Introduction to Big Data: Assignment 2

Author: Anastasiia Shvets
Group: B22-DS-02

Methodology

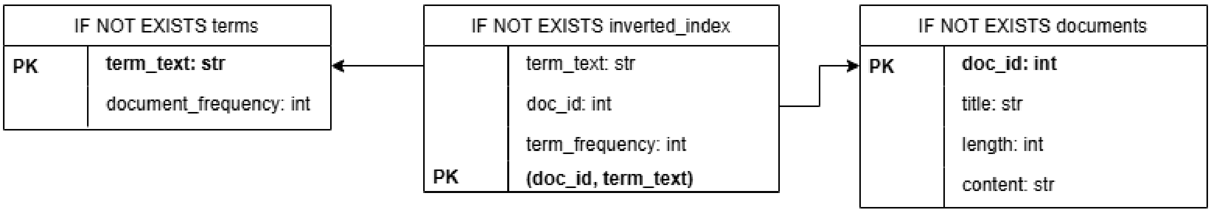
I followed the structure proposed in the github repository. However, I have made some design choices that are not present in the original solution. Here are the most import notes about my solution:

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:



IF NOT EXISTS terms	
PK	**term_text: str**
	document_frequency: int

IF NOT EXISTS inverted_index	
	term_text: str
	doc_id: int
	term_frequency: int
PK	**(doc_id, term_text)**

IF NOT EXISTS documents	
PK	**doc_id: int**
	title: str
	length: int
	content: str

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 is really simple. It gets a user query, separates it into tokens and runs the BM25 as described in the homework. When it shows 10 document titles with the best scores, as well as a corresponding score.

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

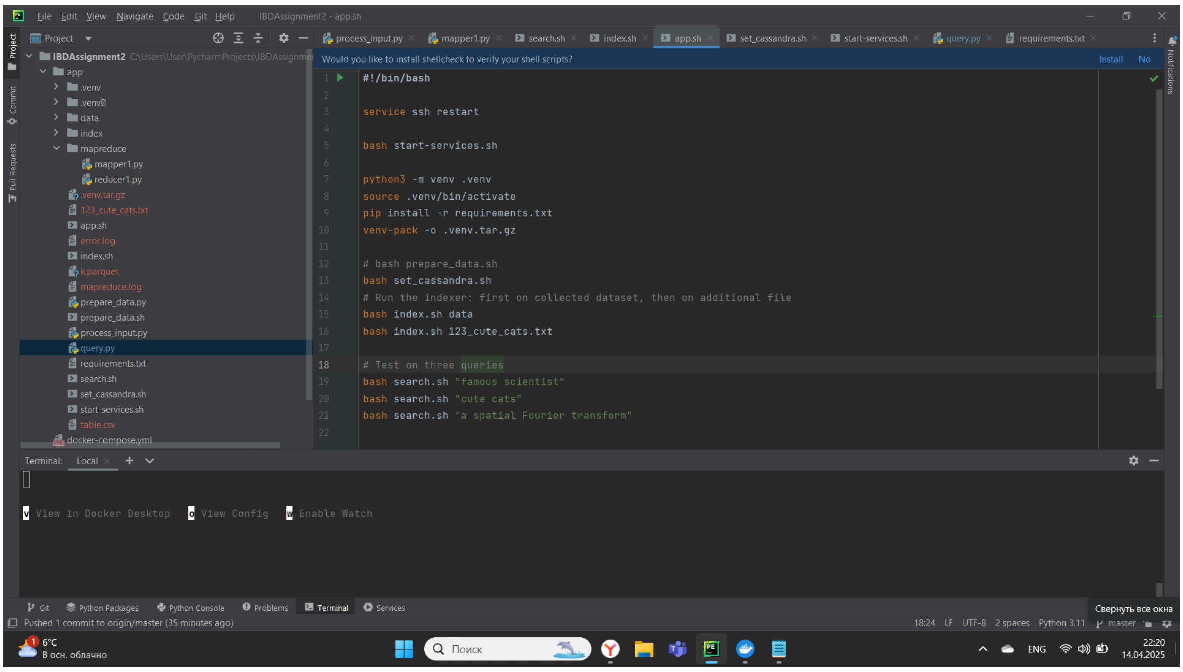
```
docker compose up
```

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:

