Track: Data Science	**Assignment 2**
Name: Sofia Gamershmidt	**Introduction to Big Data**
Email: s.gamershmidt@innopolis.university	

Report

Methodology

Data collection and preparation

I will proceed with my work with a.parquet file.
Initially, I needed to explore the docker-compose.yaml file to understand at which point and in which file the data is processed. The cluster-master container runs the /app/app.sh script. The app/app.sh runs prepare_data.sh, which runs spark-submit prepare_data.py and puts everything in the hdfs. So I need only to modify prepare_data.py in order to prepare data.

I took the code from assignment description as base, but made some modifications:

1. Add some configs(to avoid random problems with javaerrors)
2. Got rid of line .sample(fraction=100 * n / df.count(), seed=0) because of javaheap error
3. Added filter on nonempty and nonnull texts
4. A bit modified filename creation
5. Added PySpark RDD operations, fixed issues with addition of existing paths

Additionly, I faced problems with functions like df_sample.collect(). It caused also JavaHeap errors.
Everything I run manually for step by step execution and full control at the developing stage

Indexer tasks

I will create 3 tables:

1. term_freq (term, doc_id, TF)
2. doc_freq (term, DF)
3. doc_stats (doc_id, length, title)

term_freq will contain information about each term, document id for which the TF was calculated, and TF value (calculation is explained in assignment description)

doc_freq will contain term and document frequency (I will add + 1 to document frequency, because I will use in denominator in BM25 calculation and I want to avoid division by 0 problem)

doc_stats will contain doc id, number of tokens in it and doc title

To do it I am going to use 2 mappers and reducers: since they communicate via stdin, i needed to use mapper1 to process input documents and create triplets (word, doc_id, 1), then reducer aggregates it to triplets (word, doc_id, tf), mapper2 creates tuples (doc_id, title) and reducer2 deletes duplicates

Index.sh runs two Hadoop streaming jobs to generate the term frequency index and to extract document titles and launches the Cassandra insertion using app.py
Issues I faced: errors connected to existence of certain paths. The solution is to delete them before:

```
hdfs dfs -rm -r -f "$OUTPUT_PATH"
hdfs dfs -rm -r -f "$OUTPUT_TITLE_PATH"
```

Ranker tasks

search.sh runs query.py file using spark-submit. In this file I connect to cassandra and read three tables: term_freq, doc_freq, and doc_stats. Then I take user query, split it into words, and search only for those terms in index.

I calculate BM25 score using formula provided in the assignment description. I sum scores for same document and take top 10 results with highest score. I also show title of document from doc_stats table. I added +1 in denominator in previous step to avoid division by zero.

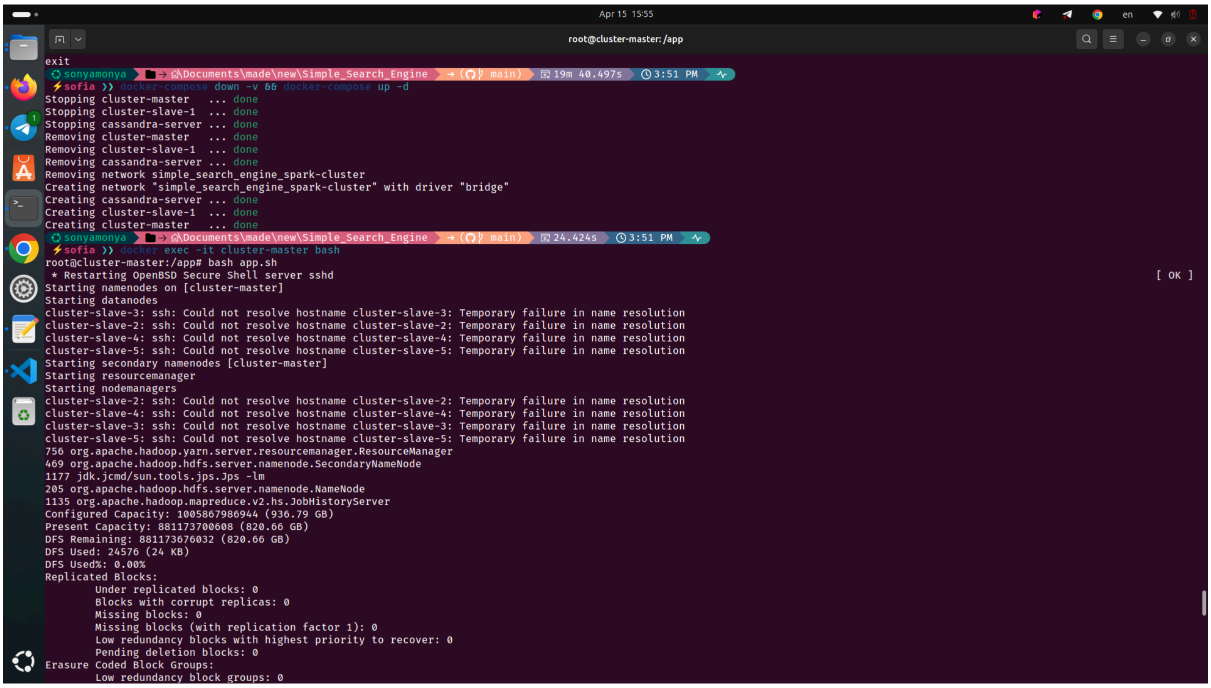
The results are in the next section
I tried this on the test sample with 10 documents only and if there are files with 0 bm25 they are not included into final answer(so there can be even 0 matches for small data sizes)

Demonstration

ATTENTION: code may fail due to cache reasons, to overcome this you need just to clear docker cache and restart the docker-compose. This will solve all problems

