

# NoSQL Assignment 2 Part B

Vishruth Vijay IMT2022507  
Siddharth Menon IMT2022001  
Sathvik S Rao IMT2022082  
Shreyas Arun Saggere IMT2022006

March 24, 2025

## Problem 4

(a)

The goal of this question was to identify the top 50 most frequently occurring words from a Wikipedia dump while eliminating stop words. The solution was built upon the WordCount2 implementation, incorporating distributed caching to filter out stop words during the mapping phase.

### Mapper (TokenizerMapper)

The Mapper class processes input text to extract meaningful words while filtering out stop words.

- **Stop Words Filtering:** The Distributed Cache is used to load stop words from an external file. These are stored in a HashSet for fast lookup.
- **Tokenization:** The input text is split using the regex pattern `[^\w']+` to isolate valid words. This regex pattern was already part of the WordCount2 code and hasn't been changed.

### Mapping Process

The Mapper performs the following steps:

1. Reads each line from the input.
2. Converts text to lowercase.
3. Splits the text into tokens and removes stop words.
4. Emits each valid word with a count of 1 in the form: `(word, 1)`

## Reducer (IntSumReducer)

The Reducer class aggregates the word counts and extracts the top 50 most frequent words.

- **Counting Words:** Each word's frequency is aggregated using a HashMap.
- **Sorting Logic:** A Priority Queue is used to efficiently track the top 50 words.
- **Final Output:** The top 50 words are written to the output.

## Reduction Process

The Reducer executes the following steps:

1. Aggregates word counts for each key.
2. Stores results in a HashMap.
3. Inserts all word-frequency pairs into a Priority Queue.
4. Extracts and writes the top 50 most frequent words.

## Driver Program

The Driver program initializes and configures the Hadoop job. The following key settings are configured:

- Input and output paths are specified.
- The stop words file is added to the Distributed Cache using:  
`job.addCacheFile(new Path(args[i]).toUri());`
- Optional case sensitivity is enabled using the argument `-casesensitive`. This was part of the WordCount2 java code, while running our hadoop task, we do not pass an argument for this, and hence runs on default non-case sensitivity.

(b)

The objective of this question was to construct a **co-occurrence word matrix** based on the **frequent words** identified in the previous question. The program employs the **Pairs Approach** to compute co-occurrences within a specified **window distance  $d$** . The solution filters out less frequent words using a precomputed stop-word list, ensuring that only significant word pairs are considered.

```

united 28742
time 28324
2009 28117
county 27588
world 27373
2007 27239
university 27012
states 26787
7 26075
8 25424
n 25423
d 24165
years 24101
people 24027
m 23814
9 23502
used 23492
war 23049
history 23038
state 22967
12 21385
uk 21373
11 21060

```

Figure 1: Top 50 words based on frequency of occurrence

### Mapper (PairsMapper)

The Mapper class reads input text and constructs co-occurrence pairs.

- **Frequent Words Filtering:** A Distributed Cache file is used to load frequent words, ensuring only significant words are included in co-occurrence calculations.
- **Adjustable Window Size:** The parameter  $d$  determines the maximum distance within which two words are considered co-occurring.
- **Pairs Generation:** Word pairs are emitted in the format `(word1, word2, 1)`.

### Mapping Process

1. The input line is tokenized into words.
2. Each word is cleaned (punctuation removal, lowercase conversion).
3. If the word is in the **frequent words set**, a co-occurrence window is created around it.
4. Co-occurring word pairs within the window distance  $d$  are emitted.

## Reducer (PairsReducer)

The Reducer class aggregates the frequency of each word pair.

- **Summing Pair Counts:** For each word pair, the frequency is summed across all input splits.
- **Final Output:** The total count of each pair is written to the output.

## Reduction Process

1. The Reducer receives key-value pairs of the form  $(word1, word2), count$ .
2. The total frequency for each pair is computed.
3. The final count is emitted as output.

## Driver Program

The Driver program configures the Hadoop job with the following settings:

- The input and output paths are specified.
- The frequent words file is added to the **Distributed Cache**.
- The window distance  $d$  is set using:

```
conf.setInt("windowDistance", windowDistance);
```

- The Mapper and Reducer classes are registered.
- The job is submitted to Hadoop for execution.

## Runtime Comparison

Window Distance ( $d$ )	Execution Time (seconds)
1	173.536
2	177.309
3	179.853
4	187.254

Table 1: Runtime Analysis for Different Window Distances

```

Reduce output records=785
Spilled Records=116954
Shuffled Maps =10000
Failed Shuffles=0
Merged Map outputs=10000
GC time elapsed (ms)=6
Total committed heap usage (bytes)=3527409664
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=10231
2025-03-24 14:59:57 INFO mapred.LocalJobRunner: Finishing task: attempt_local454038973_0001_r_000000_0
2025-03-24 14:59:57.071 INFO mapred.LocalJobRunner: reduce task executor complete.
2025-03-24 14:59:57.895 INFO mapreduce.Job: map 100% reduce 100%
2025-03-24 14:59:58.896 INFO mapreduce.Job: Job job_local454038973_0001 completed successfully
2025-03-24 14:59:59.139 INFO mapreduce.Job: Counters: 36
File System Counters
FILE: Number of bytes read=480448606574
FILE: Number of bytes written=39106287061
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=1195415722606
HDFS: Number of bytes written=10231
HDFS: Number of read operations=100260028
HDFS: Number of large read operations=0
HDFS: Number of write operations=10003
HDFS: Number of bytes read erasure-coded=0
Map-Reduce Framework
Map input records=10000
Map output records=116954
Map output bytes=1797488
Map output serialized bytes=2091396
Input split bytes=1366997
Combine input records=0
Combine output records=0
Reduce input groups=785
Reduce input records=785
Reduce shuffle bytes=2091396
Reduce input records=116954
Reduce output records=785
Spilled Records=233908
Shuffled Maps =10000
Failed Shuffles=0
Merged Map outputs=10000
GC time elapsed (ms)=5137
Total committed heap usage (bytes)=35040133644288
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=166222646
File Output Format Counters
Bytes Written=10231
real    1m45.526s
user    2m53.501s
sys     0m10.303s
hadoop@pop-os: ~
```

