

Introduction to Big Data. Assignment 2

Dzhavid Sadreddinov

April 2025

1 Methodology

1.1 System Architecture Overview

Our search engine implementation consists of three core components:

- **Indexing Pipeline:**
 - Batch processing using MapReduce
 - Cassandra for persistent storage
 - Hadoop for distributed computation
- **Search Components:**
 - Query processing and tokenization
 - BM25 ranking algorithm
 - Distributed scoring using PySpark
- **Infrastructure:**
 - Cassandra cluster for low-latency lookups
 - Spark for distributed computation
 - HDFS for document storage

1.2 Indexing Design Choices

1.2.1 Document Processing Pipeline

Implemented a two-phase MapReduce workflow:

- **Phase 1:** Document frequency counting (mapper1 \rightarrow reducer1)
- **Phase 2:** Term frequency counting (mapper2 \rightarrow reducer2)
- **Phase 3:** Document lengths evaluation (mapper3 \rightarrow reducer3)

1.2.2 Cassandra Schema Design

Optimized for read performance with denormalized tables:

```
CREATE TABLE term_frequencies (  
    document_id bigint,  
    term text,  
    frequency counter,  
    PRIMARY KEY ((document_id, term))  
);  
CREATE TABLE IF NOT EXISTS document_frequencies (  
    term text PRIMARY KEY,  
    count counter  
);  
CREATE TABLE IF NOT EXISTS document_lengths (  
    document_id bigint PRIMARY KEY,  
    length int,  
    title text  
);
```

- Store term frequencies at each document
- Store document frequencies for each term
- Store document lengths and their titles for search results

1.3 Search Implementation

1.3.1 BM25 Ranking Algorithm

Implemented the standard BM25 ranking function:

$$\text{score}(D, Q) = \sum_{q_i \in Q} \text{IDF}(q_i) \cdot \frac{f(q_i, D) \cdot (k_1 + 1)}{f(q_i, D) + k_1 \cdot (1 - b + b \cdot \frac{|D|}{\text{avgdl}})} \quad (1)$$

Where:

- $k_1 = 1.2$ (term frequency saturation)
- $b = 0.75$ (length normalization)

1.3.2 Distributed Scoring

1. Query Processing:

- Case normalization
- Tokenization using regex `\w+`

2. Term Lookup:

```

query_terms_bc = sc.broadcast(query_terms)
document_freqs_bc = sc.broadcast(document_freqs)
total_docs_bc = sc.broadcast(total_docs)
avg_doc_length_bc = sc.broadcast(avg_doc_length)
term_freqs_all_bc = sc.broadcast(term_freqs_all)

```

3. Score Calculation:

- Mapped across matching documents
- Reduce phase sums partial scores
- Top-k selection using takeOrdered

2 Guide on running

Clone the repo and run docker compose. Download zipped cassandra package to /app/cassandra.zip from here

```

git clone https://github.com/sadjava/big-data-assignment2.git
cd big-data-assignment2
docker compose up