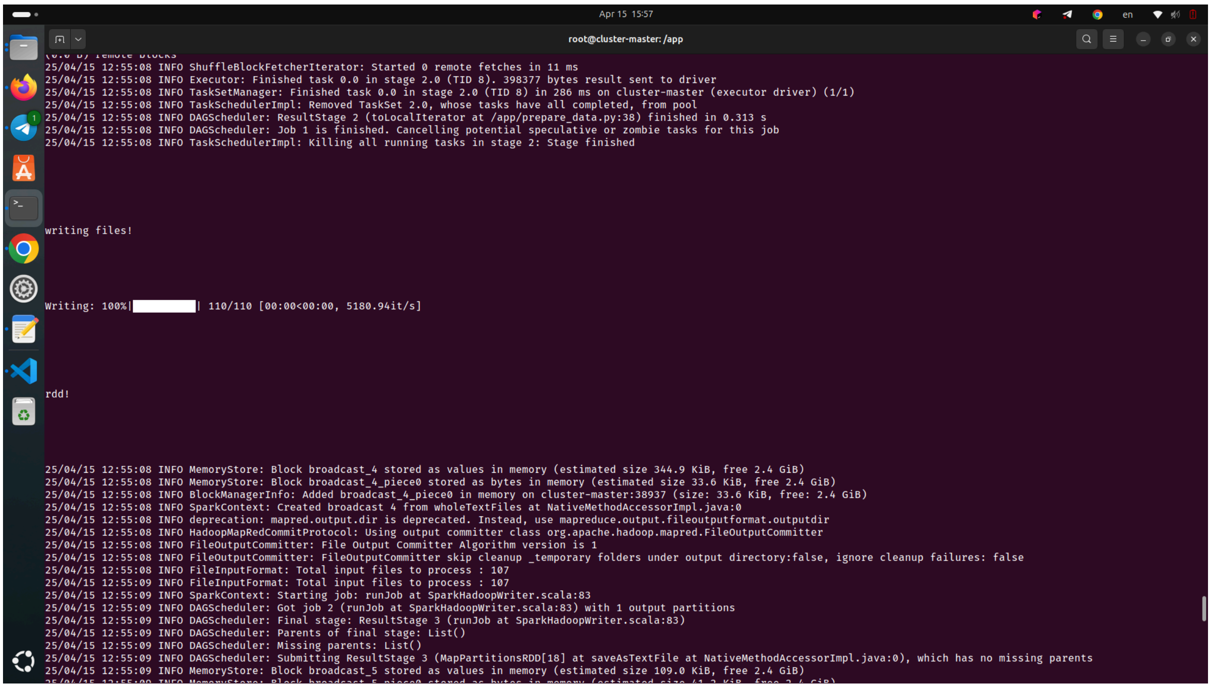
To run the code you need to print in shell:
docker-compose up

But I will run manually for demonstration
cd project/path
docker-compose up -d
docker exec -it cluster-master bash
bash app.sh

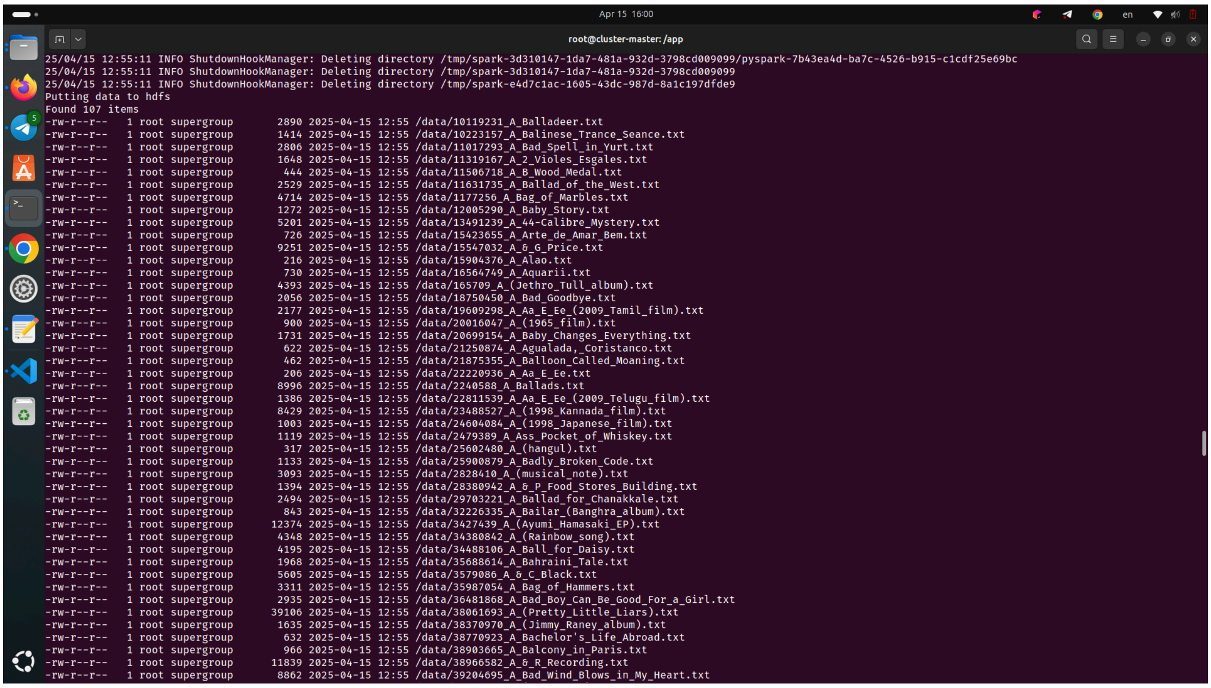


```
exit
 sonyamonya  → ⌂\Documents\made\new\Simple_Search_Engine  → ( main)  19m 40.497s  3:51 PM
 ⚡sofia ❯❯ docker-compose down -v && docker-compose up -d
Stopping cluster-master   ... done
Stopping cluster-slave-1  ... done
Stopping cassandra-server ... done
Removing cluster-master   ... done
Removing cluster-slave-1  ... done
Removing cassandra-server ... done
Removing network simple_search_engine_spark-cluster
Creating network "simple_search_engine_spark-cluster" with driver "bridge"
Creating cassandra-server ... done
Creating cluster-slave-1  ... done
Creating cluster-master   ... done
 sonyamonya  → ⌂\Documents\made\new\Simple_Search_Engine  → ( main)  24.424s  3:51 PM
 ⚡sofia ❯❯ docker exec -it cluster-master bash
root@cluster-master:/app# bash app.sh
 * Restarting OpenBSD Secure Shell server sshd                                                                                   [ OK ]
Starting namenodes on [cluster-master]
Starting datanodes
cluster-slave-3: ssh: Could not resolve hostname cluster-slave-3: Temporary failure in name resolution
cluster-slave-2: ssh: Could not resolve hostname cluster-slave-2: Temporary failure in name resolution
cluster-slave-4: ssh: Could not resolve hostname cluster-slave-4: Temporary failure in name resolution
cluster-slave-5: ssh: Could not resolve hostname cluster-slave-5: Temporary failure in name resolution
Starting secondary namenodes [cluster-master]
Starting resourcemanager
Starting nodemanagers
cluster-slave-2: ssh: Could not resolve hostname cluster-slave-2: Temporary failure in name resolution
cluster-slave-4: ssh: Could not resolve hostname cluster-slave-4: Temporary failure in name resolution
cluster-slave-3: ssh: Could not resolve hostname cluster-slave-3: Temporary failure in name resolution
cluster-slave-5: ssh: Could not resolve hostname cluster-slave-5: Temporary failure in name resolution
756 org.apache.hadoop.yarn.server.resourcemanager.ResourceManager
469 org.apache.hadoop.hdfs.server.namenode.SecondaryNameNode
1177 jdk.jcmd/sun.tools.jps.Jps -lm
205 org.apache.hadoop.hdfs.server.namenode.NameNode
1135 org.apache.hadoop.mapreduce.v2.hs.JobHistoryServer
Configured Capacity: 1005867986944 (936.79 GB)
Present Capacity: 881173700608 (820.66 GB)
DFS Remaining: 881173676032 (820.66 GB)
DFS Used: 24576 (24 KB)
DFS Used%: 0.00%
Replicated Blocks:
        Under replicated blocks: 0
        Blocks with corrupt replicas: 0
        Missing blocks: 0
        Missing blocks (with replication factor 1): 0
        Low redundancy blocks with highest priority to recover: 0
        Pending deletion blocks: 0
Erasure Coded Block Groups:
        Low redundancy block groups: 0
```