Figure 2: The figure represents the output for pairs approach ( $d=1$ )

```

Reduce output records=849
Spilled Records=272444
Shuffled Maps =10000
Failed Shuffles=0
Merged Map outputs=10000
GC time elapsed (ms)=8
Total committed heap usage (bytes)=3893886976
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=11471
2025-03-24 15:02:11,620 INFO mapred.LocalJobRunner: Finishing task: attempt_local2010567203_0001_r_000000_0
2025-03-24 15:02:11,620 INFO mapred.LocalJobRunner: reduce task executor complete.
2025-03-24 15:02:11,944 INFO mapreduce.Job: map 100% reduce 100%
2025-03-24 15:02:12,945 INFO mapreduce.Job: Job job_local2010567203_0001 completed successfully
2025-03-24 15:02:13,203 INFO mapreduce.Job: Counters: 36
File System Counters
FILE: Number of bytes read=480453560354
FILE: Number of bytes written=56821644260
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=1195415722606
HDFS: Number of bytes written=11471
HDFS: Number of read operations=100260028
HDFS: Number of large read operations=0
HDFS: Number of write operations=10003
HDFS: Number of bytes read erasure-coded=0
Map-Reduce Framework
Map input records=10000
Map output records=272444
Map input bytes=3967838
Map output serialized bytes=4572786
Input split bytes=1366997
Combine input records=0
Combine output records=0
Reduce input groups=849
Reduce shuffle bytes=4572786
Reduce input records=272444
Reduce output records=849
Spilled Records=544888
Shuffled Maps =10000
Failed Shuffles=0
Merged Map outputs=10000
GC time elapsed (ms)=6247
Total committed heap usage (bytes)=37967991144448
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=166222646
File Output Format Counters
Bytes Written=11471
real    1m46.383s
user    2m57.309s
sys     0m10.187s
hadoop@pop-os: ~
```

Figure 3: The figure represents the output for pairs approach ( $d=2$ )

```

Reduce output records=859
Spilled Records=423468
Shuffled Maps =10000
Failed Shuffles=0
Merged Map outputs=10000
GC time elapsed (ms)=4
Total committed heap usage (bytes)=3704094720
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=11851
2025-03-24 15:04:22 INFO mapred.LocalJobRunner: Finishing task: attempt_local262628610_0001_r_000000_0
2025-03-24 15:04:22,690 INFO mapred.LocalJobRunner: reduce task executor complete.
2025-03-24 15:04:22,845 INFO mapreduce.Job: map 100% reduce 100%
2025-03-24 15:04:23,845 INFO mapreduce.Job: Job job_local262628610_0001 completed successfully
2025-03-24 15:04:24,092 INFO mapreduce.Job: Counters: 36
File System Counters
FILE: Number of bytes read=480458445258
FILE: Number of bytes written=73751216831
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=1195415722606
HDFS: Number of bytes written=11851
HDFS: Number of read operations=100260028
HDFS: Number of large read operations=0
HDFS: Number of write operations=10003
HDFS: Number of bytes read erasure-coded=0
Map-Reduce Framework
Map input records=10000
Map output records=423468
Map input bytes=6103738
Map output serialized bytes=7010734
Input split bytes=1366997
Combine input records=0
Combine output records=0
Reduce input groups=859
Reduce shuffle bytes=7010734
Reduce input records=423468
Reduce output records=859
Spilled Records=846936
Shuffled Maps =10000
Failed Shuffles=0
Merged Map outputs=10000
GC time elapsed (ms)=6703
Total committed heap usage (bytes)=35381491793920
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=166222646
File Output Format Counters
Bytes Written=11851
real    1m47.867s
user    2m59.853s
sys     0m10.011s
hadoop@pop-os: ~ ]
```

Figure 4: The figure represents the output for pairs approach (d=3)

```

Reduce output records=865
Spilled Records=561864
Shuffled Maps =10000
Failed Shuffles=0
Merged Map outputs=10000
GC time elapsed (ms)=15
Total committed heap usage (bytes)=3927965696

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=12071

2025-03-24 15:06:36 INFO mapred.LocalJobRunner: Finishing task: attempt_local1525689069_0001_r_000000_0
2025-03-24 15:06:36:232 INFO mapred.LocalJobRunner: reduce task executor complete.
2025-03-24 15:06:36:757 INFO mapreduce.Job: map 100% reduce 100%
2025-03-24 15:06:37:757 INFO mapreduce.Job: Job job_local1525689069_0001 completed successfully
2025-03-24 15:06:37:987 INFO mapreduce.Job: Counters: 36

File System Counters
FILE: Number of bytes read=480462901882
FILE: Number of bytes written=89548597884
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=1195415722606
HDFS: Number of bytes written=12071
HDFS: Number of read operations=100260028
HDFS: Number of large read operations=0
HDFS: Number of write operations=10003
HDFS: Number of bytes read erasure-coded=0

Map-Reduce Framework
Map input records=10000
Map output records=561864
Map output bytes=8055322
Map output serialized bytes=9239050
Input split bytes=1366997
Combine input records=0
Combine output records=0
Reduce input groups=865
Reduce shuffle bytes=9239050
Reduce input records=561864
Reduce output records=865
Spilled Records=1123728
Shuffled Maps =10000
Failed Shuffles=0
Merged Map outputs=10000
GC time elapsed (ms)=6157
Total committed heap usage (bytes)=36702326882304

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=166222646
File Output Format Counters
Bytes Written=12071

real    1m49.005s
user    3m7.254s
sys     0m10.151s
hadoop@pop-os: ~
```

Figure 5: The figure represents the output for pairs approach (d=4)

### (c)

The goal of the question was to construct a **co-occurrence word matrix** using the **Stripe Algorithm**. The implementation considers only the **frequent words** identified in the subpart (a) and computes co-occurring word pairs within a specified **window distance  $d$** . This approach leverages a more efficient data structure compared to the Pairs Approach, reducing the number of intermediate key-value pairs transmitted.

#### Mapper (StripesMapper)

The Mapper processes the input text to identify co-occurring words within a given window distance.

- **Frequent Words Filtering:** A Distributed Cache file is loaded to retain only important words.
- **Window-Based Co-occurrence:** Each word forms a local co-occurrence matrix (stripe) with its neighbors.
- **Efficient Data Representation:** Instead of emitting individual word pairs, a single stripe (hashmap) is emitted per word.

#### Mapping Process

1. The input line is tokenized into words.
2. Words are cleaned (punctuation removal, lowercase conversion).
3. If the word is frequent, a **co-occurrence stripe** is created.
4. Co-occurrence counts are updated in the stripe.
5. The word and its stripe are emitted in the format:

```
(word1, {word2: count, word3: count, ...})
```

#### Reducer (StripesReducer)

The Reducer merges stripes from multiple mappers, ensuring that final co-occurrence counts are computed.

