Assignment No 3

P180066 Usman Ahmad

April 2, 2023

Abstract

Your abstract.

1 Introduction

A unigram model which is trained by the provided corpus to learn the probility of each word in the voculubary. Then, create the tables for insertion, deletion, substitution, and transposition of each alphabet from a-z, based on misspelled words. In addition use the language model and the error tables to calculate the probability of each candidate word that is one edit away from the misspelled word. Finally, you will choose the candidate word with the highest probability as the corrected word.

2 Assignment Task

2.1 Step1

First, the program reads in a text file called "data.txt" and trains a unigram model to build a vocabulary of words and their frequencies. This model is used to calculate the probability of a word occurring in the corpus.

2.2 Step2

the program reads in a text file called "misspellings.txt" which contains pairs of misspelled words and their correct spellings. It then creates four different dictionaries for insertion, deletion, substitution, and transposition of each letter in the alphabet.

2.3 Step3

The program then defines a function called P x given w which calculates the probability of a misspelled word given a correct word. This function uses the error models to calculate the probability of a misspelled word being produced from a correct word by performing an insert, delete, substitution, or transposition of a letter.

2.4 Step4

The generate function is defined to generate a set of candidate words that are one edit away from a given word. This function generates all possible insertions, deletions, substitutions, and transpositions of each letter in the given word and checks if each candidate word is in the vocabulary built by the unigram model. If a candidate word is in the vocabulary, it is added to the set of candidate words. The program also defines a function called "correct" which takes a misspelled word and generates a set of candidate words using the "generate" function. It then calculates the probability of each candidate word using the unigram model and the P x given w function. Finally, it returns the most probable corrected word. there is section that reads the misspelled words and calls the generate function to check the word and give the most probable corrected word and print on the console.

References