EDAN20 - Lab 0

André Frisk - an8218fr-s

August 31, 2021

The Spell checker

Firstly the preparatory work has to be done by listing all unique words and counts their amount which is inserted into a list WORDS. These words are read from the "big.txt" file. The function P() calculates the probability of each random word being the specific word.

Finding a candidate for the word

The program checks each word if its a already known word or if the current word is one or two changes away from a known word. The matched known words are then flagged as candidates to the real English word. These words that count as candidates are checked against the word probability and their changes of which is the most likely candidate to be the real word (this is done in correction() by taking the maximum probability of a word).

For example, with the description above, if we look at the word *somthing*, this word could have the candidates *something*, *seething* or maybe *smoothing* as the true word.

How to test different candidates

The program checks four different cases for a candidate, one if the word is known, one if the word itself is already correct, one where the word gets edited one time and the last case where the word gets edited two times. The candidates() function checks this. How the word is edited and checked is done in four different ways:

- Remove letter - checks versions with a missing letter in the word
- Swap two letters - checks versions when swapping two adjacent letters in the word
- Replace letter - checks versions when a letters is replaced with another in the word
- Insert letter - checks versions when a letter is added to the word

The set that is returned can be quite huge, however the program restricts to words that are *known*–that is, in the dictionary– then the set is much smaller. The list above with the different methods to find a candidate is used by edit1() and edit2() where edit1() edit the word one time while edit2() makes two edits to the word. Edit2() generates a much bigger set of possibilities than edit1() but usually only a few of them are known words.

The test code

A unit-test was provided with the spell checker to check if the program worked using assertions. It checks for example, if different types of misspelled words were corrected with the right word

in correction() and if the most common words are correctly counted by WORDS. The unit-test provides two text-files to try out the program and see statistically how many words are spelled correctly without any change of the word and if there are any unknown words. The first set test gave 75% correct and the second set test gave 68% of words that was correctly spelled.