**Week 7 – Lab Homework**

**Part 1**

Using the gender\_features function defined in Lab1 (Part 1) as a reference, define a new feature extraction function that includes features for two-letter suffixes, such as the one here:

>>> def gender\_features3(word):

return {‘suffix1’: word[-1],

‘suffix2’: word[-2]}

Keep the variables train\_names and test\_names that define the training and test set.

Make new train\_set and test\_set variables. Carry out the classification and look at the errors in the test set.

Is this classification more accurate? Can you see error examples that you could use to make new features to improve? (You don’t have to do this, just observe it.)

***Report your original accuracy on the test set and the new accuracy, and include any observations about the remaining errors.***

***Bonus:*** Make a new gender\_features function that keeps three suffix letters and report the accuracy. Be sure to make allowances if any names that are only two characters long. Alternatively, make a gender\_features function that uses the first letter and the last two letters.

**Part 2**

Re-run the movie review text classification problem using again, once with a number of words less than the 2000 from Lab3 (Part3), and then re-run with a number of words greater than 2000. Don’t perform the random shuffle again!.

***Report the accuracy numbers for each number of words used. What trends and observations do you see with the different range of number of words?***

**Homework:**

Submit a report summary that includes the results for both Part 1 and Part 2.  Include any observations and lessons learned in your report.  In addition to the report summary, upload a copy of the python code you developed as part of this lab.