# Homework 2: Practicing Spark

### Data Intensive Computing Spring 2018

**Due Date: April 26rd, 2018 at 11:59 pm**

In this homework, you’ll have hands on experience with Spark. You need to install Spark to your laptop/server. You are free to use Spark in standalone mode or in Hadoop YARN mode.

* Install Spark
* Download Shakespeare’s book in this [link](http://www.gutenberg.org/files/1524/1524-0.txt). Save it as “*hamlet*”. You’ll use this in questions 3-15
* Download the ZIP file [here](https://archive.org/details/twitter_cikm_2010) that you used in homework 1. We will use same file in this homework as well. Its size is around 405 MB.
* Unzip files “training\_set\_tweets.txt” and “training\_set\_users.txt” and upload them to HDFs if you plan to use Spark on YARN
* You’ll use Spark-Shell for questions 1-15 and write standalone Sparks application for questions 16-18

Questions

1. Launch the Spark shell. (**2pt**)
2. Make a parallel collection of Array(1, 2, 3, 4, 5)  and sum up all its

elements. (**2pt**)

1. Create an RDD named *pagecounts*from the input file hamlet (**3pt**)
2. Get the first 10 lines of hamlet (i.e., first 10 records of *pagecounts*). (**3pt**)
3. Make a more readable print of the step 4. (**3pt**)
4. Count the total records in the data set pagecounts, and confirm its correctness by comparing the result with the Bash wc  command: wc -l hamlet . (**3pt**)
5. Filter the data set *pagecounts*and return the items that have the word

*“this”*. (**5pt**)

1. Cache the new data set in memory, to avoid reading from disks. Show cached RDD in web interface (**5pt**)
2. Find 5 lines with the most number of words. Print lines and the number of words(**6pt**)
3. Count the total number words. (**3pt**)
4. Count the number of unique words. (**5pt**)
5. Count the number of each word. (**10pt**)
6. Show the jobs for Q12 in web interface (**3pt**)
7. Save the data set in a text file. (**3pt**)
8. Collect the word counts in the shell. (**4pt**)
9. Write a standalone application in Spark to find 20 the most mentions in “training\_set\_tweets.txt” file (**20pt**)
   1. Mentions starts with “@”.
   2. Multiple mentions in a tweet counts one. For example, “*I love UNR @unr @unr @unr @cse @cse*”. You need to count one “@unr” and one “@cse” from this tweet
   3. You need to strip punctuations from mentions. For example, *@unr! = !@unr = @unr*
10. Write a standalone application in Spark to find 10 retweeted users (aka twitter handle) (**20pt**)
    1. Example twitter handle is “@unr”
    2. Retweets can be identified by “RT @” filters. Simply searching for RT will lead incorrect results

## Graduate Students only

1. Write a standalone application in Spark to find 10 most tweeted users (ids of users) from Los Angeles for dates 09/16/2009 - 09/20/2009 (**30pt**)

## What to deliver

Create following files/folders and compress them in a single zip file with name <**LASTNAME>\_<NAME>\_HW2.zip** and submit on WebCampus

1. Take screenshots for Question 1-15 to a file answers1-15.pdf (Please take short screenshots. For questions with long output, please only paste your *command* and 5-10 lines of output
2. Copy the most repeated mentions along with number of occurrences to a file called “popular\_mentions.txt” file
3. Copy the most retweeted users along with number of retweets to a file called “most\_retweeted\_users.txt” file
4. Copy the most tweeted users along with number of tweets to a file called “most\_tweeted\_users.txt” file
5. Create three directories Q16, Q17, and Q18 and copy your source code for question 16, 17, and 18 into those directories. Please also include README document to describe how to run your code.