Assignment 5 (5th Sept., 2018)

Exercise 1
(Hadoop Installation) Install the Hadoop Distribution of Cloudera (http://www.cloudera.com/hadoop/) in Pseudo-Distributed Mode or use the VMWare Image provided by Cloudera to familiarize yourself with Hadoop, especially with the distributed file system HDFS and the implementation of MapReduce programs in Java. You can find a good introduction to MapReduce and Hadoop in the given literature at the end of this exercise sheet, for example.

Exercise 2 (Basic Text Operations) For the following tasks use the text file 'big.txt' https://norvig.com/big.txt as input which contains a collection of the works from "Adventures of Sherlock Holmes".

a) Implement a MapReduce program that outputs all words of the input in a sorted order. Your program should not distinguish between upper and lower case and duplicates should be preserved. Example: {To be or not to be} −→ {be be not or to to}

b) Extend your program from part (a) such that every word occurs only once in the output together with the corresponding frequency of the word. Your program should not distinguish between upper and lower case. Example: {To be or not to be} −→ {(be,2) (not,1) (or,1) (to,2)}

c) Extend your word count implementation from part (b) with an additional Combiner. Therefor you should familiarize yourself with the function of a Combiner and think about how to usefully integrate a Combiner into your implementation. Characterize advantages and disadvantages of a Combiner.

d) Implement a MapReduce program that computes the inverted index for the given input, i.e. for every word in the input it should output a list of (byte) offsets. The offset should be the byte offset of the row that contains the word. However, typical stop words should not be part of the index. Stop words are frequently occurring words like 'and' that do not have a substantial relevance. You can find a list of typical English stop words in the given file ('big.txt')