# Homework 1: Analyzing Sample Twitter data

### Data Intensive Computing Spring 2018

**Due Date: March 13, 2018 at 11:59 pm**

In this homework, you’ll write a MapReduce programs to analyze sample twitter dataset containing approximately 3.8 million tweets.

* Install Hadoop to your own server or AWS.
  + For AWS account please go to [AWS student page](https://aws.amazon.com/about-aws/whats-new/2015/05/aws-educate-students-and-educators-can-access-aws-technology-cloud-courses-training-and-collaboration-tools/) and create a student account. You may need to enter credit card information but there won’t be any charges unless you exceed free-tier usage limitations. (Student account gives you 750 hours of Amazon EC2  Linux t2.micro instance which is more than enough for this project) Please stop instances when you’re not using it. You are responsible for any charges caused by overuse.
* Download ZIP file in [here](https://archive.org/details/twitter_cikm_2010). Its size is around 405 MB.
* Unzip the file and upload “training\_set\_tweets.txt” and “training\_set\_users.txt” files to HDFS as “tweets” and “users”

Once your Hadoop cluster is up and running do the following tasks:

1. Run “jps” command and show running tasks **(5 pts)**
2. Show how many blocks created in HDFS for “tweets” file **(5 pts)**
3. Show how many map tasks are created when you try to process “tweets” file in HDFS **(10pts)**
4. Set the number of reduce tasks to 3 and show that Hadoop created 3 reduce tasks

**(10 pts)**

1. Write a MapReduce code to count the occurrences of hashtags and find the most repeated 100 hashtags. **(20 pts)**
2. Write a MapReduce code find the most tweeted 5 days. (Tweets are associated with time stamps so you need to count all the tweets posted in same days) **(20 pts)**
3. Write a MapReduce code to find the most tweeted 10 cities along with the number of tweets (“training\_set\_users.txt” file has user\_id 🡪 city relation to extract city information) **(30 pts)**

## What to deliver

Create following files/folders and compress them in a single zip file with name <**LASTNAME>\_<NAME>\_HW1.zip** and submit on WebCampus

1. Take screenshots for Question 1-4 to a file answers1-4.pdf
2. Copy the most repeated 100 hashtags along with number of occurrences to a file called “popular\_tweets.txt” file
3. Copy the most tweeted 5 days along with number of tweets to a file called “most\_tweeted\_days.txt” file
4. Copy the most tweeted 10 cities along with number of tweets to a file called “most\_tweeted\_citites.txt” file
5. Create three directories Q5, Q6, and Q7 and copy your source code for question 5, 6, and 7 into those directories.

**Cheaters will be caught and punished harshly so please do not attempt!**