected statistics from the Cassandra database.

Ranker

The ranker itself is simple, but I performed many workarounds to make it work with SparkRDD, such as trying multiple hosts for Cassandra and creating a new connection for every action. However, the logic is the following: I get all the documents, split the query into tokens and compute BM25. Then I rank the queries based on scores, and print the 10 highest.

Demonstration

How to run:

Before running the repository, make sure the “data” folder is created in the “app” folder, otherwise data preparation will not work!
Also make sure all files you are trying to rank exist.
Ensure you have k.parquet installed, or, if you want to use different file in data preparation, change file name both in prepare_data.py and prepare_data.sh. If you want to skip data preparation, comment out the line “bash prepare_data.sh” in app.sh

Go to app.sh file and put the file/directory you want to index after the “bash index.sh”

In the same file put queries you want to test after the “bash query.sh”

Make sure you have docker and docker compose installed.
Go to the project folder, and run

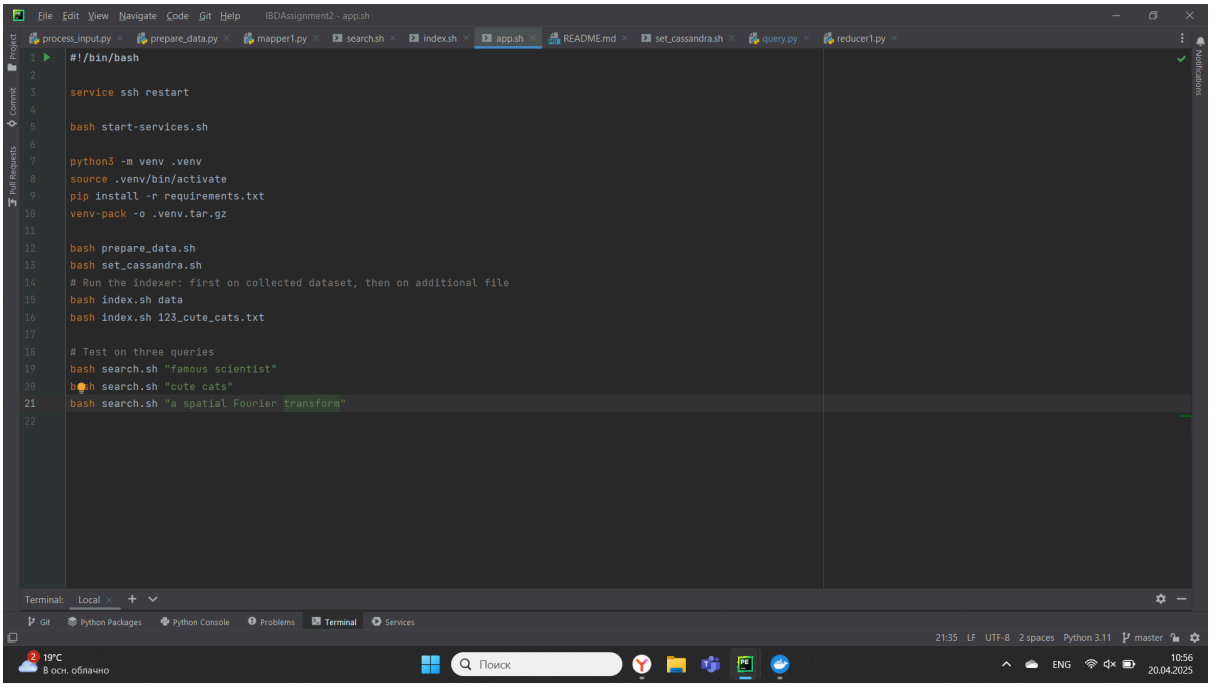
```
docker compose up
```

In the terminal.

