Maor Rocky 203246079 rocky@post.bgu.ac.il
Yoan Shimoni 317756609

Running Instructions:

- 3 Jar Files should be already in the S3 bucket named “maorrockyjars”.
- Run main.class make sure you gave the correct Ngarm String as input to StepOne

  - EMR instance will be created
  - 3 Steps are made with their unique jar (taken from the S3 bucket), and each step receives different Arguments - explicit the location in S3 for input and output.
- The resulting file will be uploaded to S3//maorrockyjars unique bucket.

Discussion about our results:

When we didn't use the combiner when we executed our program on the English Ngram we got these results:

- Reduce input groups=354048822
- Reduce shuffle bytes=10022297659
- Reduce input records=1614481045
- Reduce output records=354048822

but when we did use combiner we got:

- Input split bytes=85207
- Combine input records=2282742736
- Combine output records=455703852
- Reduce input groups=455622196
- Reduce shuffle bytes=6096796918
- Reduce input records=455703852
- Reduce output records=455622196

as we can see when we used combiner we received much fewer input records in the "reduce" stage, of stepOne of our program. The same result was achieved when we executed our program on Hebrew.

Bad allocations:

- 1850 "he """ 120159.49052026903
- 1850 "all """ 27941.64928493055
- 1900 "those """ 74113.10377664817
- 1730 """ Arise" 57.77450716784891
- 2000 who's whos NaN
- 2000 leaf whorls NaN
- 1690 "Israel """ 12.699867684953908
- 1590 5 Like 8.260170485092605
- 1670 Mr Samuel NaN
- 1720 poflefs ' 70.04989891862097

We believe that the wrong collocations were caused due to lack in the amount of data in the corpus and since Ngram is not a perfect data set.

Good allocations:

- 1930 National Survey 61690.03129382327
- 1930 Foreign Trade 58638.1637361291
- 1740 great commandment 106.41571709466305
- 1960 Spanish America 209173.41281852886
- 1960 economic theory 204631.90333819442
- 1730 Henry VIII 90.00195224313714
- 1730 great day 90.00195224313714
- 1970 Saudi Arabia 327622.9760910838
- 1970 crude oil 379908.0503371056
- 1970 Don Juan 389424.01861319365
- 1890 God forgive 57316.545345214894

LOGS

- English Ngram without combiner :

File System Counters
FILE: Number of bytes read=8904172579
FILE: Number of bytes written=19007030750
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=85207
HDFS: Number of bytes written=0
HDFS: Number of read operations=613
HDFS: Number of large read operations=0
HDFS: Number of write operations=0
S3: Number of bytes read=41256298612
S3: Number of bytes written=8869561095
S3: Number of read operations=0
S3: Number of large read operations=0
S3: Number of write operations=0
Job Counters
Killed map tasks=10
Killed reduce tasks=2
Launched map tasks=622
Launched reduce tasks=19
Other local map tasks=9
Data-local map tasks=613
Total time spent by all maps in occupied slots (ms)=4690510368
Total time spent by all reduces in occupied slots (ms)=4446727584
Total time spent by all map tasks (ms)=97718966
Total time spent by all reduce tasks (ms)=46320079
Total vcore-milliseconds taken by all map tasks=97718966
Total vcore-milliseconds taken by all reduce tasks=46320079
Total megabyte-milliseconds taken by all map tasks=150096331776
Total megabyte-milliseconds taken by all reduce

tasks=142295282688

Map-Reduce Framework
Map input records=3923370881
Map output records=1614481045
Map output bytes=38115241589
Map output materialized bytes=10022297659
Input split bytes=85207
Combine input records=0
Combine output records=0
Reduce input groups=354048822
Reduce shuffle bytes=10022297659
Reduce input records=1614481045
Reduce output records=354048822
Spilled Records=3228962090
Shuffled Maps =10421
Failed Shuffles=0
Merged Map outputs=10421
GC time elapsed (ms)=1623852
CPU time spent (ms)=75555460
Physical memory (bytes) snapshot=547685998592
Virtual memory (bytes) snapshot=2106441711616
Total committed heap usage (bytes)=488053407744
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=41256298612
File Output Format Counters
Bytes Written=8869561095

- Enlgish Ngram with combiner :