```

3 Screenshots

Running mappers and reducers to index documents:

```

cluster-master: This script includes commands to run Hadoop jobs using Hadoop Streaming to index documents
cluster-master: Input file is: /index/data
cluster-master: Checking if input exists in HDFS...
cluster-master: Running doc_frequencies job...
cluster-master: Input: /index/data
cluster-master: Output: /tmp/index/doc_frequencies
cluster-master: package36b3ar: [] [/usr/local/hadoop/share/hadoop/tools/lib/hadoop-streaming-3.3.1-jar] /tmp/streamjob2736861392653339494.jar tmp0trnull
2025-04-15 20:02:58,737 INFO client.DefaultHadoopProxyProvider: Connecting to ResourceManager at cluster-master/172.21.0.4:8032
2025-04-15 20:02:58,832 INFO client.DefaultHadoopProxyProvider: Connecting to ResourceManager at cluster-master/172.21.0.4:8032
2025-04-15 20:02:58,946 INFO mapreduce.JobResourceUploader: Disabling frasure coding for path: /tmp/hadoop-yarn/staging/root/.staging/job_1744747812493_0001
2025-04-15 20:02:59,114 INFO mapreduce.FileOutputFormat: Total input files to process = 1
2025-04-15 20:02:59,134 INFO mapreduce.JobSubmitter: number of splits=2
2025-04-15 20:02:59,203 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1744747812493_0001
2025-04-15 20:02:59,203 INFO mapreduce.JobSubmitter: Executing with tokens: []
2025-04-15 20:02:59,283 INFO conf.Configuration: resource-types.xml not found
2025-04-15 20:02:59,284 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2025-04-15 20:02:59,582 INFO impl.YarnClientImpl: Submitted application application_1744747812493_0001
2025-04-15 20:02:59,615 INFO mapreduce.Job: The url to track the job: http://cluster-master:8080/proxy/application_1744747812493_0001/
2025-04-15 20:02:59,616 INFO mapreduce.Job: Running job: job_1744747812493_0001
2025-04-15 20:03:04,666 INFO mapreduce.Job: Job job_1744747812493_0001 running in uber mode : false
2025-04-15 20:03:04,666 INFO mapreduce.Job: map 0% reduce 0%
2025-04-15 20:03:07,690 INFO mapreduce.Job: map 100% reduce 0%
cassandra@k8s:~$
cluster-master: INFO [NativeTransport-Requests-1] 2025-04-15 20:03:10,285 QueryProcessor.java:654 - Fully upgraded to at least 5.0.3
2025-04-15 20:03:21,739 INFO mapreduce.Job: map 100% reduce 100%
2025-04-15 20:03:21,746 INFO mapreduce.Job: Job job_1744747812493_0001 completed successfully
2025-04-15 20:03:21,790 INFO mapreduce.Job: Counters: 34
cluster-master:
cluster-master: File System Counters
cluster-master: FILE: Number of bytes read=2703851
cluster-master: FILE: Number of bytes written=6238971
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=3569227
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:
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)=2718
cluster-master: Total time spent by all reduces in occupied slots (ms)=11306
cluster-master: Total time spent by all map tasks (ms)=2718
cluster-master: Total time spent by all reduce tasks (ms)=21306
cluster-master: Total vcore-milliseconds taken by all map tasks=2718
cluster-master: Total vcore-milliseconds taken by all reduce tasks=11306
cluster-master: Total megabyte-milliseconds taken by all map tasks=2783232
cluster-master: Total megabyte-milliseconds taken by all reduce tasks=1157344
cluster-master: Map-Reduce Framework
cluster-master: Map Input records=1003
cluster-master: Map output records=230211

