

Big Data Assignment 2 Report

Arsenii Pavlov ars.pavlov@innopolis.university

Methodology

Data Preparation

- 1) Used a Parquet file containing Wikipedia articles. Selected 4000 documents using PySpark's `sample()` and `limit()` functions.
- 2) Each document is saved as `<doc_id>_<doc_title>.txt` in HDFS. Spaces in titles are replaced with `_`.

pipeline:

- 1) Read the Parquet file into a Spark DataFrame.
- 2) Extracted id, title, and text columns.
- 3) Saved documents to HDFS to `/data`
- 4) After that save all to `index/data` as `<doc_id>_<doc_title>_<text>`

Indexer Tasks

The indexer uses two Hadoop MapReduce pipelines and save results in Cassandra.

1) TF Calculation

Mapper1

- 1) Read input documents from HDFS.
- 2) Tokenizes text cast to lowercase and splits on non-alphanumeric characters.
- 3) Convert to key-value pairs: `<term>#<doc_id>` to 1.

Reducer1

- 1) Aggregates counts for `<term>#<doc_id>` to compute TF.
- 2) Output: `<term> <doc_id> <tf>` to `/tmp/index/pipeline1`

2) DF Calculation

Mapper2

- 1) Reads TF data from Pipeline 1.
- 2) Convert term to 1 for each unique term-document pair.

Reducer2:

- 1) Sums counts to compute DF.
- 2) Output `<term>\t<df>` to `/tmp/index/pipeline2`

3) Saving to Cassandra

Tables:

vocabulary: words and DF values
inverted_index words and TF values
documents: map document IDs and titles and text

Loading Data: app.py reads HDFS outputs and inserts data into Cassandra using batch queries.

Ranker Tasks

ranker use PySpark to compute BM25 scores for queries.

BM25 Calculation

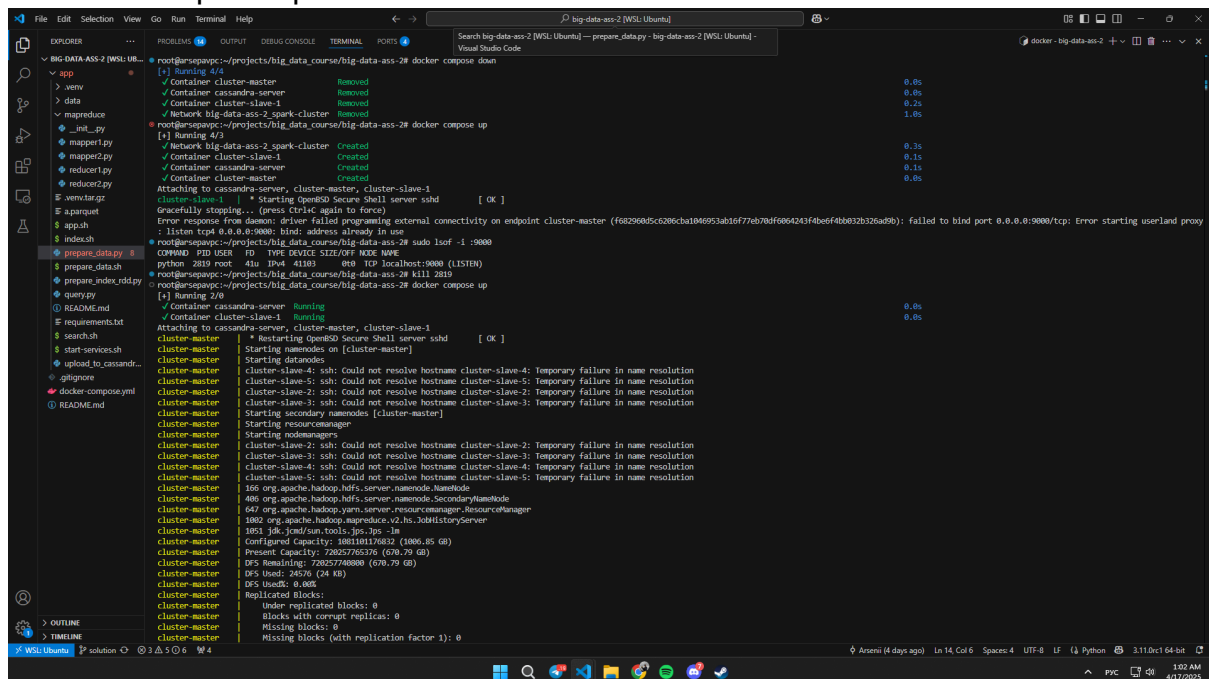
- 1) Fetch DF for query terms from vocabulary table.
- 2) Fetch TF and document lengths from inverted_index and documents.
- 3) Compute BM25 score using provided formula
- 4) Sum scores across all query terms.
- 5) Rank documents by score and return top 10.

