EDAN20 - Assignment 0
Spelling Corrector

Hugo Mattsson
hu5174ma-s

September 2024

The spell checker implemented by Norvig isn't the greatest, only achieving an accuracy of 68-75% for his English spell tests, however it's simplicity lends to an easier understanding. The main driving force for the checker is the usage frequency of a word in the English language. This probability has been approximated with a million words long corpus of different word lists, literary works and other corpora. In this corpus the most likely word to exist is 'the' with a 7% probability. More likely words will be prioritised as candidates when looking to correct misspellings.

When a word is checked the strategy is as follows:

1. The word is known from our dictionary/corpus and therefore correctly spelled.

2. The word is one edit away from one or more known words and is changed to the most probable word.

3. The word is two edits away from one or more known words and is changed to the most probable word.

4. We don't recognise the word, and neither one or two edits changes this fact, so the word is left as is.

In the above strategy list, edits were used to find suitable candidates that are known correct words. This is done by iterating over the word and, for every letter, trying to either delete it, switch its place with the next, replace it with another letter, or finally insert a new letter. This produces way to many candidates, but is easily pruned by only keeping those that exist in the dictionary. If suitable candidates were found, the most probable one is chosen as the correct spelling.