Cmpe 493 Introduction to Information Retrieval, Spring 2020
Assignment 1 - Spelling Error Correction, Due: 20/03/2020 (Friday), 17:00

In this assignment you will implement an isolated word spelling error corrector based on the noisy channel model.

You should create your dictionary and estimate a unigram language model ($P(w)$) using the provided *corpus.txt* file. This file was obtained from Peter Norvig's web site (http://norvig.com/big.txt). It contains a concatenation of several public domain books from *Project Gutenberg* as well as 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 should 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 that maximises $P(w)P(x|w)$. Note that several spelling errors involve transpositions of characters. Therefore, you should use the Damerau-Levenshtein edit distance.

You should estimate the error probability ($p(x|w)$) using the corpus.txt file as well as the *spell-errors.txt* file obtained from Peter Norvig's web site (http://norvig.com/ngrams/spell-errors.txt). Each line of the spell-errors.txt file contains a correct word followed by : and a comma and space separated list of observed misspelled versions of the word. *x denotes that the corresponding misspelled version has been observed x number of times.

Implement a second version of your spelling error corrector using add-one smoothing (Laplace smoothing with alpha = 1).

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.

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

1. Report: (i) Describe how you implemented the spelling error corrector. (ii) Provide the four confusion matrices that you computed. (iii) Provide screenshots of running your system. (iv) Report the accuracy scores you obtained for the provided test set both by using smoothing and without smoothing. (v) Investigate the errors of your system and discuss how your system can be improved.

2. Commented source code and executable.

3. Readme: Describing how to run your program step by step. I should be able to run your program using a different test set.

Honor Code: You should work individually on this assignment and all the source code should be written by you. You are NOT allowed to use any available libraries or any code written by other people.

Late Submission: You are allowed a total of 3 late days on homeworks with no late penalties applied. You can use these 3 days as you wish. For example, you can submit the first homework 2 days late, then the second homework 1 day late. In that case you will have to submit the third homework on time. After using these 3 extra days, 10 points will be deducted for each late day.