Middle East Technical University
Department of Computer Engineering

CENG 495
Cloud Computing
Spring 2017-2018 Homework 3

This homework aims to get you familiar with MapReduce paradigm. You are going to develop and deploy a MapReduce application by using Apache Hadoop Packages and Java language.

Keywords: Cloud Computing, Hadoop, Apache, MapReduce, Java

1. Apache Hadoop

- Download and install the latest stable release of Apache Hadoop. (http://hadoop.apache.org)
- You must have the required JDK version to use Hadoop.
- Configure Hadoop and environment variables to link Hadoop with Java.
- You can look at the following tutorial and use the corresponding code as a base for your work. (https://hadoop.apache.org/docs/stable/hadoop-mapreduce-client/hadoop-mapreduce-client-core/MapReduceTutorial.html)

2. Specifications

- You will implement a Java code with Hadoop environment for analyzing the input files consisting of words.
- You will be given a folder containing input text files.
- Your program will execute the following tasks:

a. Finding the number of letters in longest word and print the result. (**longest**)

b. Finding the occurrences of each letter in the input file, write vowel and consonant letters in different output files. i.e. vowels on “part-r-00000” , consonants on “part-r-00001” (**letter**)

c. Finding the occurrences of each word in the input file, write down the results in 3 different output files separated by the length of the word. i.e. length<4 => “part-r-00000”, 3<length<7 => “part-r-00001”, 7<length => “part-r-00002” (**count**)

- Only lower case letters will be used in input files. Thus, you do not need to handle uppercase letters or any other characters.
- There can be more than one input files. Your program should read all the files in the input folder.
- You can see the input and output formats on the sample input and output files. Since black-box testing will be used for grading, be sure to stick the format.
- The solution must be in Java language using the Apache Hadoop library.
- Your solutions will be evaluated automatically in Local (Standalone) Mode of Hadoop. Assuming that all of the Java files of your solution exist in the current directory, the command sequence below will be executed in order to build the solution:

```
hadoop com.sun.tools.javac.Main *.java

jar cf Hw3.jar *.class
```

- The output jar file will be tested with commands given below with different inputs.

```
hadoop jar Hw3.jar Hw3 longest input output1

hadoop jar Hw3.jar Hw3 letter input output2

hadoop jar Hw3.jar Hw3 count input output3
```

3. Submission

- In this assignment, you are expected to submit your Java code(s) to COW with a tar.gz archive file (named hw3.tar.gz) that contains all your source code files.

- The work you submit should be implemented by only you and genuine.
- We have zero tolerance policy for cheating. There is no teaming up! People involved in cheating will be punished according to the university regulations and will get 0. You can discuss design choices, but sharing code between each other or submitting third party code as a whole is strictly forbidden. In case a match is found, this will be considered as cheating.
- The deadline for homework is 21.05.2018 23:55.