Assignment 2 : Counting & Heavy Hitters in Twitter

**Post Date:** Thursday, February 28, 2013  
**Due Date:** Thursday, March 7, 2013 11:59 PM  
**TA Office Hours:** Monday, March 4, 2013 10:00 AM - 12:00 PM (North Building N021)

In this assignment, you will be identifying trends in Twitter data. We have already streamed \( \sim 10 \) millions tweets (one day) and stored them in http://www.cs.duke.edu/courses/spring13/compsci590.2/Assignment2/tweetstream.zip. Like in the previous assignment, you are advised to start working as soon as possible, and in groups. Throughout the assignment, assume that \( \delta = 0.05 \). For problems 1-3 read each tweet at most once. For problem 4 read each tweet at most twice. Bonus credit if you can solve problem 4 using a one pass mechanism.

1. **Frequencies of HashTags:** Using the first 0.5 million tweets, construct a uniform sample \( S \) of 20 hash tags from the twitter stream. For each of the hash tags in \( S \), estimate the number of times they appear in the rest of the stream within an error of \( \epsilon = 0.001 \). Plot these counts after seeing every 30 minutes of tweets.

2. **Number of Distinct HashTags:** Estimate the number of distinct hash tags in the stream \((\epsilon = 0.01)\).

3. **Heavy Hitter HashTags:** For \( \phi = 0.001 \) and \( \epsilon = 0.001 \), develop an algorithm to output a set of hash tags \( H \) such that (a) every hashtags that appears in at least a \((\phi + \epsilon)\) fraction of the tweets is in \( H \) with probability 1, and (b) with probability \((1 - \delta)\) any hashtag that appears in less than \( \phi \) fraction of the tweets does not appear in \( H \).  
   *Hint: It is OK to return only the first 15 characters of the hashtag.*

4. **Location Specific HashTags:** For this part of the assignment, consider only the tweets having non-null latitude/longitude values in the `geo` field. Assume that the latitude/longitude values span a two dimensional space. Divide this 2D space into grid cells of size 0.1 x 0.1 (round off every lat/long to the first decimal place).
   
   (a) Identify 20 grid cells with the most number of tweets.
   (b) For each of these grid cells and each \( h \in S \) (from 1), compute the fraction of tweets for that location that contain \( h \).
   (c) For each of these grid cells and each \( h \in H \) (from 3), compute the fraction of tweets for that location that contain \( h \).