COMP 330 Lab 2: Writing a Hadoop Program

Your task in this lab is to start with the word count count code from Lab 1. There three subtasks:

1) Modify the word count mapper so that the program computes counts not for all of the *words* in the corpus, but for all of 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:

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

Don't worry about bigrams that span lines; we're only concerned with bigrams on the same line. Also, as you implement this, represent bigrams as text strings exactly as I've depicted above (as strings that contain the comma-separated pairs of words, with parens). Further, order the words in a bigram lexically (according to `String.compareTo`).

2) Modify the reducer so that the program only writes out those bigrams that appear more than twenty times overall in the corpus.

3) After you do this, run your program, bring up one of your result files, and get checked off.

And remember, **SHUT DOWN YOUR CLUSTER**