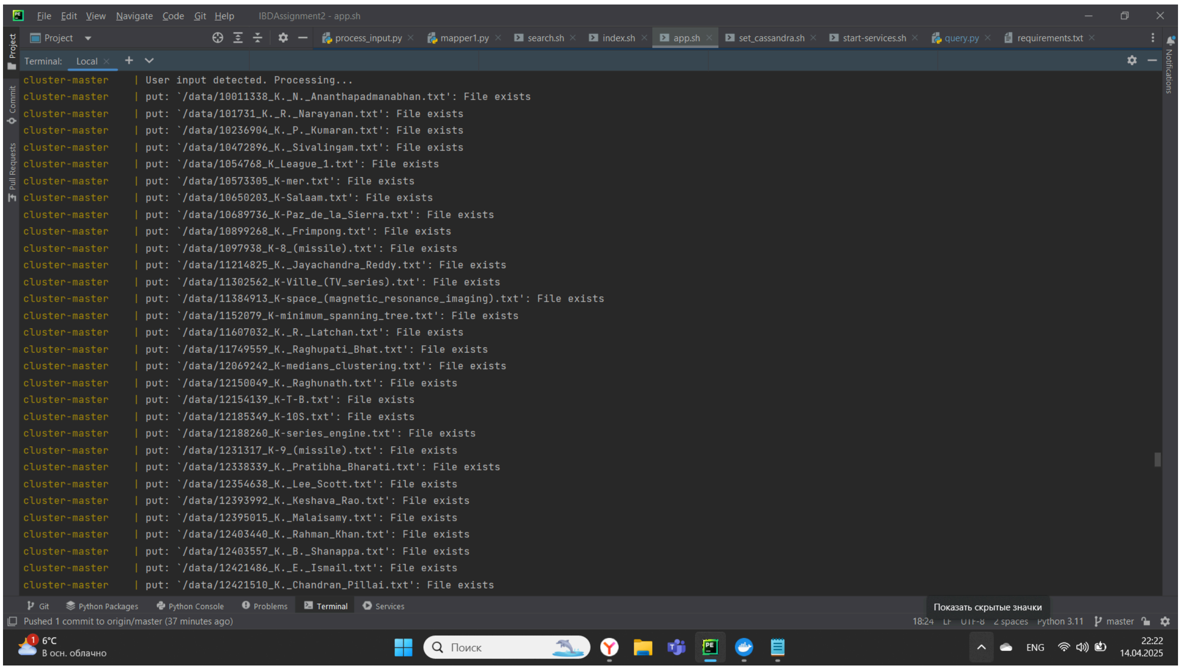


```
#!/bin/bash

service ssh restart

bash start-services.sh

python3 -m venv .venv
source .venv/bin/activate
pip install -r requirements.txt
venv-pack -o .venv.tar.gz

# bash prepare_data.sh
bash set_cassandra.sh
# Run the indexer: first on collected dataset, then on additional file
bash index.sh data
bash index.sh 123_cute_cats.txt

# Test on three queries
bash search.sh "famous scientist"
bash search.sh "cute cats"
bash search.sh "a spatial Fourier transform"
```

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

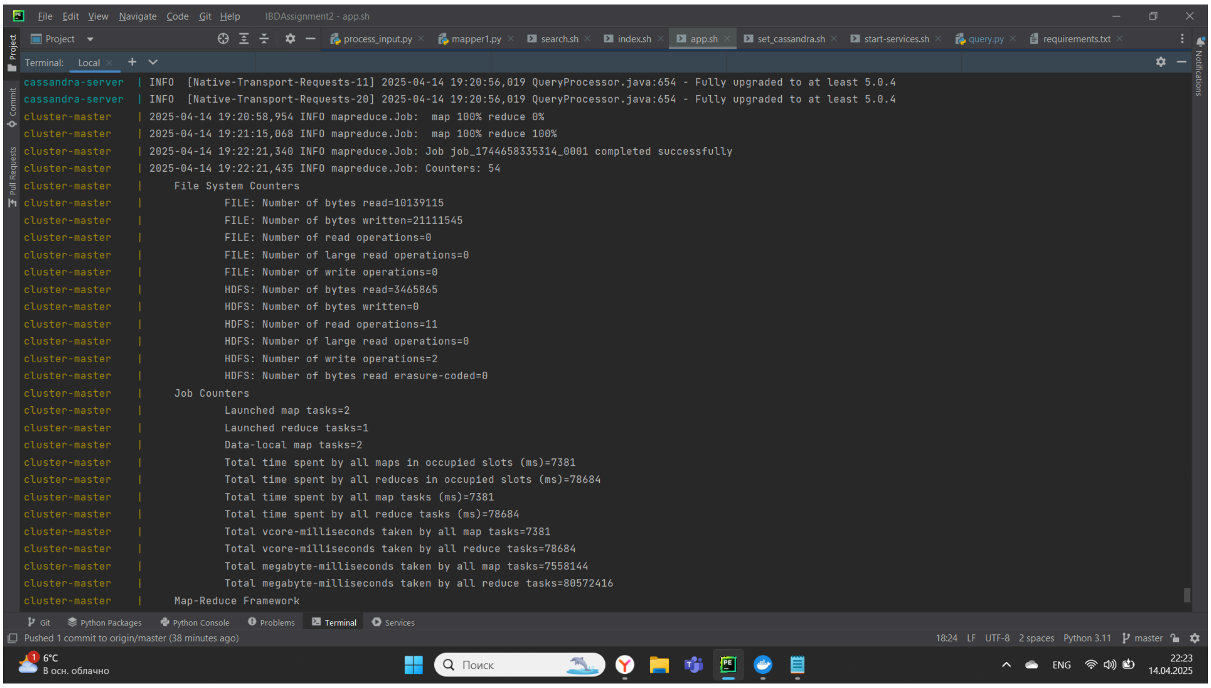
IBDAssignment2 - app.sh
Would you like to install shellcheck to verify your shell scripts?
Install
No
Pushed 1 commit to origin/master (35 minutes ago)

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:

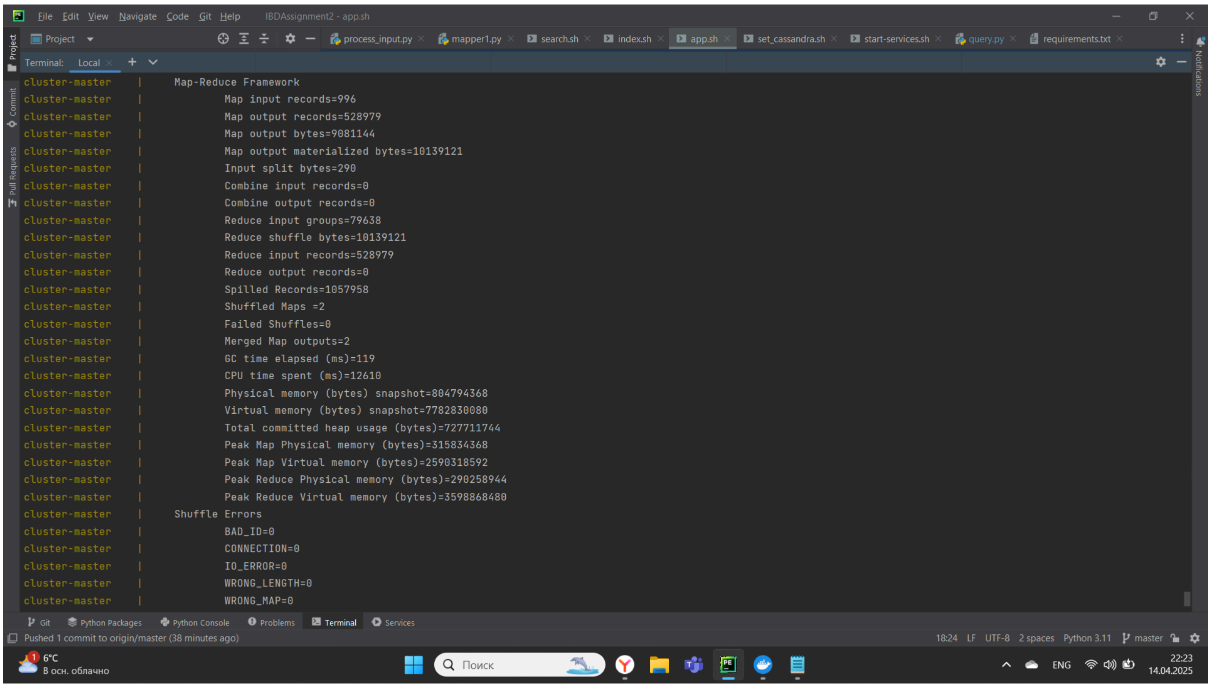


```
cluster-master    | User input detected. Processing...
cluster-master    | put: `/data/10011338_K._N._Ananthapadmanabhan.txt': File exists
cluster-master    | put: `/data/101731_K._R._Narayanan.txt': File exists
cluster-master    | put: `/data/10236904_K._P._Kumaran.txt': File exists
cluster-master    | put: `/data/10472896_K._Sivalingam.txt': File exists
cluster-master    | put: `/data/1054768_K_League_1.txt': File exists
cluster-master    | put: `/data/10573305_K-mer.txt': File exists
cluster-master    | put: `/data/10650203_K-Salaam.txt': File exists
cluster-master    | put: `/data/10689736_K-Paz_de_la_Sierra.txt': File exists
cluster-master    | put: `/data/10899268_K._Frimpong.txt': File exists
cluster-master    | put: `/data/1097938_K-8_(missile).txt': File exists
cluster-master    | put: `/data/11214825_K._Jayachandra_Reddy.txt': File exists
cluster-master    | put: `/data/11302562_K-Ville_(TV_series).txt': File exists
cluster-master    | put: `/data/11384913_K-space_(magnetic_resonance_imaging).txt': File exists
cluster-master    | put: `/data/1152079_K-minimum_spanning_tree.txt': File exists
cluster-master    | put: `/data/11607032_K._R._Latchan.txt': File exists
cluster-master    | put: `/data/11749559_K._Raghupati_Bhat.txt': File exists
cluster-master    | put: `/data/12069242_K-medians_clustering.txt': File exists
cluster-master    | put: `/data/12150049_K._Raghunath.txt': File exists
cluster-master    | put: `/data/12154139_K-T-B.txt': File exists
cluster-master    | put: `/data/12185349_K-10S.txt': File exists
cluster-master    | put: `/data/12188260_K-series_engine.txt': File exists
cluster-master    | put: `/data/1231317_K-9_(missile).txt': File exists
cluster-master    | put: `/data/12338339_K._Pratibha_Bharati.txt': File exists
cluster-master    | put: `/data/12354638_K._Lee_Scott.txt': File exists
cluster-master    | put: `/data/12393992_K._Keshava_Rao.txt': File exists
cluster-master    | put: `/data/12395015_K._Malaisamy.txt': File exists
cluster-master    | put: `/data/12403440_K._Rahman_Khan.txt': File exists
cluster-master    | put: `/data/12403557_K._B._Shanappa.txt': File exists
cluster-master    | put: `/data/12421486_K._E._Ismail.txt': File exists
cluster-master    | put: `/data/12421510_K._Chandran_Pillai.txt': File exists
```

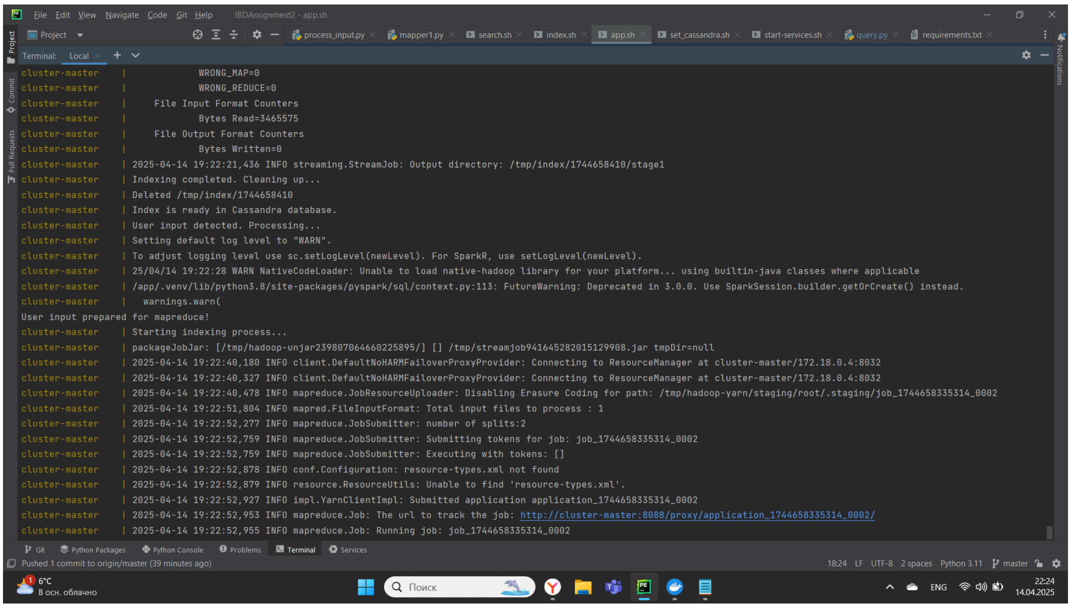
(As you can see, the files from the data folder have already been added to hadoop during data preparation.)