My results:

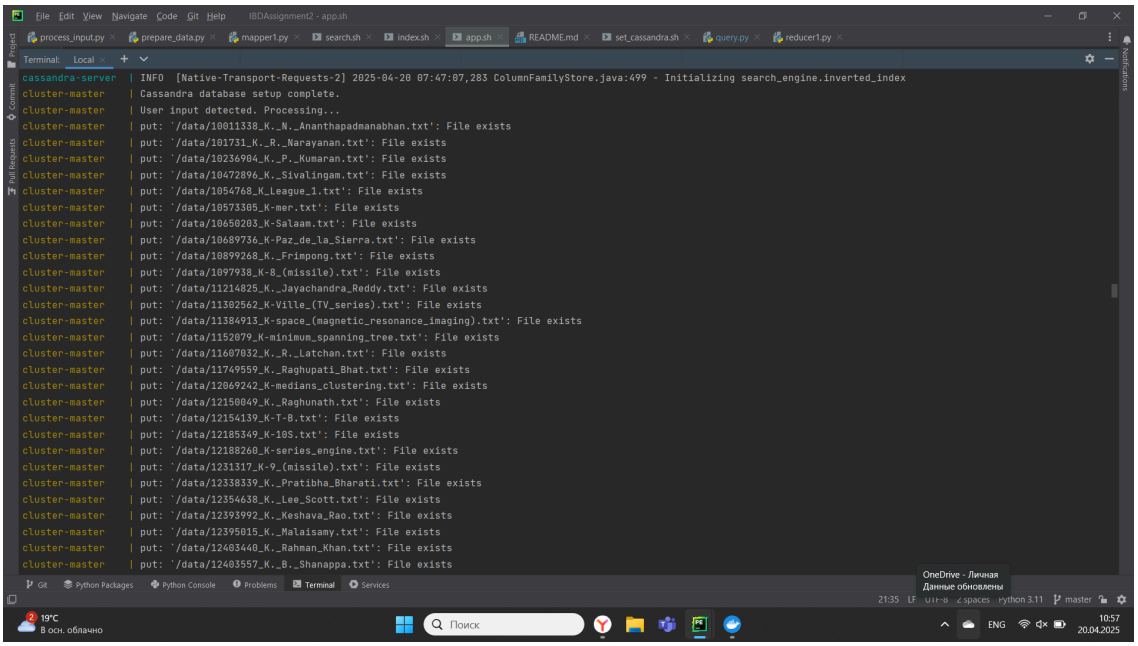
I have used a different laptop, not the same one as for Assignment 1, so the interface may look different. This laptop belongs to a person who does not have the Big Data course, since he is on Robotics, so he could not have benefitted from my solution in any way, shape or form.

So, here is my app.sh script:



As you can see, it will index both local data directory, and local 123_cute_cats.txt file.

So, here are the outputs for index.sh:



(As you can see, the files from the data folder have already been added to hadoop during data preparation.)

```
cluster-master    | 2025-04-20 07:49:03,262 INFO impl.YarnClientImpl: Submitted application application_1745135063118_0001
cluster-master    | 2025-04-20 07:49:03,305 INFO mapreduce.Job: The url to track the job: http://cluster-master:8088/proxy/application_1745135063118_0001/
cluster-master    | 2025-04-20 07:49:03,307 INFO mapreduce.Job: Running job: job_1745135063118_0001
cluster-master    | 2025-04-20 07:49:26,690 INFO mapreduce.Job: Job job_1745135063118_0001 running in uber mode : false
cluster-master    | 2025-04-20 07:49:26,698 INFO mapreduce.Job:  map 0% reduce 0%
cassandra-server  | INFO  [Native-Transport-Requests-15] 2025-04-20 07:49:31,618 QueryProcessor.java:654 - Fully upgraded to at least 5.0.4
cluster-master    | 2025-04-20 07:49:34,822 INFO mapreduce.Job:  map 100% reduce 0%
cluster-master    | 2025-04-20 07:49:53,015 INFO mapreduce.Job:  map 100% reduce 100%
cluster-master    | 2025-04-20 07:51:50,654 INFO mapreduce.Job: Job job_1745135063118_0001 completed successfully
cluster-master    | 2025-04-20 07:51:50,876 INFO mapreduce.Job: Counters: 54
cluster-master    |     File System Counters
cluster-master    |             FILE: Number of bytes read=10139115
cluster-master    |             FILE: Number of bytes written=21111545
cluster-master    |             FILE: Number of read operations=0
cluster-master    |             FILE: Number of large read operations=0
cluster-master    |             FILE: Number of write operations=0
cluster-master    |             HDFS: Number of bytes read=3465865
cluster-master    |             HDFS: Number of bytes written=0
cluster-master    |             HDFS: Number of read operations=11
cluster-master    |             HDFS: Number of large read operations=0
cluster-master    |             HDFS: Number of write operations=2
cluster-master    |             HDFS: Number of bytes read erasure-coded=0
cluster-master    |     Job Counters
cluster-master    |             Launched map tasks=2
cluster-master    |             Launched reduce tasks=1
cluster-master    |             Data-local map tasks=2
cluster-master    |             Total time spent by all maps in occupied slots (ms)=11566
cluster-master    |             Total time spent by all reduces in occupied slots (ms)=130598
cluster-master    |             Total time spent by all map tasks (ms)=11566
cluster-master    |             Total time spent by all reduce tasks (ms)=130598
cluster-master    |             Total vcore-milliseconds taken by all map tasks=11566
```

```
cluster-master    |             Total vcore-milliseconds taken by all reduce tasks=130598
cluster-master    |             Total megabyte-milliseconds taken by all map tasks=11843584
cluster-master    |             Total megabyte-milliseconds taken by all reduce tasks=133732352
cluster-master    |     Map-Reduce Framework
cluster-master    |             Map input records=996
cluster-master    |             Map output records=528979
cluster-master    |             Map output bytes=9081144
cluster-master    |             Map output materialized bytes=10139121
cluster-master    |             Input split bytes=290
cluster-master    |             Combine input records=0
cluster-master    |             Combine output records=0
cluster-master    |             Reduce input groups=79638
cluster-master    |             Reduce shuffle bytes=10139121
cluster-master    |             Reduce input records=528979
cluster-master    |             Reduce output records=0
cluster-master    |             Spilled Records=1057958