Spark

- 1) Initialize Spark and Process Query
- 2) Compute BM25 Scores
- 3) Rank and Output Results

Demonstration

To start containers run
docker compose up



It will run all containers and run data preparation indexing and 1 query

```
File Edit Selection View Go Run Terminal Help
EXPLORER
  BIG-Data-AS2 (WSL Ubuntu)
  > src
  > data
  > mapreduce
  > _init_py
  > mapper1.py
  > mapper2.py
  > reducer1.py
  > reducer2.py
  > vername.txt
  > app.jar
  > app.sh
  > index.sh
  > prepare_data.sh
  > prepare_data_idx.py
  > query.py
  > README.md
  > requirements.txt
  > search.sh
  > start-services.sh
  > upload_to_cassandra...
  > gpudrive
  > docker-compose.yml
  > README.md
  > OUTLINE
  > WSL: Ubuntu
  > solution
  > 1
  > 2
  > 3
  > 4
  > 5
  > 6
  > 7
  > 8
  > 9
  > 10
  > 11
  > 12
  > 13
  > 14
  > 15
  > 16
  > 17
  > 18
  > 19
  > 20
  > 21
  > 22
  > 23
  > 24
  > 25
  > 26
  > 27
  > 28
  > 29
  > 30
  > 31
  > 32
  > 33
  > 34
  > 35
  > 36
  > 37
  > 38
  > 39
  > 40
  > 41
  > 42
  > 43
  > 44
  > 45
  > 46
  > 47
  > 48
  > 49
  > 50
  > 51
  > 52
  > 53
  > 54
  > 55
  > 56
  > 57
  > 58
  > 59
  > 60
  > 61
  > 62
  > 63
  > 64
  > 65
  > 66
  > 67
  > 68
  > 69
  > 70
  > 71
  > 72
  > 73
  > 74
  > 75
  > 76
  > 77
  > 78
  > 79
  > 80
  > 81
  > 82
  > 83
  > 84
  > 85
  > 86
  > 87
  > 88
  > 89
  > 90
  > 91
  > 92
  > 93
  > 94
  > 95
  > 96
  > 97
  > 98
  > 99
  > 100
  > 101
  > 102
  > 103
  > 104
  > 105
  > 106
  > 107
  > 108
  > 109
  > 110
  > 111
  > 112
  > 113
  > 114
  > 115
  > 116
  > 117
  > 118
  > 119
  > 120
  > 121
  > 122
  > 123
  > 124
  > 125
  > 126
  > 127
  > 128
  > 129
  > 130
  > 131
  > 132
  > 133
  > 134
  > 135
  > 136
  > 137
  > 138
  > 139
  > 140
  > 141
  > 142
  > 143
  > 144
  > 145
  > 146
  > 147
  > 148
  > 149
  > 150
  > 151
  > 152
  > 153
  > 154
  > 155
  > 156
  > 157
  > 158
  > 159
  > 160
  > 161
  > 162
  > 163
  > 164
  > 165
  > 166
  > 167
  > 168
  > 169
  > 170
  > 171
  > 172
  > 173
  > 174
  > 175
  > 176
  > 177
  > 178
  > 179
  > 180
  > 181