- **Stripe Merging:** Partial counts from multiple stripes are aggregated.
- **Final Output Format:** The final output contains each word and its corresponding co-occurrence matrix.

## Reduction Process

1. Receives key-value pairs where each key is a word and each value is a stripe.
2. Merges all stripes for the word.
3. Outputs the final co-occurrence matrix for each word.

## Driver Program

The Driver program initializes the Hadoop job with the following configurations:

- Input and output paths are specified.
- The frequent words file is added to the **Distributed Cache**.
- The window distance  $d$  is set dynamically using:  
`conf.setInt("windowDistance", windowDistance);`
- The Mapper and Reducer classes are registered.
- The job is submitted to Hadoop for execution.

## Runtime Comparison

Window Distance ( $d$ )	Execution Time (seconds)
1	172.412
2	182.862
3	192.073
4	199.560

Table 2: Runtime Analysis for Different Window Distances

(d)

The goal of this question is to provide a comparative analysis of local aggregation techniques applied to the **Pairs and Stripes Approaches** for constructing a **co-occurrence word matrix**. The primary objective is to optimize computation by reducing the number of intermediate key-value pairs generated during the Map phase. Two levels of local aggregation are implemented:

- **Map-Function Level Aggregation:** Aggregates data within a single call to the `map()` function.
- **Map-Class Level Aggregation:** Aggregates data across multiple calls to `map()` within the same Mapper instance.

```

Reduce output records=31
Spilled Records=112819
Shuffled Maps =10000
Failed Shuffles=0
Merged Map outputs=10000
GC time elapsed (ms)=9
Total committed heap usage (bytes)=3595567104
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=7148
2025-03-24 14:47:30,272 INFO mapred.LocalJobRunner: Finishing task: attempt_local1520207428_0001_r_000000_0
2025-03-24 14:47:30,272 INFO mapred.LocalJobRunner: reduce task executor complete.
2025-03-24 14:47:30,272 INFO mapreduce.Job: map 100% reduce 100%
2025-03-24 14:47:31,273 INFO mapreduce.Job: Job job_local1520207428_0001 completed successfully
2025-03-24 14:47:31,513 INFO mapreduce.Job: Counters: 36
File System Counters
FILE: Number of bytes read=480454339353
FILE: Number of bytes written=44523860605
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=1195415722606
HDFS: Number of bytes written=7148
HDFS: Number of read operations=100260028
HDFS: Number of large read operations=0
HDFS: Number of write operations=10003
HDFS: Number of bytes read erasure-coded=0
Map-Reduce Framework
Map input records=10000
Map output records=112819
Map output bytes=2566937
Map output serialized bytes=2852575
Input split bytes=1366997
Combine input records=0
Combine output records=0
Reduce input groups=31
Reduce shuffle bytes=2852575
Reduce input records=112819
Reduce output records=31
Spilled Records=225638
Shuffled Maps =10000
Failed Shuffles=0
Merged Map outputs=10000
GC time elapsed (ms)=5030
Total committed heap usage (bytes)=35672820285440
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=166222646
File Output Format Counters
Bytes Written=7148
real    1m49.522s
user    2m52.412s
sys     0m10.457s
hadoop@pop-os: ~
```

Figure 6: The figure represents the output for stripes approach (d=1)

```

Reduce output records=31
Spilled Records=241987
Shuffled Maps =10000
Failed Shuffles=0
Merged Map outputs=10000
GC time elapsed (ms)=7
Total committed heap usage (bytes)=4108845056
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=8102
2025-03-24 14:49:47,786 INFO mapred.LocalJobRunner: Finishing task: attempt_local1025784682_0001_r_000000_0
2025-03-24 14:49:47,786 INFO mapred.LocalJobRunner: reduce task executor complete.
2025-03-24 14:49:47,786 INFO mapreduce.Job: map 100% reduce 100%
2025-03-24 14:49:48,266 INFO mapreduce.Job: Job job_local1025784682_0001 completed successfully
2025-03-24 14:49:48,498 INFO mapreduce.Job: Counters: 36
File System Counters
FILE: Number of bytes read=48046058329
FILE: Number of bytes written=668707779100
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=1195415722606
HDFS: Number of bytes written=8102
HDFS: Number of read operations=100260028
HDFS: Number of large read operations=0
HDFS: Number of write operations=10003
HDFS: Number of bytes read erasure-coded=0
Map-Reduce Framework
Map input records=10000
Map output records=241987
Map input bytes=5430544
Map output serialized bytes=5974518
Input split bytes=1366997
Combine input records=0
Combine output records=0
Reduce input groups=31
Reduce shuffle bytes=5974518
Reduce input records=241987
Reduce output records=31
Spilled Records=483974
Shuffled Maps =10000
Failed Shuffles=0
Merged Map outputs=10000
GC time elapsed (ms)=6700
Total committed heap usage (bytes)=38451518373888
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=166222646
File Output Format Counters
Bytes Written=8102
real    1m48.161s
user    3m2.862s
sys     0m10.334s
hadoop@pop-os: ~
```

Figure 7: The figure represents the output for stripes approach (d=2)

```

Reduce output records=31
Spilled Records=349828
Shuffled Maps =10000
Failed Shuffles=0
Merged Map outputs=10000
GC time elapsed (ms)=7
Total committed heap usage (bytes)=4108845056
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=8437
2025-03-24 14:52:41,025 INFO mapred.LocalJobRunner: Finishing task: attempt_local59528967_0001_r_000000_0
2025-03-24 14:52:41,025 INFO mapred.LocalJobRunner: reduce task executor complete.
2025-03-24 14:52:41,821 INFO mapreduce.Job: map 100% reduce 100%
2025-03-24 14:52:41,821 INFO mapreduce.Job: Job job_local59528967_0001 completed successfully
2025-03-24 14:52:42,044 INFO mapreduce.Job: Counters: 36
File System Counters
FILE: Number of bytes read=480466075721
FILE: Number of bytes written=85963867459
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=1195415722606
HDFS: Number of bytes written=8437
HDFS: Number of read operations=100260028
HDFS: Number of large read operations=0
HDFS: Number of write operations=10003
HDFS: Number of bytes read erasure-coded=0
Map-Reduce Framework
Map input records=10000
Map output records=349828
Map output bytes=7961103
Map output serialized bytes=8720759
Input split bytes=1366997
Combine input records=0
Combine output records=0
Reduce input groups=31
Reduce shuffle bytes=8720759
Reduce input records=349828
Reduce output records=31
Spilled Records=699656
Shuffled Maps =10000
Failed Shuffles=0
Merged Map outputs=10000
GC time elapsed (ms)=7638
Total committed heap usage (bytes)=37161486778368
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=166222646
File Output Format Counters
Bytes Written=8437
real    1m50.356s
user    3m12.073s
sys     0m10.566s
hadoop@pop-os: ~
```

