**CMPE 493 - ASSIGNMENT I REPORT**

In this assignment, we have implemented a spelling corrector based on the noisy channel model. To achieve this, we need two models which are language model and error model.

For the language model, a dictionary named ‘WORDS’ is created using the corpus.txt. It includes the tokens and the number of occurrence of them in that file. So, the language model P(c) is determined.

To form the error model, the confusion matrices are created using the errors in the spell-errors.txt as follows: the Damerau-Levenshtein edit distance between the misspelled version and the correct one is calculated for each word in the document using dynamic programming. While doing this, a backtrace matrix has stored the edit type as ‘N’, ‘I’, ‘D’, ‘S’, ‘T corresponding to “none”, “insert”, “delete”, “substitute”, “transpose” respectively. Then, by backtracking, the edits is found and added to the corresponding confusion matrix when the edit distance is minimum. Additionally, unigrams and bigrams are formed as dictionaries. Thus, the data for the error model is got.

Finally, for testing we have created the candidates for each misspelled word with one edit distance. Then, the candidate which is most likely to occur is chosen depending on the model. The results are written in a file named noisy-channel-results.txt for noisy channel model and a file named language-results.txt on the current working directory. The accuracy is computes by comparing these files with the given correct words.

While the accuracy result for the language model is 72.9166666667 %, the accuracy of the noisy channel model is 76.0416666667 %. The results of running both versions of program on the given set is below:

