IEMS 5709 Fall 2016 Homework 3

Release date: Nov 15, 2016
Due date: 23:59, Dec 5, 2016

The solution will be posted right after the deadline, so no late homework will be accepted!

Every Student MUST include the following statement, together with his/her signature in the submitted homework.

I declare that the assignment submitted on Elearning system is original except for source material explicitly acknowledged, and that the same or related material has not been previously submitted for another course. I also acknowledge that I am aware of University policy and regulations on honesty in academic work, and of the disciplinary guidelines and procedures applicable to breaches of such policy and regulations, as contained in the website http://www.cuhk.edu.hk/policy/academichonesty/.

Signed (Student_____) Date:_____

Name_____ SID_____

Submission notice:

- Submit your homework via the elearning system

General homework policies:

A student may discuss the problems with others. However, the work a student turns in must be created COMPLETELY by oneself ALONE. A student may not share ANY written work or pictures, nor may one copy answers from any source other than one's own brain.

Each student **MUST LIST** on the homework paper the **name of every person he/she has discussed or worked with**. If the answer includes content from any other source, the student **MUST STATE THE SOURCE**. Failure to do so is cheating and will result in sanctions. Copying answers from someone else is cheating even if one lists their name(s) on the homework.

If there is information you need to solve a problem but the information is not stated in the problem, try to find the data somewhere. If you cannot find it, state what data you need, make a reasonable estimate of its value, and justify any assumptions you make. You will be graded not only on whether your answer is correct, but also on whether you have done an intelligent analysis.

Q1 [20 marks]: Basic Hive Operations

You are required to use Hive in IE DIC Cluster to perform exactly the same task as that of Question 3 in Homework 2 with the same datasets. Compare the performance with Pig in terms of overall run-time and explain your observation.

Submit your output, explanation and your Hive commands/ scripts.

Hints:

- Hive will store its tables on HDFS and those locations needs to be bootstrapped:

  $ hdfs dfs -mkdir /tmp
  $ hdfs dfs -mkdir /user/hive/warehouse
  $ hdfs dfs -chmod g+w /tmp
  $ hdfs dfs -chmod g+w /user/hive/warehouse

- While working with the interactive shell (or otherwise), you should first test on a small subset of the data instead of the whole data set. Once your Hive commands/ scripts work as desired, you can then run them up on the complete data set.

Q2-Q4 [80 marks]: Spark related questions:

Q2: Sparks RDD Basics [25 marks]
Q3: Spark PageRank [30 marks]
Q4: Spark Streaming [25 marks]

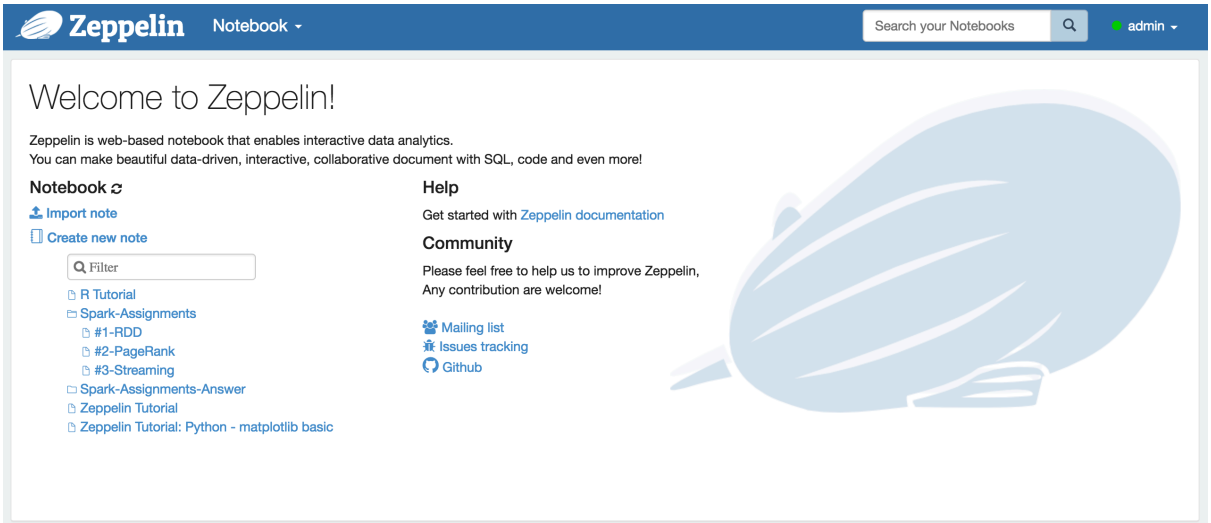
Spark Homework consists of three separate assignments: Q2, Q3 and Q4. For all Spark assignments, you will be using Zeppelin Notebook, a web-based notebook that enables interactive data analytics. Detailed description and requirements are listed in the question notebooks.