data preparation started:

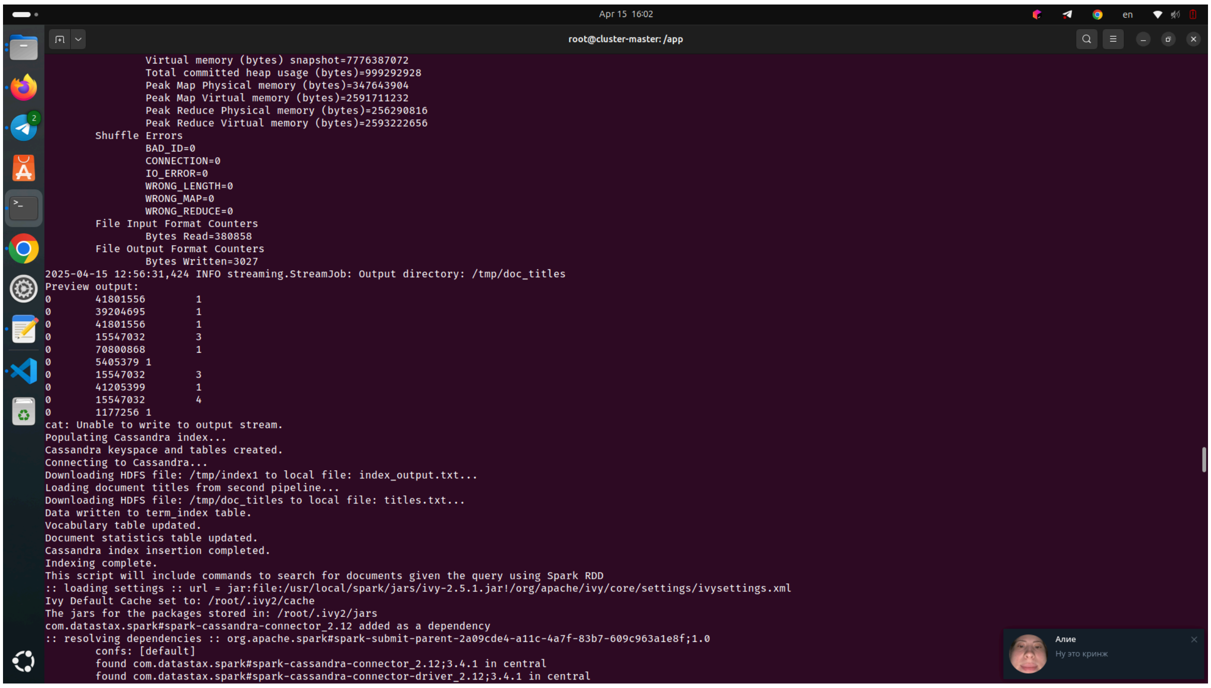


```
25/04/15 12:55:08 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 11 ms
25/04/15 12:55:08 INFO Executor: Finished task 0.0 in stage 2.0 (TID 8). 398377 bytes result sent to driver
25/04/15 12:55:08 INFO TaskSetManager: Finished task 0.0 in stage 2.0 (TID 8) in 286 ms on cluster-master (executor driver) (1/1)
25/04/15 12:55:08 INFO TaskSchedulerImpl: Removed TaskSet 2.0, whose tasks have all completed, from pool
25/04/15 12:55:08 INFO DAGScheduler: ResultStage 2 (toLocalIterator at /app/prepare_data.py:38) finished in 0.313 s
25/04/15 12:55:08 INFO DAGScheduler: Job 1 is finished. Cancelling potential speculative or zombie tasks for this job
25/04/15 12:55:08 INFO TaskSchedulerImpl: Killing all running tasks in stage 2: Stage finished




writing files!




Writing: 100%|██████████| 110/110 [00:00<00:00, 5180.94it/s]




rdd!




25/04/15 12:55:08 INFO MemoryStore: Block broadcast_4 stored as values in memory (estimated size 344.9 KiB, free 2.4 GiB)
25/04/15 12:55:08 INFO MemoryStore: Block broadcast_4_piece0 stored as bytes in memory (estimated size 33.6 KiB, free 2.4 GiB)
25/04/15 12:55:08 INFO BlockManagerInfo: Added broadcast_4_piece0 in memory on cluster-master:38937 (size: 33.6 KiB, free: 2.4 GiB)
25/04/15 12:55:08 INFO SparkContext: Created broadcast 4 from wholeTextFiles at NativeMethodAccessorImpl.java:0
25/04/15 12:55:08 INFO deprecation: mapred.output.dir is deprecated. Instead, use mapreduce.output.fileoutputformat.outputdir
25/04/15 12:55:08 INFO HadoopMapRedCommitProtocol: Using output committer class org.apache.hadoop.mapred.FileOutputCommitter
25/04/15 12:55:08 INFO FileOutputCommitter: File Output Committer Algorithm version is 1
25/04/15 12:55:08 INFO FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
25/04/15 12:55:08 INFO FileInputFormat: Total input files to process : 107
25/04/15 12:55:09 INFO FileInputFormat: Total input files to process : 107
25/04/15 12:55:09 INFO SparkContext: Starting job: runJob at SparkHadoopWriter.scala:83
25/04/15 12:55:09 INFO DAGScheduler: Got job 2 (runJob at SparkHadoopWriter.scala:83) with 1 output partitions
25/04/15 12:55:09 INFO DAGScheduler: Final stage: ResultStage 3 (runJob at SparkHadoopWriter.scala:83)
25/04/15 12:55:09 INFO DAGScheduler: Parents of final stage: List()
25/04/15 12:55:09 INFO DAGScheduler: Missing parents: List()
25/04/15 12:55:09 INFO DAGScheduler: Submitting ResultStage 3 (MapPartitionsRDD[18] at saveAsTextFile at NativeMethodAccessorImpl.java:0), which has no missing parents
25/04/15 12:55:09 INFO MemoryStore: Block broadcast_5 stored as values in memory (estimated size 109.0 KiB, free 2.4 GiB)
```



```
25/04/15 12:55:11 INFO ShutdownHookManager: Deleting directory /tmp/spark-3d310147-1da7-481a-932d-3798cd009099/pyspark-7b43ea4d-ba7c-4526-b915-c1cdf25e69bc
25/04/15 12:55:11 INFO ShutdownHookManager: Deleting directory /tmp/spark-3d310147-1da7-481a-932d-3798cd009099
25/04/15 12:55:11 INFO ShutdownHookManager: Deleting directory /tmp/spark-e4d7c1ac-1605-43dc-987d-8a1c197dfde9
Putting data to hdfs
Found 107 items
-rw-r--r--   1 root supergroup       2890 2025-04-15 12:55 /data/10119231_A_Balladeer.txt
-rw-r--r--   1 root supergroup       1414 2025-04-15 12:55 /data/10223157_A_Balinese_Trance_Seance.txt
-rw-r--r--   1 root supergroup       2806 2025-04-15 12:55 /data/11017293_A_Bad_Spell_in_Yurt.txt
-rw-r--r--   1 root supergroup       1648 2025-04-15 12:55 /data/11319167_A_2_Violes_Esgales.txt
-rw-r--r--   1 root supergroup        444 2025-04-15 12:55 /data/11506718_A_B_Wood_Medal.txt
-rw-r--r--   1 root supergroup       2529 2025-04-15 12:55 /data/11631735_A_Ballad_of_the_West.txt
