# INFO 208 Quiz #1 — Hadoop Lab Exploration Component

Your Name: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

The following Lab items should be carried out by you using your **Hortonworks HDP Sandbox** VM. Remember, you were asked to keep your HDP Sandbox VM for later use — both in this course and for continuing education beyond this course.

The purpose of this **Lab Exploration Component to Quiz #1** is to determine how well you understood the Hadoop Labs that you did in the first weeks of this course and how well you can apply the concepts and approaches you learned there to simple situations that are different but somewhat similar.

**Open up the area after each question / section** in this MS Word file below and type in (or paste) your answer(s) and your comments. **Mail** the resulting file to me at [glen.mules@sjsu.edu](mailto:glen.mules@sjsu.edu)

Your email **Subject** line should be:

**INFO 208 Quiz #1**

And the attached MS Word file sent by email should be named:

**INFO\_208\_HadoopQuiz-*YourLastName.YouFirstName*.docx**

The file naming is important as it causes my email software to sort your *Quiz Response Email* into a directory for me so that I don’t lose amongst my other daily email.

**Basic Lab** (100% of your grade for the Lab Exploration Component):

1. Locate a plain-text copy of the US Constitution. One place you can find it is: <https://www.constitution.org/cons/constitu.txt>
2. Using the **wordcount2** MapReduce program that you have **already compiled** (in Hadooop Labs #8), run **wordcount2** on the US Constitution text file. Find the top 20 non-trivial words (i.e., eliminate such words as: a, the, and, when, …

Hint: When you get your results eventually, make sure that you do not “break” words because you eliminated “a”, “on”. … in the middle of words. This requires some thinking!!

1. Remember your output directory must not exist or must be empty when you run wordcount2.
2. Explain *what you did*, *where you did it*, *how you ran it*, etc. with all the statements that you needed to use. The ***why*** you did something is even more important than ***what*** you did.

**Lab for Extra Credit** (beyond the 100% of grade for the Lab Exploration Component — for up to 20% extra). ***This is optional extra***. No knowledge of programming in Java is actually needed to complete this.

1. Find a program on the Internet that does pairings of words (called a 2-gram or bi-gram) in the text of the text files analyzed. They exist — probably programmed in Java for Hadoop/MapReduce.

Bigrams are simply *sequences of two consecutive words*. For example, the previous sentence contains the following bigrams: "Bigrams are", "are simply", "simply sequences", "sequence of", etc.

1. Using the approach that was used with wordcount2, *compile* the Java program and have it run against the Gutenberg documents that were loaded in Lab #8 *or* the US Constitution document that you downloaded above.
2. Find the most common bi-grams in your target text. The first 20 or so that do not use non-trivial / stop-words will be sufficient