

Ivan Belstov i.beltsov@innopolis.university

Methodology

Data Preparation

There are 3 files responsible for data preparation: `prepare_data.sh`, `prepare_data.py` and `prepare_index_data.py`

`prepare_data.sh` transfers a.parquet from local to hdfs, starts `prepare_data.py`, copies newly made data folder to hdfs and starts `prepare_index_data.py`

`prepare_data.py` is unchanged from initial file

`prepare_index_data.py` reads files from hdfs data folder and creates a new partition using rdd

Indexer tasks

`index.sh` – Pipeline Runner Script that orchestrates the full indexing pipeline.

Actions:

- Accepts input path (local or HDFS) and uploads local files if needed
- Runs first MapReduce job to compute raw term frequencies
- Runs second MapReduce job to generate document stats and build the inverted index
- Triggers the Python script (`app.py`) to load results into Cassandra

MapReduce Stage 1

`mapper1.py` – Tokenizes document text, emits one entry per word occurrence.

Takes tab-separated lines with `doc_id`, `title`, `text` as input.

Returns lines with `doc_id`, `term`, and `title`.

reducer1.py – Sums word counts per document to get term frequency.

Takes sorted output from mapper1.py.

Returns aggregated doc_id, term, tf, title.

MapReduce Stage 2

mapper2.py – Computes total document length and prepares structured input for indexing.

Takes output from reducer1 and returns:

DOCLEN lines with total doc length and title.

TERM lines with individual term frequencies per document.

reducer2.py – Groups all documents per term, computes document frequency (df) per term and finalizes all index entries.

Takes output from mapper2.

Output:

VOCAB entries: term → document frequency (df).

INDEX entries: term → document ID, term frequency (tf).

DOCLEN entries: document ID → length, title.

app.py – Loads structured index output into Cassandra tables

Functions:

fetch_hdfs_output(): Reads reducer2 output from HDFS.

connect_cassandra(): Connects to Cassandra and ensures keyspace exists.

create_tables(): Creates tables for vocabulary, inverted_index, and documents.

parse_and_insert(): Parses output and performs batch inserts into Cassandra.

Stored Tables:

vocabulary(term, df) is used to calculate BM25 IDF component

inverted_index(term, doc_id, tf) is used to calculate BM25 TF component

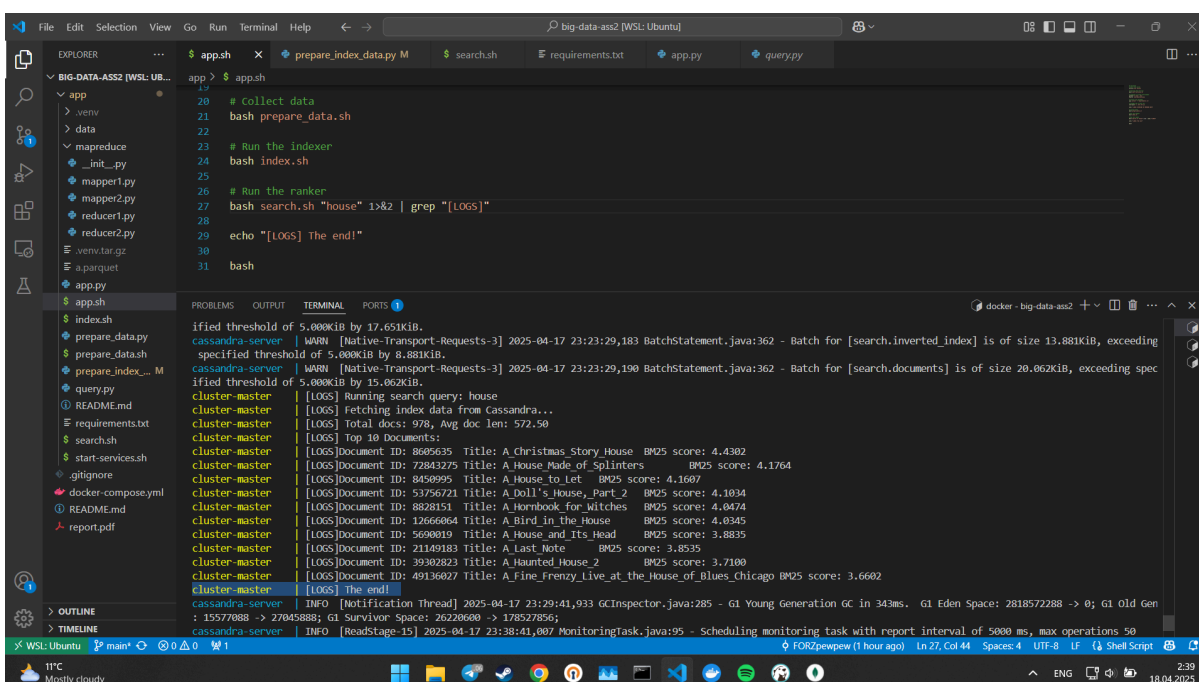
documents(doc_id, length, title): **length** is needed for normalization, **title** and **doc_id** are used as an output for search

