**CS437 Fall 2014  
Problem set 5**

1. **Map-Reduce**

For each of the following problems describe how you would solve it using map-reduce. You should explain how the input is mapped into (key, value) pairs by the map stage, i.e., specify what is the key and what is the associated value in each pair, and, if needed, how the key(s) and value(s) are computed. Then you should explain how the (key, value) pairs produced by the map stage are processed by the reduce stage to get the final answer(s). If the job cannot be done in a single map-reduce pass, describe how it would be structured into two or more map-reduce jobs with the output of the first job becoming input to the next one(s). You should just describe your solution algorithm.

(a) The input is a list of housing data where each input record contains information about a single house: (address, city, state, zip, value). The output should be the average house value in each zip code.

• What are the input and output of the map and reduce stages? How are they computed?

• What is a combine function used for? Would a combine function help with this task? If so, what are the input and output of the function and are the map and reduce stages modified? (

(b) Consider a new input with two different types of lists. One list gives voter information for every registered voter: (voter-id, name, age, zip). The other list gives disease in- formation: (zip, age, disease). For each unique pair of age and zip values, the output should give a list of names and a list of diseases for people in that zip code with that age. If a particular age/zip pair appears in one input list but not the other, then that age/zip pair can appear in the output with an empty list of names or diseases, or you can omit it from the output entirely, depending on which is easier. (Hint: the keys in a map/reduce step do not need to be single atomic values.)

• What are the input and output of the map and reduce stages? How are they computed?