NoSQL Assignment 2 – Part B

IMT2021008 – Sheikh Muteeb

IMT2021003 – Keshav Chandak

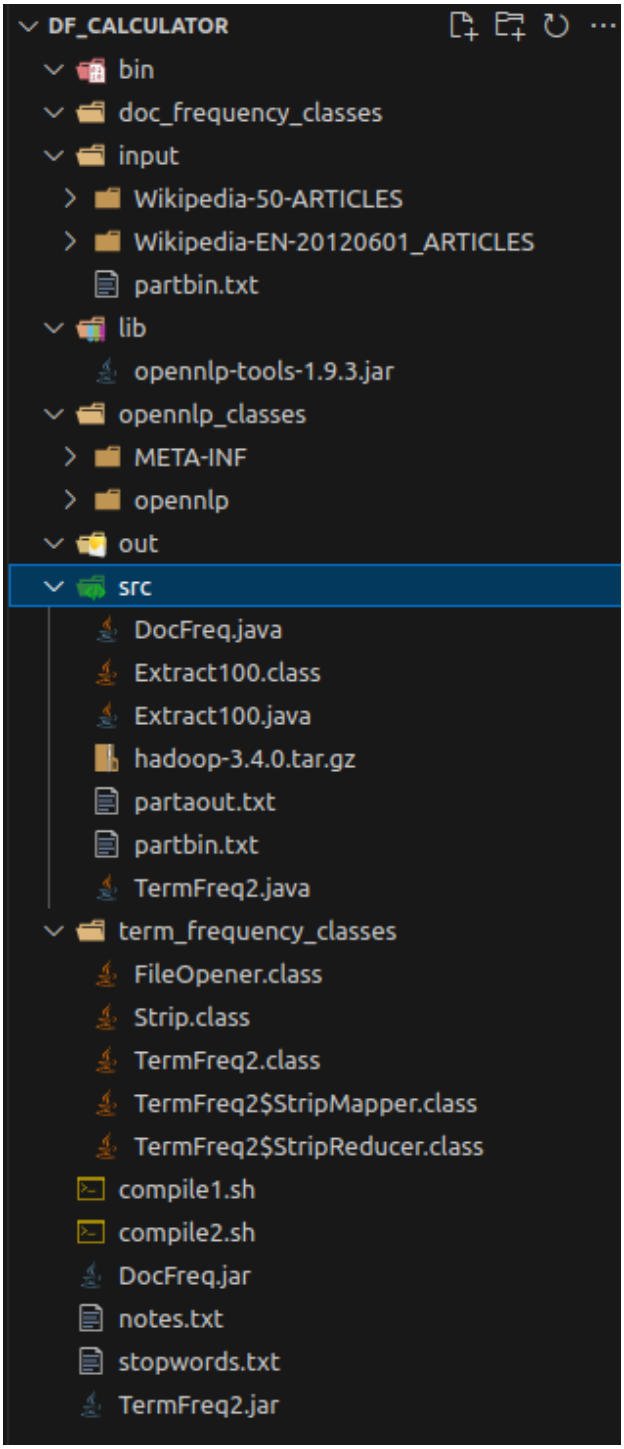
IMT2021007 – Sunny Kaushik

IMT2021076 – Devendara Rishi Nelapati

Problem 5

A

1. File Structure:



2. Outputs:

Worked on WIKI-50 and WIKI-EN-2012

```
2025-03-24 23:25:06,295 INFO mapred.Task: Final Counters for attempt_local321164020_0001_m_000048_0: Counters: 23
        File System Counters
                FILE: Number of bytes read=1458799
                FILE: Number of bytes written=2397317
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=714375
                HDFS: Number of bytes written=0
                HDFS: Number of read operations=101
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=1
                HDFS: Number of bytes read erasure-coded=0
        Map-Reduce Framework
                Map input records=1
                Map output records=232
                Map output bytes=1955
                Map output materialized bytes=2425
                Input split bytes=143
                Combine input records=0
                Spilled Records=232
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=0
                Total committed heap usage (bytes)=1700265984
        File Input Format Counters
                Bytes Read=3509
```

```
2025-03-24 23:25:06,313 INFO mapred.LocalJobRunner: map
2025-03-24 23:25:06,313 INFO mapred.Task: Task 'attempt_local321164020_0001_m_000049_0' done.
2025-03-24 23:25:06,314 INFO mapred.Task: Final Counters for attempt_local321164020_0001_m_000049_0: Counters: 23
        File System Counters
                FILE: Number of bytes read=1459347
                FILE: Number of bytes written=2399748
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=717789
                HDFS: Number of bytes written=0
                HDFS: Number of read operations=103
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=1
                HDFS: Number of bytes read erasure-coded=0
        Map-Reduce Framework
                Map input records=1
                Map output records=233
                Map output bytes=1927
                Map output materialized bytes=2399
                Input split bytes=143
                Combine input records=0
                Spilled Records=233
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=0
                Total committed heap usage (bytes)=1700265984
        File Input Format Counters
                Bytes Read=3414
```

```
        File System Counters
                FILE: Number of bytes read=72062577
                FILE: Number of bytes written=116534344
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=25893893
                HDFS: Number of bytes written=99515
                HDFS: Number of read operations=2808
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=53
                HDFS: Number of bytes read erasure-coded=0
        Map-Reduce Framework
                Map input records=50
                Map output records=32208
                Map output bytes=293631
                Map output materialized bytes=358347
                Input split bytes=7133
                Combine input records=0
                Combine output records=0
                Reduce input groups=12257
                Reduce shuffle bytes=358347
                Reduce input records=32208
                Reduce output records=12257
                Spilled Records=64416
                Shuffled Maps =50
                Failed Shuffles=0
                Merged Map outputs=50
                GC time elapsed (ms)=172
                Total committed heap usage (bytes)=67576528896
        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=717789
        File Output Format Counters
                Bytes Written=99515
protrigger99@protrigger99-Inspiron-14-5410:~/DF_Calculator$ hdfs dfs -ls /user
Found 1 items
drwxr-xr-x   - protrigger99 supergroup          0 2025-03-24 23:25 /user/protrigger99
```

```
2025-03-24 23:25:07,031 INFO mapred.Task: Task attempt_local321164020_0001_r_000000_0 is allowed to commit now