Search tasks

1. Parsing: taking input string from search.sh and splitting it to lowercase standalone words.
2. Preparing RDD from inverted index: `index_rdd = sc.parallelize(index)`
3. Filtering: `filtered = index_rdd.filter(lambda x: x[0] in query_terms)`
4. Calculating BM25 Score for each matching document using function `score()`
5. Aggregating scores by document: `doc_scores = scored.reduceByKey(lambda a, b: a + b)`
6. Getting top 10 results: `top10 = doc_scores.takeOrdered(10, key=lambda x: -x[1])`

Demonstration

All you need to do for the full demo of the assignment is to put *.parquet in /app folder(modify prepare_data.sh if the file in question is not named a.parquet) and run docker compose up, which will build 3 containers (cluster-master, cluster-slave, and cassandra). After app.sh is finished running (demonstrated by top 10 documents for query “house” and message “[LOGS] The end!” on screenshot), in a separate terminal you can write “**docker exec -it cluster-master bash**” to connect to cluster-master’s bash console for search queries using command “**bash search.sh "your query"**” or accessing hdfs using “**hdfs df**”. Alternatively “**docker exec -it cassandra-server cqlsh**” can be used to connect to cqlsh of cassandra (it uses “search” keyspace, which has 3 tables: documents, vocabulary and inverted index).



```
app.sh
19 # collect data
20 bash prepare_data.sh
21
22 # Run the indexer
23 bash index.sh
24
25 # Run the ranker
26 bash search.sh "house" 1x&2 | grep "[LOGS]"
27
28 echo "[LOGS] The end!"
29
30
31 bash
```

```
ifed threshold of 5.000KiB by 17.651KiB.
cassandra-server WARN [Native-Transport-Requests-3] 2025-04-17 23:23:29,183 BatchStatement.java:362 - Batch for [search.inverted_index] is of size 13.881KiB, exceeding specified threshold of 5.000KiB by 8.881KiB.
cassandra-server WARN [Native-Transport-Requests-3] 2025-04-17 23:23:29,190 BatchStatement.java:362 - Batch for [search.documents] is of size 20.062KiB, exceeding specified threshold of 5.000KiB by 15.062KiB.
cluster-master [LOGS] Running search query: house
cluster-master [LOGS] Fetching index data from Cassandra...
cluster-master [LOGS] Total docs: 978, Avg doc len: 572.50
cluster-master [LOGS] Top 10 documents:
cluster-master [LOGS] Document ID: 8695635 Title: A Christmas Story House BM25 score: 4.4302
cluster-master [LOGS] Document ID: 72843275 Title: A House Made of Splinters BM25 score: 4.1764
cluster-master [LOGS] Document ID: 8459995 Title: A House to Let BM25 score: 4.1607
cluster-master [LOGS] Document ID: 53756721 Title: A Doll's House, Part 2 BM25 score: 4.1034
cluster-master [LOGS] Document ID: 8828151 Title: A Hornbook for Witches BM25 score: 4.0474
cluster-master [LOGS] Document ID: 12666064 Title: A Bird in the House BM25 score: 4.0345
cluster-master [LOGS] Document ID: 5690019 Title: A House and Its Head BM25 score: 3.8835
cluster-master [LOGS] Document ID: 21140183 Title: A Last Note BM25 score: 3.8535
cluster-master [LOGS] Document ID: 39302823 Title: A Haunted House 2 BM25 score: 3.7100
cluster-master [LOGS] Document ID: 49136027 Title: A Fine Frenzy Live at the House of Blues Chicago BM25 score: 3.6602
cluster-master [LOGS] The end!
cassandra-server INFO [Notification Thread] 2025-04-17 23:29:41,933 GCInspector.java:285 - G1 Young Generation GC in 343ms. G1 Eden Space: 2818572288 -> 0; G1 Old Gen : 15577088 -> 27045888; G1 Survivor Space: 26220600 -> 178527856;
cassandra-server INFO [ReadStage-15] 2025-04-17 23:38:41,007 MonitoringTask.java:95 - Scheduling monitoring task with report interval of 5000 ms, max operations 50
```

