Introduction to Machine Learning with Apache Spark

One of the most common uses of big data is to predict what users want.  This allows Google to show you relevant ads, Amazon to recommend relevant products, and Netflix to recommend movies that you might like.  This lab will demonstrate how we can use Apache Spark to recommend movies to a user.  We will start with some basic techniques, and then use the [Spark MLlib](https://spark.apache.org/mllib/) library's Alternating Least Squares method to make more sophisticated predictions.

The lab is due Jul 3, 2015 at 00:00 UTC. There is a three day grace period for late submissions until the end of the course on Jul 6, 2015 at 00:00 UTC.

HOW TO COMPLETE THIS ASSIGNMENT

This assignment is broken up into sections with bite-sized examples for demonstrating Spark functionality for Machine Learning. For each problem, you should start by thinking about the algorithm that you will use to *efficiently* process the log in a parallel, distributed manner. This means using the various [RDD](http://spark.apache.org/docs/latest/api/python/pyspark.html#pyspark.RDD) operations along with [lambda functions](https://docs.python.org/2/tutorial/controlflow.html#lambda-expressions) that are applied at each worker.

INSTRUCTIONS

Here are the instructions.

1. Start the VM - To start the VM, from a DOS prompt (Windows) or Terminal (Mac/Linux), issue the command "vagrant up", while in the custom directory [created](https://courses.edx.org/courses/BerkeleyX/CS100.1x/1T2015/courseware/d1f293d0cb53466dbb5c0cd81f55b45b/920d3370060540c8b21d56f05c64bdda/) for this course.
2. Once the Virtual Machine is running, access the Jupyter web UI for running IPython notebooks by navigating your web browser to "<http://localhost:8001/>" (or "<http://127.0.0.1:8001/>").
3. If you have any running notebooks they **SHOULD BE** shutdown.  Only **ONE** notebook should be run at a time.  Running notebooks have a green icon to the left of the notebook name and green text to the right of the screen that says "Running".  Shutdown running notebooks by clicking the checkbox next to the notebook and then clicking the orange "Shutdown" button.
4. Download the Lab 4 IPython notebook.  Make sure that the file extension is .ipynb.  If the download adds an extension (e.g. ".txt"), rename the file so that the extension is just .ipynb.
   * Lab 4 Introduction to Machine Learning with Apache Spark <https://raw.githubusercontent.com/spark-mooc/mooc-setup/master/lab4_machine_learning_student.ipynb>. You can view this lab exercise online [here](http://nbviewer.ipython.org/github/spark-mooc/mooc-setup/blob/master/lab4_machine_learning_student.ipynb).
5. Upload the IPython notebook.  This process was explained during "[Setting up the Course Software Environment](https://courses.edx.org/courses/BerkeleyX/CS100.1x/1T2015/courseware/d1f293d0cb53466dbb5c0cd81f55b45b/920d3370060540c8b21d56f05c64bdda/)" in week one.
6. In the **Lab 4 Introduction to Machine Learning with Apache Spark** notebook, please follow the instructions in the notebook and replace <FILL IN> sections with your solution.  After you confirm that your code passes all of the tests while running in the VM, please export it as a **python file (.py)**and submit it to the autograder server.  The submission process is the same as in the previous week for Lab 3 - *make sure you submit to the correct autograder.*  In the next module, the instructions are provided again for your convenience.
7. An outline of what will be covered in the notebook is included below.
8. When you have submitted successfully, you can shutdown the VM by issuing the command "vagrant halt".

INTRODUCTION TO MACHINE LEARNING WITH APACHE SPARK

This exercise consists of 3 parts and quiz questions:

* *Part 1*: Basic Recommendations
* *Part 2*: Collaborative Filtering
* *Part 3*: Predictions for Yourself (this is part where you will enter your own ratings and see what movies are recommended for you)

After you complete the lab, make sure you answer the quiz questions in the next section.

**Please do not post your programming exercises in publicly visible repositories, such as GitHub.**