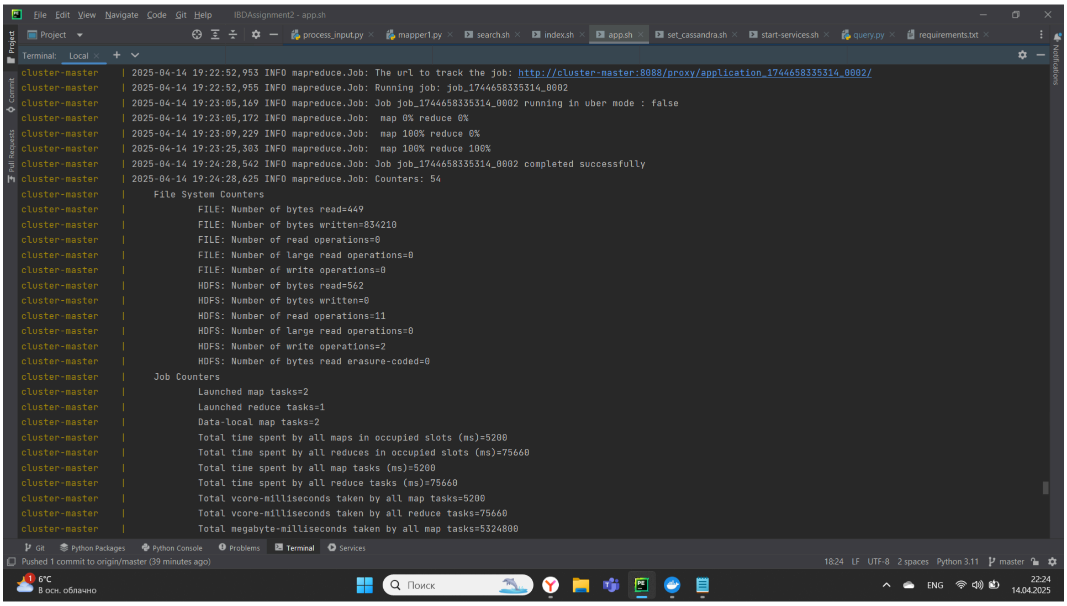
```
cassandra-server  | INFO  [Native-Transport-Requests-11] 2025-04-14 19:20:56,019 QueryProcessor.java:654 - Fully upgraded to at least 5.0.4
cassandra-server  | INFO  [Native-Transport-Requests-20] 2025-04-14 19:20:56,019 QueryProcessor.java:654 - Fully upgraded to at least 5.0.4
cluster-master    | 2025-04-14 19:20:58,954 INFO mapreduce.Job:  map 100% reduce 0%
cluster-master    | 2025-04-14 19:21:15,068 INFO mapreduce.Job:  map 100% reduce 100%
cluster-master    | 2025-04-14 19:22:21,340 INFO mapreduce.Job: Job job_1744658335314_0001 completed successfully
cluster-master    | 2025-04-14 19:22:21,435 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)=7381
cluster-master    |             Total time spent by all reduces in occupied slots (ms)=78684
cluster-master    |             Total time spent by all map tasks (ms)=7381
cluster-master    |             Total time spent by all reduce tasks (ms)=78684
cluster-master    |             Total vcore-milliseconds taken by all map tasks=7381
cluster-master    |             Total vcore-milliseconds taken by all reduce tasks=78684
cluster-master    |             Total megabyte-milliseconds taken by all map tasks=7558144
cluster-master    |             Total megabyte-milliseconds taken by all reduce tasks=80572416
cluster-master    |     Map-Reduce Framework
```