Figure 8: The figure represents the output for stripes approach (d=3)

```

Reduce output records=31
Spilled Records=422262
Shuffled Maps =10000
Failed Shuffles=0
Merged Map outputs=10000
GC time elapsed (ms)=7
Total committed heap usage (bytes)=3503292416
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=8634
2025-03-24 14:55:00 INFO mapred.LocalJobRunner: Finishing task: attempt_local136493926_0001_r_000000_0
2025-03-24 14:55:00:230 INFO mapred.LocalJobRunner: reduce task executor complete.
2025-03-24 14:55:00:239 INFO mapreduce.Job: map 100% reduce 100%
2025-03-24 14:55:01:230 INFO mapreduce.Job: Job job_local136493926_0001 completed successfully
2025-03-24 14:55:01:483 INFO mapreduce.Job: Counters: 36
File System Counters
FILE: Number of bytes read=48047006819
FILE: Number of bytes written=100176370160
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=1195415722606
HDFS: Number of bytes written=8634
HDFS: Number of read operations=100260028
HDFS: Number of large read operations=0
HDFS: Number of write operations=10003
HDFS: Number of bytes read erasure-coded=0
Map-Reduce Framework
Map input records=10000
Map output records=422262
Map output bytes=90172434
Map output serialized bytes=10716958
Input split bytes=1366997
Combine input records=0
Combine output records=0
Reduce input groups=31
Reduce shuffle bytes=10716958
Reduce input records=422262
Reduce output records=31
Spilled Records=844524
Shuffled Maps =10000
Failed Shuffles=0
Merged Map outputs=10000
GC time elapsed (ms)=5612
Total committed heap usage (bytes)=34162382209024
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=166222646
File Output Format Counters
Bytes Written=8634
real    1m48.452s
user    2m59.560s
sys     0m10.047s
hadoop@pop-os: ~
```

Figure 9: The figure represents the output for stripes approach (d=4)

The runtime is analyzed for different window distances  $d = \{1, 2, 3, 4\}$ . The implementation introduces an **aggregationLevel** parameter to toggle between different aggregation strategies.

## Pairs Approach with Aggregation

The **PairsAggregation** program builds upon the traditional **Pairs Approach** that was implemented in (b), introducing local aggregation before emitting word pairs.

### Mapper (PairsMapper)

- **Function-Level Aggregation:** Uses a `HashMap` within each call to `map()` to store word pair counts before emitting them.
- **Class-Level Aggregation:** Stores counts in a `HashMap` across multiple calls to `map()`, only emitting pairs during `cleanup()`.

### Reducer (PairsReducer)

The Reducer sums all emitted word pair counts and outputs the final co-occurrence matrix.

## Stripes Approach with Aggregation

The **StripesAggregation** program optimizes the **Stripes Approach** that was implemented in (c) by introducing local aggregation.

### Mapper (StripesMapper)

- **Function-Level Aggregation:** A `MapWritable` (`hashmap`) is used within each call to `map()`.
- **Class-Level Aggregation:** A single `MapWritable` is maintained across multiple calls to `map()` and emitted in `cleanup()`.

### Reducer (StripesReducer)

Receives and merges multiple stripes to generate the final co-occurrence matrix.

Window Distance ( $d$ )	Pairs (Function-Level)	Pairs (Class-Level)	Stripes (Function-Level)	Stripes (Class-Level)
1	8m:41s	8m:41s	6m:9s	6m:35s
2	6m:38sec	6m:50s	6m:39s	6m:37s
3	6m:42sec	6m:43s	6m:43s	6m:43s
4	6m:46sec	6m:51s	6m:59s	6m:57s

Table 3: Runtime Analysis for Different Window Distances and Aggregation Levels

```

hadoop1@siddharth-VivoBook-ASUSLaptop-X415JAB-X415JA: /home/siddharth/Downloads/nosql/assn2_4_d_part2
hadoop1@siddharth-VivoBook-ASUSLaptop-X415J... x hadoop1@siddharth-VivoBook-ASUSLaptop-X415J... x hadoop1@siddharth-VivoBook-ASUSLaptop-X415J... x siddharth@siddharth-VivoBook-ASUSLaptop-X415J... x

FILE: Number of bytes read=449252496436
FILE: Number of bytes written=31247662336
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=1195415722666
HDFS: Number of bytes written=10231
HDFS: Number of read operations=100240026
HDFS: Number of large read operations=0
HDFS: Number of write operations=10063
HDFS: Number of bytes read erasure-coded=0
Map-Reduce Framework
  Map input records=10000
  Map output records=55572
  Map output bytes=821728
  Map output materialized bytes=992872
  Input split bytes=1044997
  Combine input records=0
  Combine output records=0
  Reduce input groups=785
  Reduce shuffle bytes=992872
  Reduce input records=55572
  Reduce output records=785
  Spilled Records=11114
  Shuffled Maps =10000
  Failed Shuffles=0
  Merged Map outputs=10000
  GC time elapsed (ms)=3399
  Total committed heap usage (bytes)=41202554503168
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=166222646
File Output Format Counters
  Bytes Written=10231

real    3m11.849s
user    8m41.215s
sys     0m20.065s

```

Figure 10: Pairs Approach with function level aggregation (d=1)

```

hadoop1@siddharth-VivoBook-ASUSLaptop-X415JAB-X415JA: /home/siddharth/Downloads/nosql/assn2_4_d_part2
hadoop1@siddharth-VivoBook-ASUSLaptop-X415J... x hadoop1@siddharth-VivoBook-ASUSLaptop-X415J... x hadoop1@siddharth-VivoBook-ASUSLaptop-X415J... x siddharth@siddharth-VivoBook-ASUSLaptop-X415J... x

FILE: Number of bytes read=449252496436
FILE: Number of bytes written=31247662336
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=1195415722666
HDFS: Number of bytes written=10231
HDFS: Number of read operations=100240026
HDFS: Number of large read operations=0
HDFS: Number of write operations=10063
HDFS: Number of bytes read erasure-coded=0
Map-Reduce Framework
  Map input records=10000
  Map output records=55572
  Map output bytes=821728
  Map output materialized bytes=992872
  Input split bytes=1044997
  Combine input records=0
  Combine output records=0
  Reduce input groups=785
  Reduce shuffle bytes=992872
  Reduce input records=55572
  Reduce output records=785
  Spilled Records=11114
  Shuffled Maps =10000
  Failed Shuffles=0
  Merged Map outputs=10000
  GC time elapsed (ms)=1981
  Total committed heap usage (bytes)=41217018560512
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=166222646
File Output Format Counters
  Bytes Written=10231

real    2m18.195s
user    6m38.749s
sys     0m16.233s

