CPSC 4330 Big Data Analytics
Winter 2025

Homework 3

Due: 10:55 am, Friday, Feb. 28

There are two problems in this homework.

1. Problem 1 (55 points)

Write a Spark application that computes the total number of reviews and the average rate of each product. The output needs to be sorted based on the product_id in ascending order. This is the same task you did earlier in homework 1 where you wrote a MapReduce Java program to run on Hadoop. The logic is the same, but this time you will need to write a Spark application.

The data that you can use to test your program is the same as in homework 1. That is to say, you can test your code on the sample file (Gift_Cards.csv). If that looks good, you can upload your application to AWS, scale out worker nodes and run on a slightly larger data set (review_data.zip) which is available on Canvas.

2. Problem 2 (5 points each)

The file “College_2015_16.csv” (posted on Canvas) contains the following fields:

- Unique ID
- Name
- City
- State
- Zip
- Admission rate
- Average SAT score
- Enrollment
- CostA
- CostP

The last two columns are the cost of public and private universities. If one is non-null, the other should be null. If both are null, that's a missing value. If both are non-null, use either value.

Write the PySpark code to solve the following problems. Your solution should be contained in a notebook hw3_2.ipynb.

1. Convert the lines in the file to a tuple of fields, and only keep these attributes: ID, name, state, enrollment, and cost, where cost is either costA or costP as above. If enrollment cannot be converted to an int, set it to null.

2. Find how many records were filtered due to the invalid number of fields in the data (the file has 10 fields).

3. Find how many records are there from the state of California?

4. What percentage of the records have a non-null enrollment?

5. What is the name and cost of the 5 most expensive universities?

6. Find the number of universities in each state.

7. Find the total number of enrollments in each state.

8. Find the average enrollment for each state.

9. Another file “College_2017_18.csv” (posted on Canvas) has the college enrollment data of year 2017 - 2018. Write code to calculate percent of change in enrollment from year 2015 – 2016 to the year 2017 – 2018.

Submission

Use Canvas to submit the following:

- For problem 1
  - The Spark application (hw3_1.py)
  - All your output files of running your code against the files in review_data.zip

- For problem 2
  - The file hw3_2.ipynb containing your solution