Here is a demonstration of how to copy part-00000 from hdfs to local container storage to use as a file for local indexing:

```
root@cluster-master:/app
pewpew@pewpew:/mnt/c/Users/gachi$ docker exec -it cluster-master bash
root@cluster-master:/app# ls
README.md  app.py  data  index.sh  prepare_data.sh  query.py  search.sh
a.parquet  app.sh  index.sh  prepare_data.py  prepare_index_data.py  requirements.txt  start-services.sh
root@cluster-master:/app# hdfs dfs -ls /index/data
Found 2 items
-rw-r--r-- 1 root supergroup 0 2025-04-17 23:22 /index/data/_SUCCESS
-rw-r--r-- 1 root supergroup 3428372 2025-04-17 23:22 /index/data/part-00000
root@cluster-master:/app# hdfs dfs -get /index/data/part-00000 /app
root@cluster-master:/app# ls
README.md  app.py  data  index.sh  part-00000  prepare_data.sh  prepare_index_data.py  query.py  requirements.txt  start-services.sh  search.sh
root@cluster-master:/app#
```

And by running (or modifying the corresponding lines in app.sh) ***“bash index.sh part-00000”***, you can index the local file, but the file should be a result of prepare_data.sh. A script that processes a.parquet (you can change the name of the processed file in bash script yourself if needed and run it, resulting part-00000 file will always be at hdfs’ index/data folder).

```
root@cluster-master:/app
a.parquet  app.sh  index.sh  part-00000  prepare_data.sh  query.py  search.sh
root@cluster-master:/app# bash index.sh part-00000
[LOGS]: Copying local file to HDFS...
Deleted /tmp/index/step1
Deleted /tmp/index/step2
2025-04-18 00:12:08,279 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
packageJobJar: [mapreduce/mapper1.py, mapreduce/reducer1.py, /tmp/hadoop-unjar7600348615261405526/] [] /tmp/streamjob302328634418309930
4.jar tmpDir=null
2025-04-18 00:12:11,553 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at cluster-master/172.18.0.4:8032
2025-04-18 00:12:11,725 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at cluster-master/172.18.0.4:8032
2025-04-18 00:12:11,871 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/root/.staging/j
ob_1744931561078_0011
2025-04-18 00:12:12,474 INFO mapred.FileInputFormat: Total input files to process : 1
2025-04-18 00:12:12,919 INFO mapreduce.JobSubmitter: number of splits:2
2025-04-18 00:12:13,431 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1744931561078_0011