-rw-r--r--   1 root supergroup       4714 2025-04-15 12:55 /data/1177256_A_Bag_of_Marbles.txt
-rw-r--r--   1 root supergroup       1272 2025-04-15 12:55 /data/12005290_A_Baby_Story.txt
-rw-r--r--   1 root supergroup       5201 2025-04-15 12:55 /data/13491239_A_44-Calibre_Mystery.txt
-rw-r--r--   1 root supergroup        726 2025-04-15 12:55 /data/15423655_A_Arte_de_Amar_Bem.txt
-rw-r--r--   1 root supergroup       9251 2025-04-15 12:55 /data/15547032_A_&_G_Price.txt
-rw-r--r--   1 root supergroup        216 2025-04-15 12:55 /data/15904376_A_Alao.txt
-rw-r--r--   1 root supergroup        730 2025-04-15 12:55 /data/16564749_A_Aquarii.txt
-rw-r--r--   1 root supergroup       4393 2025-04-15 12:55 /data/165709_A_(Jethro_Tull_album).txt
-rw-r--r--   1 root supergroup       2056 2025-04-15 12:55 /data/18750450_A_Bad_Goodbye.txt
-rw-r--r--   1 root supergroup       2177 2025-04-15 12:55 /data/19609298_A_Aa_E_Ee_(2009_Tamil_film).txt
-rw-r--r--   1 root supergroup        900 2025-04-15 12:55 /data/20016047_A_(1965_film).txt
-rw-r--r--   1 root supergroup       1731 2025-04-15 12:55 /data/20699154_A_Baby_Changes_Everything.txt
-rw-r--r--   1 root supergroup        622 2025-04-15 12:55 /data/21250874_A_Agualada,_Coristanco.txt
-rw-r--r--   1 root supergroup        462 2025-04-15 12:55 /data/21875355_A_Balloon_Called_Moaning.txt
-rw-r--r--   1 root supergroup        206 2025-04-15 12:55 /data/22220936_A_Aa_E_Ee.txt
-rw-r--r--   1 root supergroup       8996 2025-04-15 12:55 /data/2240588_A_Ballads.txt
-rw-r--r--   1 root supergroup       1386 2025-04-15 12:55 /data/22811539_A_Aa_E_Ee_(2009_Telugu_film).txt
-rw-r--r--   1 root supergroup       8429 2025-04-15 12:55 /data/23488527_A_(1998_Kannada_film).txt
-rw-r--r--   1 root supergroup       1003 2025-04-15 12:55 /data/24604084_A_(1998_Japanese_film).txt
-rw-r--r--   1 root supergroup       1119 2025-04-15 12:55 /data/2479389_A_Ass_Pocket_of_Whiskey.txt
-rw-r--r--   1 root supergroup        317 2025-04-15 12:55 /data/25602480_A_(hangul).txt
-rw-r--r--   1 root supergroup       1133 2025-04-15 12:55 /data/25900879_A_Badly_Broken_Code.txt
-rw-r--r--   1 root supergroup       3093 2025-04-15 12:55 /data/2828410_A_(musical_note).txt
-rw-r--r--   1 root supergroup       1394 2025-04-15 12:55 /data/28380942_A_&_P_Food_Stores_Building.txt
-rw-r--r--   1 root supergroup       2494 2025-04-15 12:55 /data/29703221_A_Ballad_for_Chanakkale.txt
-rw-r--r--   1 root supergroup        843 2025-04-15 12:55 /data/32226335_A_Bailar_(Banghra_album).txt
-rw-r--r--   1 root supergroup      12374 2025-04-15 12:55 /data/3427439_A_(Ayumi_Hamasaki_EP).txt
-rw-r--r--   1 root supergroup       4348 2025-04-15 12:55 /data/34380842_A_(Rainbow_song).txt
-rw-r--r--   1 root supergroup       4195 2025-04-15 12:55 /data/34488106_A_Ball_for_Daisy.txt
-rw-r--r--   1 root supergroup       1968 2025-04-15 12:55 /data/35688614_A_Bahraini_Tale.txt
-rw-r--r--   1 root supergroup       5605 2025-04-15 12:55 /data/3579086_A_&_C_Black.txt
-rw-r--r--   1 root supergroup       3311 2025-04-15 12:55 /data/35987054_A_Bag_of_Hammers.txt
-rw-r--r--   1 root supergroup       2935 2025-04-15 12:55 /data/36481868_A_Bad_Boy_Can_Be_Good_For_a_Girl.txt
-rw-r--r--   1 root supergroup      39106 2025-04-15 12:55 /data/38061693_A_(Pretty_Little_Liars).txt
-rw-r--r--   1 root supergroup       1635 2025-04-15 12:55 /data/38370970_A_(Jimmy_Raney_album).txt
-rw-r--r--   1 root supergroup        632 2025-04-15 12:55 /data/38770923_A_Bachelor's_Life_Abroad.txt
-rw-r--r--   1 root supergroup        966 2025-04-15 12:55 /data/38903665_A_Balcony_in_Paris.txt
-rw-r--r--   1 root supergroup      11839 2025-04-15 12:55 /data/38966582_A_&_R_Recording.txt
-rw-r--r--   1 root supergroup       8862 2025-04-15 12:55 /data/39204695_A_Bad_Wind_Blows_in_My_Heart.txt
```