cluster-master    |             Shuffled Maps =2
cluster-master    |             Failed Shuffles=0
cluster-master    |             Merged Map outputs=2
cluster-master    |             GC time elapsed (ms)=190
cluster-master    |             CPU time spent (ms)=15300
cluster-master    |             Physical memory (bytes) snapshot=865804288
cluster-master    |             Virtual memory (bytes) snapshot=7773130752
cluster-master    |             Total committed heap usage (bytes)=758120448
cluster-master    |             Peak Map Physical memory (bytes)=329252864
cluster-master    |             Peak Map Virtual memory (bytes)=2590117888
cluster-master    |             Peak Reduce Physical memory (bytes)=310681600
cluster-master    |             Peak Reduce Virtual memory (bytes)=3587436544
cluster-master    |     Shuffle Errors
cluster-master    |             BAD_ID=0
```

```
cluster-master    |             BAD_ID=0
cluster-master    |             CONNECTION=0
cluster-master    |             IO_ERROR=0
cluster-master    |             WRONG_LENGTH=0
cluster-master    |             WRONG_MAP=0
cluster-master    |             WRONG_REDUCE=0
cluster-master    |     File Input Format Counters
cluster-master    |             Bytes Read=3465575
cluster-master    |     File Output Format Counters
cluster-master    |             Bytes Written=0
cluster-master    | 2025-04-20 07:51:50,877 INFO streaming.StreamJob: Output directory: /tmp/index/1745135292/stage1
cluster-master    | Indexing completed. Cleaning up...
cluster-master    | Deleted /tmp/index/1745135292
cluster-master    | Index is ready in Cassandra database.
cluster-master    | User input detected. Processing...
cluster-master    | Setting default log level to "WARN".
cluster-master    | To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
cluster-master    | 25/04/20 07:52:04 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
cluster-master    | /app/.venv/lib/python3.8/site-packages/pyspark/sql/context.py:113: FutureWarning: Deprecated in 3.0.0. Use SparkSession.builder.getOrCreate() instead.
cluster-master    |   warnings.warn(
User input prepared for mapreduce!
cluster-master    | Starting indexing process...
cluster-master    | packageJobJar: [/tmp/hadoop-unjar7576191271368134830/] [] /tmp/streamjob2606570591249577378.jar tmpDir=null
cluster-master    | 2025-04-20 07:52:27,940 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at cluster-master/172.18.0.4:8032
cluster-master    | 2025-04-20 07:52:28,123 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at cluster-master/172.18.0.4:8032
cluster-master    | 2025-04-20 07:52:28,339 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/root/.staging/job_1745135063118_0002
cluster-master    | 2025-04-20 07:52:48,262 INFO mapred.FileInputFormat: Total input files to process : 1
cluster-master    | 2025-04-20 07:52:48,742 INFO mapreduce.JobSubmitter: number of splits:2
cluster-master    | 2025-04-20 07:52:49,396 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1745135063118_0002
cluster-master    | 2025-04-20 07:52:49,396 INFO mapreduce.JobSubmitter: Executing with tokens: []
cluster-master    | 2025-04-20 07:52:49,651 INFO conf.Configuration: resource-types.xml not found
```

```
cluster-master | 2025-04-20 07:52:49,867 INFO mapreduce.Job: The url to track the job: http://cluster-master:8088/proxy/application_1745135063118_0002/
cluster-master | 2025-04-20 07:52:49,870 INFO mapreduce.Job: Running job: job_1745135063118_0002
cluster-master | 2025-04-20 07:53:08,229 INFO mapreduce.Job: Job job_1745135063118_0002 running in uber mode : false
cluster-master | 2025-04-20 07:53:08,272 INFO mapreduce.Job:  map 0% reduce 0%
