INTRNLP S17
Poblete, Brian Michael
Rivera, Louie IV Y.
Soliven, Adrienne Francesca O.

PE#3: Spelling Correction Write-Up

The text that we used for the corpus of this spell-check program was taken from the American National Corpus [1]. For this spell-check program, we decided to limit our corpus to formal English. This means that our corpus does not take into consideration slang words and colloquialisms. Thus, we decided to use text files that include court transcripts, debates, government documents, technical papers, movie scripts, journal articles, excerpts from nonfiction literature, and news articles.

The code used was obtained from Norvig.com [2] with some edits to the algorithm added by the group. The minimum edit distance function will return a set of words that has a minimum of 1 distance. The words returned are already edited, which means that function already deleted, inserted, substitute and transposed, regardless whether the resulted combinations or edits are existing words in the english language that's why the function known is used to check if the subset of the words exist within the dictionary. The corpus was also used to calculate the probability of what the corrected word is based on Bayes' theorem [2]. The edits to the algorithm included the implementation of Laplace/ Add-1 smoothing [3] to the original code from Norvig. Another modification was that it also checked whether the word has no error correction.

As for the limitations of this program, the correction suggestions are limited to the text in our corpus. An example is the word "they're". "They're" is a word that is not in our vocabulary and no other words are close to be a potential correct word. The solution given by the algorithm is that the correction to the word is the word itself. This solution is observed in the spellcheck of software some keyboards (e.g., Google's Gboard). The program is also not the most efficient considering how the minimum edit distance function will return a large amount of edited words and has to go through everyone one of them to check if it exist within the dictionary.

References

[1] http://www.anc.org/data/masc/downloads/data-download/

[2] https://norvig.com/spell-correct.html

[3] https://lazyprogrammer.me/probability-smoothing-for-natural-language-processing/