Assignment 1 hy562

Pachilakis Michalis

Exercise 1

1) The 10 more frequent word with their frequencies are in ascending order:

1.	42028	was
2.	53414	that
3.	59911	it
4.	60418	in
5.	72242	i
6.	89702	a
7.	93324	to
8.	99513	of
9.	151989	and
10.	185805	the

Exercise 2

a) In the base case where I run the exercise 1 without any modification it takes 1 min 43 sec for the first job and 31 sec for the second

   1. When running with 10 reducers it takes 3 min and 8 sec the first job and 1 min 38 sec the second.
   2. When using a combiner it takes 1 min and 27 the first job and 22 sec the second. This happens because the reducer doesn’t need to combine the mapper’s output and it just iterate the list which is already created by the combiner.
   3. When using compression the first job takes 1min 40sec and the second 21 sec. The time is reduced a little because the file which the reducer needs to read is less bytes than if it was uncompressed, so it can be transferred faster. We would see a better result if the file corpus was larger (gigabytes).
   4. When running with 50 reducers it takes 8 min 43 sec the first job and 6 min 17 sec the second. This

happens because the mapper’s output needs to be split between the reducers and then the output should be combined again. Also because the machine used has limited resources (1 core) the 50 reducers can’t work in parallel efficiently.

b) 2. In the document corpus there are 55.642 unique words. The counters revealing this information is the *Reduce input groups* and the *Reduce Output Records.*