2025-04-18 00:12:13,431 INFO mapreduce.JobSubmitter: Executing with tokens: []
2025-04-18 00:12:13,571 INFO conf.Configuration: resource-types.xml not found
2025-04-18 00:12:13,571 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2025-04-18 00:12:13,647 INFO impl.YarnClientImpl: Submitted application application_1744931561078_0011
2025-04-18 00:12:13,684 INFO mapreduce.Job: The url to track the job: http://cluster-master:8088/proxy/application_1744931561078_0011/
2025-04-18 00:12:13,686 INFO mapreduce.Job: Running job: job_1744931561078_0011
2025-04-18 00:12:18,773 INFO mapreduce.Job: Job job_1744931561078_0011 running in uber mode : false
2025-04-18 00:12:18,774 INFO mapreduce.Job: map 0% reduce 0%
2025-04-18 00:12:22,814 INFO mapreduce.Job: map 100% reduce 0%
2025-04-18 00:12:27,839 INFO mapreduce.Job: map 100% reduce 100%
2025-04-18 00:12:27,846 INFO mapreduce.Job: Job job_1744931561078_0011 completed successfully
2025-04-18 00:12:27,921 INFO mapreduce.Job: Counters: 54
File System Counters
FILE: Number of bytes read=23446595
FILE: Number of bytes written=47722557
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
```

```
root@cluster-master:/app
Shuffled Maps =2
Failed Shuffles=0
Merged Map outputs=2
GC time elapsed (ms)=84
CPU time spent (ms)=3320
Physical memory (bytes) snapshot=812937216
Virtual memory (bytes) snapshot=7710928896
Total committed heap usage (bytes)=710934528
Peak Map Physical memory (bytes)=323239936
Peak Map Virtual memory (bytes)=2570698752
Peak Reduce Physical memory (bytes)=233410560
Peak Reduce Virtual memory (bytes)=2572029952

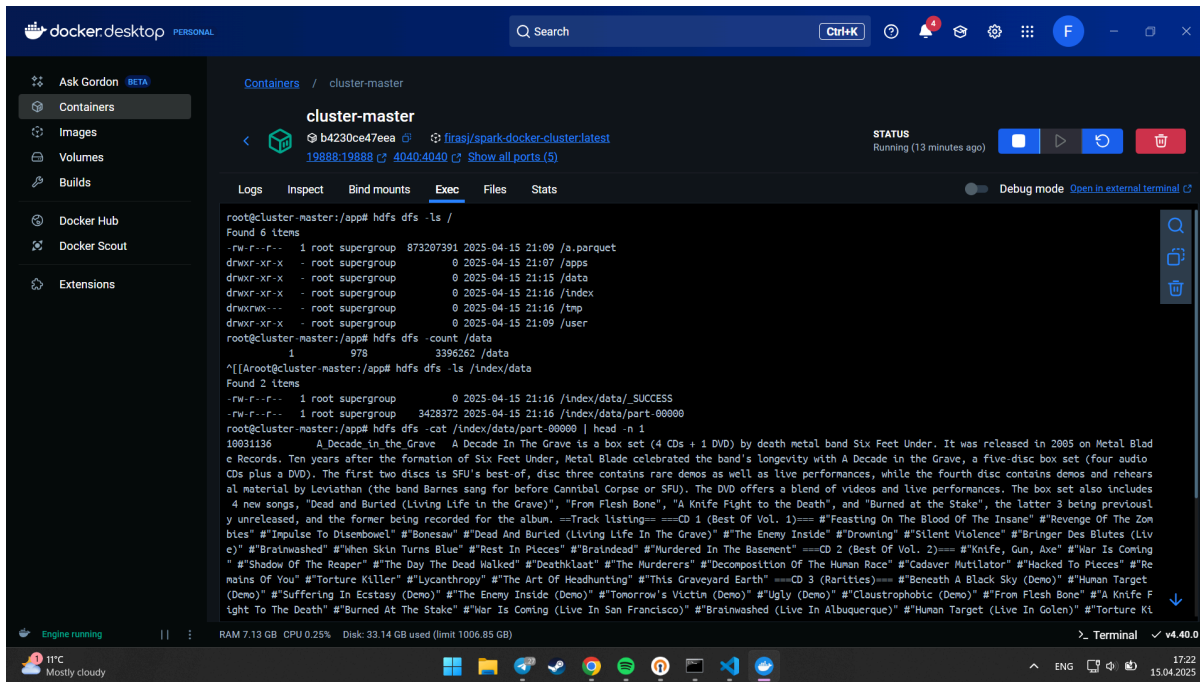
Shuffle Errors
BAD_ID=0
CONNECTION=0
IO_ERROR=0
WRONG_LENGTH=0
WRONG_MAP=0
WRONG_REDUCE=0

File Input Format Counters
Bytes Read=9974954
File Output Format Counters
Bytes Written=6395816
2025-04-18 00:12:48,685 INFO streaming.StreamJob: Output directory: /tmp/index/step2
root@cluster-master:/app# hdfs dfs -ls /tmp/index
Found 2 items
drwxr-xr-x - root supergroup          0 2025-04-18 00:12 /tmp/index/step1
drwxr-xr-x - root supergroup          0 2025-04-18 00:12 /tmp/index/step2
root@cluster-master:/app# hdfs dfs -ls /tmp/index/step2
Found 2 items
-rw-r--r-- 1 root supergroup          0 2025-04-18 00:12 /tmp/index/step2/_SUCCESS
-rw-r--r-- 1 root supergroup 6395816 2025-04-18 00:12 /tmp/index/step2/part-000000
root@cluster-master:/app#
```

