Individual Assignment 3, Spring 2019, Due: April 11 by 11:59PM

*Individual Category C – Question 1*

*Individual Category A – Question 2*

Submission:

Please put all of your files (code and output) into a single folder, ZIP it, and then submit via Blackboard. Save your ZIP file as LastName-FirstName.ZIP. (If you have issues uploading to Blackboard, then email it to me directly. Email both my Mason and GMAIL.)

Emails: Joe.Wilck@mason.wm.edu Joe.Wilck@gmail.com

Grade:

This assignment will count as 15% of your course grade. Questions 1 and 2 will count 50% each.

Question 1:

Data: You will also need the ml-100k data (Lesson 2)

Create a Spark (Scala) program that uses Data Frames (i.e., Spark SQL) like we did in Lessons 16, 18, etc. that calculates the average rating for every movie with at least 100 ratings. Sort the output so that it includes average ratings from smallest to largest. The output should be formatted

<Movie Name>”,”<Average>”,”<Number of Ratings> (or something similar).

[Hint: You may need to start this one from scratch, since it combines a number of concepts from the examples we went over in prior lessons.]

*Note, this is the same question that was asked on the Lesson 7 (Group Assignment 1) and the Individual Assignment 2 (Question 1). But you need to do it using Spark SQL. By now we all know that the first entry in this file should be The Cable Guy and the last entry should be A Close Shave.*

Include your Scala code and your output.