Cmpe 493 Introduction to Information Retrieval, Spring 2017
Assignment 1 - Spelling Error Correction, Due: 31/03/2017 (Friday), 17:00

In this assignment you will implement a simple spelling error corrector based on the noisy channel model (discussed in lecture 4 - IR-Lec4.pdf). Before you start this assignment, I suggest you to read the article by Peter Norvig on “How to Write a Spelling Corrector” available at http://norvig.com/spell-correct.html.

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* 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 most likely word from the dictionary using the noisy channel model. 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 can use the *corpus.txt* file to learn the language model ($P(w)$). You can use the provided *spell-errors.txt* file, which includes Peter Norvig’s spelling error list to learn the channel model (i.e., error model $P(x|w)$). The file includes on each line a correct word followed by a list of misspelled versions. *“*n”* means the corresponding misspelled version was observed n times in a corpus used to compile the misspellings.

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 on this test set and write it in your report. Compute and compare the accuracy when you only use the language model and the accuracy when you use both the language model and the channel model.

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

1. Report: Report the accuracy results obtained by both versions of your spelling corrector on the provided test set. Include *screenshots* of running both versions of your program on the 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.

Late Submission: You are allowed a total of 3 late days (including weekends) 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 remaining homeworks on time. After using these 3 extra days, 10 points will be deducted for each late day.