CS 696 Intro to Big Data
Fall Semester, 2016
Assignment 3

San Diego State University -- This page last updated 9/11/17

Using Spark
Due Oct 8 23:59

In this assignment you are to use Scala and Spark. Turn in a Jupyter notebook that uses the Apache-Toree kernel. In problem 2-7 your code needs to use Spark and be able to run on a cluster using multiple worker nodes to the computations.

1. Write a function that will put N doubles into a file. The doubles need to be normally distributed with mean 0 and standard deviation 1. The function should have two arguments: N and the full name of the file (ie includes path to file location).

Create a file with 50,000 doubles using the function from problem 1. This file will be used for the next several problems. It is best if you put the file in the current directory to avoid paths that do not exist on other machines.

2. Read the file created in #2 into an RDD and compute the mean and standard deviation of the doubles in the file. Work on the RDD, that is do not convert the RDD to a DataFrame or Dataset.. You are to use Spark code to compute the values as we want this to run on a cluster using multiple machines. So the pure Scala code you used in assignment will not work.

3. Repeat #3 but using a DataFrame instead of RDD. Here work on the DataFrame not an RDD.

4. Using a DataFrame create a random sample of about 100 elements of the file created in #2 and compute the mean of the sample.

5. Create a file of 100 normally distributed doubles. Read the doubles from the file into an RDD. Using the RDD create a sliding window of size 20 and compute the mean of each window.

The following problems deal with dwell times on websites. That is how long people stay on a web page. It uses files you download from the assignment page on the course website.

6. The file “multiple-sites.tsv” contains two columns: site and dwell-time. Using Spark compute the average dwell time for each site.

7. The file “multiple-sites.tsv” contains two columns: date and dwell-time. Using Spark compute the following:

    1. The average dwell time each hour
    2. The average dwell time per day of week
    3. The average dwell time on week-days (Monday - Friday)
    4. Average dwell time on the weekend.

8. Do the average dwell times computed in #7 indicate any difference in users behavior?

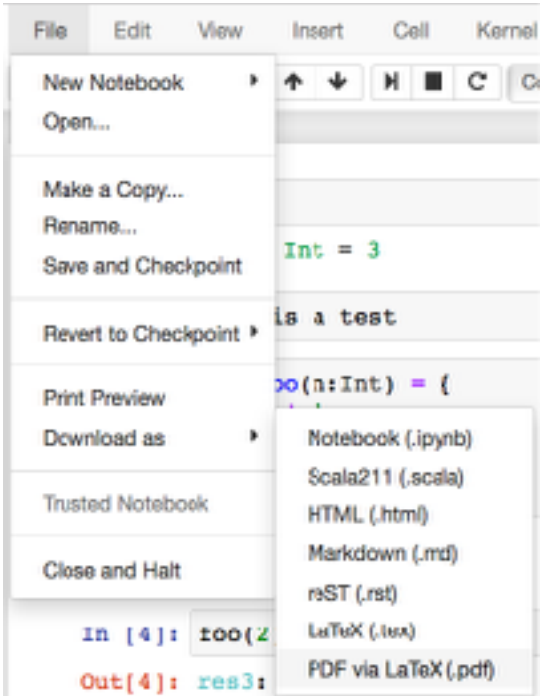
Grading

Each problem is worth 10 points.

What to turn in

You are to turn in a Jupyter [Jupyter-scala] notebook containing the answers to the questions above. Since Jupyter notebooks can contain text and code, before each problem indicate which problem it is in text, not in code comment.

To turn in your assignment download your Jupyter notebook as an Notebook (.ipynb). See image below. This will allow me to run your assignment in Jupyter. Do not download it as a Scala file (.ji) as this will not run in Jupyter and removes all the text (markdown).



Once you have downloaded the assignment zip it up and then upload the zip file to the course portal.

Late Penalty

An assignment turned in 1-7 days late, will lose 5% of the total value of the assignment per day late. The eight day late the penalty will be 40% of the assignment, the ninth day late the penalty will be 60%, after the ninth day late the penalty will be 90%. Once a solution to an assignment has been posted or discussed in class, the assignment will no longer be accepted. Late penalties are always rounded up to the next integer value.