

ASSIGNMENT 4

In this assignment, you will implement some of the design patterns that we went over in class.

1. Implement the "in-mapper" combining design pattern for the Wordcount problem using the 6 files that you downloaded for the assignment 1. You have to create an associative array, known as Hashmap in Java, and maintain states across all documents in a map task.

You also need to report the running time for this and compare it with the running time of using a simple combiner i.e. the code that was given to you in lab 3

2. We saw in class how the stripes approach can be used to compute co-occurrence frequency of words. You will use that approach to find the co-occurrence frequency of all words from the 6 files that you downloaded.

Since there are very large number of words, you will limit the words by the following two criteria:

- Remove all stop words in the mapper program
- Have a command line argument that limits words by their length. For example, if the parameter is 5, you will only consider words that are of length 5.