Computational Social Science
APAM E4990
Spring 2013

Homework 1

1. In-memory descriptive statistics

In this exercise you'll investigate the impact of inventory size on customer satisfaction for the 10M ratings MovieLens[1] dataset discussed in class, producing the equivalent of Figure 2 from the "Anatomy of the Long Tail" paper[2] for these data.

Specifically, for the subset of users who rated at least 10 movies, produce a plot that shows the fraction of users satisfied (vertical axis) as a function of inventory size (horizontal axis). We will define "satisfied" as follows: an individual user is satisfied p% of the time at inventory of size k if at least p% of the movies they rated are contained in the top k most popular movies. As in the paper, produce one curve for the 100% user satisfaction level and another for 90%—do not, however, bother implementing the null model.

You may use R, Python, or MATLAB with standard libraries and dependencies. (Please clear other languages or tools with us beforehand.)

2. Local streaming statistics

When working with a new dataset, we typically begin by collecting the salient statistics—mean, median, etc.—to gain some intuition about the data. In this exercise you'll develop a simple streaming tool to compute these statistics within subgroups of the data.

Specifically, your goal is to write a script that takes as input a text file with two tab-separated columns—the first a key that represents the group and the second a numeric value for that observation—and output:

- the minimum value for each key,
- the median value for each key,
- the average value for each key,
- and the maximum value for each key.

You may assume that the data have been pre-grouped, so that all observations for a given key appear in consecutive lines. Your script should stream through the data exactly once and should avoid loading the entire dataset in to memory.

For instance, suppose the input is the movie ids and ratings from the MovieLens dataset in question 1. This script would then compute the minimum, median, average, and maximum rating for each movie, outputing the statistics for each movie on a separate line. For example, this input:

```
4       1
4       2
4       2
4       4
2       4
2       2
3       3
3       2
3       1
1       1
```

[1] http://www.grouplens.org/system/files/ml-10m-README.html

[2] http://doi.acm.org/10.1145/1718487.1718513

should produce the following output:

```
4        1       2.00         2.25        4
2        2       3.00         3.00        4
3        1       2.00         2.00        3
1        1       1.00         1.00        1
```

where the columns give the key, minimum, median, average, and maximum for each group. Your script should take the input filename as a command line argument and produce the appropriate output on standard output.

Example input and output along with a solution template is available here: `http://github.com/jhofman/css2013/tree/master/homework/homework_01`.

3. Counting scenarios

You are given a dataset of phone calls between pairs of people, listing the caller, callee, time of phone call and duration of the phone call (in seconds), a snapshot is given below:

```
2125550123    2125559876    Wed Feb 13 19:27:47 EST 2013    123
6465550123    4155559876    Tue Feb 19 11:35:09 EST 2013    1
4155550912    2125550123    Mon Apr 9 23:33:59 PST 2012     679
2125559876    2125550123    Wed Feb 13 19:07:47 EST 2013    509
...
```

Here the first line represents a phone call lasting slightly over two minutes, the second just a quick 1 second call, etc. Your task is to compute for each pair of phone numbers the total amount of time the parties spent on the phone to each other (regardless of who called whom).

- Suppose your dataset is the call log of a small town of 100,000 people each of whom calls 50 people on average. Please describe how you would compute the statistics.
- Suppose your dataset is a call log of a large town of 10,000,000 people, each of whom calls 100 people on average. Please describe how you would compute the statistics.
- Suppose the dataset is a call log of a nation of 300,000,000 people, each of whom calls 200 people on average. Please describe how you would compute the statistics.

In writing your descriptions above, you don't need to provide actual working code, but please provide enough detail that someone can easily implement your approach. What differences are there between the three different approaches? What prompted you to answer these questions differently?