**NLP Lab 7**

Dixitha Kasturi

[dkasturi@syr.edu](mailto:dkasturi@syr.edu)

**Topic: Text Classification**

**Part 1**

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,
2. 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.

For this question I took the 2 cases with 2 and 3 features respectively.

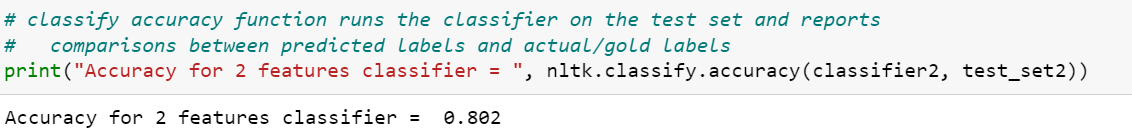
1. For 2 features the accuracy was 80.2%

The features function was modifed as follows:

Graphical user interface, text, application

Description automatically generated

The accuracy was as follows:



1. When I tried to run the classifier, I initially got the following error, I suspected that the out of bounds index meant that there were names which were only of length 2

Graphical user interface, text, application

Description automatically generated

To verify if my suspicion is correct, I calculated the lengths of each word and put them in a list and got the frequency of each of those lengths, unsurprisingly, there were 19 words with length = 2

Text

Description automatically generated with medium confidence

I printed out the list of those names :

Text

Description automatically generated

To train my classifier model with 3 features, I excluded out the above 2 letter words/names and further confirmed that only 3 and above length words are present in the new list to give to the model:

Graphical user interface, text, application, email

Description automatically generated

The accuracy then was 78%



**Part 2**

I chose 1350 and 5050 as my common words length for this question. I did play around with the numbers and observed that as the number is closer to the 2000 mark, the accuracy did not vary a lot or it was the same. But if random shuffling was performed the accuracies varied.

For 1350 count the accuracy was 73%

For 5050 count the accuracy was 79%

Text

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text, application

Description automatically generated