  > 182
  > 183
  > 184
  > 185
  > 186
  > 187
  > 188
  > 189
  > 190
  > 191
  > 192
  > 193
  > 194
  > 195
  > 196
  > 197
  > 198
  > 199
  > 200
  > 201
  > 202
  > 203
  > 204
  > 205
  > 206
  > 207
  > 208
  > 209
  > 210
  > 211
  > 212
  > 213
  > 214
  > 215
  > 216
  > 217
  > 218
  > 219
  > 220
  > 221
  > 222
  > 223
  > 224
  > 225
  > 226
  > 227
  > 228
  > 229
  > 230
  > 231
  > 232
  > 233
  > 234
  > 235
  > 236
  > 237
  > 238
  > 239
  > 240
  > 241
  > 242
  > 243
  > 244
  > 245
  > 246
  > 247
  > 248
  > 249
  > 250
  > 251
  > 252
  > 253
  > 254
  > 255
  > 256
  > 257
  > 258
  > 259
  > 260
  > 261
  > 262
  > 263
  > 264
  > 265
  > 266
  > 267
  > 268
  > 269
  > 270
  > 271
  > 272
  > 273
  > 274
  > 275
  > 276
  > 277
  > 278
  > 279
  > 280
  > 281
  > 282
  > 283
  > 284
  > 285
  > 286
  > 287
  > 288
  > 289
  > 290
  > 291
  > 292
  > 293
  > 294
  > 295
  > 296
  > 297
  > 298
  > 299
  > 300
  > 301
  > 302
  > 303
  > 304
  > 305
  > 306
  > 307
  > 308
  > 309
  > 310
  > 311
  > 312
  > 313
  > 314
  > 315
  > 316
  > 317
  > 318
  > 319
  > 320
  > 321
  > 322
  > 323
  > 324
  > 325
  > 326
  > 327
  > 328
  > 329
  > 330
  > 331
  > 332
  > 333
  > 334
  > 335
  > 336
  > 337
  > 338
  > 339
  > 340
  > 341
  > 342
  > 343
  > 344
  > 345
  > 346
  > 347
  > 348
  > 349
  > 350
  > 351
  > 352
  > 353
  > 354
  > 355
  > 356
  > 357
  > 358
  > 359
  > 360
  > 361
  > 362
  > 363
  > 364
  > 365
  > 366
  > 367
  > 368
  > 369
  > 370
  > 371
  > 372
  > 373
  > 374
  > 375
  > 376
  > 377
  > 378
  > 379
  > 380
  > 381
  > 382
  > 383
  > 384
  > 385
  > 386
  > 387
  > 388
  > 389
  > 390
  > 391
  > 392
  > 393
  > 394
  > 395
  > 396
  > 397
  > 398
  > 399
  > 400
  > 401
  > 402
  > 403
  > 404
  > 405
  > 406
  > 407
  > 408
  > 409
  > 410
  > 411
  > 412
  > 413
  > 414
  > 415
  > 416
  > 417
  > 418
  > 419
  > 420
  > 421
  > 422
  > 423
  > 424
  > 425
  > 426
  > 427
  > 428
  > 429
  > 430
  > 431
  > 432
  > 433
  > 434
  > 435
  > 436
  > 437
  > 438
  > 439

