CSCI 404 - Assignment 2 - Josh Loehr & Robin Cosbey - essay.pdf

To generate the candidate list, find all words of edit distance one from the input word. Add the input word to the candidate list in case the user did not make a mistake. Take the intersection with the dictionary set to filter out non-words. In this case, 'acress' would be filtered out. We next want to find the candidate word which maximizes the probability of that word given the misspelled input word. To do this, apply Bayes' Theorem such that this probability becomes the product of probabilities involving both language and error model components. The language model is the probability of that word appearing in the corpus; this is calculated by the frequency of that word divided by the total number of tokens. The error model involves the combination of the confusion matrices for each edit that is accounted for; these edits are insertion, deletion, substitution, and/or transposition. Once those models have been applied to the product of probabilities the optimal candidate can be selected. We can also use a keyboard layout confusion matrix, n-gram context analysis, and part of speech analysis if so desired. The above procedure applied to 'acress' would be as follows:

1. Generate the candidate list of length one edits, remove non-words
C = {actress, cress, caress, access, across, acres}

2. Find the probability of each candidate word given the typo: $P(c_i \mid w)$

3. This probability can be found by applying Bayes' Theorem: $P(c_i) * P(w \mid c_i)$

4. For $P(c_i)$, use the language model which finds the frequency of each c_i within the corpus and divides by the total number of tokens in the corpus

5. The error model uses a confusion matrix corresponding to the edit which each candidate word derives from

6. The highest probability candidate turns out to be 'actress' at 36.44%