2025-03-24 23:25:07,043 INFO output.FileOutputCommitter: Saved output of task 'attempt_local321164020_0001_r_000000_0' to hdfs://localhost:9000/user/protrigger99/output
2025-03-24 23:25:07,043 INFO mapred.LocalJobRunner: reduce > reduce
2025-03-24 23:25:07,043 INFO mapred.Task: Task 'attempt_local321164020_0001_r_000000_0' done.
2025-03-24 23:25:07,043 INFO mapred.Task: Final Counters for attempt_local321164020_0001_r_000000_0: Counters: 30
        File System Counters
                FILE: Number of bytes read=2177347
                FILE: Number of bytes written=2757801
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=717789
                HDFS: Number of bytes written=99515
                HDFS: Number of read operations=108
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=3
                HDFS: Number of bytes read erasure-coded=0
        Map-Reduce Framework
                Combine input records=0
                Combine output records=0
                Reduce input groups=12257
                Reduce shuffle bytes=358347
                Reduce input records=32208
                Reduce output records=12257
                Spilled Records=32208
                Shuffled Maps =50
                Failed Shuffles=0
                Merged Map outputs=50
                GC time elapsed (ms)=0
                Total committed heap usage (bytes)=1700265984
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Output Format Counters
                Bytes Written=99515
2025-03-24 23:25:07,044 INFO mapred.LocalJobRunner: Finishing task: attempt_local321164020_0001_r_000000_0
2025-03-24 23:25:07,044 INFO mapred.LocalJobRunner: reduce task executor complete.
2025-03-24 23:25:07,848 INFO mapreduce.Job:  map 100% reduce 100%
2025-03-24 23:25:07,849 INFO mapreduce.Job: Job job_local321164020_0001 completed successfully
2025-03-24 23:25:07,872 INFO mapreduce.Job: Counters: 36
```

```
protrigger99@protrigger99-Inspiron-14-5410:~/DF_Calculator$ hadoop jar ./DocFreq.jar DocFreq input/Wikipedia-50-ARTICLES output
2025-03-24 23:22:24,147 INFO impl.MetricsConfig: Loaded properties from hadoop-metrics2.properties
2025-03-24 23:22:24,235 INFO impl.MetricsSystemImpl: Scheduled Metric snapshot period at 10 second(s).
2025-03-24 23:22:24,235 INFO impl.MetricsSystemImpl: JobTracker metrics system started
2025-03-24 23:22:24,352 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
2025-03-24 23:22:24,518 INFO input.FileInputFormat: Total input files to process : 50
2025-03-24 23:22:24,541 INFO mapreduce.JobSubmitter: number of splits:50
2025-03-24 23:22:24,634 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local1600834101_0001
2025-03-24 23:22:24,634 INFO mapreduce.JobSubmitter: Executing with tokens: []
2025-03-24 23:22:24,784 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
2025-03-24 23:22:24,785 INFO mapreduce.Job: Running job: job_local1600834101_0001
2025-03-24 23:22:24,785 INFO mapred.LocalJobRunner: OutputCommitter set in config null
2025-03-24 23:22:24,791 INFO output.PathOutputCommitterFactory: No output committer factory defined, defaulting to FileOutputCommitterFactory
2025-03-24 23:22:24,793 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2025-03-24 23:22:24,793 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
2025-03-24 23:22:24,793 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter
2025-03-24 23:22:24,853 INFO mapred.LocalJobRunner: Waiting for map tasks
2025-03-24 23:22:24,853 INFO mapred.LocalJobRunner: Starting task: attempt_local1600834101_0001_m_000000_0
2025-03-24 23:22:24,870 INFO output.PathOutputCommitterFactory: No output committer factory defined, defaulting to FileOutputCommitterFactory
2025-03-24 23:22:24,870 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2025-03-24 23:22:24,870 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
2025-03-24 23:22:24,887 INFO mapred.Task:  Using ResourceCalculatorProcessTree : [ ]
2025-03-24 23:22:24,891 INFO mapred.MapTask: Processing split: hdfs://localhost:9000/user/protrigger99/input/Wikipedia-50-ARTICLES/113147.txt:0+59262
2025-03-24 23:22:24,923 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
2025-03-24 23:22:24,924 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