```

Figure 11: Pairs Approach with function level aggregation (d=2)

```

hadoop1@siddharth-VivoBook-ASUSLaptop-X415JAB-X415JA: /home/siddharth/Downloads/nosql/assn2_4_d_part2
hadoop1@siddharth-VivoBook-ASUSLaptop-X415J... x hadoop1@siddharth-VivoBook-ASUSLaptop-X415J... x hadoop1@siddharth-VivoBook-ASUSLaptop-X415J... x siddharth@siddharth-VivoBook-ASUSLaptop-X415J... x

FILE: Number of bytes read=44925398448
FILE: Number of bytes written=36066876416
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=1195415722606
HDFS: Number of bytes written=11471
HDFS: Number of read operations=100240026
HDFS: Number of large read operations=0
HDFS: Number of write operations=10063
HDFS: Number of bytes read erasure-coded=0
Map-Reduce Framework
  Map input records=10000
  Map output records=101396
  Map output bytes=1474086
  Map output materialized bytes=1736878
  Input split bytes=1044997
  Combine input records=0
  Combine output records=0
  Reduce input groups=849
  Reduce shuffle bytes=1736878
  Reduce input records=101396
  Reduce output records=849
  Spilled Records=202792
  Shuffled Maps =10000
  Failed Shuffles=0
  Merged Map outputs=10000
  GC time elapsed (ms)=2089
  Total committed heap usage (bytes)=41177116049408
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=166222646
File Output Format Counters
  Bytes Written=11471

real    2m18.701s
user    6m42.406s
sys     0m15.794s

```

Figure 12: Pairs Approach with function level aggregation (d=3)

```

hadoop1@siddharth-VivoBook-ASUSLaptop-X415JAB-X415JA: /home/siddharth/Downloads/nosql/assn2_4_d_part2
hadoop1@siddharth-VivoBook-ASUSLaptop-X415J... x hadoop1@siddharth-VivoBook-ASUSLaptop-X415J... x hadoop1@siddharth-VivoBook-ASUSLaptop-X415J... x siddharth@siddharth-VivoBook-ASUSLaptop-X415J... x

FILE: Number of bytes read=449255481448
FILE: Number of bytes written=40833626987
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=1195415722606
HDFS: Number of bytes written=11851
HDFS: Number of read operations=100240026
HDFS: Number of large read operations=0
HDFS: Number of write operations=10063
HDFS: Number of bytes read erasure-coded=0
Map-Reduce Framework
  Map input records=10000
  Map output records=145783
  Map output bytes=2133812
  Map output materialized bytes=2485378
  Input split bytes=1044997
  Combine input records=0
  Combine output records=0
  Reduce input groups=859
  Reduce shuffle bytes=2485378
  Reduce input records=145783
  Reduce output records=859
  Spilled Records=291566
  Shuffled Maps =10000
  Failed Shuffles=0
  Merged Map outputs=10000
  GC time elapsed (ms)=2064
  Total committed heap usage (bytes)=41190084837376
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=166222646
File Output Format Counters
  Bytes Written=11851

real    2m19.750s
user    6m46.187s
sys     0m15.857s

```

Figure 13: Pairs Approach with function level aggregation (d=4)

```

hadoop1@siddharth-VivoBook-ASUSLaptop-X415JAB-X415JA: /home/siddharth/Downloads/nosql/assn2_4_d_part2
hadoop1@siddharth-VivoBook-ASUSLaptop-X415J... x hadoop1@siddharth-VivoBook-ASUSLaptop-X415J... x hadoop1@siddharth-VivoBook-ASUSLaptop-X415J... x siddharth@siddharth-VivoBook-ASUSLaptop-X415J... x

FILE: Number of bytes read=449252496436
FILE: Number of bytes written=31247602336
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=1195415722606
HDFS: Number of bytes written=10231
HDFS: Number of read operations=100240026
HDFS: Number of large read operations=0
HDFS: Number of write operations=10063
HDFS: Number of bytes read erasure-coded=0
Map-Reduce Framework
  Map input records=10000
  Map output records=55572
  Map output bytes=821728
  Map output materialized bytes=992872
  Input split bytes=1044997
  Combine input records=0
  Combine output records=0
  Reduce input groups=785
  Reduce shuffle bytes=992872
  Reduce input records=55572
  Reduce output records=785
  Spilled Records=11114
  Shuffled Maps =10000
  Failed Shuffles=0
  Merged Map outputs=10000
  GC time elapsed (ms)=3399
  Total committed heap usage (bytes)=41202554503168
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=166222646
File Output Format Counters
  Bytes Written=10231

real    3m11.849s
user    8m41.215s
sys     0m20.065s

```

Figure 14: Pairs Approach with class level aggregation (d=1)

```

hadoop1@siddharth-VivoBook-ASUSLaptop-X415JAB-X415JA: /home/siddharth/Downloads/nosql/assn2_4_d_part2
hadoop1@siddharth-VivoBook-ASUSLaptop-X415J... x hadoop1@siddharth-VivoBook-ASUSLaptop-X415J... x hadoop1@siddharth-VivoBook-ASUSLaptop-X415J... x siddharth@siddharth-VivoBook-ASUSLaptop-X415J... x

FILE: Number of bytes read=449253984448
FILE: Number of bytes written=360666876416
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=1195415722606
HDFS: Number of bytes written=11471
HDFS: Number of read operations=100240026
HDFS: Number of large read operations=0
HDFS: Number of write operations=10063
HDFS: Number of bytes read erasure-coded=0
Map-Reduce Framework
  Map input records=10000
  Map output records=101396
  Map output bytes=1474086
  Map output materialized bytes=1736878
  Input split bytes=1044997
  Combine input records=0
  Combine output records=0
  Reduce input groups=849
  Reduce shuffle bytes=1736878
  Reduce input records=101396
  Reduce output records=849
  Spilled Records=202792
  Shuffled Maps =10000
  Failed Shuffles=0
  Merged Map outputs=10000
  GC time elapsed (ms)=2172
  Total committed heap usage (bytes)=41216544604160
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=166222646
File Output Format Counters
  Bytes Written=11471

real    2m20.827s
user    6m50.706s
sys     0m15.664s

