## HU Extension Assignment 03 E63 Big Data Analytics

### Handed out: 09/15/2017 Due by 11:59 AM EST on Saturday, 09/23/2017

### Use either VMWare Workstation or VMWare Fusion

<https://e5.onthehub.com/WebStore/Welcome.aspx?ws=4185a0dc-d0d1-e511-9416-b8ca3a5db7a1&vsro=8>

If you feel comfortable with a virtualization technology different from VMWare, please be free to use it. Rather than creating a VM with CentOS, you can create the VM with a different flavor of Linux you more familiar with or consider better.

**Problem 1**. Create your own Virtual Machine with a Linux operating system. The lecture notes speak about CentOS. You are welcome to work with another Linux OS. When creating the VM, create an administrative user. Call that user whatever you feel like. Please record the password of the new user. Once the VM is created transfer the attached text file Ulysses10.txt to the home of new user. You can do it using scp (secure copy command) or email. Examine the version of Java, Python and Scala on your VM. If any of those versions is below requirements for Spark 2.2 install proper version. Set JAVA\_HOME environmental variable. Set your PATH environmental variable properly, so that you can invoke: java, sbt and python commands from any directory on your system. [20%]

**Problem 2**. Install Spark 2.2 on your VM. Make sure that pyspark is also installed. Demonstrate that you can successfully open spark-shell and that you can eliminate most of WARNing messages. [15%]

**Problem 3**. Find the number of lines in the text file ulysses10.txt that contain word “afternoon” or “night” or “morning”. In this problem use RDD API. Do this in two ways, first create a lambda function which will test whether a line contains any one of those 3 words. Second, create a named function in the language of choice that returns TRUE if a line passed to it contains any one of those three words. Demonstrate that the count is the same. Use pyspark and Spark Python API. If convenient you are welcome to implement this problem in any other language: Scala, Java or R. [15%]

**Problem 4.** Implement the above task, finding the number of lines with one of those three words in file ulysses10.txt using Dataset/DataFrame API. Again, use the language of your choice. [20%]

**Problem 5**. Create a standalone Python script that will count all words in file ulysses10.txt. You are expected to produce a single number. Do it using RDD API. If convenient, you are welcome to implement this problem in other languages: Scala, Java or R. [%15]

**Problem 6.** Create a standalone Python script that will count all words in file ulysses10.txt. You are expected to produce a single number. Do it using Dataset/DataFrame API. If convenient, you are welcome to implement this problem in other languages: Scala, Java or R. [%15]

Please, describe every step of your work and present all intermediate and final results in a Word document. Please, copy past text version of all essential command and snippets of results into the Word document with explanations of the purpose of those commands. We cannot retype text that is in JPG images. Please, always submit a separate copy of the original, working scripts and/or class files you used. Sometimes we need to run your code and retyping is too costly. Please include in your MS Word document only relevant portions of the console output or output files. Sometime either console output or the result file is too long and including it into the MS Word document makes that document too hard to read. PLEASE COPY Snippets of your Code but DO NOT EMBED files into your MS Word document. For issues and comments visit the class Discussion Board on Piazza.