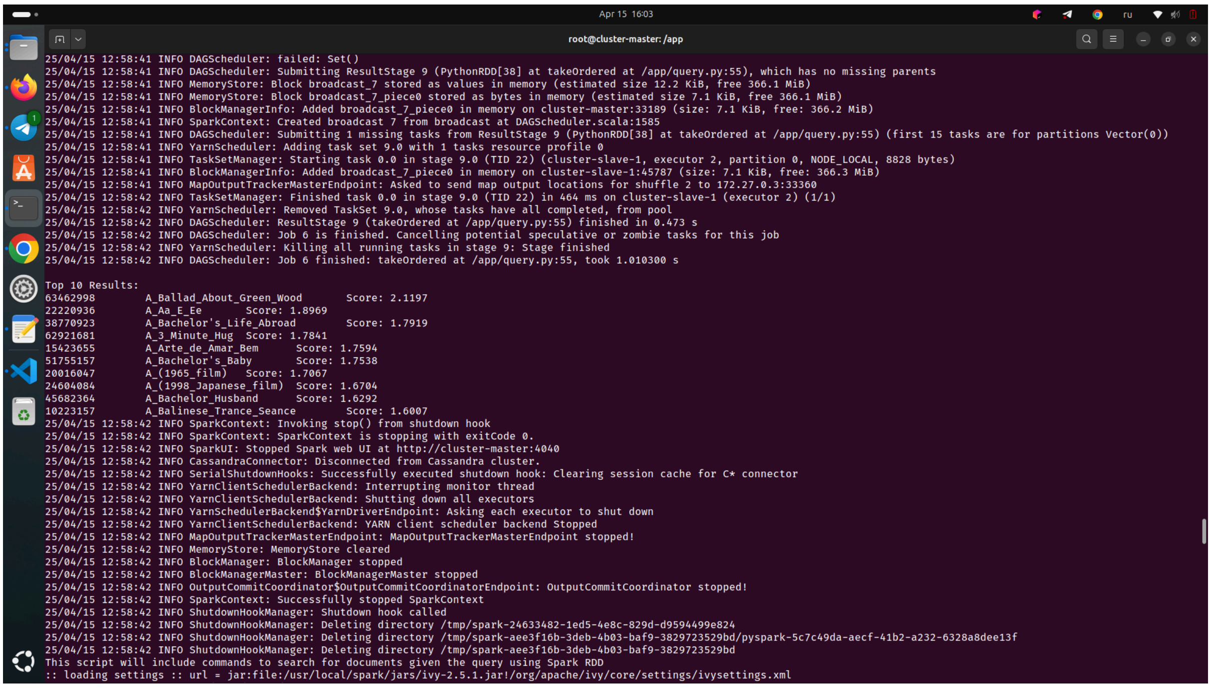
indexing



```
                Virtual memory (bytes) snapshot=7776387072
                Total committed heap usage (bytes)=999292928
                Peak Map Physical memory (bytes)=347643904
                Peak Map Virtual memory (bytes)=2591711232
                Peak Reduce Physical memory (bytes)=256290816
                Peak Reduce Virtual memory (bytes)=2593222656
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=380858
        File Output Format Counters
                Bytes Written=3027
2025-04-15 12:56:31,424 INFO streaming.StreamJob: Output directory: /tmp/doc_titles
Preview output:
0       41801556        1
0       39204695        1
0       41801556        1
0       15547032        3
0       70800868        1
0       5405379 1
0       15547032        3
0       41205399        1
0       15547032        4
0       1177256 1
cat: Unable to write to output stream.
Populating Cassandra index...
Cassandra keyspace and tables created.
Connecting to Cassandra...
Downloading HDFS file: /tmp/index1 to local file: index_output.txt...
Loading document titles from second pipeline...
Downloading HDFS file: /tmp/doc_titles to local file: titles.txt...
Data written to term_index table.
Vocabulary table updated.
Document statistics table updated.
Cassandra index insertion completed.
Indexing complete.
This script will include commands to search for documents given the query using Spark RDD
:: 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-2a09cde4-a11c-4a7f-83b7-609c963a1e8f;1.0
        confs: [default]
        found com.datastax.spark#spark-cassandra-connector_2.12;3.4.1 in central
        found com.datastax.spark#spark-cassandra-connector-driver_2.12;3.4.1 in central
```