```

Figure 15: Pairs Approach with class level aggregation (d=2)

```

hadoop1@siddharth-VivoBook-ASUSLaptop-X415JAB-X415JA: /home/siddharth/Downloads/nosql/assn2_4_d_part2
hadoop1@siddharth-VivoBook-ASUSLaptop-X415J... x hadoop1@siddharth-VivoBook-ASUSLaptop-X415J... x hadoop1@siddharth-VivoBook-ASUSLaptop-X415J... x siddharth@siddharth-VivoBook-ASUSLaptop-X415J... x

FILE: Number of bytes read=44925546148
FILE: Number of bytes written=40833626987
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=1195415722606
HDFS: Number of bytes written=11851
HDFS: Number of read operations=100240026
HDFS: Number of large read operations=0
HDFS: Number of write operations=10063
HDFS: Number of bytes read erasure-coded=0
Map-Reduce Framework
  Map input records=10000
  Map output records=145783
  Map output bytes=2133812
  Map output materialized bytes=2485378
  Input split bytes=1044997
  Combine input records=0
  Combine output records=0
  Reduce input groups=859
  Reduce shuffle bytes=2485378
  Reduce input records=145783
  Reduce output records=859
  Spilled Records=291566
  Shuffled Maps =10000
  Failed Shuffles=0
  Merged Map outputs=10000
  GC time elapsed (ms)=2492
  Total committed heap usage (bytes)=41231612641280
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=166222646
File Output Format Counters
  Bytes Written=11851

real    2m18.052s
user    6m43.269s
sys     0m16.146s

```

Figure 16: Pairs Approach with class level aggregation (d=3)

```

hadoop1@siddharth-VivoBook-ASUSLaptop-X415JAB-X415JA: /home/siddharth/Downloads/nosql/assn2_4_d_part2
hadoop1@siddharth-VivoBook-ASUSLaptop-X415J... x hadoop1@siddharth-VivoBook-ASUSLaptop-X415J... x hadoop1@siddharth-VivoBook-ASUSLaptop-X415J... x siddharth@siddharth-VivoBook-ASUSLaptop-X415J... x

FILE: Number of bytes read=449256732952
FILE: Number of bytes written=44931445467
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=1195415722606
HDFS: Number of bytes written=12071
HDFS: Number of read operations=100240026
HDFS: Number of large read operations=0
HDFS: Number of write operations=10063
HDFS: Number of bytes read erasure-coded=0
Map-Reduce Framework
  Map input records=10000
  Map output records=183432
  Map output bytes=2684266
  Map output materialized bytes=31111130
  Input split bytes=1044997
  Combine input records=0
  Combine output records=0
  Reduce input groups=865
  Reduce shuffle bytes=31111130
  Reduce input records=183432
  Reduce output records=865
  Spilled Records=366864
  Shuffled Maps =10000
  Failed Shuffles=0
  Merged Map outputs=10000
  GC time elapsed (ms)=2048
  Total committed heap usage (bytes)=41222192234496
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=166222646
File Output Format Counters
  Bytes Written=12071

real    2m21.929s
user    6m51.476s
sys     0m16.721s

```

Figure 17: Pairs Approach with class level aggregation (d=4)

```

hadoop1@siddharth-VivoBook-ASUSLaptop-X415JAB-X415JA: /home/siddharth/Downloads/nosql/assn2_4_d_part2
hadoop1@siddharth-VivoBook-ASUSLaptop-X415J... x hadoop1@siddharth-VivoBook-ASUSLaptop-X415J... x hadoop1@siddharth-VivoBook-ASUSLaptop-X415J... x siddharth@siddharth-VivoBook-ASUSLaptop-X415J... x

FILE: Number of bytes read=449251298811
FILE: Number of bytes written=28763535402
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=1195415722606
HDFS: Number of bytes written=7584
HDFS: Number of read operations=100240026
HDFS: Number of large read operations=0
HDFS: Number of write operations=10063
HDFS: Number of bytes read erasure-coded=0
Map-Reduce Framework
  Map input records=10000
  Map output records=8423
  Map output bytes=581648
  Map output materialized bytes=659086
  Input split bytes=1044997
  Combine input records=0
  Combine output records=0
  Reduce input groups=29
  Reduce shuffle bytes=659086
  Reduce input records=8423
  Reduce output records=29
  Spilled Records=16846
  Shuffled Maps =10000
  Failed Shuffles=0
  Merged Map outputs=10000
  GC time elapsed (ms)=1895
  Total committed heap usage (bytes)=41206627172352
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=166222646
File Output Format Counters
  Bytes Written=7584

real    2m10.170s
user    6m9.858s
sys     0m15.340s

```

Figure 18: Stripes Approach with function level aggregation (d=1)

```

hadoop1@siddharth-VivoBook-ASUSLaptop-X415JAB-X415JA: /home/siddharth/Downloads/nosql/assn2_4_d_part2
hadoop1@siddharth-VivoBook-ASUSLaptop-X415J... x hadoop1@siddharth-VivoBook-ASUSLaptop-X415J... x hadoop1@siddharth-VivoBook-ASUSLaptop-X415J... x siddharth@siddharth-VivoBook-ASUSLaptop-X415J... x

FILE: Number of bytes read=449251785853
FILE: Number of bytes written=30161569680
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=1195415722606
HDFS: Number of bytes written=8092
HDFS: Number of read operations=100240026
HDFS: Number of large read operations=0
HDFS: Number of write operations=10063
HDFS: Number of bytes read erasure-coded=0
Map-Reduce Framework
  Map input records=10000
  Map output records=9613
  Map output bytes=821838
  Map output materialized bytes=902607
  Input split bytes=1044997
  Combine input records=0
  Combine output records=0
  Reduce input groups=29
  Reduce shuffle bytes=982607
  Reduce input records=9613
  Reduce output records=29
  Spilled Records=19226
  Shuffled Maps =10000
  Failed Shuffles=0
  Merged Map outputs=10000
  GC time elapsed (ms)=2335
  Total committed heap usage (bytes)=41173915795456
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=166222646
File Output Format Counters
  Bytes Written=8092

real    2m19.932s
user    6m39.387s
sys     0m16.729s

