CSE 586 Information Retrieval, Spring 2019
Assignment 1 - Spelling Error Correction, Due: 29/03/2019 (Friday), 06:00

In this assignment you will implement a simple spelling error corrector. Before you start this assignment, I suggest you to read the article by Peter Norvig on "How to Write a Spelling Cor- rector" available at http://norvig.com/spell-correct.html. The spelling error corrector that you will implement will be simpler than the one described in this article.

First, you should create a dictionary containing the English words and their frequencies by using the provided *corpus.txt* file. This file was obtained from Peter Norvig's web site. It contains a concatenation of several public domain books from *Project Gutenberg* and lists of most frequent words from *Wiktionary* and the *British National Corpus*. You can assume that the words in the corpus.txt file are spelled correctly. In order to create your dictionary, you will need to tokenize the file and perform case-folding. You can predict the correct spelling of a misspelled word by generating all the words whose edit distances to the word are 1 and select the one with the highest frequency in the corpus.txt file. Note that several spelling errors involve transpositions of characters. Therefore, you should use the Damerau-Levenshtein edit distance, where the valid operations are defined as insertion, deletion, and substitution of a single character, or transposition of two adjacent characters.

You may use any programming language of your choice. Your program should take a file containing a list of misspelled words (one word per line) as input, and produce a file with the predicted correct spellings of these words (one word per line) as output. If your program can not produce predictions for any of the words in the input file, the corresponding lines in the output file should be printed as blank lines.

A list of 384 misspelled words (*test-words-misspelled.txt*) and their corresponding correct spellings (*test-words-correct.txt*) are provided for you to test your program. Compute the accuracy of your program for this test set and write it in your report.

Suggest and implement an extension for your spelling corrector. Provide the accuracy results obtained by the enhanced algorithm in your report. The extension that you suggest can be a simple heuristic such as assuming that the first letter of a word is always spelled correctly.

Submission: You should submit a *".zip"* file named as YourNameSurname.zip containing the following files:

1. Report: Describe the extension that you propose for the baseline spelling corrector. Report the accuracy results obtained by both versions of your spelling corrector on the provided test set.
2. Source code and executable: Commented source code and executables of both versions of your spelling corrector.
3. Readme: Describing how to run your program. I should be able to run your program using a different test set.