Spell Correction Report

Jilsa Chandarana
chandarj@uwindsor.ca
University of Windsor
Windsor, Ontario, Canada

1 INTRODUCTION

The aim of the project is to develop a spell correction system using minimum edit distance. For edit distance, we can use Levenshtein distance which allows the operations such as insertion, deletion and substitution. [7] [1] To test our system, we have used Birkbeck corpus [4] which has 809 pairs of misspelt words along with the correct spelling. Our task also includes the calculation of success at k which is whether the correct word is found in the top-k. The implementation of the project can be found on GitHub. [2]

2 PROPOSED SOLUTION

We propose a solution using the online platform Google Colab which is a tool to execute python programs in an easy and efficient way. We have created 5 files which are

main.ipynb Main Python notebook

dictionary.py Returns a list of all required English words using the library nltk.

birkbeck.py Returns the key-value pair of incorrect and correct English words from the Birkbeck corpus, provided the birkbeck.txt.

mindist.py Returns the minimum distance of a word from all the words provided in the specified list.

successatk.py Returns s@k.

2.1 Preprocessing

The number of words in our dictionary is 236736. That gives us an opportunity to experiment with a different subset of the dictionary to see the improvement in the success score. Since considering all the words takes too much time, we have chosen the following cases for that.

(1) All words with the same initial
(2) All words with the same length as the correct word
(3) All words with the same initial and the same length as the correct word

[3] [4]

2.2 Edit Distance Calculation

The Levenshtein distance in `mindist.py` is calculated using Dynamic Programming. [3] It's a RAM extensive process to calculate the distance of all the words from the dictionary, so to overcome the excessive RAM usage, we temporarily stored the distance values in the secondary storage and once we get the list of top k elements, we removed them.

2.3 Parallel Processing

As the volume of our dictionary is large, sequential execution i.e. calculating the edit distance of one word at a time can consume more time. The system is implemented on the Google Colab platform which provides two CPUs [5] so to utilize both of them, we've used the concept of multiprocessing. [6]

3 RESULT

Measure	Case 1	Case 2	Case 3
Success at 1	0.8232	0.8393	0.8430
Success at 5	0.8368	0.8516	0.8566
Success at 10	0.8368	0.8516	0.8566

4 ANALYSIS

We have used the `PyTrec Eval` to calculate the average of success@1, success@5 and success@10. [1]. We can see that the success rate with the same length is moderately higher than the success rate with the same initial. Although the success rate doesn't increase much in the final case as compared to the previous one. [5]

5 CONCLUSION

From the given project of spell correction using Wordnet and Birkbeck corpus, we experienced the effects of various factors such as the dictionary word selections and multiprocessing. We calculated the success rate for three different cases and we believe that the results can be improved by using better pre-processing techniques.

REFERENCES

[1] bandpooja. 2022. *Assignment 1.* Retrieved January 27, 2023 from https://github.com/bandpooja/Assignment_1/blob/master/assignment1.ipynb
[2] Aditeya Baral. 2021. *PyDictionary.* Retrieved January 27, 2023 from https://github.com/aditeyabaral/pydictionary
[3] Dan Jurafsky. 2023. *Minimum Edit Distance.* Retrieved January 27, 2023 from https://web.stanford.edu/class/cs124/lec/med.pdf
[4] Roger Mitton. 1980. *Birkbeck spelling error corpus.* Retrieved January 27, 2023 from https://ota.bodleian.ox.ac.uk/repository/xmlui/handle/20.500.12024/0643
[5] Privalov Vladimir. 2019. *Hardware exploration on Google Colab and Kaggle platforms.* Retrieved January 27, 2023 from https://vovaprivalov.medium.com/hardware-exploration-on-google-colab-and-kaggle-platforms-576bf51c54e#:~:text=There%20are%20two%20CPUs%20%40%202.2GHz%20available.&text=Output%20on%20Google%20Colab%20%E2%80%94%2049GB%20in%20total.
[6] Privalov Vladimir. 2023. *multiprocessing — Process-based parallelism.* Retrieved January 27, 2023 from https://docs.python.org/3/library/multiprocessing.html
[7] Inc. Wikimedia Foundation. 2022. *Edit distance.* Retrieved January 27, 2023 from https://en.wikipedia.org/wiki/Edit_distance

[1] We have used the NLTK library over PyDictionary because the PyDictionary also uses the NLTK at its backend so why complicate things instead of using the base? [2]

[2] https://github.com/JILSA212/Spell-Correction

[3] The basic shortcoming of the same initial subset is that if the first letter has an error, we will not be able to find the correct word in the dictionary.

[4] We need to consider the length of the correct word instead of the incorrect one because if the misspelt word has a different number of alphabets then the dictionary will not have the correct word.

[5] The reference from the previous assignment was taken because we were unable to find proper documentation for pytrec eval.