Assignment 5: Spark Practice

Special Topics in Computer Science: Big Data Systems
CS-UH 3260 Spring 2023
MAX 10 points

This assignment contains data processing tasks to get you to practice writing Spark code. While we cover many examples in class, this assignment encourages you to read the documentation to navigate your way through answering these tasks.

1. System installation (not graded):

- Install Spark on your machine, and make sure you can start a shell in local mode:
  - **pyspark** *(python)*
  - **spark-shell** *(or, if you wish to write code in scala)*
- Contact me (or Sara Mumtaz for MS Windows) if you need help with installation.

2. Description

This assignment uses the "Titanic" dataset (from https://www.kaggle.com/c/titanic/data). You can check the description on Kaggle to understand the columns' content. Google Drive

First, load the Titanic dataset into a PySpark DataFrame and display the first five rows.
>> **titanic_df = spark.read.csv("titanic.csv", header=True, inferSchema=True)**

1. Count and retrieve 10 passengers who survived.
2. **Add a column** called "IsChild" that contains "True" if the passenger's age is less than 18, and "False" otherwise.
3. Group the DataFrame by the "Pclass" (Ticket class) column and count the number of passengers in each class.
4. Rename the "Pclass" column to "PassengerClass" using the alias function.
5. Sort the DataFrame by the "Age" column in descending order.
6. Calculate the average age of passengers in each passenger class using the groupBy and agg functions.
7. Find the top 5 passengers with the highest fare.
8. Create a new DataFrame that contains the total fare collected per embarkation point (C = Cherbourg, Q = Queenstown, S = Southampton). Then, join this DataFrame with the Titanic dataset on the "Embarked" column.
9. Calculate the survival rate per **class and gender**.
10. Save the modified DataFrame titanic_df to parquet format **and read it back.**

🎯 Deliverable and grading:

- Return all the tasks in a single file, we will run the code in our shell for evaluation.
- Each task is worth **1 point.**

Notes:

- You are free to perform these tasks using both PySpark DataFrames, or Spark SQL.
- This assignment aims to get you to explore Dataframes, you should try your best effort to read the documentation (it's important), and seek help on Slack.