Big Data Lab Answer Sheet.

Please complete this answer sheet and turn it in at the beginning of class on the due date posted in LEARN.

Part I

|  |  |
| --- | --- |
| **Part 1:** | Answer |
| **1**  (4 pts) | The JSON data was organized by the years decreasing and then a decreasing count for each cause. The states were also listed going from highest count to lowest count. |
| **2**  (4 pts) | The XML data was organized by the alphabetical order of the states. Also, the causes were put in alphabetical order aligned with the states. And the it is in decreasing by year starting at 2016 |
| **3**  (4 pts) | Alabama: 2755 Alaska: 439  I think these numbers are so different mainly due to the difference in population of the states. |
| **4**  (6 pts) | I just looked thorough the website and got to the data set of “Demographics by zip code”. I prefer the XML format because I think it is organized more efficiently and is easier to analyze data with it. |
| **Part 2:** | Answer |
| **5**  (2 pts) | The mapper emits every key separately along with its value in it. |
| **6**  (2 pts) | The mapper uses each line of the poem as a piece of the data. |
| **7**  (2 pts) | This would run faster as the input becomes larger because it will be easier and more efficient for the mapper to emit one line at a time. |
| **8**  (2 pts) | By removing the space inside the parenthesis, the mapper emits one line at a time with its value instead of one word at a time from the line. |
| **9**  (2 pts) | The function takes the piece of data and produces its emitted values using the coding. In the code, first, it defines a list variables of the broken-up data based off a provided character, Also, it defines a list variable that keeps track of the times that a word appears in the line. Then, it counts the amount of time a word shows up and it is presented as value depending on the count. |
| **10**  (2 pts) | Each reducer uses the output from the mapper as the input, and then the word is emitted along with the number of times it appears in the dataset. |
| **11**  (5 pts) | This approach still works because it still searches for each time that a word shows up from the mapper, but not the count given by the mapper. |
| **12**  (15 pts) |  |
| **13**  (20 pts) |  |