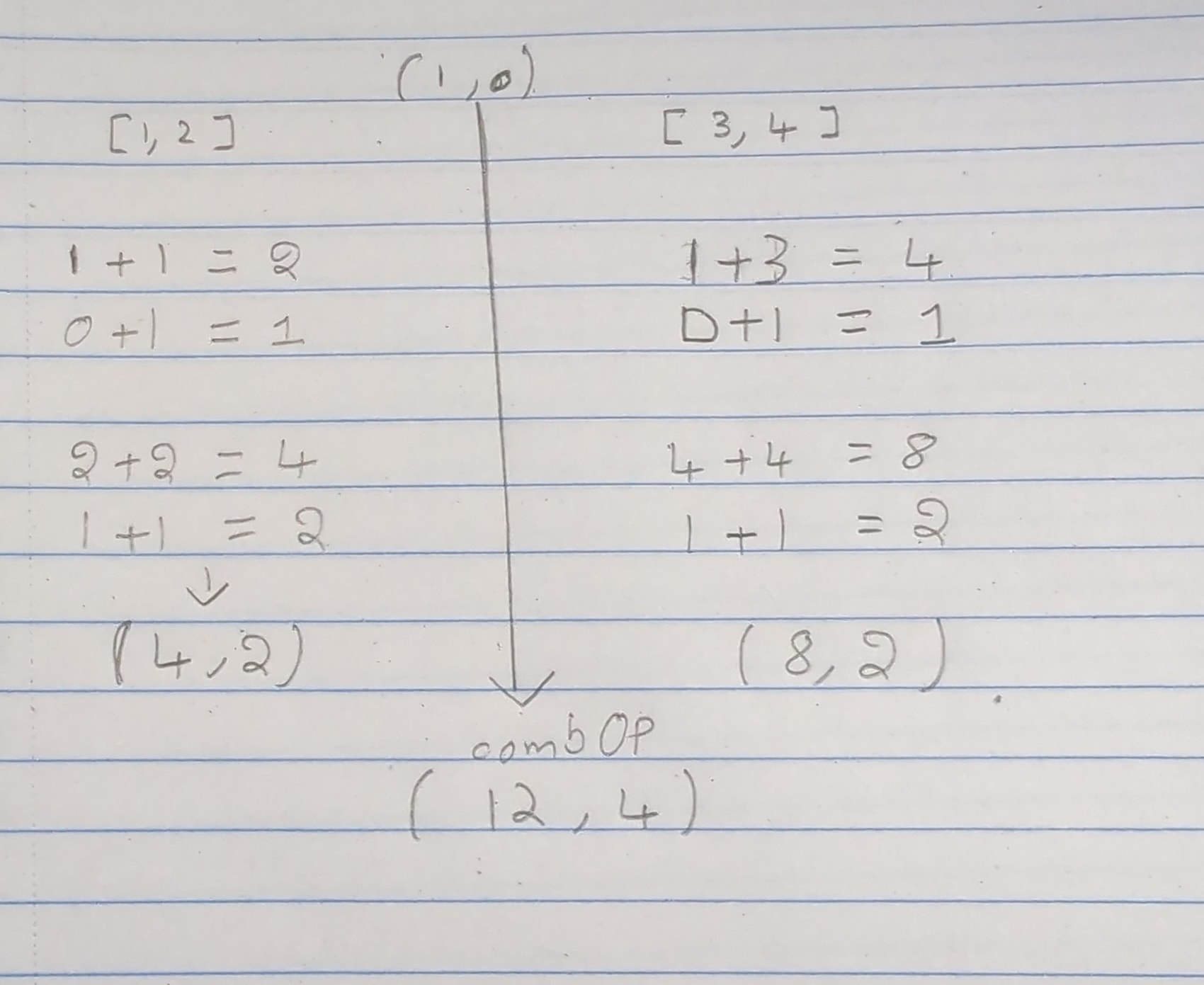
|  |  |  |
| --- | --- | --- |
| **UMBC – CSEE Department** |  | Dr. Waleed Youssef |
| **Data Science Program** |  | youssef1@umbc.edu |
| **Spring 2023** |  |  |
| **DATA 603 – Big Data Platforms** | | |
|  |  |  |
| **Homework #6 – Spark Programming** | | |
|  |  |  |

**Questions:**

1. **[10 points]** Simulate the Aggregate() example in the slides, with initial value of (1,0), What would be the expected results?
   1. List is [1, 2, 3, 4]
   2. Partition is 2
   3. Initial zeroValue is (1, 0)



1. **[15 points]** Write a Spark program that reads your browser history file, then displays the top 5 websites you visited in the last week?



1. **[25 points]** Implement a spark program that performs the following:
   1. Reads the posted text file for a book named “Applied Data Science.txt”
   2. Read the text file into an RDD, and then perform actions and transformations on the RDD
   3. Displays the most used 5 words of length greater than 5 characters in the file *(ensure you result is not case sensitive, so the word “Data” and “data” are the same and should be counted as same word)*
   4. The output should be like this:

**The most used words in the Applied Data Science textbook are:**

**<<word1>> occurred <<n1>> times**

**<<word2>> occurred <<n2>> times**

**<<word3>> occurred <<n3>> times**

**<<word4>> occurred <<n4>> times**

**<<word5>> occurred <<n5>> times**