```

The screenshot displays the Docker Desktop interface for a multi-container application running on a Windows Subsystem for Linux (WSL) Ubuntu environment. The interface is divided into several panels:

- Explorer:** Shows a file tree for a project named 'big-data-ass-2'. The tree includes directories like 'app', 'data', 'mapreduce', and 'prepare_data.py'. The 'prepare_data.py' file is currently selected.
- Problems:** This panel is empty, indicating no errors or warnings.
- Output:** This panel shows the output of the selected container, 'cluster-master'. The output displays the progress of a Hadoop MapReduce job, including the number of map and reduce tasks completed, and the final result of the job.
- Debug Console:** This panel is empty.
- Terminal:** This panel is empty.
- Ports:** This panel is empty.

The main area displays the output of the 'cluster-master' container, which is running a Hadoop MapReduce job. The output shows the progress of the job, including the number of map and reduce tasks completed, and the final result of the job.

```

cluster-master | [ 9753361, 'A Division (New York City Subway)']
cluster-master | 9753361 A Division (New York City Subway)
cluster-master | [ 9797096, 'A Good Man (1941 Film)']
cluster-master | 9797096 A Good Man (1941 Film)
cluster-master | [ 9848347, 'A City in Winter']
cluster-master | 9848347 A City in Winter
cluster-master | [ 9847946, 'A Hard Day's Night (Grey's Anatomy)']
cluster-master | 9847946 A Hard Day's Night (Grey's Anatomy)
cluster-master | [ 9848866, 'A Big 10-8 Place']
cluster-master | 9848866 A Big 10-8 Place
cluster-master | [ 9869812, 'A Dream (Common song)']
cluster-master | 9869812 A Dream (Common song)
cluster-master | [ 9870217, 'A Date with Luyu']
cluster-master | 9870217 A Date with Luyu
cluster-master | [ 9883859, 'A Holiday Romance']
cluster-master | 9883859 A Holiday Romance
cluster-master | [ 9892055, 'A Diary for Timothy']
cluster-master | 9892055 A Diary for Timothy
cluster-master | [ 9897801, 'A Grand Night for Singing']
cluster-master | 9897801 A Grand Night for Singing
cluster-master | [ 9914910, 'A Day Without a Mexican']
cluster-master | 9914910 A Day Without a Mexican
cluster-master | [ 9919932, 'A Family Affair (musical)']
cluster-master | 9919932 A Family Affair (musical)
cluster-master | [ 9938278, 'A Band Called David']
cluster-master | 9938278 A Band Called David
cluster-master | [ 9997241, 'A Day of Renew']
cluster-master | 9997241 A Day of Renew
cluster-master | [ 9995207, 'A King and His Movie']
cluster-master | 9995207 A King and His Movie
cluster-master | [ 9995276, 'A Book of Human Language']
cluster-master | 9995276 A Book of Human Language
cluster-master | [ 9995276, 'A Good Enough Day']
cluster-master | 9995276 A Good Enough Day
cluster-master | [ 9995276, 'A Day to Remember']
cluster-master | 9995276 A Day to Remember
cluster-master | [ 9993015, 'A Guy Like Me']
cluster-master | 9993015 A Guy Like Me
cluster-master | Deleted /tmp/index/pipeline2
cluster-master | Deleted /tmp/index/pipeline2
cluster-master | BIG DATA APP: loaded to cassandra!!!!!!!
cluster-master | BIG DATA APP: Indexing completed successfully!
cassandra-server | INFO [Native-Transport-Requests-1] 2025-04-16 22:10:55,565 QueryProcessor-Java:654 - Fully upgraded to at least 5.0.3
cluster-master | BIG DATA APP: Search results:
cluster-master | BIG DATA APP: doc_id: 693378, title: A Cyborg Manifesto, score: 5.732855077774739
cluster-master | BIG DATA APP: doc_id: 31812710, title: A Friend in London, score: 5.732855077774739
cluster-master | BIG DATA APP: doc_id: 28476897, title: A Z @ Shantai All, score: 5.35864739256424
cluster-master | BIG DATA APP: doc_id: 407392, title: A Connection Yankee in King Arthur's Court, score: 5.35864739256424
cluster-master | BIG DATA APP: doc_id: 68888414, title: A House Divided (Person of Interest), score: 5.35864739256424
cluster-master | BIG DATA APP: doc_id: 69227029, title: A Z @ Shantai All, score: 5.35864739256424
cluster-master | BIG DATA APP: doc_id: 37688054, title: A Coffee in Berlin, score: 5.016248193852897
cluster-master | BIG DATA APP: doc_id: 6444283, title: A Doomsday Like Any Other, score: 5.016248193852897
cluster-master | BIG DATA APP: doc_id: 62572557, title: A City Under Siege, score: 5.016248193852897
cluster-master | BIG DATA APP: doc_id: 28022556, title: A Christmas Carol (Doctor Who), score: 5.016248193852897
  
```

The bottom status bar indicates the system is running on WSL Ubuntu, with 3.11GB of memory and 64-bit architecture.