cluster-master | 2025-04-20 07:53:14,511 INFO mapreduce.Job:  map 100% reduce 0%
cluster-master | 2025-04-20 07:53:31,606 INFO mapreduce.Job:  map 100% reduce 100%
cluster-master | 2025-04-20 07:55:21,109 INFO mapreduce.Job: Job job_1745135063118_0002 completed successfully
cluster-master | 2025-04-20 07:55:21,269 INFO mapreduce.Job: Counters: 54
cluster-master |     File System Counters
cluster-master |             FILE: Number of bytes read=449
cluster-master |             FILE: Number of bytes written=834213
cluster-master |             FILE: Number of read operations=0
cluster-master |             FILE: Number of large read operations=0
cluster-master |             FILE: Number of write operations=0
cluster-master |             HDFS: Number of bytes read=562
cluster-master |             HDFS: Number of bytes written=0
cluster-master |             HDFS: Number of read operations=11
cluster-master |             HDFS: Number of large read operations=0
cluster-master |             HDFS: Number of write operations=2
cluster-master |             HDFS: Number of bytes read erasure-coded=0
cluster-master |     Job Counters
cluster-master |             Launched map tasks=2
cluster-master |             Launched reduce tasks=1
cluster-master |             Data-local map tasks=2
cluster-master |             Total time spent by all maps in occupied slots (ms)=6529
cluster-master |             Total time spent by all reduces in occupied slots (ms)=123183
cluster-master |             Total time spent by all map tasks (ms)=6529
cluster-master |             Total time spent by all reduce tasks (ms)=123183
cluster-master |             Total vcore-milliseconds taken by all map tasks=6529
cluster-master |             Total vcore-milliseconds taken by all reduce tasks=123183
cluster-master |             Total megabyte-milliseconds taken by all map tasks=6685696
```

```
cluster-master |             Total megabyte-milliseconds taken by all map tasks=6685696
cluster-master |             Total megabyte-milliseconds taken by all reduce tasks=126139392
cluster-master |     Map-Reduce Framework
cluster-master |             Map input records=1
cluster-master |             Map output records=35
cluster-master |             Map output bytes=373
cluster-master |             Map output materialized bytes=455
cluster-master |             Input split bytes=290
cluster-master |             Combine input records=0
cluster-master |             Combine output records=0
cluster-master |             Reduce input groups=29
cluster-master |             Reduce shuffle bytes=455
cluster-master |             Reduce input records=35
cluster-master |             Reduce output records=0
cluster-master |             Spilled Records=70
cluster-master |             Shuffled Maps =2
cluster-master |             Failed Shuffles=0
cluster-master |             Merged Map outputs=2
cluster-master |             GC time elapsed (ms)=179
cluster-master |             CPU time spent (ms)=10970
cluster-master |             Physical memory (bytes) snapshot=844423168
cluster-master |             Virtual memory (bytes) snapshot=7772311552
cluster-master |             Total committed heap usage (bytes)=747110400
cluster-master |     File Input Format Counters
cluster-master |             Bytes Read=272
cluster-master |     File Output Format Counters
cluster-master |             Bytes Written=0
cluster-master | 2025-04-20 07:55:21,269 INFO streaming.StreamJob: Output directory: /tmp/index/1745135544/stage1
cluster-master | Indexing completed. Cleaning up...
cluster-master | Deleted /tmp/index/1745135544
cluster-master | Index is ready in Cassandra database.
```

