

COMP 330/543 Lab #2: Modifying Hadoop Code

You should begin by opening the WordCount code from Lab 1. There are 3 subtasks you need to complete:

- 1) Modify the WordCount mapper so that the program computes counts not for all the *words* in the corpus, but for all the *bigrams* in the corpus. Bigrams are pairs of words that appear one after another. Consider the sentence:

```
This is a really cool sentence
```

The bigrams in this sentence are:

```
(this, is)
(is, a)
(a, really)
(really, cool)
(cool, sentence)
```

Don't worry about bigrams that span lines. We are only concerned with bigrams on the same line. Represent bigrams as text strings exactly as depicted above (as strings that contain the comma-separated pairs of words, with parens).

- 2) Modify the reducer so that the program only writes out those bigrams that appear more than twenty times overall in the groups.
- 3) After you do this, run your program, copy the first 20 lines from one of your files and submit them to Canvas.

As usual, **REMEMBER TO SHUT DOWN YOUR CLUSTER.**