**SI 618 Fall 2017 Lab 4**

In this lab, we will practice SparkSQL. Put your NFL team jerseys on because we will look into NFL data (Go Hawks!).

hdfs:///var/si618f17/NFLPlaybyPlay2015.json

(1 points) Load the json file and register it as a table.

(3 points) For each team, identify how able their offense is in progressing the ball compared to their opponent in each game (measured through total yards gained). Let’s call this *delta-yards*. For example, imagine a hypothetical match between Seattle Seahawks (SEA) and Green Bay Packers (GB). Let’s say total yards for Seahawks that game was 500 and for Packers it was 400. Then, the delta-yards for Seahawks for that game would be 100 and delta-yards for Packers would be -100. Provide the sorted list of teams as a function of mean of this measure across all games (let’s call this *mean-delta-yards*). You output should include teamname <tab> mean-delta-yards in decreasing order of mean-delta-yards. Save it as si618\_f17\_lab4\_output\_1\_UNIQUENAME.tsv. Your output should match lab4\_desired\_output\_1.tsv.

(3 points) Some teams have better running backs while others have better receivers. Similarly, some quarterbacks are accurate passers while others are more mobile and hence able to run the ball. So teams vary in how much they rely on the “run” vs “pass” game. You can learn this from this data. For each team estimate the run to pass (PlayType) ratio. And output teamname <tab> run-to-pass-ratio where teams are sorted in increasing order of run-to-pass-ratio.

Save it as si618\_f17\_lab4\_output\_2\_UNIQUENAME.tsv. Your output should match lab4\_desired\_output\_2.tsv.

(3 points) Identify the most penalized 10 players in the league. Output results where each row is of the form: player-name <tab> teamname <tab> number of penalties.

The lines should be sorted in a decreasing order of number of penalties and then an alphabetical order of teams. Save it as si618\_f17\_lab4\_output\_3\_UNIQUENAME.tsv. Your output should match lab4\_desired\_output\_3.tsv.

You MUST use SparkSQL to do this homework. Other solutions will not get any credit

**What to submit:**

Submit a zip file named si618\_f17\_lab4\_youruniqname.zip containing your Python source code file, si618\_f17\_lab4\_uniquename.py, and 3 tsv files.