Warn messages from cassandra due to batch size is higher than threshold by a couple of KiB(but files are still added to cassandra)

```
File Edit Selection View Go Run Terminal Help
big-data-ass2 [WSL: Ubuntu]
EXPLORER
  app
  .env
  data
  mapreduce
  .venv.tar.gz
  a.parquet
  app.py
  app.sh
  index.sh
  prepare_data.py
  prepare_data.sh
  prepare_index_data...
  query.py
  README.md
  requirements.txt
  search.sh
  start-services.sh
  .gitignore
  docker-compose.yml
  README.md
PROBLEMS
OUTPUT
TERMINAL
  [LOGS]
  1 of 1
  cassandra-server | WARN | [Native-Transport-Requests-5] 2025-04-15 21:17:23,855 BatchStatement.java:362 - Batch for [search.inverted_index] is of size 19.515KiB, exceeding specified threshold of 5.000KiB by 14.515KiB.
  cassandra-server | WARN | [Native-Transport-Requests-4] 2025-04-15 21:17:23,865 BatchStatement.java:362 - Batch for [search.inverted_index] is of size 19.878KiB, exceeding specified threshold of 5.000KiB by 14.878KiB.
  cassandra-server | WARN | [Native-Transport-Requests-1] 2025-04-15 21:17:23,876 BatchStatement.java:362 - Batch for [search.vocabulary] is of size 18.066KiB, exceeding specified threshold of 5.000KiB by 13.066KiB.
  cassandra-server | WARN | [Native-Transport-Requests-5] 2025-04-15 21:17:23,887 BatchStatement.java:362 - Batch for [search.inverted_index] is of size 19.878KiB, exceeding specified threshold of 5.000KiB by 14.878KiB.
  cassandra-server | WARN | [Native-Transport-Requests-5] 2025-04-15 21:17:23,898 BatchStatement.java:362 - Batch for [search.vocabulary] is of size 18.066KiB, exceeding specified threshold of 5.000KiB by 13.066KiB.
  cassandra-server | WARN | [Native-Transport-Requests-1] 2025-04-15 21:17:23,906 BatchStatement.java:362 - Batch for [search.inverted_index] is of size 19.845KiB, exceeding specified threshold of 5.000KiB by 14.845KiB.
  cassandra-server | WARN | [Native-Transport-Requests-1] 2025-04-15 21:17:23,917 BatchStatement.java:362 - Batch for [search.vocabulary] is of size 18.066KiB, exceeding specified threshold of 5.000KiB by 13.066KiB.
  cassandra-server | WARN | [Native-Transport-Requests-2] 2025-04-15 21:17:23,926 BatchStatement.java:362 - Batch for [search.inverted_index] is of size 19.951KiB, exceeding specified threshold of 5.000KiB by 14.951KiB.
  cassandra-server | WARN | [Native-Transport-Requests-5] 2025-04-15 21:17:23,938 BatchStatement.java:362 - Batch for [search.vocabulary] is of size 18.066KiB, exceeding specified threshold of 5.000KiB by 13.066KiB.
  cassandra-server | WARN | [Native-Transport-Requests-3] 2025-04-15 21:17:23,945 BatchStatement.java:362 - Batch for [search.inverted_index] is of size 19.871KiB, exceeding specified threshold of 5.000KiB by 14.871KiB.
  cassandra-server | WARN | [Native-Transport-Requests-4] 2025-04-15 21:17:23,957 BatchStatement.java:362 - Batch for [search.vocabulary] is of size 18.066KiB, exceeding specified threshold of 5.000KiB by 13.066KiB.
  cassandra-server | WARN | [Native-Transport-Requests-2] 2025-04-15 21:17:23,965 BatchStatement.java:362 - Batch for [search.inverted_index] is of size 19.837KiB, exceeding specified threshold of 5.000KiB by 14.837KiB.
  cassandra-server | WARN | [Native-Transport-Requests-1] 2025-04-15 21:17:23,978 BatchStatement.java:362 - Batch for [search.documents] is of size 28.402KiB, exceeding specified threshold of 5.000KiB by 23.402KiB.
  cassandra-server | WARN | [Native-Transport-Requests-5] 2025-04-15 21:17:23,988 BatchStatement.java:362 - Batch for [search.documents] is of size 28.507KiB, exceeding specified threshold of 5.000KiB by 23.507KiB.
  cassandra-server | WARN | [Native-Transport-Requests-1] 2025-04-15 21:17:23,998 BatchStatement.java:362 - Batch for [search.documents] is of size 28.208KiB, exceeding specified threshold of 5.000KiB by 23.208KiB.
  cassandra-server | WARN | [Native-Transport-Requests-5] 2025-04-15 21:17:24,006 BatchStatement.java:362 - Batch for [search.vocabulary] is of size 9.395KiB, exceeding specified threshold of 5.000KiB by 4.395KiB.
  cassandra-server | WARN | [Native-Transport-Requests-5] 2025-04-15 21:17:24,012 BatchStatement.java:362 - Batch for [search.inverted_index] is of size 9.926KiB, exceeding specified threshold of 5.000KiB by 4.926KiB.
  cassandra-server | WARN | [Native-Transport-Requests-1] 2025-04-15 21:17:24,019 BatchStatement.java:362 - Batch for [search.documents] is of size 25.777KiB, exceeding specified threshold of 5.000KiB by 20.777KiB.
  cluster-master | [LOGS] The end
```