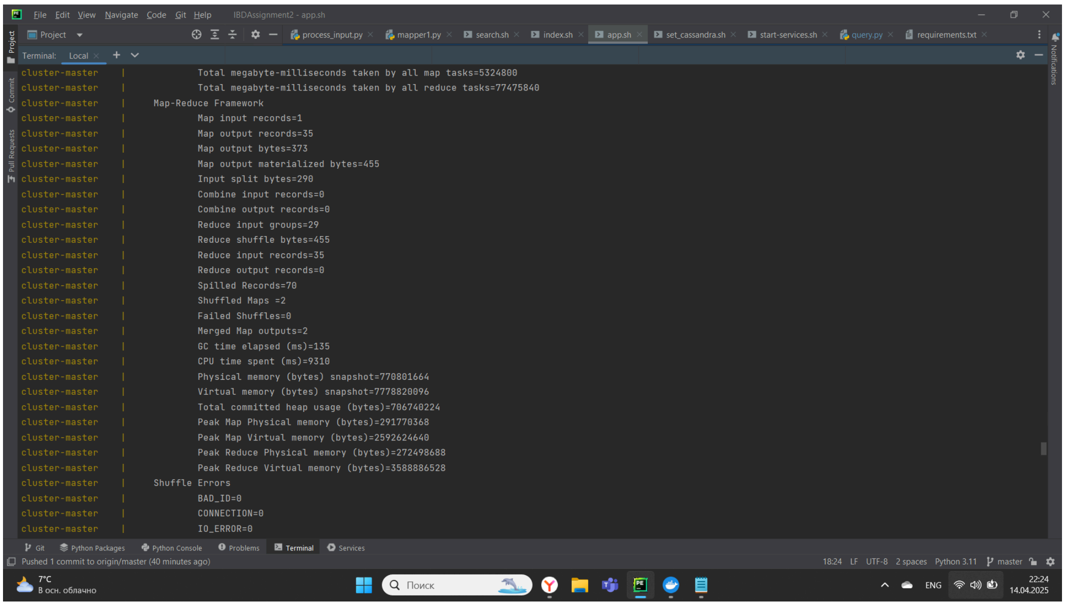
```
cluster-master    |     Map-Reduce Framework
cluster-master    |             Map input records=996
cluster-master    |             Map output records=528979
cluster-master    |             Map output bytes=9081144
cluster-master    |             Map output materialized bytes=10139121
cluster-master    |             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)=119
cluster-master    |             CPU time spent (ms)=12610
cluster-master    |             Physical memory (bytes) snapshot=804794368
cluster-master    |             Virtual memory (bytes) snapshot=7782830080
cluster-master    |             Total committed heap usage (bytes)=727711744
cluster-master    |             Peak Map Physical memory (bytes)=315834368
cluster-master    |             Peak Map Virtual memory (bytes)=2590318592
cluster-master    |             Peak Reduce Physical memory (bytes)=290258944
cluster-master    |             Peak Reduce Virtual memory (bytes)=3598868480
cluster-master    |     Shuffle Errors
cluster-master    |             BAD_ID=0
cluster-master    |             CONNECTION=0
cluster-master    |             IO_ERROR=0
cluster-master    |             WRONG_LENGTH=0
cluster-master    |             WRONG_MAP=0
```