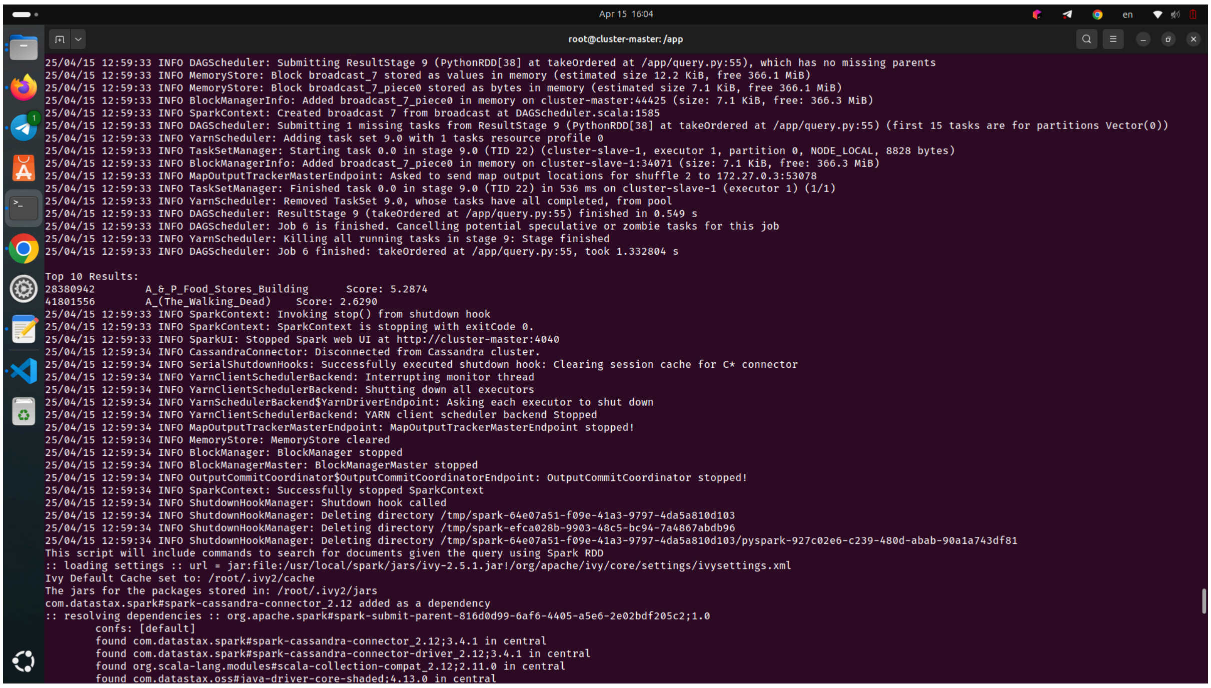
Алие
Ну это кринж

“Film” search



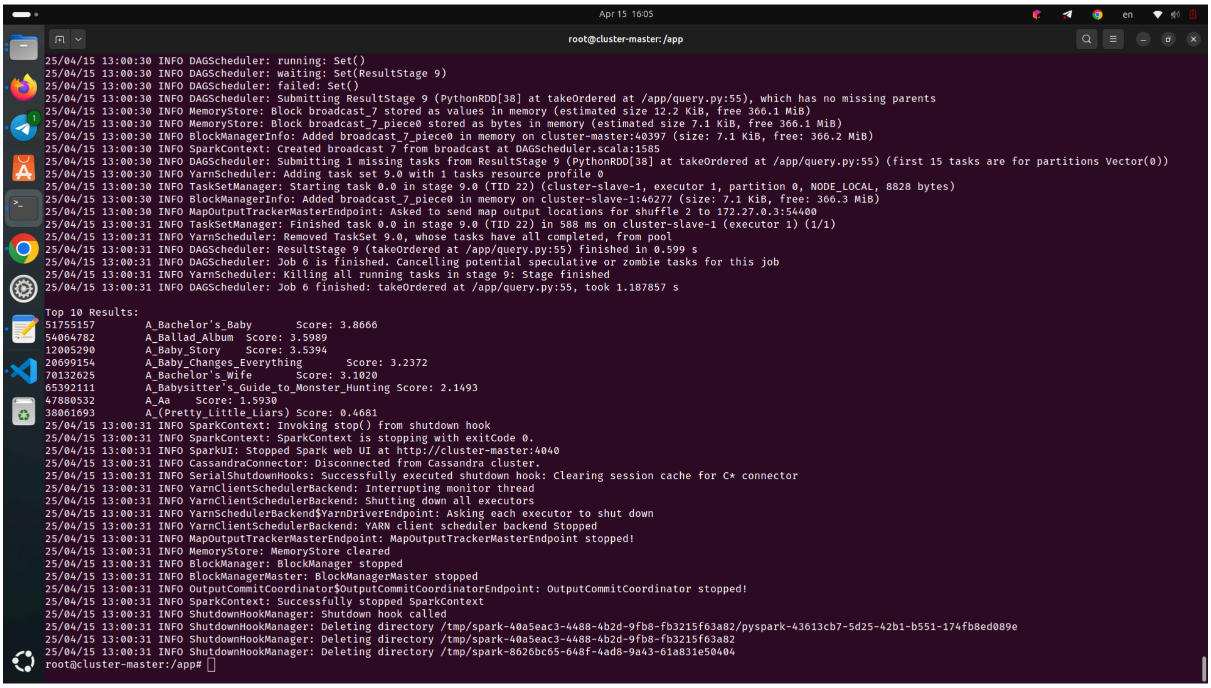
```
25/04/15 12:58:41 INFO DAGScheduler: failed: Set()
25/04/15 12:58:41 INFO DAGScheduler: Submitting ResultStage 9 (PythonRDD[38] at takeOrdered at /app/query.py:55), which has no missing parents
25/04/15 12:58:41 INFO MemoryStore: Block broadcast_7 stored as values in memory (estimated size 12.2 KiB, free 366.1 MiB)
25/04/15 12:58:41 INFO MemoryStore: Block broadcast_7_piece0 stored as bytes in memory (estimated size 7.1 KiB, free 366.1 MiB)
25/04/15 12:58:41 INFO BlockManagerInfo: Added broadcast_7_piece0 in memory on cluster-master:33189 (size: 7.1 KiB, free: 366.2 MiB)
25/04/15 12:58:41 INFO SparkContext: Created broadcast 7 from broadcast at DAGScheduler.scala:1585
25/04/15 12:58:41 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 9 (PythonRDD[38] at takeOrdered at /app/query.py:55) (first 15 tasks are for partitions Vector(0))
25/04/15 12:58:41 INFO YarnScheduler: Adding task set 9.0 with 1 tasks resource profile 0
25/04/15 12:58:41 INFO TaskSetManager: Starting task 0.0 in stage 9.0 (TID 22) (cluster-slave-1, executor 2, partition 0, NODE_LOCAL, 8828 bytes)
25/04/15 12:58:41 INFO BlockManagerInfo: Added broadcast_7_piece0 in memory on cluster-slave-1:45787 (size: 7.1 KiB, free: 366.3 MiB)
25/04/15 12:58:41 INFO MapOutputTrackerMasterEndpoint: Asked to send map output locations for shuffle 2 to 172.27.0.3:33360
25/04/15 12:58:42 INFO TaskSetManager: Finished task 0.0 in stage 9.0 (TID 22) in 464 ms on cluster-slave-1 (executor 2) (1/1)
25/04/15 12:58:42 INFO YarnScheduler: Removed TaskSet 9.0, whose tasks have all completed, from pool
25/04/15 12:58:42 INFO DAGScheduler: ResultStage 9 (takeOrdered at /app/query.py:55) finished in 0.473 s
25/04/15 12:58:42 INFO DAGScheduler: Job 6 is finished. Cancelling potential speculative or zombie tasks for this job
25/04/15 12:58:42 INFO YarnScheduler: Killing all running tasks in stage 9: Stage finished
25/04/15 12:58:42 INFO DAGScheduler: Job 6 finished: takeOrdered at /app/query.py:55, took 1.010300 s

Top 10 Results:
63462998        A_Ballad_About_Green_Wood       Score: 2.1197
22220936        A_Aa_E_Ee       Score: 1.8969
38770923        A_Bachelor's_Life_Abroad        Score: 1.7919
62921681        A_3_Minute_Hug  Score: 1.7841
15423655        A_Arte_de_Amar_Bem      Score: 1.7594
51755157        A_Bachelor's_Baby       Score: 1.7538
20016047        A_(1965_film)   Score: 1.7067
24604084        A_(1998_Japanese_film)  Score: 1.6704
45682364        A_Bachelor_Husband      Score: 1.6292
10223157        A_Balinese_Trance_Seance        Score: 1.6007
25/04/15 12:58:42 INFO SparkContext: Invoking stop() from shutdown hook
25/04/15 12:58:42 INFO SparkContext: SparkContext is stopping with exitCode 0.
25/04/15 12:58:42 INFO SparkUI: Stopped Spark web UI at http://cluster-master:4040
25/04/15 12:58:42 INFO CassandraConnector: Disconnected from Cassandra cluster.
25/04/15 12:58:42 INFO SerialShutdownHooks: Successfully executed shutdown hook: Clearing session cache for C* connector
25/04/15 12:58:42 INFO YarnClientSchedulerBackend: Interrupting monitor thread
25/04/15 12:58:42 INFO YarnClientSchedulerBackend: Shutting down all executors
25/04/15 12:58:42 INFO YarnSchedulerBackend$YarnDriverEndpoint: Asking each executor to shut down
25/04/15 12:58:42 INFO YarnClientSchedulerBackend: YARN client scheduler backend Stopped
25/04/15 12:58:42 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
25/04/15 12:58:42 INFO MemoryStore: MemoryStore cleared
25/04/15 12:58:42 INFO BlockManager: BlockManager stopped
25/04/15 12:58:42 INFO BlockManagerMaster: BlockManagerMaster stopped
25/04/15 12:58:42 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
25/04/15 12:58:42 INFO SparkContext: Successfully stopped SparkContext
25/04/15 12:58:42 INFO ShutdownHookManager: Shutdown hook called
25/04/15 12:58:42 INFO ShutdownHookManager: Deleting directory /tmp/spark-24633482-1ed5-4e8c-829d-d9594499e824
25/04/15 12:58:42 INFO ShutdownHookManager: Deleting directory /tmp/spark-aee3f16b-3deb-4b03-baf9-3829723529bd/pyspark-5c7c49da-aecf-41b2-a232-6328a8dee13f
25/04/15 12:58:42 INFO ShutdownHookManager: Deleting directory /tmp/spark-aee3f16b-3deb-4b03-baf9-3829723529bd
This script will include commands to search for documents given the query using Spark RDD
:: loading settings :: url = jar:file:/usr/local/spark/jars/ivy-2.5.1.jar!/org/apache/ivy/core/settings/ivysettings.xml
```