File System Counters
FILE: Number of bytes read=4361734073
FILE: Number of bytes written=10539177881
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=85207
HDFS: Number of bytes written=0
HDFS: Number of read operations=613
HDFS: Number of large read operations=0
HDFS: Number of write operations=0
S3: Number of bytes read=41256092393
S3: Number of bytes written=12030420965
S3: Number of read operations=0
S3: Number of large read operations=0
S3: Number of write operations=0
Job Counters
Killed map tasks=7
Killed reduce tasks=2
Launched map tasks=619
Launched reduce tasks=19
Other local map tasks=6
Data-local map tasks=613
Total time spent by all maps in occupied slots (ms)=6073840992
Total time spent by all reduces in occupied slots (ms)=4505981568
Total time spent by all map tasks (ms)=126538354
Total time spent by all reduce tasks (ms)=46937308
Total vcore-milliseconds taken by all map tasks=126538354
Total vcore-milliseconds taken by all reduce tasks=46937308
Total megabyte-milliseconds taken by all map tasks=194362911744

```
        Total megabyte-milliseconds taken by all reduce tasks=144191410176
    Map-Reduce Framework
        Map input records=3923370881
        Map output records=2282742736
        Map output bytes=49510086806
        Map output materialized bytes=6096796918
        Input split bytes=85207
        Combine input records=2282742736
        Combine output records=455703852
        Reduce input groups=455622196
        Reduce shuffle bytes=6096796918
        Reduce input records=455703852
        Reduce output records=455622196
        Spilled Records=911407704
        Shuffled Maps =10421
        Failed Shuffles=0
        Merged Map outputs=10421
        GC time elapsed (ms)=1072974
        CPU time spent (ms)=85382950
        Physical memory (bytes) snapshot=535559536640
        Virtual memory (bytes) snapshot=2106464882688
        Total committed heap usage (bytes)=474499514368
    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=41256092393
    File Output Format Counters
        Bytes Written=12030420965
```

Hebrew Ngram without combiner:

File System Counters

- FILE: Number of bytes read=305457746
- FILE: Number of bytes written=620431354
- FILE: Number of read operations=0
- FILE: Number of large read operations=0
- FILE: Number of write operations=0
- HDFS: Number of bytes read=3267
- HDFS: Number of bytes written=0
- HDFS: Number of read operations=27
- HDFS: Number of large read operations=0
- HDFS: Number of write operations=0
- S3: Number of bytes read=1191479911
- S3: Number of bytes written=136029
- S3: Number of read operations=0
- S3: Number of large read operations=0
- S3: Number of write operations=0

Job Counters

- Killed map tasks=7
- Killed reduce tasks=3
- Launched map tasks=32
- Launched reduce tasks=20
- Other local map tasks=4
- Data-local map tasks=28
- Total time spent by all maps in occupied slots (ms)=131787552
- Total time spent by all reduces in occupied slots (ms)=910629888
- Total time spent by all map tasks (ms)=2745574
- Total time spent by all reduce tasks (ms)=9485728
- Total vcore-milliseconds taken by all map tasks=2745574
- Total vcore-milliseconds taken by all reduce tasks=9485728

Total megabyte-milliseconds taken by all map tasks=4217201664
Total megabyte-milliseconds taken by all reduce tasks=29140156416
Map-Reduce Framework
Map input records=26820728
Map output records=26820728
Map output bytes=1226813042
Map output materialized bytes=309373159
Input split bytes=3267
Combine input records=0
Combine output records=0
Reduce input groups=26820728
Reduce shuffle bytes=309373159
Reduce input records=26820728
Reduce output records=3300
Spilled Records=53641456
Shuffled Maps =459
Failed Shuffles=0
Merged Map outputs=459
GC time elapsed (ms)=28970
CPU time spent (ms)=1138690
Physical memory (bytes) snapshot=31156477952
Virtual memory (bytes) snapshot=168498491392
Total committed heap usage (bytes)=28149547008
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=1191479911
File Output Format Counters
Bytes Written=136029

Hebrew Ngram with combiner:

File System Counters

FILE: Number of bytes read=415126494
FILE: Number of bytes written=806381727
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=5304
HDFS: Number of bytes written=0
HDFS: Number of read operations=39
HDFS: Number of large read operations=0
HDFS: Number of write operations=0
S3: Number of bytes read=2581890414
S3: Number of bytes written=852838653
S3: Number of read operations=0
S3: Number of large read operations=0
S3: Number of write operations=0

Job Counters

Failed reduce tasks=1
Killed map tasks=1
Launched map tasks=39
Launched reduce tasks=18
Data-local map tasks=39
Total time spent by all maps in occupied slots (ms)=465069696
Total time spent by all reduces in occupied slots (ms)=236084832
Total time spent by all map tasks (ms)=9688952
Total time spent by all reduce tasks (ms)=2459217
Total vcore-milliseconds taken by all map tasks=9688952

Total vcore-milliseconds taken by all reduce tasks=2459217
Total megabyte-milliseconds taken by all map tasks=14882230272
Total megabyte-milliseconds taken by all reduce tasks=7554714624
Map-Reduce Framework
Map input records=252069581
Map output records=156120895
Map output bytes=3906789053
Map output materialized bytes=391346017
Input split bytes=5304
Combine input records=156120895
Combine output records=28714865
Reduce input groups=28712228
Reduce shuffle bytes=391346017
Reduce input records=28714865
Reduce output records=28712228
Spilled Records=65791597
Shuffled Maps =663
Failed Shuffles=0
Merged Map outputs=663
GC time elapsed (ms)=80336
CPU time spent (ms)=5674200
Physical memory (bytes) snapshot=49977454592
Virtual memory (bytes) snapshot=208271519744
Total committed heap usage (bytes)=44851789824
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=2581890414
File Output Format Counters
Bytes Written=852838653