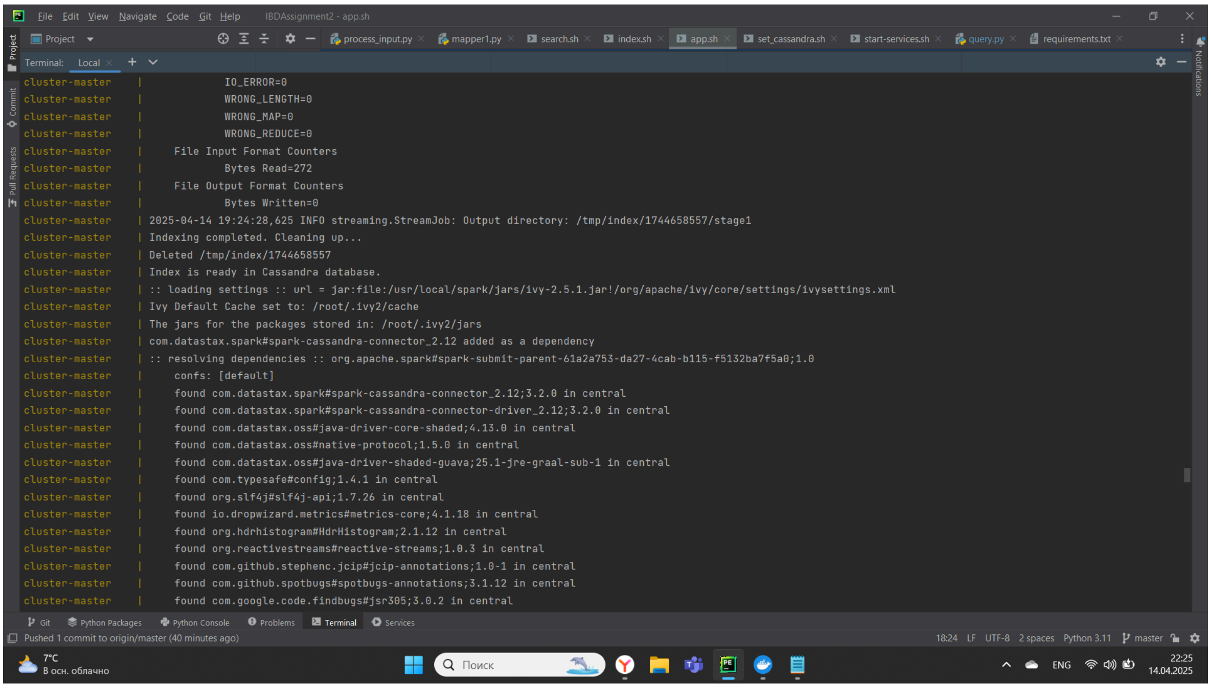
```
cluster-master    |             WRONG_MAP=0
cluster-master    |             WRONG_REDUCE=0
cluster-master    |     File Input Format Counters
cluster-master    |             Bytes Read=3465575
cluster-master    |     File Output Format Counters
cluster-master    |             Bytes Written=0
cluster-master    | 2025-04-14 19:22:21,436 INFO streaming.StreamJob: Output directory: /tmp/index/1744658410/stage1
cluster-master    | Indexing completed. Cleaning up...
cluster-master    | Deleted /tmp/index/1744658410
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    | 25/04/14 19:22:28 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-unjar2398070646602258955/] [] /tmp/streamjob941645282015129908.jar tmpDir=null
cluster-master    | 2025-04-14 19:22:40,180 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at cluster-master/172.18.0.4:8032
cluster-master    | 2025-04-14 19:22:40,327 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at cluster-master/172.18.0.4:8032
cluster-master    | 2025-04-14 19:22:40,478 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/root/.staging/job_1744658335314_0002
cluster-master    | 2025-04-14 19:22:51,804 INFO mapred.FileInputFormat: Total input files to process : 1
cluster-master    | 2025-04-14 19:22:52,277 INFO mapreduce.JobSubmitter: number of splits:2
cluster-master    | 2025-04-14 19:22:52,759 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1744658335314_0002
cluster-master    | 2025-04-14 19:22:52,759 INFO mapreduce.JobSubmitter: Executing with tokens: []
cluster-master    | 2025-04-14 19:22:52,878 INFO conf.Configuration: resource-types.xml not found
cluster-master    | 2025-04-14 19:22:52,879 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
cluster-master    | 2025-04-14 19:22:52,927 INFO impl.YarnClientImpl: Submitted application application_1744658335314_0002
cluster-master    | 2025-04-14 19:22:52,953 INFO mapreduce.Job: The url to track the job: http://cluster-master:8088/proxy/application_1744658335314_0002/
cluster-master    | 2025-04-14 19:22:52,955 INFO mapreduce.Job: Running job: job_1744658335314_0002
```



```
cluster-master    | 2025-04-14 19:22:52,953 INFO mapreduce.Job: The url to track the job: http://cluster-master:8088/proxy/application_1744658335314_0002/
cluster-master    | 2025-04-14 19:22:52,955 INFO mapreduce.Job: Running job: job_1744658335314_0002
cluster-master    | 2025-04-14 19:23:05,169 INFO mapreduce.Job: Job job_1744658335314_0002 running in uber mode : false
cluster-master    | 2025-04-14 19:23:05,172 INFO mapreduce.Job:  map 0% reduce 0%
cluster-master    | 2025-04-14 19:23:09,229 INFO mapreduce.Job:  map 100% reduce 0%
cluster-master    | 2025-04-14 19:23:25,303 INFO mapreduce.Job:  map 100% reduce 100%
cluster-master    | 2025-04-14 19:24:28,542 INFO mapreduce.Job: Job job_1744658335314_0002 completed successfully
cluster-master    | 2025-04-14 19:24:28,625 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=834210
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)=5200
cluster-master    |             Total time spent by all reduces in occupied slots (ms)=75660
cluster-master    |             Total time spent by all map tasks (ms)=5200
cluster-master    |             Total time spent by all reduce tasks (ms)=75660
cluster-master    |             Total vcore-milliseconds taken by all map tasks=5200
cluster-master    |             Total vcore-milliseconds taken by all reduce tasks=75660
cluster-master    |             Total megabyte-milliseconds taken by all map tasks=5324800
```



```
cluster-master    |             Total megabyte-milliseconds taken by all map tasks=5324800
cluster-master    |             Total megabyte-milliseconds taken by all reduce tasks=77475840
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)=135
cluster-master    |             CPU time spent (ms)=9310
cluster-master    |             Physical memory (bytes) snapshot=770801664
cluster-master    |             Virtual memory (bytes) snapshot=7778820096
cluster-master    |             Total committed heap usage (bytes)=706740224
cluster-master    |             Peak Map Physical memory (bytes)=291770368
cluster-master    |             Peak Map Virtual memory (bytes)=2592624640
cluster-master    |             Peak Reduce Physical memory (bytes)=272498688
cluster-master    |             Peak Reduce Virtual memory (bytes)=3588886528
cluster-master    |     Shuffle Errors
cluster-master    |             BAD_ID=0
cluster-master    |             CONNECTION=0
cluster-master    |             IO_ERROR=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=272
cluster-master    |     File Output Format Counters
cluster-master    |             Bytes Written=0
cluster-master    | 2025-04-14 19:24:28,625 INFO streaming.StreamJob: Output directory: /tmp/index/1744658557/stage1
cluster-master    | Indexing completed. Cleaning up...
cluster-master    | Deleted /tmp/index/1744658557
cluster-master    | Index is ready in Cassandra database.
cluster-master    | :: loading settings :: url = jar:file:/usr/local/spark/jars/ivy-2.5.1.jar!/org/apache/ivy/core/settings/ivysettings.xml
cluster-master    | Ivy Default Cache set to: /root/.ivy2/cache
cluster-master    | The jars for the packages stored in: /root/.ivy2/jars
cluster-master    | com.datastax.spark#spark-cassandra-connector_2.12 added as a dependency
cluster-master    | :: resolving dependencies :: org.apache.spark#spark-submit-parent-61a2a753-da27-4cab-b115-f5132ba7f5a0;1.0
cluster-master    |     confs: [default]
cluster-master    |     found com.datastax.spark#spark-cassandra-connector_2.12;3.2.0 in central
cluster-master    |     found com.datastax.spark#spark-cassandra-connector-driver_2.12;3.2.0 in central
cluster-master    |     found com.datastax.oss#java-driver-core-shaded;4.13.0 in central
cluster-master    |     found com.datastax.oss#native-protocol;1.5.0 in central
cluster-master    |     found com.datastax.oss#java-driver-shaded-guava;25.1-jre-graal-sub-1 in central
cluster-master    |     found com.typesafe#config;1.4.1 in central
cluster-master    |     found org.slf4j#slf4j-api;1.7.26 in central
cluster-master    |     found io.dropwizard.metrics#metrics-core;4.1.18 in central
cluster-master    |     found org.hdrhistogram#HdrHistogram;2.1.12 in central
cluster-master    |     found org.reactivestreams#reactive-streams;1.0.3 in central
cluster-master    |     found com.github.stephenc.jcip#jcip-annotations;1.0-1 in central
cluster-master    |     found com.github.spotbugs#spotbugs-annotations;3.1.12 in central
cluster-master    |     found com.google.code.findbugs#jsr305;3.0.2 in central
```

As you can see, both files were successfully indexed!