As you can see, both files were successfully indexed!

Now for the ranker part. I have created three test queries:

`"famous scientist"`

`"cute cats"`

`"a spatial Fourier transform"`

I expected the first query to output some scientist, the second to rank the cute_cats file at the top, and the last one to rank K-space file high, since it is a quote from it. I got the following results:

```
cluster-master  | 25/04/20 07:55:41 INFO TaskSchedulerImpl: Removed TaskSet 0.0, whose tasks have all completed, from pool
cluster-master  | 25/04/20 07:55:41 INFO DAGScheduler: ResultStage 0 (takeOrdered at /app/query.py:124) finished in 3.556 s
cluster-master  | 25/04/20 07:55:41 INFO DAGScheduler: Job 0 is finished. Cancelling potential speculative or zombie tasks for this job
cluster-master  | 25/04/20 07:55:41 INFO TaskSchedulerImpl: Killing all running tasks in stage 0: Stage finished
cluster-master  | 25/04/20 07:55:41 INFO DAGScheduler: Job 0 finished: takeOrdered at /app/query.py:124, took 3.629020 s
cluster-master  |
cluster-master  | Query: famous scientist
cluster-master  |
cluster-master  | Top 10 relevant documents:
cluster-master  | --------------------------
cluster-master  | 1. K. G. Markose.txt (ID: 38645005, Score: 13.3694)
cluster-master  | 2. K. Christopher Garcia.txt (ID: 52035641, Score: 12.2767)
cluster-master  | 3. K. Sundaram.txt (ID: 24917114, Score: 0.0000)
cluster-master  | 4. K. Rangaraj.txt (ID: 50667854, Score: 0.0000)
cluster-master  | 5. K. G. Paulose.txt (ID: 53430244, Score: 0.0000)
cluster-master  | 6. K. G. Ossiannilsson.txt (ID: 47948151, Score: 0.0000)
cluster-master  | 7. K-Jee (kickboxer).txt (ID: 65816655, Score: 0.0000)
cluster-master  | 8. K. T. Turner.txt (ID: 73440944, Score: 0.0000)
cluster-master  | 9. K. Reuben Mark.txt (ID: 47655625, Score: 0.0000)
cluster-master  | 10. K. V. Iyer.txt (ID: 60188825, Score: 0.0000)
cluster-master  | 25/04/20 07:55:41 INFO SparkContext: SparkContext is stopping with exitCode 0.
cluster-master  | 25/04/20 07:55:41 INFO SparkUI: Stopped Spark web UI at http://cluster-master:4040
cluster-master  | 25/04/20 07:55:41 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
cluster-master  | 25/04/20 07:55:41 INFO MemoryStore: MemoryStore cleared
cluster-master  | 25/04/20 07:55:41 INFO BlockManager: BlockManager stopped
cluster-master  | 25/04/20 07:55:41 INFO BlockManagerMaster: BlockManagerMaster stopped
cluster-master  | 25/04/20 07:55:41 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
cluster-master  | 25/04/20 07:55:41 INFO SparkContext: Successfully stopped SparkContext
cluster-master  | 25/04/20 07:55:42 INFO ShutdownHookManager: Shutdown hook called
cluster-master  | 25/04/20 07:55:42 INFO ShutdownHookManager: Deleting directory /tmp/spark-6d39d9f6-7ea6-437a-a000-8d29531f3195/pyspark-e013fc2c-d500-43cd-861a-2b5ad56ab168
cluster-master  | 25/04/20 07:55:42 INFO ShutdownHookManager: Deleting directory /tmp/spark-234a41a9-8b93-478e-bbf0-fc27bb2c65b3
```

```
cluster-master  | 25/04/20 07:55:57 INFO DAGScheduler: ResultStage 0 (takeOrdered at /app/query.py:124) finished in 3.763 s
cluster-master  | 25/04/20 07:55:57 INFO DAGScheduler: Job 0 is finished. Cancelling potential speculative or zombie tasks for this job
cluster-master  | 25/04/20 07:55:57 INFO TaskSchedulerImpl: Killing all running tasks in stage 0: Stage finished
cluster-master  | 25/04/20 07:55:57 INFO DAGScheduler: Job 0 finished: takeOrdered at /app/query.py:124, took 3.840724 s
cluster-master  |
cluster-master  | Query: cute cats
cluster-master  |
cluster-master  | Top 10 relevant documents:
cluster-master  | --------------------------
cluster-master  | 1. cute cats.txt (ID: 123, Score: 23.1824)
cluster-master  | 2. K. Sundaram.txt (ID: 24917114, Score: 0.0000)
cluster-master  | 3. K. Rangaraj.txt (ID: 50667854, Score: 0.0000)
cluster-master  | 4. K. G. Paulose.txt (ID: 53430244, Score: 0.0000)
cluster-master  | 5. K. G. Ossiannilsson.txt (ID: 47948151, Score: 0.0000)
cluster-master  | 6. K-Jee (kickboxer).txt (ID: 65816655, Score: 0.0000)
cluster-master  | 7. K. T. Turner.txt (ID: 73440944, Score: 0.0000)
cluster-master  | 8. K. Reuben Mark.txt (ID: 47655625, Score: 0.0000)