The screenshot shows a VS Code interface with a terminal window. The terminal is running a Docker container named 'cluster-master' and executing a search for 'computer mouse' in a dataset. The search results are displayed in the terminal output, showing a list of documents with their IDs, titles, and scores.

```
root@arsenapwpc:/projects/big_data_course/big_data-ass-2# docker exec -it cluster-master bash
Error response from daemon: No such container: cluster-master
root@arsenapwpc:/projects/big_data_course/big_data-ass-2# docker exec -it cluster-master bash
root@cluster-master:/app
root@cluster-master:/app# bash search.sh "computer mouse"
BIG_DATA_APP: Search result:
BIG_DATA_APP: doc_id: 34313572, title: A Computer Animated Hand, score: 6.585286884876252
BIG_DATA_APP: doc_id: 484442, title: A Big's Life, score: 6.2922388152483
BIG_DATA_APP: doc_id: 48227028, title: A P.J. Shaker's All, score: 6.156023791738874
BIG_DATA_APP: doc_id: 38279, title: A Commentary on the UNIX Operating System, score: 5.646888975768635
BIG_DATA_APP: doc_id: 26481662, title: A Glitch Is a Glitch, score: 5.646888975768635
BIG_DATA_APP: doc_id: 67426266, title: A Grandchild's Guide to Using Grandpa's Computer, score: 5.42188572673821
BIG_DATA_APP: doc_id: 48347188, title: A Line in the Sand (video game), score: 5.42188572673821
BIG_DATA_APP: doc_id: 7868751, title: A Christmas Carol (2006 film), score: 5.42188572673821
BIG_DATA_APP: doc_id: 714551, title: A Deepness in the Sky, score: 5.42188572673821
BIG_DATA_APP: doc_id: 12588454, title: A K Peters, score: 5.42188572673821
root@cluster-master:/app# bash search.sh "pony"
BIG_DATA_APP: Search result:
BIG_DATA_APP: doc_id: 41578816, title: A Brony Tale, score: 8.587232234038157
BIG_DATA_APP: doc_id: 52933824, title: A Hearth's Warming Tail, score: 8.178116499881996
BIG_DATA_APP: doc_id: 76818626, title: A Carousel for Missoula, score: 7.433893399148645
BIG_DATA_APP: doc_id: 35578102, title: A Canterlot Wedding, score: 7.156026936698497
BIG_DATA_APP: doc_id: 938675, title: A Light in the Attic, score: 6.368912832620887
BIG_DATA_APP: doc_id: 2332186, title: A House, score: 6.368912832620887
BIG_DATA_APP: doc_id: 23311725, title: A Fool's Paradise, score: 6.368912832620887
BIG_DATA_APP: doc_id: 28182945, title: A Good Git-Together, score: 4.778684624465665
BIG_DATA_APP: doc_id: 24114650, title: A Hero for Handy, score: 4.778684624465665
BIG_DATA_APP: doc_id: 11631735, title: A Ballad of the West, score: 4.778684624465665
root@cluster-master:/app# bash search.sh "apple tree"
BIG_DATA_APP: Search result:
BIG_DATA_APP: doc_id: 113226, title: A Charlie Brown Christmas, score: 7.1838281879572255
BIG_DATA_APP: doc_id: 1293854, title: A Charlie Brown Thanksgiving, score: 6.73483851289899
BIG_DATA_APP: doc_id: 381971, title: A Christmas Gift for You from Phil Spector, score: 6.73483851289899
BIG_DATA_APP: doc_id: 3145259, title: A Fifth of Beethoven, score: 6.861154966888909
BIG_DATA_APP: doc_id: 2294422, title: A Great Day to Care, score: 6.861154966888909
BIG_DATA_APP: doc_id: 28238318, title: A Divine Looking Glass, score: 6.861154966888909
BIG_DATA_APP: doc_id: 33618926, title: A Chinese Ghost Story II, score: 5.387871888967919
BIG_DATA_APP: doc_id: 143425, title: A Day in the Life, score: 5.387871888967919
BIG_DATA_APP: doc_id: 66889766, title: A Breath of French Air, score: 5.387871888967919
BIG_DATA_APP: doc_id: 15344871, title: A Celtic Requiem, score: 5.387871888967919
root@cluster-master:/app#
```

The results are actually good it can be seen in the example with computer mouse titles (at the top they connected to computer)