```

```
Activities Terminal ap15 23:08 sadjava@katana: ~/Documents/innobybigdata/assignment2/app sadjava@katana: ~/Documents/innobybigdata/assignment2/app sadjava@katana: ~/Documents/innobybigdata/assignment2/app Running term_frequencies Job... Input: /index/data Output: /tmp/index/term_frequencies packageJobJar: [] [/usr/local/hadoop/share/hadoop/tools/lib/hadoop-streaming-3.3.1-jar] [/tmp/streamjob286142658813681212.jar tmpDir=null] 2025-04-15 20:03:23,741 INFO client.DefaultHadoopMallowerProxyProvider: Connecting to ResourceManager at cluster-master/172.21.0.4:8032 2025-04-15 20:03:23,827 INFO client.DefaultHadoopMallowerProxyProvider: Connecting to ResourceManager at cluster-master/172.21.0.4:8032 2025-04-15 20:03:23,950 INFO mapreduce.JobResourceUploader: Disabling frasure coding for path: [/tmp/hadoop-yarn/staging/root/.staging/job_1744747012493_0002 2025-04-15 20:03:24,133 INFO mapreduce.JobSubmitter: Total input files to process: 1 2025-04-15 20:03:24,189 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1744747012493_0002 2025-04-15 20:03:24,189 INFO mapreduce.JobSubmitter: Executing with tokens: [] 2025-04-15 20:03:24,274 INFO conf.Configuration: resource-types.xml not found 2025-04-15 20:03:24,274 INFO resource.ResourceUtils: Unable to find /resource-types.xml' 2025-04-15 20:03:24,310 INFO Impl.VarnClientImpl: Submitted application application_1744747012493_0002 2025-04-15 20:03:24,334 INFO mapreduce.Job: The url to track the job: http://cluster-master:8088/proxy/application_1744747012493_0002/ 2025-04-15 20:03:24,335 INFO mapreduce.Job: Running job: job_1744747012493_0002 2025-04-15 20:03:31,387 INFO mapreduce.Job: job_1744747012493_0002 running in uber mode: false 2025-04-15 20:03:31,388 INFO mapreduce.Job: map 0% reduce 0% 2025-04-15 20:03:34,413 INFO mapreduce.Job: map 100% reduce 0% 2025-04-15 20:03:39,400 INFO mapreduce.Job: map 100% reduce 70% 2025-04-15 20:03:56,475 INFO mapreduce.Job: map 100% reduce 72% 2025-04-15 20:04:02,493 INFO mapreduce.Job: map 100% reduce 74% 2025-04-15 20:04:08,509 INFO mapreduce.Job: map 100% reduce 76% 2025-04-15 20:04:16,523 INFO mapreduce.Job: map 100% reduce 78% 2025-04-15 20:04:20,538 INFO mapreduce.Job: map 100% reduce 80% 2025-04-15 20:04:26,553 INFO mapreduce.Job: map 100% reduce 83% 2025-04-15 20:04:32,569 INFO mapreduce.Job: map 100% reduce 83% 2025-04-15 20:04:38,586 INFO mapreduce.Job: map 100% reduce 85% 2025-04-15 20:04:44,601 INFO mapreduce.Job: map 100% reduce 87% 2025-04-15 20:04:50,616 INFO mapreduce.Job: map 100% reduce 89% 2025-04-15 20:04:56,627 INFO mapreduce.Job: map 100% reduce 91% 2025-04-15 20:05:02,638 INFO mapreduce.Job: map 100% reduce 93% 2025-04-15 20:05:08,651 INFO mapreduce.Job: map 100% reduce 95% 2025-04-15 20:05:14,665 INFO mapreduce.Job: map 100% reduce 96% 2025-04-15 20:05:20,680 INFO mapreduce.Job: map 100% reduce 98% 2025-04-15 20:05:26,690 INFO mapreduce.Job: map 100% reduce 100% 2025-04-15 20:05:27,700 INFO mapreduce.Job: job_1744747012493_0002 completed successfully 2025-04-15 20:05:27,739 INFO mapreduce.Job: Counters: 54 Cluster-master File System Counters Cluster-master