For the final project I will be working alone. My proposal for my final project is to re-build the spelling checker but with a few modifications. Misspellings are very common especially when typing quickly. At the same time misspelled words are very unprofessional, and they detract from the quality of content in which they appear. For these reasons and more, an automated spelling checker is a very useful and tool.

The original spelling checker assignment fell short in several ways. An obvious way in which my original submission fell short was that it did not take in to account the probability of switching certain letters being greater than the probability of switching others (for example it is more reasonable to swap 'r' and 't' than it is to swap 'r' and 'm'). Another way in which the original submission fell short was that it did not perform any stemming. This could lead to the issue of, for example, having the word 'swim' was in your dictionary, but marking the perfectly reasonable word 'swimmingly' as misspelled.

A third way in which I aim to improve the original submission is by adding a language model. If the words 'tan' and 'ran' are both in the dictionary used for the original assignment, then the sentence “He tan ten miles today.” would not raise any flags, whereas a human reader would guess that the intention was “He ran ten miles today.” Another way in which a language model will improve this spell checker is by suggesting words that are not only similar to the misspelled word, but also fit in the same context.

To generate both my dictionary of words and my language model, I intend to use a subset of Wikipedia. A data dump is available at this link: http://dumps.wikimedia.org/enwiki/latest/. My motivation for using Wikipedia is the thought that it will encompass more colloquial words than other corpora because anyone can edit it. A potential downside is that Wikipedia itself might have spelling errors, but I believe this to be a potential issue with any corpus as colloquial as the data I will be using.