Parakh Jaggi
Andrew Walker
NLP

Project 2 Write Up

Corpuses Used

A major problem we had during the development of our project was finding a big enough corpus to create N-Grams off of. One option was to use a small corpus to optimize run time. Using a small corpus reduced our algorithms run time by about 50%, however accuracy went down to about 10%. Using a small corpus resulted in a lot of inaccuracies when it comes to detecting misspelt words and suggesting corrected words. We decided to emphasize accuracy, so we used the following corpuses:

- nltk.brown
- nltk.state_union
- nltk.words
- nltk.punkt
- nltk.wordnet
- nltk.gutenberg
- nltk.twitter_samples

Algorithm

The first thing we do in our algorithm is establish a list of valid words using the nltk corpuses. We then go word per word in the given string and check if it is a valid word. If misspelt words are found, we do multiple steps to get the correct word. The first thing we do is collect 2-grams using the Group 3's N-Gram program. We then create potential candidates based off the edit distance. We then test to see if any words are in both or one of the potential candidate lists. If

a word is in both the N-grams and the edit distance list, then we suggest the word as a potential correction.