2025-03-24 23:22:24,924 INFO mapred.MapTask: soft limit at 83886080
2025-03-24 23:22:24,924 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
2025-03-24 23:22:24,924 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
2025-03-24 23:22:24,926 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuffer
2025-03-24 23:22:25,122 INFO mapred.LocalJobRunner:
2025-03-24 23:22:25,124 INFO mapred.MapTask: Starting flush of map output
2025-03-24 23:22:25,124 INFO mapred.MapTask: Spilling map output
2025-03-24 23:22:25,124 INFO mapred.MapTask: bufstart = 0; bufend = 17835; bufvoid = 104857600
2025-03-24 23:22:25,124 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 26206900(104827600); length = 7497/6553600
2025-03-24 23:22:25,143 INFO mapred.MapTask: Finished spill 0
2025-03-24 23:22:25,155 INFO mapred.Task: Task:attempt_local1600834101_0001_m_000000_0 is done. And is in the process of committing
2025-03-24 23:22:25,158 INFO mapred.LocalJobRunner: map
2025-03-24 23:22:25,158 INFO mapred.Task: Task 'attempt_local1600834101_0001_m_000000_0' done.
2025-03-24 23:22:25,168 INFO mapred.Task: Final Counters for attempt_local1600834101_0001_m_000000_0: Counters: 23
       File System Counters
              FILE: Number of bytes read=1304931
              FILE: Number of bytes written=2064862
              FILE: Number of read operations=0
              FILE: Number of large read operations=0
              FILE: Number of write operations=0
              HDFS: Number of bytes read=59262
              HDFS: Number of bytes written=0
              HDFS: Number of read operations=5
              HDFS: Number of large read operations=0
              HDFS: Number of write operations=1
              HDFS: Number of bytes read erasure-coded=0
```

```
        File Input Format Counters
                Bytes Read=3414
2025-03-24 23:25:06,314 INFO mapred.LocalJobRunner: Finishing task: attempt_local321164020_0001_m_000049_0
2025-03-24 23:25:06,314 INFO mapred.LocalJobRunner: map task executor complete.
2025-03-24 23:25:06,318 INFO mapred.LocalJobRunner: Waiting for reduce tasks
2025-03-24 23:25:06,318 INFO mapred.LocalJobRunner: Starting task: attempt_local321164020_0001_r_000000_0
2025-03-24 23:25:06,324 INFO output.PathOutputCommitterFactory: No output committer factory defined, defaulting to FileOutputCommitterFactory
2025-03-24 23:25:06,324 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2025-03-24 23:25:06,324 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
2025-03-24 23:25:06,324 INFO mapred.Task:  Using ResourceCalculatorProcessTree : [ ]
2025-03-24 23:25:06,326 INFO mapred.ReduceTask: Using ShuffleConsumerPlugin: org.apache.hadoop.mapreduce.task.reduce.Shuffle@7f47e05c
2025-03-24 23:25:06,327 WARN impl.MetricsSystemImpl: JobTracker metrics system already initialized!
2025-03-24 23:25:06,341 INFO reduce.MergeManagerImpl: MergerManager: memoryLimit=2568277248, maxSingleShuffleLimit=642069312, mergeThreshold=1695063040, ioSortFactor=10, memToMemMergeOutputsThreshold=10
2025-03-24 23:25:06,343 INFO reduce.EventFetcher: attempt_local321164020_0001_r_000000_0 Thread started: EventFetcher for fetching Map Completion Events
2025-03-24 23:25:06,367 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of map attempt_local321164020_0001_m_000024_0 decomp: 6349 len: 6353 to MEMORY
2025-03-24 23:25:06,369 INFO reduce.InMemoryMapOutput: Read 6349 bytes from map-output for attempt_local321164020_0001_m_000024_0
2025-03-24 23:25:06,370 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 6349, inMemoryMapOutputs.size() -> 1, commitMemory -> 0, usedMemory ->6349
2025-03-24 23:25:06,372 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of map attempt_local321164020_0001_m_000025_0 decomp: 5631 len: 5635 to MEMORY
2025-03-24 23:25:06,373 INFO reduce.InMemoryMapOutput: Read 5631 bytes from map-output for attempt_local321164020_0001_m_000025_0
2025-03-24 23:25:06,373 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 5631, inMemoryMapOutputs.size() -> 2, commitMemory -> 6349, usedMemory ->11980
2025-03-24 23:25:06,374 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of map attempt_local321164020_0001_m_000000_0 decomp: 21587 len: 21591 to MEMORY
2025-03-24 23:25:06,374 INFO reduce.InMemoryMapOutput: Read 21587 bytes from map-output for attempt_local321164020_0001_m_000000_0
2025-03-24 23:25:06,374 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 21587, inMemoryMapOutputs.size() -> 3, commitMemory -> 11980, usedMemory ->33567
2025-03-24 23:25:06,375 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of map attempt_local321164020_0001_m_000026_0 decomp: 5527 len: 5531 to MEMORY
2025-03-24 23:25:06,375 INFO reduce.InMemoryMapOutput: Read 5527 bytes from map-output for attempt_local321164020_0001_m_000026_0
2025-03-24 23:25:06,375 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 5527, inMemoryMapOutputs.size() -> 4, commitMemory -> 33567, usedMemory ->39094
2025-03-24 23:25:06,376 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of map attempt_local321164020_0001_m_000001_0 decomp: 17815 len: 17819 to MEMORY
2025-03-24 23:25:06,376 INFO reduce.InMemoryMapOutput: Read 17815 bytes from map-output for attempt_local321164020_0001_m_000001_0
2025-03-24 23:25:06,376 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 17815, inMemoryMapOutputs.size() -> 5, commitMemory -> 39094, usedMemory ->56909
2025-03-24 23:25:06,377 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of map attempt_local321164020_0001_m_000027_0 decomp: 5561 len: 5565 to MEMORY
2025-03-24 23:25:06,377 INFO reduce.InMemoryMapOutput: Read 5561 bytes from map-output for attempt_local321164020_0001_m_000027_0
2025-03-24 23:25:06,377 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 5561, inMemoryMapOutputs.size() -> 6, commitMemory -> 56909, usedMemory ->62470
2025-03-24 23:25:06,378 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of map attempt_local321164020_0001_m_000021_0 decomp: 6307 len: 6311 to MEMORY
2025-03-24 23:25:06,378 INFO reduce.InMemoryMapOutput: Read 6307 bytes from map-output for attempt_local321164020_0001_m_000021_0
2025-03-24 23:25:06,378 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 6307, inMemoryMapOutputs.size() -> 7, commitMemory -> 62470, usedMemory ->68777
2025-03-24 23:25:06,379 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of map attempt_local321164020_0001_m_000047_0 decomp: 3248 len: 3252 to MEMORY
2025-03-24 23:25:06,379 INFO reduce.InMemoryMapOutput: Read 3248 bytes from map-output for attempt_local321164020_0001_m_000047_0
2025-03-24 23:25:06,379 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 3248, inMemoryMapOutputs.size() -> 8, commitMemory -> 68777, usedMemory ->72025
2025-03-24 23:25:06,380 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of map attempt_local321164020_0001_m_000022_0 decomp: 6615 len: 6619 to MEMORY
2025-03-24 23:25:06,380 INFO reduce.InMemoryMapOutput: Read 6615 bytes from map-output for attempt_local321164020_0001_m_000022_0
2025-03-24 23:25:06,380 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 6615, inMemoryMapOutputs.size() -> 9, commitMemory -> 72025, usedMemory ->78640
2025-03-24 23:25:06,381 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of map attempt_local321164020_0001_m_000048_0 decomp: 2421 len: 2425 to MEMORY
2025-03-24 23:25:06,382 INFO reduce.InMemoryMapOutput: Read 2421 bytes from map-output for attempt_local321164020_0001_m_000048_0
2025-03-24 23:25:06,382 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 2421, inMemoryMapOutputs.size() -> 10, commitMemory -> 78640, usedMemory ->81061
2025-03-24 23:25:06,383 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of map attempt_local321164020_0001_m_000023_0 decomp: 5954 len: 5958 to MEMORY
2025-03-24 23:25:06,383 INFO reduce.InMemoryMapOutput: Read 5954 bytes from map-output for attempt_local321164020_0001_m_000023_0
2025-03-24 23:25:06,384 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 5954, inMemoryMapOutputs.size() -> 11, commitMemory -> 81061, usedMemory ->87015
2025-03-24 23:25:06,385 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of map attempt_local321164020_0001_m_000049_0 decomp: 2395 len: 2399 to MEMORY
2025-03-24 23:25:06,385 INFO reduce.InMemoryMapOutput: Read 2395 bytes from map-output for attempt_local321164020_0001_m_000049_0
```

3. Execution Time:

Execution start time: 3:35:17,574

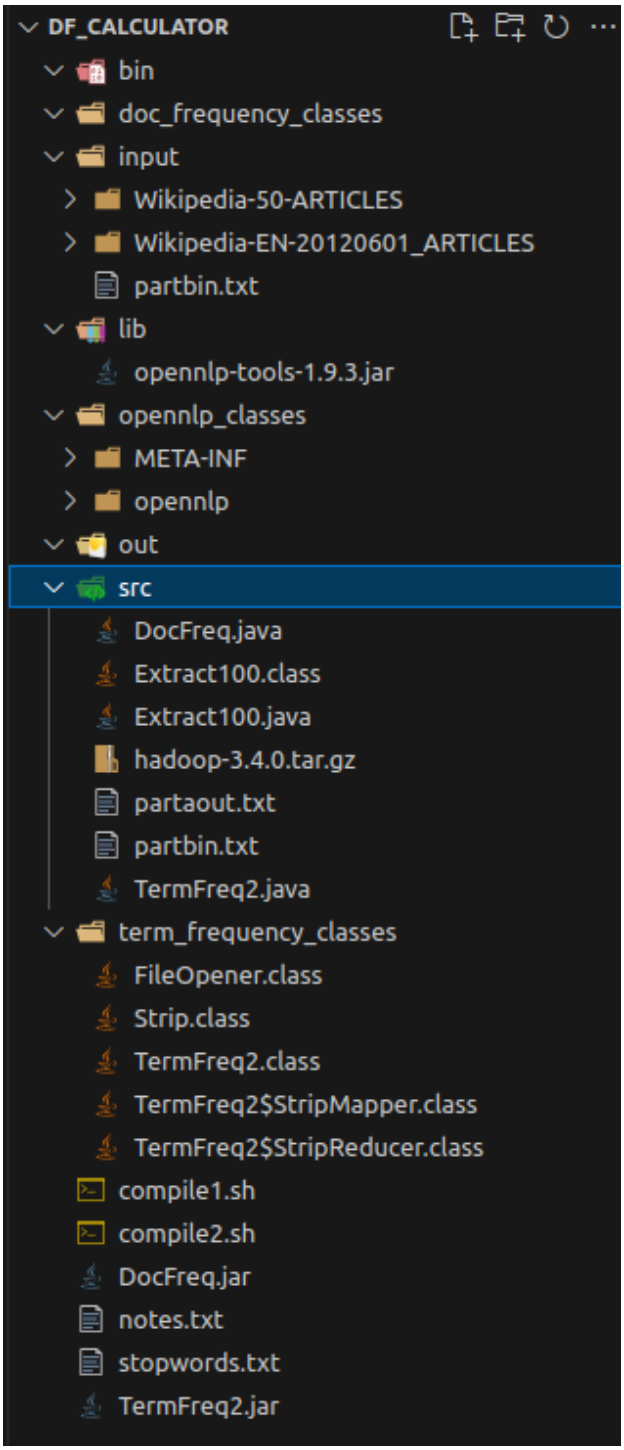
Execution end time and successful completion: 3:35:17:611

```
protrigger99@protrigger99-Inspiron-14-5410:~/DF_Calculator$ mapred job -list
2025-03-25 03:35:17,574 INFO impl.MetricsConfig: Loaded properties from hadoop-metrics2.properties
2025-03-25 03:35:17,611 INFO impl.MetricsSystemImpl: Scheduled Metric snapshot period at 10 second(s).
2025-03-25 03:35:17,611 INFO impl.MetricsSystemImpl: JobTracker metrics system started
```

Total Execution Time: **10 seconds.**

B

1. File Structure:



2. Outputs:

Output of Extract100.java which is the input for our Term Frequency Calculator:

```
1 d        50
2 h        50
3 re       50
4 l        50
5 referenc         49
6 m        49
7 wi       49
8 s        48
9 cy       48
10 ll      48
11 w       47
12 g       47
13 n       47
14 r       47
15 le      46
16 c       46
17 po      46
18 i       45
19 p       45
20 hd      44
21 e       44
22         44
23 4       44
24 quot    44
25 1       42
26 3       42
```

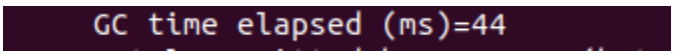
Outputs of TermFreq2.java:

```
2025-03-25 00:28:37,799 INFO mapred.LocalJobRunner:
2025-03-25 00:28:37,800 INFO mapred.MapTask: Starting flush of map output
2025-03-25 00:28:37,800 INFO mapred.MapTask: Spilling map output
2025-03-25 00:28:37,800 INFO mapred.MapTask: bufstart = 0; bufend = 411; bufvoid = 104857600
2025-03-25 00:28:37,800 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 26214396(104857584); length = 1/6553600
2025-03-25 00:28:37,806 INFO mapred.MapTask: Finished spill 0
2025-03-25 00:28:37,818 INFO mapred.Task: Task:attempt_local693412938_0001_m_000000_0 is done. And is in the process of committing
2025-03-25 00:28:37,820 INFO mapred.LocalJobRunner: map
2025-03-25 00:28:37,820 INFO mapred.Task: Task 'attempt_local693412938_0001_m_000000_0' done.
2025-03-25 00:28:37,830 INFO mapred.Task: Final Counters for attempt_local693412938_0001_m_000000_0: Counters: 23
        File System Counters
                FILE: Number of bytes read=1308604
                FILE: Number of bytes written=2049013
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=62459
                HDFS: Number of bytes written=0
                HDFS: Number of read operations=18
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=1
                HDFS: Number of bytes read erasure-coded=0
        Map-Reduce Framework
                Map input records=1
                Map output records=1
                Map output bytes=411
                Map output materialized bytes=421
                Input split bytes=143
                Combine input records=0
                Spilled Records=1
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=11
                Total committed heap usage (bytes)=367525888
        File Input Format Counters
                Bytes Read=59262
2025-03-25 00:28:37,830 INFO mapred.LocalJobRunner: Finishing task: attempt_local693412938_0001_m_000000_0
2025-03-25 00:28:37,830 INFO mapred.LocalJobRunner: Starting task: attempt_local693412938_0001_m_000001_0
2025-03-25 00:28:37,831 INFO output.PathOutputCommitterFactory: No output committer factory defined, defaulting to FileOutputCommitterFactory
2025-03-25 00:28:37,831 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2025-03-25 00:28:37,831 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
2025-03-25 00:28:37,832 INFO mapred.Task:  Using ResourceCalculatorProcessTree : [ ]
2025-03-25 00:28:37,833 INFO mapred.MapTask: Processing split: hdfs://localhost:9000/user/protrigger99/input/Wikipedia-50-ARTICLES/21197.txt:0+43114
2025-03-25 00:28:37,856 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
2025-03-25 00:28:37,856 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
2025-03-25 00:28:37,856 INFO mapred.MapTask: soft limit at 83886080
2025-03-25 00:28:37,856 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
2025-03-25 00:28:37,856 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
2025-03-25 00:28:37,856 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuffer
```

```
2025-03-25 00:28:38,719 INFO mapred.LocalJobRunner:
2025-03-25 00:28:38,719 INFO mapred.MapTask: Starting flush of map output
2025-03-25 00:28:38,719 INFO mapred.MapTask: Spilling map output
2025-03-25 00:28:38,719 INFO mapred.MapTask: bufstart = 0; bufend = 411; bufvoid = 104857600
2025-03-25 00:28:38,719 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 26214396(104857584); length = 1/6553600
2025-03-25 00:28:38,720 INFO mapred.MapTask: Finished spill 0
2025-03-25 00:28:38,720 INFO mapred.Task: Task:attempt_local693412938_0001_m_000032_0 is done. And is in the process of committing
2025-03-25 00:28:38,722 INFO mapred.LocalJobRunner: map
2025-03-25 00:28:38,722 INFO mapred.Task: Task 'attempt_local693412938_0001_m_000032_0' done.
2025-03-25 00:28:38,722 INFO mapred.Task: Final Counters for attempt_local693412938_0001_m_000032_0: Counters: 23
        File System Counters
                FILE: Number of bytes read=1435784
                FILE: Number of bytes written=2063496
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=621122
                HDFS: Number of bytes written=0
                HDFS: Number of read operations=82
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=1
                HDFS: Number of bytes read erasure-coded=0
        Map-Reduce Framework
                Map input records=1
                Map output records=1
                Map output bytes=411