FILE: Number of bytes read=1063656 Cluster-master FILE: Number of bytes written=2283884 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=3368227 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 Running doc_lengths Job... Input: /index/data Output: /tmp/index/doc_lengths packageJobJar: [] [/usr/local/hadoop/share/hadoop/tools/lib/hadoop-streaming-3.3.1-jar] [/tmp/streamjob3956875151262072054.jar tmpDir=null] 2025-04-15 20:05:29,677 INFO client.DefaultHadoopMallowerProxyProvider: Connecting to ResourceManager at cluster-master/172.21.0.4:8032 2025-04-15 20:05:29,763 INFO client.DefaultHadoopMallowerProxyProvider: Connecting to ResourceManager at cluster-master/172.21.0.4:8032 2025-04-15 20:05:29,875 INFO mapreduce.JobResourceUploader: Disabling frasure coding for path: [/tmp/hadoop-yarn/staging/root/.staging/job_1744747012493_0003 2025-04-15 20:05:30,033 INFO mapreduce.JobSubmitter: Total input files to process: 1 2025-04-15 20:05:30,053 INFO mapreduce.JobSubmitter: number of splits=2 2025-04-15 20:05:30,184 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1744747012493_0003 2025-04-15 20:05:30,184 INFO mapreduce.JobSubmitter: Executing with tokens: [] 2025-04-15 20:05:30,183 INFO conf.Configuration: resource-types.xml not found 2025-04-15 20:05:30,184 INFO resource.ResourceUtils: Unable to find /resource-types.xml' 2025-04-15 20:05:30,214 INFO Impl.VarnClientImpl: Submitted application application_1744747012493_0003 2025-04-15 20:05:30,236 INFO mapreduce.Job: The url to track the job: http://cluster-master:8088/proxy/application_1744747012493_0003/ 2025-04-15 20:05:30,236 INFO mapreduce.Job: Running job: job_1744747012493_0003 2025-04-15 20:05:37,296 INFO mapreduce.Job: job_1744747012493_0003 running in uber mode: false 2025-04-15 20:05:37,296 INFO mapreduce.Job: map 0% reduce 0% 2025-04-15 20:05:40,321 INFO mapreduce.Job: map 100% reduce 0% 2025-04-15 20:05:44,330 INFO mapreduce.Job: map 100% reduce 100% 2025-04-15 20:05:45,345 INFO mapreduce.Job: job_1744747012493_0003 completed successfully 2025-04-15 20:05:45,386 INFO mapreduce.Job: Counters: 14 Cluster-master File System Counters Cluster-master FILE: Number of bytes read=37656 Cluster-master FILE: Number of bytes written=905669 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=3368227 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)=2436 Cluster-master Total time spent by all reduce in occupied slots (ms)=1461 Cluster-master Total time spent by all map tasks (ms)=2436 Cluster-master Total time spent by all reduce tasks (ms)=1461 Cluster-master Total vcore-milliseconds taken by all map tasks=2436 Cluster-master Total vcore-milliseconds taken by all reduce tasks=1461 Cluster-master Total megabyte-milliseconds taken by all map tasks=248464 Cluster-master Total megabyte-milliseconds taken by all reduce tasks=1490604 Cluster-master Map-Reduce Framework Cluster-master Map input records=1003 Cluster-master Map output records=997 Cluster-master Map output bytes=16566 Cluster-master Map output materialized bytes=37662 Cluster-master Input splits=292 Cluster-master Combine input records=0
```