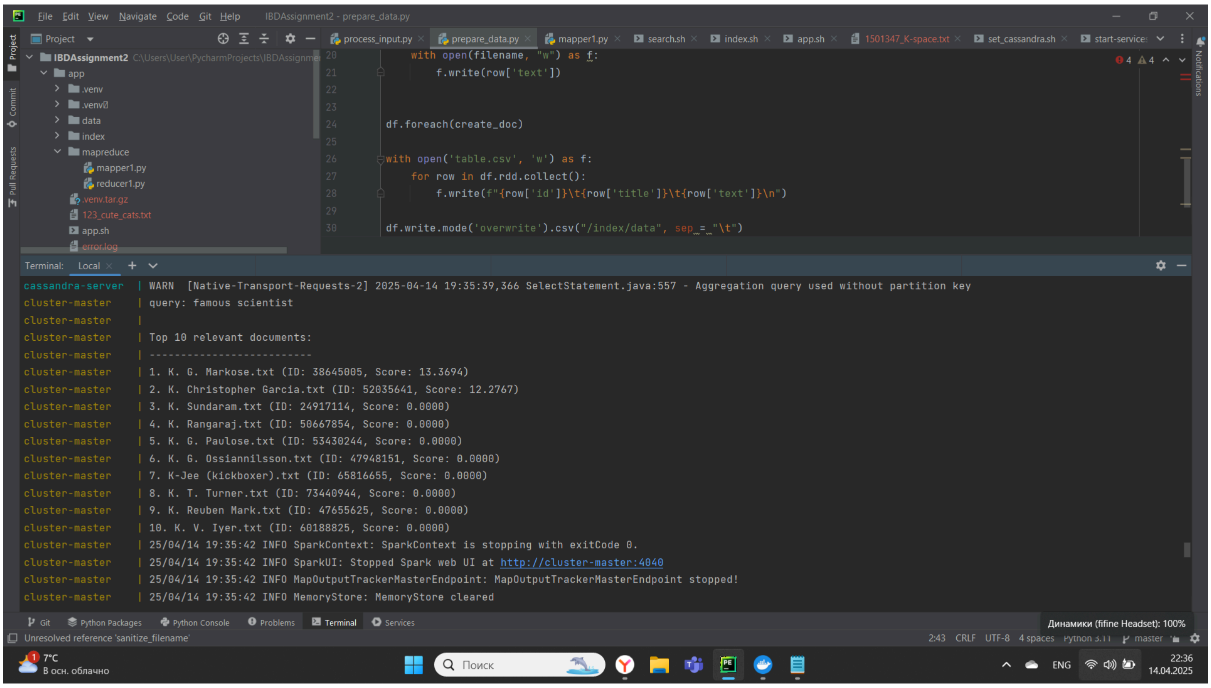
Now for the ranker part. I have created three test queries:

"famous scientist"

"cute cats"

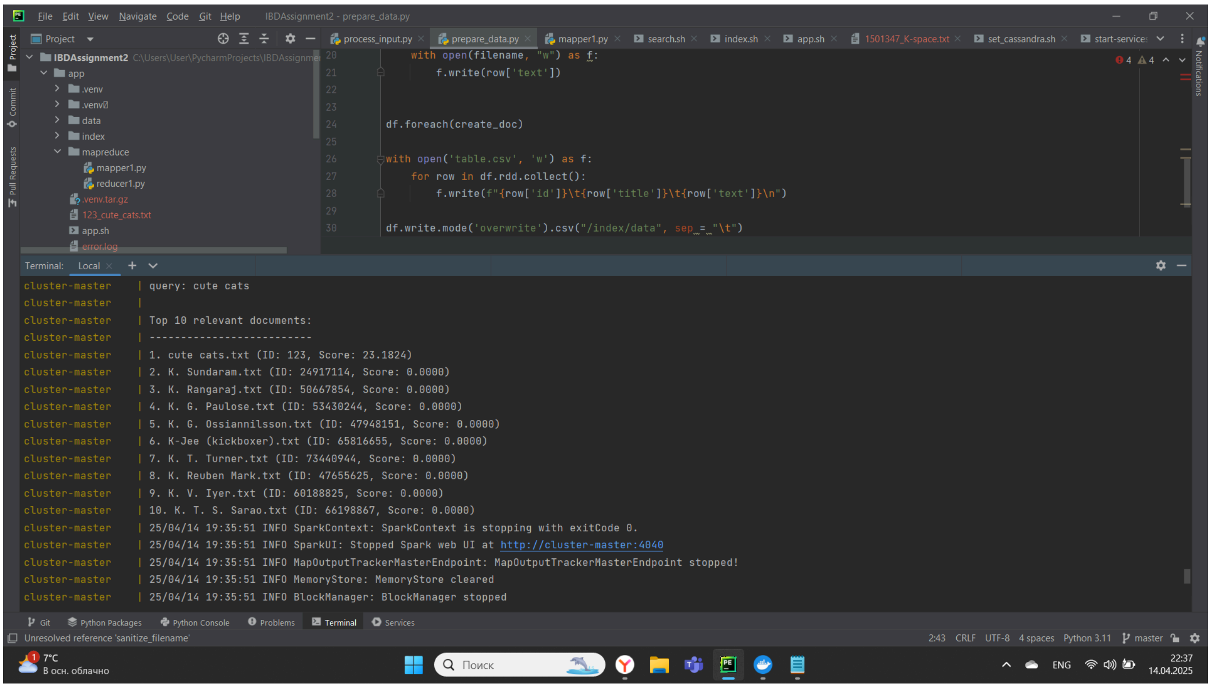
"a spatial Fourier transform"

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:


IBDAssignment2 - prepare_data.py
Terminal: Local
Unresolved reference 'sanitize_filename'

```python
        with open(filename, "w") as f:
            f.write(row['text'])


df.foreach(create_doc)

with open('table.csv', 'w') as f:
    for row in df.rdd.collect():
        f.write(f"{row['id']}\t{row['title']}\t{row['text']}\n")

df.write.mode('overwrite').csv("/index/data", sep = "\t")
```

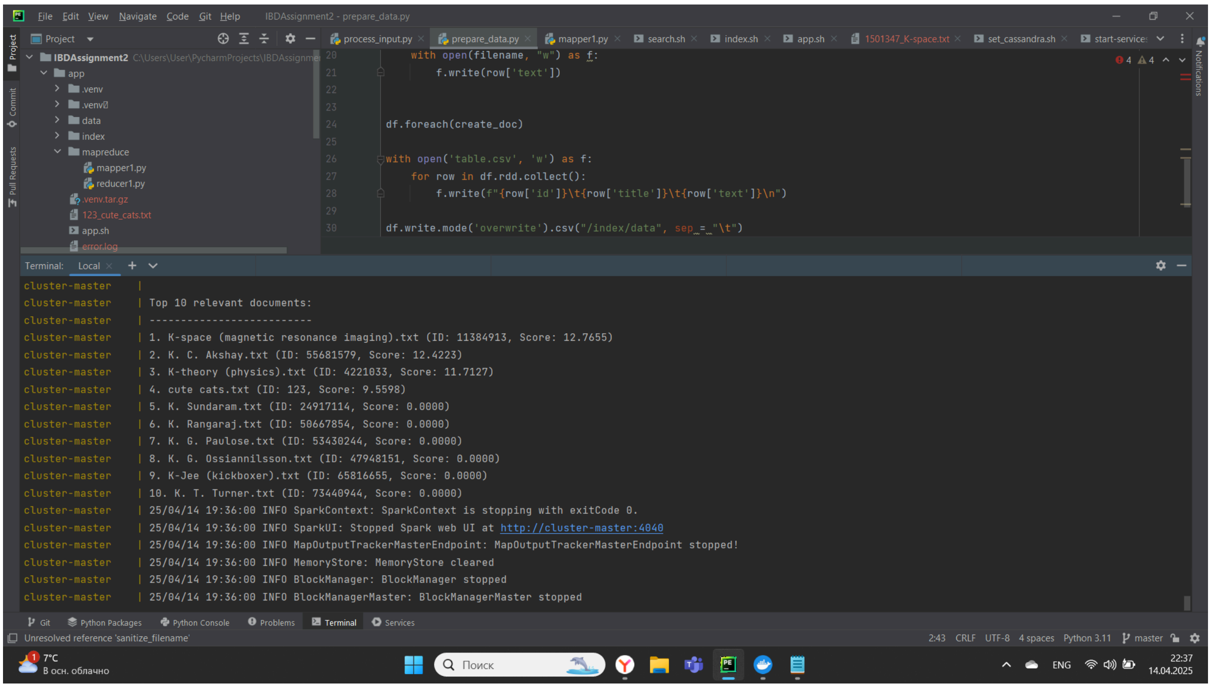
```
cassandra-server  | WARN  [Native-Transport-Requests-2] 2025-04-14 19:35:39,366 SelectStatement.java:557 - Aggregation query used without partition key
cluster-master    | query: famous scientist
cluster-master    |
cluster-master    | Top 10 relevant documents:
cluster-master    | -------------------------
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    | 25/04/14 19:35:42 INFO SparkContext: SparkContext is stopping with exitCode 0.
cluster-master    | 25/04/14 19:35:42 INFO SparkUI: Stopped Spark web UI at http://cluster-master:4040
cluster-master    | 25/04/14 19:35:42 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
cluster-master    | 25/04/14 19:35:42 INFO MemoryStore: MemoryStore cleared
```



```
    with open(filename, "w") as f:
        f.write(row['text'])


df.foreach(create_doc)

with open('table.csv', 'w') as f:
    for row in df.rdd.collect():
        f.write(f"{row['id']}\t{row['title']}\t{row['text']}\n")

df.write.mode('overwrite').csv("/index/data", sep = "\t")
```

```
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    | 25/04/14 19:35:51 INFO SparkContext: SparkContext is stopping with exitCode 0.
cluster-master    | 25/04/14 19:35:51 INFO SparkUI: Stopped Spark web UI at http://cluster-master:4040
cluster-master    | 25/04/14 19:35:51 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
cluster-master    | 25/04/14 19:35:51 INFO MemoryStore: MemoryStore cleared
cluster-master    | 25/04/14 19:35:51 INFO BlockManager: BlockManager stopped
```

Unresolved reference 'sanitize_filename'



```
    with open(filename, "w") as f:
        f.write(row['text'])


df.foreach(create_doc)

with open('table.csv', 'w') as f:
    for row in df.rdd.collect():
        f.write(f"{row['id']}\t{row['title']}\t{row['text']}\n")

df.write.mode('overwrite').csv("/index/data", sep = "\t")
```

```
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    | 25/04/14 19:36:00 INFO SparkContext: SparkContext is stopping with exitCode 0.
cluster-master    | 25/04/14 19:36:00 INFO SparkUI: Stopped Spark web UI at http://cluster-master:4040
cluster-master    | 25/04/14 19:36:00 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
cluster-master    | 25/04/14 19:36:00 INFO MemoryStore: MemoryStore cleared
cluster-master    | 25/04/14 19:36:00 INFO BlockManager: BlockManager stopped
cluster-master    | 25/04/14 19:36:00 INFO BlockManagerMaster: BlockManagerMaster stopped
```

Unresolved reference 'sanitize_filename'

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.