                Map output materialized bytes=421
                Input split bytes=143
                Combine input records=0
                Spilled Records=1
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=0
                Total committed heap usage (bytes)=1114112000
        File Input Format Counters
                Bytes Read=8326
2025-03-25 00:28:38,722 INFO mapred.LocalJobRunner: Finishing task: attempt_local693412938_0001_m_000032_0
2025-03-25 00:28:38,722 INFO mapred.LocalJobRunner: Starting task: attempt_local693412938_0001_m_000033_0
2025-03-25 00:28:38,723 INFO output.PathOutputCommitterFactory: No output committer factory defined, defaulting to FileOutputCommitterFactory
2025-03-25 00:28:38,723 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2025-03-25 00:28:38,723 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
2025-03-25 00:28:38,723 INFO mapred.Task:  Using ResourceCalculatorProcessTree : [ ]
2025-03-25 00:28:38,724 INFO mapred.MapTask: Processing split: hdfs://localhost:9000/user/protrigger99/input/Wikipedia-50-ARTICLES/115846.txt:0+7907
2025-03-25 00:28:38,729 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
2025-03-25 00:28:38,730 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
2025-03-25 00:28:38,730 INFO mapred.MapTask: soft limit at 83886080
2025-03-25 00:28:38,730 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
2025-03-25 00:28:38,730 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
2025-03-25 00:28:38,730 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuffer
```

```
protrigger99@protrigger99-Inspiron-14-5410:~/DF_Calculator$ hadoop jar ./TermFreq2.jar TermFreq2 input/Wikipedia-50-ARTICLES output input/stopwords.txt
2025-03-25 00:28:36,840 INFO impl.MetricsConfig: Loaded properties from hadoop-metrics2.properties
2025-03-25 00:28:36,924 INFO impl.MetricsSystemImpl: Scheduled Metric snapshot period at 10 second(s).
2025-03-25 00:28:36,924 INFO impl.MetricsSystemImpl: JobTracker metrics system started
2025-03-25 00:28:37,035 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
2025-03-25 00:28:37,203 INFO input.FileInputFormat: Total input files to process : 50
2025-03-25 00:28:37,225 INFO mapreduce.JobSubmitter: number of splits:50
2025-03-25 00:28:37,313 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local693412938_0001
2025-03-25 00:28:37,313 INFO mapreduce.JobSubmitter: Executing with tokens: []
2025-03-25 00:28:37,532 INFO mapred.LocalDistributedCacheManager: Localized hdfs://localhost:9000/user/protrigger99/input/stopwords.txt as file:/tmp/hadoop-protrigger99/mapred/local/job_local693412938_0001_7c68f
cbd-6277-41fa-b9bb-76cea1de91a3/stopwords.txt
2025-03-25 00:28:37,603 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
2025-03-25 00:28:37,603 INFO mapreduce.Job: Running job: job_local693412938_0001
2025-03-25 00:28:37,604 INFO mapred.LocalJobRunner: OutputCommitter set in config null
2025-03-25 00:28:37,608 INFO output.PathOutputCommitterFactory: No output committer factory defined, defaulting to FileOutputCommitterFactory
2025-03-25 00:28:37,609 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2025-03-25 00:28:37,609 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
2025-03-25 00:28:37,610 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter
2025-03-25 00:28:37,678 INFO mapred.LocalJobRunner: Waiting for map tasks
2025-03-25 00:28:37,678 INFO mapred.LocalJobRunner: Starting task: attempt_local693412938_0001_m_000000_0
2025-03-25 00:28:37,696 INFO output.PathOutputCommitterFactory: No output committer factory defined, defaulting to FileOutputCommitterFactory
2025-03-25 00:28:37,696 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2025-03-25 00:28:37,697 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
2025-03-25 00:28:37,709 INFO mapred.Task:  Using ResourceCalculatorProcessTree : [ ]
2025-03-25 00:28:37,713 INFO mapred.MapTask: Processing split: hdfs://localhost:9000/user/protrigger99/input/Wikipedia-50-ARTICLES/113147.txt:0+59262
2025-03-25 00:28:37,746 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
2025-03-25 00:28:37,746 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
2025-03-25 00:28:37,746 INFO mapred.MapTask: soft limit at 83886080
2025-03-25 00:28:37,746 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
2025-03-25 00:28:37,746 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
2025-03-25 00:28:37,749 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuffer
```

