The goal of this assignment is to have you perform some more complicated tasks using the MapReduce framework. In this assignment you will be using a smaller (toy) data set so that you can easily check your results by loading all the data into memory (e.g., using R).

Background

I recommend MrJob as a Map-Reduce library. Here is a helpful tutorial on MrJob.
https://pythonhosted.org/mrjob/guides/quickstart.html

Additionally, there is useful information in the Python and Hadoop book. (If you haven't already, you should download a copy of this book. It is free from the publisher's website.)
http://www.oreilly.com/programming/free/hadoop-with-python.csp

Data Set

This assignment uses three data files.

States : Name, Abbreviation, Area (Sq. Miles), Population
Electricity : State, Price per Kilowatt Hour
Colleges: This file has a description of each field in the first line of the file

Note: You may find it useful to remove the first line of the colleges file before you use it as input for a Map-Reduce job.

Exercises

Write Map-Reduce applications (in Python or Java) that allow you to answer the following questions.

Note: For some of these problems, it may be helpful to run multiple map-reduce jobs. (That is, pipe the output of the first map-reduce job into a text file and use it as input for the next map reduce job.) If you find it convenient to use multiple Python files to implement your solution, please make sure to include text that makes it clear how the files fit together. (Comments at the top of your code file might be a useful way to do this.)

1. Calculate the largest, smallest, and average (mean) population for a state. Calculate the largest, smallest, and average (mean) area for a state.

2. Calculate the variance in electricity prices among the states.

3. Use linear regression to fit the following simple model

   Population = Area * <alpha> + <beta>

   That is, find <alpha> and <beta> that minimize the squared residuals when the state data is represented using this model

4. Which of the following linear models is a better fit for the electricity data

   Electricity Price = Area * <alpha> + <beta>

   Or

   Electricity Price = Population * <alpha> + <beta>

5. Obtain a random sample of approximately 100 colleges, in which each college is equally likely to appear in the sample.

6. Obtain a random sample of approximately 100 colleges, in which:
   - Each public college is equally likely to be sampled and each private college is equally likely to be sampled.
   - The sample is weighted so that in expectation there are the same number of public and private colleges in the sample