3.1 Results of search

A film little game

```
Activities Terminal sadjava@katana: ~/Documents/nnn/bigdata/assignment2/app
cluster-master 25/04/15 20:31:42 INFO BlockManagerInfo: Added broadcast_1_python on disk on cluster-slave-1:45361 (size: 15.0 B)
cluster-master 25/04/15 20:31:42 INFO TaskSetManager: Finished task 1.0 in stage 0.0 (TID 1) in 1045 ms on cluster-slave-1 (executor 2) (1/2)
cluster-master 25/04/15 20:31:42 INFO PythonAccumulatorV2: Connected to AccumulatorServer at host: 127.0.0.1 port: 41713
cluster-master 25/04/15 20:31:42 INFO TaskSetManager: Finished task 0.0 in stage 0.0 (TID 0) in 1078 ms on cluster-slave-1 (executor 1) (2/2)
cluster-master 25/04/15 20:31:42 INFO VarnScheduler: Removed TaskSet 0.0, whose tasks have all completed, from pool.
cluster-master 25/04/15 20:31:42 INFO DAGScheduler: ResultStage 0 (takeOrdered at /app/query.py#1) finished in 1.111 s
cluster-master 25/04/15 20:31:42 INFO DAGScheduler: Job 0 is finished. Cancelling potential speculative or zombie tasks for this job
cluster-master 25/04/15 20:31:42 INFO VarnScheduler: Killing all running tasks in stage 0: Stage finished
cluster-master 25/04/15 20:31:42 INFO DAGScheduler: Job 0 finished. takeOrdered at /app/query.py#1, Took 1.149421 s
cluster-master 25/04/15 20:31:42 INFO SparkContext: SparkContext is stopping with exitCode 0.
cluster-master 25/04/15 20:31:42 INFO SparkUI: Stopped Spark web UI at http://cluster-master:8080
cluster-master 25/04/15 20:31:42 INFO VarnClientSchedulerBackend: Interrupting monitor thread
cluster-master 25/04/15 20:31:42 INFO VarnClientSchedulerBackend: Shutting down all executors
cluster-master 25/04/15 20:31:42 INFO VarnClientSchedulerBackend$VarDriverEndpoint: Asking each executor to shut down
cluster-master 25/04/15 20:31:42 INFO VarnClientSchedulerBackend: VARN client scheduler backend stopped
cluster-master 25/04/15 20:31:42 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
cluster-master 25/04/15 20:31:42 INFO MemoryStore: MemoryStore cleared
cluster-master 25/04/15 20:31:42 INFO BlockManager: BlockManager stopped
cluster-master 25/04/15 20:31:42 INFO BlockManagerMaster: BlockManagerMaster stopped
cluster-master 25/04/15 20:31:42 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
cluster-master 25/04/15 20:31:42 INFO SparkContext: Successfully stopped SparkContext
cluster-master Top 10 relevant documents:
cluster-master 1. Document ID: 3693795, title: A Football Life, BM25 Score: -1.2977
cluster-master 2. Document ID: 5544129, title: A History of the Modern World, BM25 Score: -1.3041
cluster-master 3. Document ID: 4808509, title: A Case de sette Heures, BM25 Score: -1.3107
cluster-master 4. Document ID: 1212586, title: A Crystal Christmas, BM25 Score: -1.3197
cluster-master 5. Document ID: 5685778, title: A Fistful of Perill, BM25 Score: -1.3258
cluster-master 6. Document ID: 1746781, title: A Double Dose of Soul, BM25 Score: -1.3299
cluster-master 7. Document ID: 3933348, title: A Bucketful of Soul, BM25 Score: -1.3312
cluster-master 8. Document ID: 4636051, title: A Bag Full of Blues, BM25 Score: -1.3321
cluster-master 9. Document ID: 2686618, title: A Beautiful Exchange, BM25 Score: -1.3374
cluster-master 10. Document ID: 1175383, title: A Journal of a Plague Year (abun), BM25 Score: -1.3416
cluster-master 25/04/15 20:31:43 INFO ShutdownHookManager: Shutdown hook called
cluster-master 25/04/15 20:31:43 INFO ShutdownHookManager: Deleting directory /tmp/spark-d087086-a27-4a93-9731-f512628cc47
cluster-master 25/04/15 20:31:43 INFO ShutdownHookManager: Deleting directory /tmp/spark-d42a2a3d-9e15-4d75-b317-808c1fe129e
cluster-master 25/04/15 20:31:43 INFO ShutdownHookManager: Deleting directory /tmp/spark-4d87185-9102-4315-ae1-4d4564084cf
cluster-master This script will include commands to search for documents given the query using Spark RDB
cluster-master 25/04/15 20:31:45 WARN NativeCodeLoader: Unable to load native-heap library for your platform... using builtin-java classes where applicable
cassandra-server WARN [Native-Transport-Requests-1] 2025-04-15 20:31:45,095 selectStatement: java557 - Aggregation query used without partition key
cluster-master 25/04/15 20:31:45 INFO SparkContext: Hunting Spark version 3.1.4
cluster-master 25/04/15 20:31:45 INFO SparkContext: OS Info Linux, 6.8.0-57-generic, amd64
cluster-master 25/04/15 20:31:45 INFO SparkContext: Java version 1.8.0_442
cluster-master 25/04/15 20:31:45 INFO ResourceUtils: =====
cluster-master 25/04/15 20:31:45 INFO ResourceUtils: No custom resources configured for spark.driver.
cluster-master 25/04/15 20:31:45 INFO SparkContext: Submitted application: BM25 Ranker
cluster-master 25/04/15 20:31:45 INFO ResourceProfile: Default ResourceProfile created, executor resources: Map(cores -> name: cores, amount: 1, script: , vendor: , memory -> name: memory, amount: 10
24, script: , vendor: , offheap -> name: offheap, amount: 0, script: , vendor: ), task resources: Map(cpus -> name: cpus, amount: 1.0)
cluster-master 25/04/15 20:31:45 INFO ResourceProfileManager: Limiting resource to cpus at 1 tasks per executor
cluster-master 25/04/15 20:31:45 INFO ResourceProfileManager: Added ResourceProfile id: 0
cluster-master 25/04/15 20:31:45 INFO SecurityManager: Changing view acls to: root
cluster-master 25/04/15 20:31:45 INFO SecurityManager: Changing modify acls to: root
cluster-master 25/04/15 20:31:45 INFO SecurityManager: Changing view acls groups to:
cluster-master 25/04/15 20:31:45 INFO SecurityManager: Changing modify acls groups to:
```

