Assignment 2:
Counting Distinct Elements in Practice

Wojtek Kowalczyk

Motivation

- There are several memory-efficient algorithms for approximate counting of distinct elements in data streams. We will focus on 3 algorithms:

  - The one described in the textbook (MMDS) (*does it really work?*)
  - Probabilistic Counting (Flajolet-Martin) and
  - LogLog Counting (Durand-Flajolet)

- To apply these algorithms <u>in practice</u>, one should know the relation between:

  - the magnitude of the expected count (**N**)
  - the amount of the required memory (**M: number of bytes**)
  - **Relative Approximation Error**:

    RAE = abs(true_count - estimated_count)/true_count

- The algorithms and theoretical bounds are described in papers

Your Task

- Study all three algorithms

- Implement them

- Run multiple experiments to experimentally establish the relation between:

  - the magnitude of the expected count (**N**)
  - the amount of the required memory (**M: number of bytes**)
  - Relative Approximation Error **(RAE)**

- **Write a report with your own description of both algorithms, their theoretical properties (according to the papers) and your own experimental results. Are your results consistent with theory?**

Your Task in more details:

- Study the relevant papers on:
    - Probabilistic Counting (Flajolet-Martin)
    - LogLog Counting (Durand-Martin)
    - Slides 24-26, 29-31 from *aofa07.pdf*

Starting point:
http://blog.notdot.net/2012/09/Dam-Cool-Algorithms-Cardinality-Estimation

- Implement these 3 algorithms (as you like – the don't have to be fast/efficient)

- Run multiple experiments (for various numbers of distinct elements and setups) to establish/verify the trade-off between the error and the amount of required memory. (required memory="memory required by your algorithm, assuming the most memory-efficient implementation")

- Write a report

The report should:

- Describe the algorithms, together with their theoretical properties
in an accessible way

- Design/describe your own "experimental setup"

- Perform experiments and document your findings

- Summarize findings and make conclusions in a form of a Practical Guide:

How to count distinct elements in limited memory?

Moreover:

- Treat it as your own research project. In particular:

  - make your own choices concerning the experimental setup

  - decide yourself how much of "theoretical estimates" you want to use

  - decide yourself how you want to present the "practical guide"

  - (optionally): suggest a concrete application of approximate counting, e.g., to cybersecurity, marketing, databases, ...

Deadline: Monday, 16 October, 23:59
Submit both the report and the scripts in the same way as you did with A1
(this time 'A1' should be replaced with 'A2')