Assignment 2 : Counting & Heavy Hitters in Twitter

Post Date: Friday, September 18, 2015
Due Date: Friday, September 25, 2015 11:59 PM
TA Office Hours: Wednesday, September 9, 2015 4:30 PM - 6:00 PM (LSRC D309)

In this assignment, you will be identifying trends in Twitter data. We have already streamed 10 millions tweets (one day) and stored them in `https://www.cs.duke.edu/courses/fall15/compsci590.4/assignment2/tweetstream.zip` (2.1G). The meaning of the fields of a tweet can be found at Tweets Page. Throughout the assignment, assume that $\delta = 0.05$. For problems 1-4 read each tweet at most once. For problem 5 read each tweet at most twice. Bonus credit if you can solve problem 5 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 $\epsilon = 0.001$ with probability $1 - \delta$. Plot these counts after seeing every 30 minutes of tweets (counts v.s. time). (2 points)

2. (BONUS) Given the uniform samples of 15 students from Q1, $S_1, ..., S_{15}$, what is the probability for one of the samples S_i have the same hash tag twice? (2 points)

3. *Number of Distinct HashTags*: Estimate the number of distinct hash tags F' in the stream, such that $F/c \leq F' \leq cF$, where F is the true number of distinct hash tags in the stream, with probability $1 - \delta$. (Take $c = 5$) (2 points)

4. *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 ϕ fraction of the tweets does not appear in H. Hint: It is OK to return only the first 15 characters of the hashtag. (3 points)

5. *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). (3 points)

   (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 within an error of $\epsilon = 0.001$ with probability $1 - \delta$.

   (c) For each of these grid cells and each $h \in H$ (from 4), compute the fraction of tweets for that location that contain h within an error of $\epsilon = 0.001$ with probability $1 - \delta$.

Like in the previous assignment, you are advised to start working as soon as possible. However, this assignment must be your own **individual** work. Please submit a PDF report to SAKAI, answering the 5 questions above, and outlining the method used to obtain the answers.