Big Data Computing
Lab - I

August 3, 2021

I. Word Count

Input: An unstructured text file with multiple lines

Output: Number of occurrences of each word, appearing in the text file.

Input: *sample.txt*

Welcome to the world of Hadoop. The introduction to Hadoop starts with the traditional wordcount program.

Output: (Welcome,1) (to, 2) (the, 2) (world, 1) (The, 1) (introduction,1) (starts, 1) (traditional, 1) (wordcount, 1) (program, 1) (of, 1) (Hadoop, 2) (with, 1)

II. Youtube data analysis

You will be provided with a text file, from “Dataset for Statistics and Social Network of YouTube Videos”, where the data are separated by '\t' in the file and comprises of following columns:

Column Name	Description
video ID	an 11-digit string, which is unique
uploader	a string of the video uploader's username
age	an integer number of days between the date when the video was uploaded and Feb.15, 2007 (YouTube's establishment)
category	a string of the video category chosen by the uploader

length	an integer number of the video length
views	an integer number of the views
rate	a float number of the video rate
ratings	an integer number of the ratings
comments	an integer number of the comments
related IDs	up to 20 strings of the related video IDs

A. You are provided with the source code of one (Map-Reduce program) which performs the following task:

a. **Top5_categories.java** : Finds out the top 5 categories with maximum number of videos uploaded.

The instructions for implementing the task is similar to the wordcount program provided to you with the first assignment except for the last line to view the output where you need to pass the following instruction in command line:

$ bin/hadoop fs -cat /youtube/outputdata/part-r-00000 | sort –n –k2 –r | head –n5

Note: sort will sort the data, **–n** means sorting numerically, **–k2** means second column, **–r** is for recursive operation and **head –n5** means to bring the first 5 values after sorting.

(Provided you make the youtube/outputdata directory structure on HDFS).