Data Mining

Homework 5

Due: 8/6/2014, 23:59.

You must hand in the homeworks electronically and before the due date and time. Check the web page for instructions about collaboration, about being late, and about handing in the homework.

Hand in:

- The `.py` files with the source code.
- Instruction to execute them.
- The output files generated, and output at the screen.

Problem 1. In this problem you are requested to implement the streaming algorithms that we did in class. We will implement them on data generated by twitter. Your algorithms should be able to obtain the data and compute their estimates online.

1. Implement the Flajolet–Martin estimate for counting distinct elements (F_0). Obtain ℓ independent estimates and combine them using the median of the average technique.
2. Implement the Alon–Matias–Szegedy algorithm for estimating the second moment (F_2). Obtain ℓ independent estimates and combine the by taking the average.
3. Create programs for calculating F_0 and F_2 without the streaming model (shell commands can be useful).

First test your algorithms by trying on the dataset at:
`http://code.google.com/p/crush-tools/downloads/detail?name=access_log.tar.gz`
by considering the frequencies of the various IPs. Experiment and report results for different values of ℓ and different group sizes (for the Flajolet–Martin schema).

Then apply your algorithms on twitter data obtained by the twitter streaming API, as they are generated. Save also the data on disk so that you can verify your algorithms.

For each case, for the different values of ℓ and group sizes (for F_0) you should report in two tables, one for F_0 and one for F_2:

- the number of records
- the values of F_0 (or F_2) returned by your streaming algorithm
- the true values F_0 (or F_2)
- the absolute and relative errors
- the value of ℓ
- the group size (for the Flajolet–Martin schema)