CS 2204 Lab 5

your name here (please print):

your student ID number here:

WARM UP:

1. Create a subdirectory called 'lab5' under your home directory.

2. Copy the file `othello.html` into this directory. This file is either available from the directory location ~cs2204/othello.html or at the URL: http://courses.cs.vt.edu/~cs2204/fall2005/assignments/othello.html. If you are copying the file from the URL location, the WRONG way to save this file onto your `lab5` directory is to block and copy the text with your mouse (or the 'select all' option) and paste it into a file in, say `emacs`. The RIGHT way to save this file is to right click on the link and use the 'save link as' option. If you do it the first way, your copy-pasting might destroy the line formatting (e.g., one line could be broken down into multiple lines) and then your answers to the questions below will be incorrect.

QUESTIONS TO ANSWER:

Using `sed`, we are going to restore the pristine version of Shakespeare's classic, by removing the modern HTML tags in the `othello.html` file and printing the result onto standard output. You must create a single file and put `sed` commands to do the following in that file. Then, invoke the file using the `-f` option to perform your edits. Write your `sed` commands in the space provided.

1. (3 points) Remove all the HTML tags in the file. For our purposes, a HTML tag is any text starting with a `<` and ending with a `>` character. You may assume that the tags do not span multiple lines. Keep in mind that, if a line contains both tags and regular text, only the tags get deleted. However, the HTML tag `<BR>` (can also be present as `<br>`) must be treated differently. Replace this tag with a newline, else the resulting output will be difficult to read.

2. (3 points) Delete the preamble text that talks about HTML, XML, and statutory warnings about computer technologies. The first line of the file must be 'The Tragedy of Othello, the Moor of Venice.'

Here's some more post processing to do after your `sed` session.

1. (2 points) `spell` is a UNIX spell checker. Do a `man spell` to see how it works. Then, run a `spell` on your output and count how many words are flagged as incorrect (Shakespeare would be very upset at the output!). Write the number here.

2. (2 points) Let us do a frequency count of word usage by Shakespeare, taking *Othello* as a representative sample of his writing. `cat` your cleaned-up version of *Othello* into the following UNIX pipe:

   ```
   tr -cs A-Za-z '\012' | sort | uniq | sort +0nr +1d
   ```

   What type of words appear at the top of output? What types of words appear at the bottom? What does this tell you about Shakespeare's writing style? For more fun, goto `amazon.com`, search for *Othello*, peruse Amazon's entry for this work, and look for what are called SIPs and CAPs. Finally, when you have enjoyed yourself, explain in simple English what the above commands in the pipe are doing.