“Food” search



```
25/04/15 12:59:33 INFO DAGScheduler: Submitting ResultStage 9 (PythonRDD[38] at takeOrdered at /app/query.py:55), which has no missing parents
25/04/15 12:59:33 INFO MemoryStore: Block broadcast_7 stored as values in memory (estimated size 12.2 KiB, free 366.1 MiB)
25/04/15 12:59:33 INFO MemoryStore: Block broadcast_7_piece0 stored as bytes in memory (estimated size 7.1 KiB, free 366.1 MiB)
25/04/15 12:59:33 INFO BlockManagerInfo: Added broadcast_7_piece0 in memory on cluster-master:44425 (size: 7.1 KiB, free: 366.3 MiB)
25/04/15 12:59:33 INFO SparkContext: Created broadcast 7 from broadcast at DAGScheduler.scala:1585
25/04/15 12:59:33 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 9 (PythonRDD[38] at takeOrdered at /app/query.py:55) (first 15 tasks are for partitions Vector(0))
25/04/15 12:59:33 INFO YarnScheduler: Adding task set 9.0 with 1 tasks resource profile 0
25/04/15 12:59:33 INFO TaskSetManager: Starting task 0.0 in stage 9.0 (TID 22) (cluster-slave-1, executor 1, partition 0, NODE_LOCAL, 8828 bytes)
25/04/15 12:59:33 INFO BlockManagerInfo: Added broadcast_7_piece0 in memory on cluster-slave-1:34071 (size: 7.1 KiB, free: 366.3 MiB)
25/04/15 12:59:33 INFO MapOutputTrackerMasterEndpoint: Asked to send map output locations for shuffle 2 to 172.27.0.3:53078
25/04/15 12:59:33 INFO TaskSetManager: Finished task 0.0 in stage 9.0 (TID 22) in 536 ms on cluster-slave-1 (executor 1) (1/1)
25/04/15 12:59:33 INFO YarnScheduler: Removed TaskSet 9.0, whose tasks have all completed, from pool
25/04/15 12:59:33 INFO DAGScheduler: ResultStage 9 (takeOrdered at /app/query.py:55) finished in 0.549 s
25/04/15 12:59:33 INFO DAGScheduler: Job 6 is finished. Cancelling potential speculative or zombie tasks for this job
25/04/15 12:59:33 INFO YarnScheduler: Killing all running tasks in stage 9: Stage finished
25/04/15 12:59:33 INFO DAGScheduler: Job 6 finished: takeOrdered at /app/query.py:55, took 1.332804 s

Top 10 Results:
28380942        A_&_P_Food_Stores_Building      Score: 5.2874
41801556        A_(The_Walking_Dead)    Score: 2.6290
25/04/15 12:59:33 INFO SparkContext: Invoking stop() from shutdown hook
25/04/15 12:59:33 INFO SparkContext: SparkContext is stopping with exitCode 0.
25/04/15 12:59:33 INFO SparkUI: Stopped Spark web UI at http://cluster-master:4040
25/04/15 12:59:34 INFO CassandraConnector: Disconnected from Cassandra cluster.
25/04/15 12:59:34 INFO SerialShutdownHooks: Successfully executed shutdown hook: Clearing session cache for C* connector
25/04/15 12:59:34 INFO YarnClientSchedulerBackend: Interrupting monitor thread
25/04/15 12:59:34 INFO YarnClientSchedulerBackend: Shutting down all executors
25/04/15 12:59:34 INFO YarnSchedulerBackend$YarnDriverEndpoint: Asking each executor to shut down
25/04/15 12:59:34 INFO YarnClientSchedulerBackend: YARN client scheduler backend Stopped
25/04/15 12:59:34 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
25/04/15 12:59:34 INFO MemoryStore: MemoryStore cleared
25/04/15 12:59:34 INFO BlockManager: BlockManager stopped
25/04/15 12:59:34 INFO BlockManagerMaster: BlockManagerMaster stopped
25/04/15 12:59:34 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
25/04/15 12:59:34 INFO SparkContext: Successfully stopped SparkContext
25/04/15 12:59:34 INFO ShutdownHookManager: Shutdown hook called
25/04/15 12:59:34 INFO ShutdownHookManager: Deleting directory /tmp/spark-64e07a51-f09e-41a3-9797-4da5a810d103
25/04/15 12:59:34 INFO ShutdownHookManager: Deleting directory /tmp/spark-efca028b-9903-48c5-bc94-7a4867abdb96
25/04/15 12:59:34 INFO ShutdownHookManager: Deleting directory /tmp/spark-64e07a51-f09e-41a3-9797-4da5a810d103/pyspark-927c02e6-c239-480d-abab-90a1a743df81
This script will include commands to search for documents given the query using Spark RDD
:: 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-816d0d99-6af6-4405-a5e6-2e02bdf205c2;1.0
        confs: [default]
        found com.datastax.spark#spark-cassandra-connector_2.12;3.4.1 in central
        found com.datastax.spark#spark-cassandra-connector-driver_2.12;3.4.1 in central
        found org.scala-lang.modules#scala-collection-compat_2.12;2.11.0 in central
        found com.datastax.oss#java-driver-core-shaded;4.13.0 in central
```