Discovering America

```
Activities Terminal sadjava@katana: ~/Documents/nnn/bigdata/assignment2/app
cluster-master 25/04/15 20:32:02 INFO BlockManagerInfo: Added broadcast_2_placed in memory on cluster-slave-1:135467 (size: 269.0 B, free: 366.3 MB)
cluster-master 25/04/15 20:32:02 INFO BlockManagerInfo: Added broadcast_2_python on disk on cluster-slave-1:135467 (size: 15.0 B)
cluster-master 25/04/15 20:32:02 INFO BlockManagerInfo: Added broadcast_1_placed in memory on cluster-slave-1:44885 (size: 269.0 B, free: 366.3 MB)
cluster-master 25/04/15 20:32:02 INFO BlockManagerInfo: Added broadcast_1_python on disk on cluster-slave-1:44885 (size: 15.0 B)
cluster-master 25/04/15 20:32:02 INFO BlockManagerInfo: Added broadcast_1_placed in memory on cluster-slave-1:135467 (size: 226.0 B, free: 366.3 MB)
cluster-master 25/04/15 20:32:02 INFO BlockManagerInfo: Added broadcast_1_python on disk on cluster-slave-1:135467 (size: 15.0 B)
cluster-master 25/04/15 20:32:02 INFO BlockManagerInfo: Added broadcast_1_placed in memory on cluster-slave-1:44885 (size: 269.0 B, free: 366.3 MB)
cluster-master 25/04/15 20:32:02 INFO BlockManagerInfo: Added broadcast_1_placed in memory on cluster-slave-1:44885 (size: 15.0 B)
cluster-master 25/04/15 20:32:02 INFO BlockManagerInfo: Added broadcast_2_placed in memory on cluster-slave-1:135467 (size: 226.0 B, free: 366.3 MB)
cluster-master 25/04/15 20:32:02 INFO BlockManagerInfo: Added broadcast_2_python on disk on cluster-slave-1:135467 (size: 15.0 B)
cluster-master 25/04/15 20:32:02 INFO TaskSetManager: Finished task 1.0 in stage 0.0 (TID 1) in 1166 ms on cluster-slave-1 (executor 1) (1/2)
cluster-master 25/04/15 20:32:02 INFO PythonAccumulatorV2: Connected to AccumulatorServer at host: 127.0.0.1 port: 45175
cluster-master 25/04/15 20:32:02 INFO TaskSetManager: Finished task 0.0 in stage 0.0 (TID 0) in 1192 ms on cluster-slave-1 (executor 2) (2/2)
cluster-master 25/04/15 20:32:02 INFO VarnScheduler: Removed TaskSet 0.0, whose tasks have all completed, from pool.
cluster-master 25/04/15 20:32:02 INFO DAGScheduler: ResultStage 0 (takeOrdered at /app/query.py#1) finished in 1.120 s
cluster-master 25/04/15 20:32:02 INFO DAGScheduler: Job 0 is finished. Cancelling potential speculative or zombie tasks for this job
cluster-master 25/04/15 20:32:02 INFO VarnScheduler: Killing all running tasks in stage 0: Stage finished
cluster-master 25/04/15 20:32:02 INFO DAGScheduler: Job 0 finished. takeOrdered at /app/query.py#1, Took 1.287919 s
cluster-master 25/04/15 20:32:02 INFO SparkContext: SparkContext is stopping with exitCode 0.
cluster-master 25/04/15 20:32:02 INFO SparkUI: Stopped Spark web UI at http://cluster-master:8080
cluster-master 25/04/15 20:32:02 INFO VarnClientSchedulerBackend: Interrupting monitor thread
cluster-master 25/04/15 20:32:02 INFO VarnClientSchedulerBackend: Shutting down all executors
cluster-master 25/04/15 20:32:02 INFO VarnClientSchedulerBackend$VarDriverEndpoint: Asking each executor to shut down
cluster-master 25/04/15 20:32:02 INFO VarnClientSchedulerBackend: VARN client scheduler backend stopped
cluster-master 25/04/15 20:32:02 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
cluster-master 25/04/15 20:32:02 INFO MemoryStore: MemoryStore cleared
cluster-master 25/04/15 20:32:02 INFO BlockManager: BlockManager stopped
cluster-master 25/04/15 20:32:02 INFO BlockManagerMaster: BlockManagerMaster stopped
cluster-master 25/04/15 20:32:02 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
cluster-master 25/04/15 20:32:02 INFO SparkContext: Successfully stopped SparkContext
cluster-master Top 10 relevant documents:
cluster-master 1. Document ID: 4193185, title: A Arma Escalante, BM25 Score: 5.7640
cluster-master 2. Document ID: 7686991, title: A Far Sunset, BM25 Score: 5.3882
cluster-master 3. Document ID: 2474151, title: A Beast With Ten Heads, BM25 Score: 4.7899
cluster-master 4. Document ID: 2878659, title: A Captive in the Land, BM25 Score: 4.2857
cluster-master 5. Document ID: 11041026, title: A Dona do Pedaço, BM25 Score: 4.1301
cluster-master 6. Document ID: 1451309, title: A Devilish Homicide, BM25 Score: 3.9252
cluster-master 7. Document ID: 867426, title: A Burnt-Out Case, BM25 Score: 3.8100
cluster-master 8. Document ID: 1746072, title: A Gun For Sale, BM25 Score: 3.7812
cluster-master 9. Document ID: 4240349, title: A Fairly Odd Summer, BM25 Score: 3.6250
cluster-master 10. Document ID: 5206493, title: A Linguistic Atlas of Early Middle English, BM25 Score: 3.0185
cluster-master 25/04/15 20:32:03 INFO ShutdownHookManager: Shutdown hook called
cluster-master 25/04/15 20:32:03 INFO ShutdownHookManager: Deleting directory /tmp/spark-5d90864-fab3-4d08-af09-af11f7139ec
cluster-master 25/04/15 20:32:03 INFO ShutdownHookManager: Deleting directory /tmp/spark-5d90864-fab3-4d08-af09-af11f7139ec
cluster-master 25/04/15 20:32:03 INFO ShutdownHookManager: Deleting directory /tmp/spark-15f2c20f-c1f0-41cd-9f5d-ffcaab80b970
cluster-master 25/04/15 20:32:03 INFO ShutdownHookManager: Deleting directory /tmp/spark-9a4301ad-099b-4139-aa4e-2dc11231b0d8
cluster-master This script will include commands to search for documents given the query using Spark RDB
cluster-master 25/04/15 20:32:04 WARN NativeCodeLoader: Unable to load native-heap library for your platform... using builtin-java classes where applicable
cassandra-server WARN [Native-Transport-Requests-10] 2025-04-15 20:32:04,214 selectStatement: java557 - Aggregation query used without partition key
cluster-master 25/04/15 20:32:05 INFO SparkContext: Hunting Spark version 3.1.4
cluster-master 25/04/15 20:32:05 INFO SparkContext: OS Info Linux, 6.8.0-57-generic, amd64
cluster-master 25/04/15 20:32:05 INFO SparkContext: Java version 1.8.0_442
cluster-master 25/04/15 20:32:05 INFO ResourceUtils: =====
cluster-master 25/04/15 20:32:05 INFO ResourceUtils: No custom resources configured for spark.driver.
```

Merry Christmas

- Positive scores indicate better term specificity
- Top scoring documents likely contain both "america" and "discovering"
- Score distribution shows better discrimination (5.76 to 3.61)
- **Anomalies:**
 - Portuguese title "A Arma Escarlata" ranking highest suggests:
 - * Possible metadata issues in indexing
 - * May contain English text about discovery despite Portuguese title
 - Linguistic Atlas appears relevant but ranks last in top 10

4.3 Query: "Merry christmas"

- **Top Result:** "A Kiss of Shadows" (Score: 8.5458)
- **Notable Patterns:**
 - Highest absolute scores among all queries
 - Christmas-themed titles rank appropriately:
 - * "A Christmas Cornucopia" (2nd)
 - * "A Christmas Melody" (3rd)
 - * "A Christmas Carol" (8th)
- **Surprises:**
 - Non-Christmas title "A Kiss of Shadows" ranking first suggests:
 - * Possible term frequency dominance in document text
 - * Metadata/text discrepancy in indexing
 - * "merry" may match other contexts (e.g., "merry making")