CPSC 4330 Big Data Analytics
Winter 2021

Homework 3

Due: 6:00 pm, Wednesday, Feb. 24th

In this homework, you will write a Spark application that computes the total number of Amazon 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 (sample_us.tsv). If that looks good, you can get one actual data set file to test on. You can still use the file https://s3.amazonaws.com/amazon-reviews-pds/tsv/amazon_reviews_us_Electronics_v1_00.tsv.gz (about 667 MB) to test your program, the first 5 lines and the last 5 lines of the output file are shown below.

```
('0141186178', 5.0, 1)
('0303532572', 4.5, 2)
('043964383X', 5.0, 1)
('0511189877', 4.3877551020408161, 49)
('0528881469', 3.1470588235294117, 34)
```

the first 5 lines of the output file

```
('B016OF0IDI', 5.0, 1)
('B0188YEWQM', 1.0, 1)
('B019IOIOK6', 5.0, 1)
('B01MZ2Z4UF', 3.0, 1)
('BT008V9J9U', 3.0, 2)
```

the last 5 lines of the output file

Now you can upload your Spark application to AWS, scale out worker nodes and run on the larger data sets. You can analyze the same datasets used in homework 1. That is to say, for the input argument of your job, you can type s3://amazon-reviews-pds/tsv/amazon_reviews_us_Books_v1_0*.tsv.gz

When your job is done, remember to make sure your cluster is terminated.

Submission

Use Canvas to submit the following:

- The Spark application (.py)
- All your output files on the three review files related to books (You can compress the files if that's convenient)
    - Be sure you submit the output of running on the three review files as input – don't submit the output from testing on a single data file by mistake.