```

Figure 19: Stripes Approach with function level aggregation (d=2)

```

hadoop1@siddharth-VivoBook-ASUSLaptop-X415JAB-X415JA: /home/siddharth/Downloads/nosql/assn2_4_d_part2
hadoop1@siddharth-VivoBook-ASUSLaptop-X415J... x hadoop1@siddharth-VivoBook-ASUSLaptop-X415J... x hadoop1@siddharth-VivoBook-ASUSLaptop-X415J... x siddharth@siddharth-VivoBook-ASUSLaptop-X415J... x

FILE: Number of bytes read=449252157801
FILE: Number of bytes written=31185670117
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=1195415722606
HDFS: Number of bytes written=8281
HDFS: Number of read operations=100240026
HDFS: Number of large read operations=0
HDFS: Number of write operations=10063
HDFS: Number of bytes read erasure-coded=0
Map-Reduce Framework
  Map input records=10000
  Map output records=9815
  Map output bytes=1086347
  Map output materialized bytes=1088581
  Input split bytes=1044997
  Combine input records=0
  Combine output records=0
  Reduce input groups=29
  Reduce shuffle bytes=1088581
  Reduce input records=9815
  Reduce output records=29
  Spilled Records=19638
  Shuffled Maps =10000
  Failed Shuffles=0
  Merged Map outputs=10000
  GC time elapsed (ms)=2402
  Total committed heap usage (bytes)=41228913606656
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=166222646
File Output Format Counters
  Bytes Written=8281

real    2m18.928s
user    6m43.947s
sys     0m16.425s

```

Figure 20: Stripes Approach with function level aggregation (d=3)

```

hadoop1@siddharth-VivoBook-ASUSLaptop-X415JAB-X415JA: /home/siddharth/Downloads/nosql/assn2_4_d_part2
hadoop1@siddharth-VivoBook-ASUSLaptop-X415J... x hadoop1@siddharth-VivoBook-ASUSLaptop-X415J... x hadoop1@siddharth-VivoBook-ASUSLaptop-X415J... x siddharth@siddharth-VivoBook-ASUSLaptop-X415J... x

FILE: Number of bytes read=449252378701
FILE: Number of bytes written=317733094401
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=1195415722606
HDFS: Number of bytes written=8428
HDFS: Number of read operations=100240026
HDFS: Number of large read operations=0
HDFS: Number of write operations=10063
HDFS: Number of bytes read erasure-coded=0
Map-Reduce Framework
  Map input records=10000
  Map output records=9882
  Map output bytes=1115931
  Map output materialized bytes=1199031
  Input split bytes=1044997
  Combine input records=0
  Combine output records=0
  Reduce input groups=29
  Reduce shuffle bytes=1199681
  Reduce input records=9882
  Reduce output records=29
  Spilled Records=19764
  Shuffled Maps =10000
  Failed Shuffles=0
  Merged Map outputs=10000
  GC time elapsed (ms)=2189
  Total committed heap usage (bytes)=41186827960320
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=166222646
File Output Format Counters
  Bytes Written=8428

real    2m24.810s
user    6m59.629s
sys     0m16.713s

```

Figure 21: Stripes Approach with function level aggregation (d=4)

```

hadoop1@siddharth-VivoBook-ASUSLaptop-X415JAB-X415JA: /home/siddharth/Downloads/nosql/assn2_4_d_part2
hadoop1@siddharth-VivoBook-ASUSLaptop-X415J... x hadoop1@siddharth-VivoBook-ASUSLaptop-X415J... x hadoop1@siddharth-VivoBook-ASUSLaptop-X415J... x siddharth@siddharth-VivoBook-ASUSLaptop-X415J... x

FILE: Number of bytes read=449251298811
FILE: Number of bytes written=28694388488
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=1195415722606
HDFS: Number of bytes written=7584
HDFS: Number of read operations=100240026
HDFS: Number of large read operations=0
HDFS: Number of write operations=10063
HDFS: Number of bytes read erasure-coded=0
Map-Reduce Framework
  Map input records=10000
  Map output records=8423
  Map output bytes=581648
  Map output materialized bytes=659086
  Input split bytes=1044997
  Combine input records=0
  Combine output records=0
  Reduce input groups=29
  Reduce shuffle bytes=659086
  Reduce input records=8423
  Reduce output records=29
  Spilled Records=16846
  Shuffled Maps =10000
  Failed Shuffles=0
  Merged Map outputs=10000
  GC time elapsed (ms)=2017
  Total committed heap usage (bytes)=41195315134464
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=166222646
File Output Format Counters
  Bytes Written=7584

real    2m15.775s
user    6m35.681s
sys     0m15.953s

```

Figure 22: Stripes Approach with class level aggregation (d=1)

```

hadoop1@siddharth-VivoBook-ASUSLaptop-X415JAB-X415JA: /home/siddharth/Downloads/nosql/assn2_4_d_part2
hadoop1@siddharth-VivoBook-ASUSLaptop-X415J... x hadoop1@siddharth-VivoBook-ASUSLaptop-X415J... x hadoop1@siddharth-VivoBook-ASUSLaptop-X415J... x siddharth@siddharth-VivoBook-ASUSLaptop-X415J... x

FILE: Number of bytes read=44925129853
FILE: Number of bytes written=3012695625
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=1195415722606
HDFS: Number of bytes written=8092
HDFS: Number of read operations=100240026
HDFS: Number of large read operations=0
HDFS: Number of write operations=10063
HDFS: Number of bytes read erasure-coded=0
Map-Reduce Framework
  Map input records=10000
  Map output records=9613
  Map output bytes=821838
  Map output materialized bytes=902607
  Input split bytes=1044997
  Combine input records=0
  Combine output records=0
  Reduce input groups=29
  Reduce shuffle bytes=982607
  Reduce input records=9613
  Reduce output records=29
  Spilled Records=19226
  Shuffled Maps =10000
  Failed Shuffles=0
  Merged Map outputs=10000
  GC time elapsed (ms)=1989
  Total committed heap usage (bytes)=41201491247104
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=166222646
File Output Format Counters
  Bytes Written=8092

real    2m16.834s
user    6m37.215s
sys     0m16.391s