“Baby” search



```
25/04/15 13:00:30 INFO DAGScheduler: running: Set()
25/04/15 13:00:30 INFO DAGScheduler: waiting: Set(ResultStage 9)
25/04/15 13:00:30 INFO DAGScheduler: failed: Set()
25/04/15 13:00:30 INFO DAGScheduler: Submitting ResultStage 9 (PythonRDD[38] at takeOrdered at /app/query.py:55), which has no missing parents
25/04/15 13:00:30 INFO MemoryStore: Block broadcast_7 stored as values in memory (estimated size 12.2 KiB, free 366.1 MiB)
25/04/15 13:00:30 INFO MemoryStore: Block broadcast_7_piece0 stored as bytes in memory (estimated size 7.1 KiB, free 366.1 MiB)
25/04/15 13:00:30 INFO BlockManagerInfo: Added broadcast_7_piece0 in memory on cluster-master:40397 (size: 7.1 KiB, free: 366.2 MiB)
25/04/15 13:00:30 INFO SparkContext: Created broadcast 7 from broadcast at DAGScheduler.scala:1585
25/04/15 13:00:30 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 9 (PythonRDD[38] at takeOrdered at /app/query.py:55) (first 15 tasks are for partitions Vector(0))
25/04/15 13:00:30 INFO YarnScheduler: Adding task set 9.0 with 1 tasks resource profile 0
25/04/15 13:00:30 INFO TaskSetManager: Starting task 0.0 in stage 9.0 (TID 22) (cluster-slave-1, executor 1, partition 0, NODE_LOCAL, 8828 bytes)
25/04/15 13:00:30 INFO BlockManagerInfo: Added broadcast_7_piece0 in memory on cluster-slave-1:46277 (size: 7.1 KiB, free: 366.3 MiB)
25/04/15 13:00:30 INFO MapOutputTrackerMasterEndpoint: Asked to send map output locations for shuffle 2 to 172.27.0.3:54400
25/04/15 13:00:31 INFO TaskSetManager: Finished task 0.0 in stage 9.0 (TID 22) in 588 ms on cluster-slave-1 (executor 1) (1/1)
25/04/15 13:00:31 INFO YarnScheduler: Removed TaskSet 9.0, whose tasks have all completed, from pool
25/04/15 13:00:31 INFO DAGScheduler: ResultStage 9 (takeOrdered at /app/query.py:55) finished in 0.599 s
25/04/15 13:00:31 INFO DAGScheduler: Job 6 is finished. Cancelling potential speculative or zombie tasks for this job
25/04/15 13:00:31 INFO YarnScheduler: Killing all running tasks in stage 9: Stage finished
25/04/15 13:00:31 INFO DAGScheduler: Job 6 finished: takeOrdered at /app/query.py:55, took 1.187857 s

Top 10 Results:
51755157        A_Bachelor's_Baby       Score: 3.8666
54064782        A_Ballad_Album  Score: 3.5989
12005290        A_Baby_Story    Score: 3.5394
20699154        A_Baby_Changes_Everything       Score: 3.2372
70132625        A_Bachelor's_Wife       Score: 3.1020
65392111        A_Babysitter's_Guide_to_Monster_Hunting Score: 2.1493
47880532        A_Aa    Score: 1.5930
38061693        A_(Pretty_Little_Liars) Score: 0.4681
25/04/15 13:00:31 INFO SparkContext: Invoking stop() from shutdown hook
25/04/15 13:00:31 INFO SparkContext: SparkContext is stopping with exitCode 0.
25/04/15 13:00:31 INFO SparkUI: Stopped Spark web UI at http://cluster-master:4040
25/04/15 13:00:31 INFO CassandraConnector: Disconnected from Cassandra cluster.
25/04/15 13:00:31 INFO SerialShutdownHooks: Successfully executed shutdown hook: Clearing session cache for C* connector
25/04/15 13:00:31 INFO YarnClientSchedulerBackend: Interrupting monitor thread
25/04/15 13:00:31 INFO YarnClientSchedulerBackend: Shutting down all executors
25/04/15 13:00:31 INFO YarnSchedulerBackend$YarnDriverEndpoint: Asking each executor to shut down
25/04/15 13:00:31 INFO YarnClientSchedulerBackend: YARN client scheduler backend Stopped
25/04/15 13:00:31 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
25/04/15 13:00:31 INFO MemoryStore: MemoryStore cleared
25/04/15 13:00:31 INFO BlockManager: BlockManager stopped
25/04/15 13:00:31 INFO BlockManagerMaster: BlockManagerMaster stopped
25/04/15 13:00:31 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
25/04/15 13:00:31 INFO SparkContext: Successfully stopped SparkContext
25/04/15 13:00:31 INFO ShutdownHookManager: Shutdown hook called
25/04/15 13:00:31 INFO ShutdownHookManager: Deleting directory /tmp/spark-40a5eac3-4488-4b2d-9fb8-fb3215f63a82/pyspark-43613cb7-5d25-42b1-b551-174fb8ed089e
25/04/15 13:00:31 INFO ShutdownHookManager: Deleting directory /tmp/spark-40a5eac3-4488-4b2d-9fb8-fb3215f63a82
25/04/15 13:00:31 INFO ShutdownHookManager: Deleting directory /tmp/spark-8626bc65-648f-4ad8-9a43-61a831e50404
root@cluster-master:/app# 
```

As you can notice, the search finds relevant items even if the search query is not in the title of the article, for some words the search may output empty results if there are no relevant items.