To use Zeppelin Notebook, you have to login first at http://dicvm1.ie.cuhk.edu.hk:8080/
Both username and password are initialized to be your student ID. Unfortunately, in this system only the administrator can change passwords. If you want to change your password, please encrypt your password with MD5 encoder at http://www.xorbin.com/tools/md5-hash-calculator and send me the encrypted codes along with your student ID.

For example, if your password is iems5709

Just send me the MD5 hash which is e823880b8121c44cd06b167fde580e82 in this case. I won't be able to know what your real password is, neither will others.

After login, you will see a page like this:



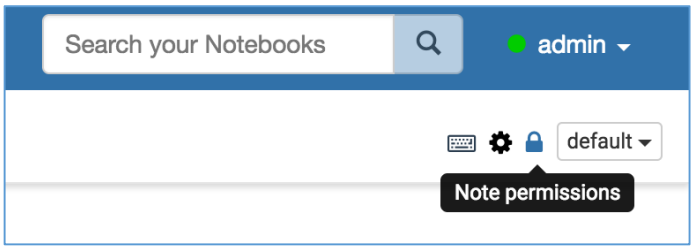
Your question notebooks are in the Spark-Assignments folder:

1. RDD Basics
2. PageRank
3. Spark Streaming

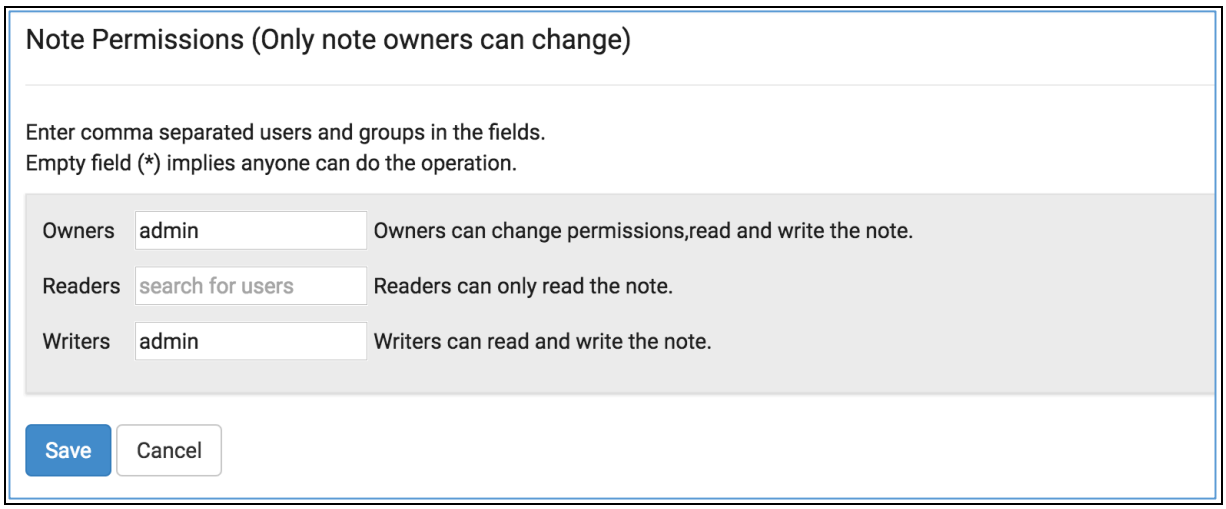
Note that you are only authorized to read these question notebooks. To finish your homework, you need to create a notebook of your own and write your codes and answers there. You are free to use the codes given in the question notebooks.

Although there you can see there is a Spark-Assignments-Answer folder, you will not be able to open the notebooks inside until the due day.

Another important thing to keep in mind is, Zeppelin allow the notebook owners to edit the permission of each notebook. By click the "lock" icon at the top-right corner (see below) after you open your notebook,



You can let specific users (in “username”) to read or see your notebook by modifying the corresponding permission (read and write) at the following cell that pops up after click



Therefore, do remember to modify your read and write permission so as to keep yourself from plagiarism. Just fill your student ID the above three blanks (Owners, Readers, Writers)!