3.Execution Time:



Total Execution Time: **4.4 seconds.**

4.Conclusions: Document Frequency (DF) Calculation using MapReduce

Execution Time Comparison:

- The execution time of the MapReduce job is influenced by the number of documents and the overhead of processing stopwords and stemming. (~10 seconds for WIKI-50)
- The job successfully filters stopwords and applies the Porter Stemmer, but these operations introduce additional computational costs.

Mapper vs. Reducer Workload:

- The **Mapper** handles a significant preprocessing workload, including tokenization, stopword removal, and stemming.
- The **Reducer** aggregates term frequencies across multiple documents, producing the final document frequency counts.

Scaling with Data Size:

- The execution time increases as the number of documents grows, but the system scales effectively due to Hadoop's distributed processing.
- However, **shuffling and sorting in the reducer phase** contribute to increased execution time, especially for high-frequency words.

Performance Considerations:

- Using **Combiners** can help optimize performance by reducing the amount of intermediate data shuffled across the network.
- The **choice of partitioning strategy** can influence load balancing, preventing skewed reducers.

Overall Takeaways:

- **Stopword filtering and stemming improve data quality** but add computation time.
- **Scaling is generally effective**, though reducer-side processing can become a bottleneck.
- **Optimizations like combiners and efficient partitioning** can improve performance.