Map Reduce Tutorials - #2 The Group By Operator

Mappers and Reducers

Here's a quick but comprehensive introduction to the idea of splitting tasks into a MapReduce model. The four important functions involved are:

```
Map (the mapper function)
EmitIntermediate(the intermediate key,value pairs emitted by the mapper functions)
Reduce (the reducer function)
Emit (the final output, after summarization from the Reduce functions)
```

We provide you with a single system, single thread version of a basic MapReduce implementation.

Task

The intial input is a number of lines with pairs of cities and their states, in the form: [city],[state] The required output is to group these records by the state - and to finally output, the list of cities in each state, in the format shown. The code for the MapReduce class, reading and splitting the text, parts related to IO etc. has already been provided. However certain parts of the mapper and reducer functions are incomplete. You need to replace the questionmarks (?). Your task is to fill up these question marks appropriately, such that the program works, and outputs the list of cities in each state, in lexicographical order, separated by commas.
Also, this program outputs certain information to the error stream. This information has been logged to help beginners gain a better understanding of the the intermediate steps in a map-reduce process.

Languages Supported
Currently, we provide the base code in Ruby, Python and Java.

Input Format

A list of tab separated pairs of states and cities/towns in those states. We have already written the input handling code to read in this data. This will then be converted to an intermediate format with JSONs of key, value pairs. The key contains [state name]-[country] (e.g. Washington-USA, Kerala-India) The value contains the name of one particular city or town in that state.

Output Format

JSONs of key, value pairs. Again, the output handling part has already been provided in the tempate code. The Key contains [state name]-[country] (e.g. Washington-USA, Kerala-India) **The value contains the entire list of the names of cities or towns in that state, sorted in lexicographical order.** The entities in this list, will naturally be confined to only those cities/towns provided in the input data.

Sample Input

```
Washington-USA  Seattle
Maharashtra-India   Pune
Maharashtra-India   Mumbai
Washington-USA  Spokane
Washington-USA  Olympia
Tamil Nadu-India    Chennai
Tamil Nadu-India    Coimbatore
California-USA  San Francisco
```

California-USA Los Angeles

Sample Output

{"key":"Washington-USA","value":"Olympia,Seattle,Spokane"}
{"key":"Maharashtra-India","value":"Mumbai,Pune"}
{"key":"Tamil Nadu-India","value":"Chennai,Coimbatore"}
{"key":"California-USA","value":"Los Angeles,San Francisco"}