This screenshot shows contents of hdfs / folder, number of documents in /data folder which is 978, contents of /index/data folder and the first line of /index/data/part-00000



Next screenshots show cassandra db tables, their contents and count of rows (global count is expensive without partitions, but there's no workaround :c)

Note the number of documents in documents table: 978, which is well over 100 specified in assignment.

```
pepew@pepew:/mnt/c/Users/gachi
cqlsh:search> describe tables;

documents inverted_index vocabulary

cqlsh:search> select count(*) from documents;

count
-----
978

(1 rows)

Warnings :
Aggregation query used without partition key

cqlsh:search> select * from documents limit 5;

doc_id | length | title
-----+-----+-----
10230685 | 115 | A_Dead_Sinking_Story
27568194 | 567 | A_Hero_Ain't_Nothing_but_a_Sandwich_(film)
39710446 | 330 | A_Little_Bit_of_Luck
38294693 | 338 | A_Change_Is_Gonna_Come_(Jack_McDuff_album)
51794980 | 306 | A_Family_Secret_(Upstairs,_Downstairs)

(5 rows)

cqlsh:search> select count(*) from inverted_index;

count
-----
244375

(1 rows)

Warnings :
Aggregation query used without partition key

cqlsh:search> _
```

```
pepew@pepew:/mnt/c/Users/gachi
(1 rows)

Warnings :
Aggregation query used without partition key

cqlsh:search> select * from inverted_index limit 5;

term | doc_id | tf
-----+-----+-----
dobson | 13633480 | 1
bessus | 12000397 | 3
ix | 19789501 | 1
ix | 32497421 | 1
ix | 67078438 | 2

(5 rows)

cqlsh:search> select count(*) from vocabulary;

count
-----
39630

(1 rows)

Warnings :
Aggregation query used without partition key

cqlsh:search> select * from vocabulary limit 5;

term | df
-----+-----
dobson | 1
bessus | 1
ix | 4
await | 2
libertad | 1

(5 rows)

cqlsh:search> _
```

