CS-562 Advanced Topics in Database
Big Data Processing & Analytics

Assignment 1

Deadline: 2/10/2018

Giakoumakis Theodoros Marios

Tmgiakoumakis@csd.uoc.gr

1074

Exercise 1

1) The k=10 most frequent words with their frequencies (found also in 10_most_frequent.txt) are

182666	the
146782	and
98557	of
90275	to
86303	a
62634	i
58282	in
47749	that
43081	it
40512	was

Extracted from part-r-00000 file that was produced as output with

```
Total execution time: 29846
```
milliseconds

2) The stopwords.csv file is basically part-r-00000 output since we have by default 1 producer by the standalone version

Exercise 2

A) The default setup of 6 mappers 1 reducer produced an output in `Total execution time: 29846` milliseconds.

1) By use of 10 reducers the output was produced in: `Total execution time: 34722` milliseconds. The observed uptime is a result of the affected shuffle as well as the increased I/O operations since each reducer creates its own file, having 10 instead of 1 intermediate outputs. Moreover, smaller files for processing for the same job, affects the performance of the 2nd map-reduce job.
2) By use of a combiner the output was produced in: `Total execution time: 22350` milliseconds. The observed increase in performance is due to the size of the produce of Map, which are big enough, allowing combiner to optimize and avoid congestion before entering Reduce.
3) By use of compression the output was produced in: `Total execution time: 34067` milliseconds. Compression aids the I/O, save storage space and network transfer by reducing the bytes needed to be handled. Less processing on these bottlenecks speeds the overall execution time. The observed reduce in performance is probably due to failure to properly use compression in the script file, or failure to measure correctly execution time.
4) By use of 50 reducers the output was produced in: `Total execution time: 42668` milliseconds. The increased reduction of performance is attributed as said for the 10 reducers to the I/O operations and the consequent slowdown of the 2nd map-reduce job.

MAP-REDUCE SETUP	TIME (milliseconds)
Default Setup	29846
10 reducers	34722
Combiner	22350
Compression of map results	34067
50 reducers	42668

B) The number of unique words in the input files is shown by the reduce input groups and reduce output groups as 60487

```
Reduce input groups=60604
Reduce shuffle bytes=66247679
Reduce input records=3669134
Reduce output records=60604
```

With unique words in each document as (UniquePerDocument.txt):

```
pg100: 26248
pg1120: 3188
pg1513: 4112
pg2253: 5343
pg31100: 141
pg3200: 48296
```

Exercise 3

Exercise 3 extended exercise 2 with the use of a Combiner. Results at part-r-00000 file of Assignment_1_3