AMPA E4900: Modeling Social Data

Spring 2019

modelingsocialdata.org

Homework 1

The first problem explores various counting techniques, the second involves some command line and R counting exercises, and the third looks at the impact of inventory size on customer satisfaction for the MovieLens data. Details are in the README.md file for each problem.

Your code and a brief report with your results are to be submitted electronically in one zipped (or tarball-ed) file through the CourseWorks site. All code should be contained in plain text files and should produce the exact results you provide in your writeup. Code should be written in bash / R and should not have complex dependencies on non-standard libraries. The report should simply present your answers to the questions in an organized format as either a plain text or pdf file. All work should be your own and done individually.

1. 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 with each other (regardless of who called whom).

1. 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.
2. Suppose your dataset is a call log of a large city of 10,000,000 people, each of whom calls 100 people on average. Please describe how you would compute the statistics.

3. 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? Would you use an in-memory or streaming approach? A single machine or multiple machines?

2. Counting exercises

This problem contains counting exercises on the command line using Unix tools (e.g., `cut`, `sort`, `uniq`, `wc`, `grep`, and `awk`) and in R using `dplyr` to explore CitiBike usage data. Use the download script to get data for August 2016 and fill in solutions in the `citibike.sh` and `citibike.R` files.

3. The long tail

In this problem you'll investigate the impact of inventory size on customer satisfaction for the 10M ratings MovieLens dataset discussed in class, producing the equivalent of Figure 2 from the Anatomy of the Long Tail 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 (shown in the dashed lines).

Use the download script to get the ratings data and add your solution to `eccentricity.R`. Tip: Use a function from the `readr` package to load the data and name columns.