```

Figure 23: Stripes Approach with class level aggregation (d=2)

```

hadoop1@siddharth-VivoBook-ASUSLaptop-X415JAB-X415JA: /home/siddharth/Downloads/nosql/assn2_4_d_part2
hadoop1@siddharth-VivoBook-ASUSLaptop-X415J... x hadoop1@siddharth-VivoBook-ASUSLaptop-X415J... x hadoop1@siddharth-VivoBook-ASUSLaptop-X415J... x siddharth@siddharth-VivoBook-ASUSLaptop-X415J... x

FILE: Number of bytes read=449252157801
FILE: Number of bytes written=31185678117
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=1195415722686
HDFS: Number of bytes written=8281
HDFS: Number of read operations=100240026
HDFS: Number of large read operations=0
HDFS: Number of write operations=10003
HDFS: Number of bytes read erasure-coded=0
Map-Reduce Framework
  Map input records=10000
  Map output records=9815
  Map output bytes=1006347
  Map output materialized bytes=1088581
  Input split bytes=1044997
  Combine input records=0
  Combine output records=0
  Reduce input groups=29
  Reduce shuffle bytes=1088581
  Reduce input records=9815
  Reduce output records=29
  Spilled Records=19638
  Shuffled Maps =10000
  Failed Shuffles=0
  Merged Map outputs=10000
  GC time elapsed (ms)=2132
  Total committed heap usage (bytes)=41185330593792
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=166222646
File Output Format Counters
  Bytes Written=8281

real    2m18.681s
user    6m43.595s
sys     0m16.460s

```

Figure 24: Stripes Approach with class level aggregation (d=3)

```

FILE: Number of bytes written=44896892812
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=1195415722686
HDFS: Number of bytes written=12071
HDFS: Number of read operations=100240026
HDFS: Number of large read operations=0
HDFS: Number of write operations=10003
HDFS: Number of bytes read erasure-coded=0
Map-Reduce Framework
  Map input records=10000
  Map output records=183432
  Map output bytes=2604264
  Map output materialized bytes=3111130
  Input split bytes=1044997
  Combine input records=0
  Combine output records=0
  Reduce input groups=865
  Reduce shuffle bytes=3111130
  Reduce input records=183432
  Reduce output records=865
  Spilled Records=366864
  Shuffled Maps =10000
  Failed Shuffles=0
  Merged Map outputs=10000
  GC time elapsed (ms)=2056
  Total committed heap usage (bytes)=41285658288128
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=166222646
File Output Format Counters
  Bytes Written=12071

real    2m24.668s
user    6m57.435s
sys     0m16.554s
All Hadoop jobs completed!

```

Figure 25: Stripes Approach with class level aggregation (d=4)

## Problem 5

The objective of this problem is to compute the **Term Frequency-Inverse Document Frequency (TF-IDF)** scores for stemmed words across a collection of 10,000 Wikipedia articles.

### (a) Document Frequency Computation

In this step, Map-Reduce is used to determine the number of documents in which each term appears.

#### Mapper

- Reads input data, converts it to lowercase, and tokenizes using whitespaces.
- Removes special characters and applies Porter Stemmer from OpenNLP.
- Excludes stopwords listed in *stopwords.txt*.
- Emits key-value pairs in the format <term, docID>.

#### Reducer

- Receives a key (term) and a list of document IDs where the term appears.
- Computes Document Frequency (DF) for each term.
- Outputs key-value pairs in the format <term, DF>.
- We store the top 100 words with the highest DF in HDFS.

### (b) TF-IDF Computation

This step computes the Term Frequency-Inverse Document Frequency (TF-IDF) scores using Map-Reduce.

#### Setup

Before execution of the Map Reduce Job, the top 100 words and their document frequencies (DF) are loaded from a cached file(Distributed Cache) into a map structure for quick lookup.

#### Mapper

- Reads each document, converts it to lowercase, and tokenizes it using \s+.
- Removes special characters and applies Porter Stemmer for word normalization.
- Counts word occurrences only if they appear in the top 100 words list.
- Emits key-value pairs <filename, word-frequency map>.

```

File System Counters
  FILE: Number of bytes read=1139149659492
  FILE: Number of bytes written=12767772368364
  FILE: Number of read operations=0
  FILE: Number of large read operations=0
  FILE: Number of write operations=0
  HDFS: Number of bytes read=1195410492083
  HDFS: Number of bytes written=13768827
  HDFS: Number of read operations=100150017
  HDFS: Number of large read operations=0
  HDFS: Number of write operations=10003
  HDFS: Number of bytes read erasure-coded=0
Map-Reduce Framework
  Map input records=10000
  Map output records=7190796
  Map output bytes=67995719
  Map output materialized bytes=82441135
  Input split bytes=1046997
  Combine input records=0
  Combine output records=0
  Reduce input groups=853695
  Reduce shuffle bytes=82441135
  Reduce input records=7190796
  Reduce output records=853695
  Spilled Records=14381592
  Shuffled Maps =10000
  Failed Shuffles=0
  Merged Map outputs=10000
  GC time elapsed (ms)=3711
  Total committed heap usage (bytes)=41218755002368
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=166222646
File Output Format Counters
  Bytes Written=13768827

real    5m8.486s
user    12m46.410s
sys     0m34.474s
hadoop1@siddharth-VivoBook-ASUSLaptop-X415JAB-X415JA:/home/siddharth/Downloads/nosql/assn2_B$ █

```

Figure 26: Run time for 5(a)

## Reducer

- Aggregates term frequencies from all mappers.
- Retrieves the DF values for words from the cached file.
- Computes TF-IDF using the formula:

$$TFIDF = TF \times \log \left( \frac{Total\ Documents}{DF + 1} \right) \quad (1)$$

- Outputs <filename, word-TFIDF> pairs.

```

File System Counters
  FILE: Number of bytes read=1139805429854
  FILE: Number of bytes written=748137144828
  FILE: Number of read operations=0
  FILE: Number of large read operations=0
  FILE: Number of write operations=0
HDFS: Number of bytes read=1266985819109
HDFS: Number of bytes written=12542998
HDFS: Number of read operations=150225024
HDFS: Number of large read operations=0
HDFS: Number of write operations=10003
HDFS: Number of bytes read erasure-coded=0

Map-Reduce Framework
  Map input records=10000
  Map output records=10000
  Map output bytes=4637228
  Map output materialized bytes=4735757
  Input split bytes=1046997
  Combine input records=0
  Combine output records=0
  Reduce input groups=10000
  Reduce shuffle bytes=4735757
  Reduce input records=10000
  Reduce output records=412307
  Spilled Records=20000
  Shuffled Maps =10000
  Failed Shuffles=0
  Merged Map outputs=10000
  GC time elapsed (ms)=4131
  Total committed heap usage (bytes)=41227623858176

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=146222646
File Output Format Counters
  Bytes Written=12542998

real    3m8.940s
user    8m15.845s
sys     0m23.723s
hadoop1@siddharth-VivoBook-ASUSLaptop-X415JAB-X415JA:/home/siddharth/Downloads/nosql/assn2_5B$ 

```

Figure 27: Run time for 5(b)