CS671
HW2 Part B
Ankit Modi (10104)

Unigram frequencies were determined and mapped to their probabilities.
Smoothing factor used was 0.5
A function was written to calculate the levenshtein distance between two words.
Further 2gram and 3gram models were used.

For every wrong word in the list of wrong words given, an attempt was made to generate 3 suggestions of correct words. But because of a gloomy data set, results were not very encouraging.

Though some of the words like 'अधिघरियों' was correctly converted into 'अधिकारीयों'.

Only those words were considered who are at a distance of 2 or less from the test word.
Further, phonetic similarity was incorporated into the system. But again, due to the size of data set, we were getting very few words at a distance of 2. Among them, having phonetic similarity had a very bleak chance, so again results shady.
Same was with the case of applying bigram model as all the bigrams were not present in the data set, so their probability was really low.

Conclusion :The spell checker falls behind with in terms of results given by aspell.