

## Nekrasov Dmitrii, DS-02 BigData Assignment 2

### Methodology:

#### explanation of design choices and approaches in implementing the search engine

Total execution time of 'docker compose up' is 40min on ryzen 5 5600.

Search for 1 query takes 30+-15 seconds, booting spark takes a large portion of this time.

In my map reduce scheme, i use only 1 mapper and 1 reducer.

In mapper, i map each word to document\_id it was in, then sort words giving:

```
3214 word1 100
3214 word1 100
2345 word1 200
2342 word2 100
9079 word2 300
3685 word3 400
```

In reducer, i have  $O(1)$  space complexity, writing to .output1 lines for each table entry:

```
print (f'word_within_one_document_count {search_word}
{search_doc} {search_word_count_within_doc}')
print (f'document_frequency {search_word}
{search_word_count_across_docs}' )
print (f'doc_id_len {doc_id} {doc_len}')
```

This is achieved with dynamic updates of search variables (search\_word, search\_doc).  
Output files weights 17110539bytes=16.3 Megabytes for 1000 documents.

In query step i use spark cassandra driver with only tables reading and bm25 fraction calculations.

### Demonstration

successful indexing of 1000 documents:

```
1 { "username": "dmitriinekrasov", "key": "6639b454322a1af72c76ab5fec5d5a081" }
```

File Path	Document Content
/data/929265_A_Chance_to_Cut_Is_a_Chance_to_Cure.txt	
/data/9413554_A_Haunting_Curse.txt	
/data/961187_A_Hangover_You_Don't_Deserve.txt	
/data/9704239_A_Contention_for_Honor_and_Riches.txt	
/data/9847946_A_Hard_Day's_Night_(Grey's_Anatomy).txt	
/data/9869812_A_Dream_(Common_song).txt	
/data/9870217_A_Date_with_Luyu.txt	
/data/9919932_A_Family_Affair_(musical).txt	
/data/9947241_A_Day_of_Renew.txt	
/data/9965276_A_Book_of_Human_Language.txt	
/index/data/ SUCCESS	
/index/data/part-00000-606c4303-1c20-43a9-820f-c270c22bfaea-c000.csv	
/9983283 A Good Enough Day is the second album by Canadian singer-songwriter Royal Wood, released in 2007 on Dead Daisy Records. --Track listing-- # "A Good Enough Day" # "Juliet" # "Safe Haven" # "A Mirror Without" # "I'm So Glad" # "Siren" # "In the Garden" # "Step Back" # "Forever Were Tied" # "About You" # "Acting Crazy (It's a Breakdown)" # "Silently" --References-- Category:2007 albums Category:Royal Wood albumsdone data preparation!	
Mapreduce jobs using hadoop streaming to index documents	
rm: /user/root/tmp/document_frequency_word_within_one_doc.output1: No such file or directory	
packageJobJar: [/tmp/hadoop-unjar7207236020291380026/] [] /tmp/streamjob3200698681963298325.jar tmpDir=null	
2025-04-12 09:40:03,508 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at cluster-master/172.21.0.5:8032	
2025-04-12 09:40:03,654 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at cluster-master/172.21.0.5:8032	
2025-04-12 09:40:03,867 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/root/.staging/job_1744450172070_0001	

successful execution of the search engine on some queries:

here query is The Dave Clark Five, Catch Us If You Can and bm25@10 didnt found actual document

```
1 from pyspark.sql import SparkSession
2 from pyspark.sql.functions import col, log, sum as spark_sum, lit
```

```
cluster-master | 25/04/12 10:05:55 INFO YarnScheduler: Killing all running tasks in stage 47: Stage finished
cluster-master | 25/04/12 10:05:55 INFO DAGScheduler: Job 26 finished: showString at NativeMethodAccessorImpl.java:0, took 0.098998 s
cluster-master | 25/04/12 10:05:55 INFO CodeGenerator: Code generated in 7.434826 ms
cluster-master | +-----+-----+
cluster-master | |document_id|bm25_score|
cluster-master | +-----+-----+
cluster-master | |8|27.372882739766126|
cluster-master | |5|22.50498356698747|
cluster-master | |3|21.50219987548129|
cluster-master | |0|18.581866750140037|
cluster-master | |1|16.86256915202236|
cluster-master | |4|14.01678391322488|
cluster-master | |25027416|13.204714893740602|
cluster-master | |34583915|12.98921239911669|
cluster-master | |26866318|12.375835111099198|
cluster-master | |2|11.563806771332395|
cluster-master | +-----+-----+
cluster-master | only showing top 10 rows
cluster-master |
```

here query is big data course study spark map reduce, documents indeed have those words in content

```
# bash search.sh big data course study spark map reduce
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
con.datastax.spark#spark-cassandra-connector_2.12 added as a dependency
:: resolving dependencies :: org.apache.spark#spark-submit-parent-f05daa61-1e28-4829-9c92-9f5d4afc603c;1.0
  confs: [default]
  found con.datastax.spark#spark-cassandra-connector_2.12;3.2.0 in central
  found con.datastax.spark#spark-cassandra-connector-driver_2.12;3.2.0 in central
  found con.datastax.spark#java-driver-core-shaded;4.13.0 in central
  found con.datastax.spark#native-protocol;1.5.0 in central
  found con.datastax.spark#java-driver-shaded-guava;25.1-jre-graal-sub-1 in central
  found con.typesafe#config;1.4.1 in central
  found org.slf4j#slf4j-api;1.7.26 in central
  found io.dropwizard.metrics#metrics-core;4.1.18 in central
  found org.hdrhistogram#HdrHistogram;2.1.12 in central
  found org.reactivestreams#reactive-streams;1.0.3 in central
  found con.github.stephenc.jcip#jcip-annotations;1.0-1 in central
  found con.github.sspotbugs#spotbugs-annotations;3.1.12 in central
  found con.google.code.findbugs#jsr305;3.0.2 in central
  found con.datastax.spark#java-driver-napper-runtime;4.13.0 in central
  found con.datastax.spark#java-driver-query-builder;4.13.0 in central
  found org.apache.commons#commons-lang3;3.10 in central
  found con.thoughtworks.paranamer#paranamer;2.8 in central
  found org.scala-lang#scala-reflect;2.12.11 in central
  :: resolution report :: resolve 436ms :: artifacts dl 15ms
  :: modules in use:
    con.datastax.spark#java-driver-core-shaded;4.13.0 from central in [default]
    con.datastax.spark#java-driver-napper-runtime;4.13.0 from central in [default]
    con.datastax.spark#java-driver-query-builder;4.13.0 from central in [default]
    con.datastax.spark#java-driver-shaded-guava;25.1-jre-graal-sub-1 from central in [default]
    con.datastax.spark#native-protocol;1.5.0 from central in [default]
    con.datastax.spark#spark-cassandra-connector-driver_2.12;3.2.0 from central in [default]
    con.datastax.spark#spark-cassandra-connector_2.12;3.2.0 from central in [default]
    con.github.sspotbugs#spotbugs-annotations;3.1.12 from central in [default]
```