Now it is time for 3 back-to-back search query demonstration:

Not all queries return 10 documents since it is possible that there are less than 10 relevant documents. A really good example of this is the first query “quantum physics” which returned only 1 book, but the book (A Brief History Of Time by Stephen Hawking) exactly matches the given query. Second query “how to pass big data course” has 6 words, but most of them likely don’t contribute much for BM25 score ([how, to, big] are definitely on a much more popular side of a vocabulary), and other terms didn’t find that many exact matches, so there are numerous documents returned, but their score is low. Lastly, the

query “geography” found only 5 relevant documents, but most of them had great scores, and the top one even featured the query in the title.

```
root@cluster-master:/app# docker exec -it cluster-master bash
pewpew@pewpew:/mnt/c/Users/gachi$ docker exec -it cluster-master bash
root@cluster-master:/app# bash search.sh "quantum physics"
[LOGS] Running search query: quantum physics
[LOGS] Fetching index data from Cassandra...
[LOGS] Total docs: 978, Avg doc len: 572.50
[LOGS] Top 10 Documents:
[LOGS] Document ID: 67227 Title: A_Brief_History_of_Time BM25 score: 10.1139
root@cluster-master:/app# bash search.sh "how to pass big data course"
[LOGS] Running search query: how to pass big data course
[LOGS] Fetching index data from Cassandra...
[LOGS] Total docs: 978, Avg doc len: 572.50
[LOGS] Top 10 Documents:
[LOGS] Document ID: 55915239 Title: A_Hairdresser's_Experience_in_High_Life BM25 score: 3.9680
[LOGS] Document ID: 6602969 Title: A_Is_for_Atom BM25 score: 3.8650
[LOGS] Document ID: 16456979 Title: A_Christmas_Carol_(1982_film) BM25 score: 3.7424
[LOGS] Document ID: 8303847 Title: A_Kid's_Guide_to_Giving BM25 score: 3.5506
[LOGS] Document ID: 57279816 Title: A_Book_of_American_Martyrs BM25 score: 3.4707
[LOGS] Document ID: 60121915 Title: A_Brief_History_of_Everyone_Who_Ever_Lived BM25 score: 3.3221
[LOGS] Document ID: 71344775 Title: A_Death_in_Bed_No._12 BM25 score: 3.2860
[LOGS] Document ID: 23204598 Title: A_Girl's_Tears BM25 score: 3.2423
[LOGS] Document ID: 5446503 Title: A_House_on_a_Street_in_a_Town_I'm_From BM25 score: 3.2337
[LOGS] Document ID: 15643468 Title: A_Beautiful_Sunset BM25 score: 3.2320
root@cluster-master:/app# bash search.sh "geography"
[LOGS] Running search query: geography
[LOGS] Fetching index data from Cassandra...
[LOGS] Total docs: 978, Avg doc len: 572.50
[LOGS] Top 10 Documents:
[LOGS] Document ID: 64552861 Title: A_Geography_of_Blood BM25 score: 9.4949
[LOGS] Document ID: 72327259 Title: A_Historical_Atlas_of_Tibet BM25 score: 5.8406
[LOGS] Document ID: 8191051 Title: A_History_of_the_Life_and_Voyages_of_Christopher_Columbus BM25 score: 4.5821
[LOGS] Document ID: 5003381 Title: A_Dying_Light_in_Corduba BM25 score: 3.8762
[LOGS] Document ID: 2166202 Title: A_Latin_Dictionary BM25 score: 3.7378
root@cluster-master:/app#
```