cluster-master  | 9. K. V. Iyer.txt (ID: 60188825, Score: 0.0000)
cluster-master  | 10. K. T. S. Sarao.txt (ID: 66198867, Score: 0.0000)
cluster-master  | 25/04/20 07:55:57 INFO SparkContext: SparkContext is stopping with exitCode 0.
cluster-master  | 25/04/20 07:55:57 INFO SparkUI: Stopped Spark web UI at http://cluster-master:4040
cluster-master  | 25/04/20 07:55:57 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
cluster-master  | 25/04/20 07:55:57 INFO MemoryStore: MemoryStore cleared
cluster-master  | 25/04/20 07:55:57 INFO BlockManager: BlockManager stopped
cluster-master  | 25/04/20 07:55:57 INFO BlockManagerMaster: BlockManagerMaster stopped
cluster-master  | 25/04/20 07:55:57 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
cluster-master  | 25/04/20 07:55:57 INFO SparkContext: Successfully stopped SparkContext
cluster-master  | 25/04/20 07:55:58 INFO ShutdownHookManager: Shutdown hook called
cluster-master  | 25/04/20 07:55:58 INFO ShutdownHookManager: Deleting directory /tmp/spark-9c4682b2-78ee-498c-a4b1-4b0a57842700
cluster-master  | 25/04/20 07:55:58 INFO ShutdownHookManager: Deleting directory /tmp/spark-9c4682b2-78ee-498c-a4b1-4b0a57842700/pyspark-f65a89ea-2613-4526-8768-9a2bde64ab91
cluster-master  | 25/04/20 07:55:58 INFO ShutdownHookManager: Deleting directory /tmp/spark-e68f583c-1ce9-4d86-8ad9-7e2d40736de0
```

```
cluster-master  | 25/04/20 08:04:20 INFO DAGScheduler: Job 0 finished: takeOrdered at /app/query.py:124, took 3.773540 s
cluster-master  |
cluster-master  | Query: a spatial Fourier transform
cluster-master  |
cluster-master  | Top 10 relevant documents:
cluster-master  | --------------------------
cluster-master  | 1. K-space (magnetic resonance imaging).txt (ID: 11384913, Score: 12.7655)
cluster-master  | 2. K. C. Akshay.txt (ID: 55681579, Score: 12.4223)
cluster-master  | 3. K-theory (physics).txt (ID: 4221033, Score: 11.7127)
cluster-master  | 4. cute cats.txt (ID: 123, Score: 9.5598)
cluster-master  | 5. K. Sundaram.txt (ID: 24917114, Score: 0.0000)
cluster-master  | 6. K. Rangaraj.txt (ID: 50667854, Score: 0.0000)
cluster-master  | 7. K. G. Paulose.txt (ID: 53430244, Score: 0.0000)
cluster-master  | 8. K. G. Ossiannilsson.txt (ID: 47948151, Score: 0.0000)
cluster-master  | 9. K-Jee (kickboxer).txt (ID: 65816655, Score: 0.0000)
cluster-master  | 10. K. T. Turner.txt (ID: 73440944, Score: 0.0000)
cluster-master  | 25/04/20 08:04:20 INFO SparkContext: SparkContext is stopping with exitCode 0.
cluster-master  | 25/04/20 08:04:20 INFO SparkUI: Stopped Spark web UI at http://cluster-master:4040
cluster-master  | 25/04/20 08:04:20 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
cluster-master  | 25/04/20 08:04:20 INFO MemoryStore: MemoryStore cleared
cluster-master  | 25/04/20 08:04:20 INFO BlockManager: BlockManager stopped
cluster-master  | 25/04/20 08:04:20 INFO BlockManagerMaster: BlockManagerMaster stopped
cluster-master  | 25/04/20 08:04:20 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
cluster-master  | 25/04/20 08:04:20 INFO SparkContext: Successfully stopped SparkContext
cluster-master  | 25/04/20 08:04:21 INFO ShutdownHookManager: Shutdown hook called
cluster-master  | 25/04/20 08:04:21 INFO ShutdownHookManager: Deleting directory /tmp/spark-64834403-e2c7-465f-9087-9b9e8dfa5214
cluster-master  | 25/04/20 08:04:21 INFO ShutdownHookManager: Deleting directory /tmp/spark-64834403-e2c7-465f-9087-9b9e8dfa5214/pyspark-3b077090-1465-43e0-a838-a2c5ccc38bbf
cluster-master  | 25/04/20 08:04:21 INFO ShutdownHookManager: Deleting directory /tmp/spark-78e7c752-5467-4f7b-8983-ab326ced2022
cluster-master exited with code 0

v View in Docker Desktop   o View Config   w Enable Watch
```

(I had to re-run for the last one, it didn’t show up the first time for some reason)

The results for queries 2 and 3 were as expected, however for 1 the top ranked documents are K. G. Markose (an Indian singer) and K. Christofer Garcia (scientist, expectedly). I think the reason for the singer appearing in the results is the word “famous”.
Overall, I am pleased with the results. Most of the queries work as expected, and even unexpected results have logical reasons.