ITNPBD7 Assignment 2019

Movie Review Data

Your task in completing this assignment is to analyse some movie review data. The data are contained in a file called ratedReviews.txt, which you can download from the module assignments page on Canvas. The file contains the text of movie reviews, each followed by a score from 1 to 9 indicating a rating given to the movie. Each review and score pair are on a single line of the file and the scores are separated from the reviews by a tab character. Consequently, you can examine the file in Excel if you want.

Your task is to write the Map/Reduce code in Java needed to process the movie reviews in a way that discovers the most common word used in reviews with each possible rating. The result will be 9 words: one for each score, which represent the most common word used in reviews with each rating score. You should make sure the stop words provided are not counted.

You will submit a written report, detailing your design and the results you found. You will also be asked to submit a Java file containing your code.

Step 1, HDFS – 20 Marks

Before you write any code, you will need to copy the data onto your own space in HDFS. In your report, give details of how HDFS stores data such as this (assume the file is much bigger than it really is for the purpose of your description). This section should be around half a page long, plus a diagram. Describe what HDFS is for, the architecture it uses, and the roles of different nodes in the cluster. Document the hdfs commands you used to create a directory for the data and place it there. Make sure everything you put here, including the diagram, is your own work. Do not copy anything from other sources.

Step 2, Design – 20 Marks

Now consider the Map/Reduce design you will implement. You know there are only nine different scores associated with the movie reviews and a larger (but unknown) number of different words used in those reviews. Consider and compare two different choices you could make to implement the given task. What keys and values will the mapper emit? Consider how much data will be moved across the network in each of your two designs. Also consider how many different reducers will be used in each case. Finally, choose one of the two designs to implement and justify your choice.

Step 3, Implement – 60 Marks

Using the MovieReview.java file provided on the assignment page in Canvas as a starting point, modify this code to produce the results requested above. This code is just a renamed version of the original WordCount.java file from your practicals and will need significant changes to meet the

desired requirements (including changing some of the types of the Key/Value pairs) . It is supplied with the file *TestMoveReview.java* that you can use with the Hadoop simulator *mochadoop* to check your logic on the smaller set of data found in shortReviews.txt. Your final output should however be produced from the full *ratedReviews.txt* file that should be run on the Hadoop server.

There is a list of words that should not be counted in the reviews – they are given in the file *exclude.txt*, which you can download from the Canvas assignment page. This list should be loaded into your program from the file, not hard coded into the MovieReview code.

You should now implement your design in Java using the Hadoop API that we have been using in class. Your code should find the most commonly used word (excluding those in the exclusion list) for movies with each rating from 1 to 9. You should allow the exclusion list to be supplied as a cache file (do not hard code the list into your Java). Make sure you implement a mapper, a combiner (if your design allows it) and a reducer.

Submission Details

Please write up your work in a report and submit it via Canvas, clearly noting your 7 digit student ID number on the front of your report *but **do not** provide your name*. Additionally, please submit your MovieReview.java file in the same way and ensure that your code is very well commented and that you have put your 7 digit ID number at the top of this file in the commented area. Make sure your report also contains the results you got when you ran the code – that is the most common word for each rating category when using the full set of data. The deadline for submission is Friday 29th of March at 4pm.