ENSF 612: Quiz 1

Marks: 20

Duration: 48 hours

The quiz is worth 5% of course marks

Instructions:

1. This assignment must be completed individually. You cannot copy from others or consult with others. Any identification of such unauthorized actions may result in 0 grade for this assignment.
2. This is a take home assignment. You can finish it in the next 48 hours.
3. You have until Oct 23, 7 AM MDT to submit this.
4. You will submit solution in a PDF file and upload it in D2L under Quiz 1.

You are given the records of user's reviews of products of a consumer product company. Each review is of the form <UserId, ReviewTime, ReviewText, ReviewRating, NumberOfProductBought>. The columns are described below.

1. UserId = the Id of the user who submitted the review
2. ReviewTime = the time of review in timestamp (e.g., UTC timestamp)
3. ReviewText = the textual contents of the provided review
4. ReviewRating = a 5 star rating
5. NumberOfProductBought = Total number of products bought

Task 1 (Marks 10).

Design a Spark transformation analytics that will add the following additional columns per review.

1. SentimentPolarity
2. UserCountry
3. UserCity
4. ReviewYear
5. ReviewMonth
6. ReviewDay

Assume that you have access to the following function that you can access via Spark

1. getSentiment(ReviewText) will return 'p' for positive, 'n' for negative' and 'o' for neutral polarity found in the ReviewText
2. getUserCountry(UserId) will return country of UserId
3. getUserCity(UserId) will return city of country where UserId is resided
4. getYear(ReviewTime) will return year of the ReviewTime
5. getMonth(ReviewTime) will return month of the ReviewTime
6. getDay(ReviewTime) will return day of a week (e.g., Monday, Tuesday, etc.) of the ReviewTime

Task 2 (Marks 5).

Design a Spark transformation action that will return the total number of products bought by all the consumers (hint: you first need to get a list of NumberOfProductBought before you do the transformation)

Task 3 (Marks 5).

Design Spark transformation pipeline, one each for the following requirement:

1. The average overall ReviewRating per user (Marks 1)
2. The total number of Reviews by:
    a. Overall positive, negative, and neutral polarity (Marks 2)
    b. By country and by user (Marks 2)