



## Project 1 reviews

## Print Options:

- ☒ Include Questions & Answers ☒ Include Comments ☒ Include All Reviews ☒ Include File Info

[Print](#)

## HAONAN LI'S PEERMARK REVIEW OF ANONYMOUS'S PAPER (100% COMPLETED)

### ASSIGNED QUESTIONS

1. Briefly summarise what the author has done

The author chooses two spelling correcting methods and applies them to a specific task. Through the comparison of two methods. the author find that Soundex algorithm can help improve the recall of edit distance match.

2. Indicate what you think the author has done well, and why

The author makes a good jobs as analysis and describing the dataset used in the task, which may have influence on the experiments. The method used in this report is novel and meaningful. The author uses global edit distance and Soundex + global edit distance algorithm to make the compare experiments and indicates that Soundex can help improve the recall of the system but do less help on precision.

3. Indicate what you think could have been improved, and why

First of all, the results are not convincible and the evaluation method might be misleading, especially the recall results. I totally disagree with the author's compute method for recall, which can get a higher score than 82%. The description of how to get the result is also ambiguous. There are evaluation on edit distance method where distance equals to 0, but what is the predicts when there is not such a results?

As for format, use a table to show dataset statistical rather than screenshot might be better. Besides, the organization of should be adjusted. For example, the introduction of evaluation indicators should be moved to Evaluation part, in this way, the whole report may seems more consistent and reasonable. For writing, I do not think you reference [3] and [5] is a good way. Use footnote instead of cite will be better.

### COMMENTS LIST

No comments added

### SUBMITTED FILE INFO

file name	kt_Project1-version2.pdf
file size	396.72K

"PROJECT1" BY ANONYMOUS

# COMP90049 Knowledge Technologies

## Project 1:

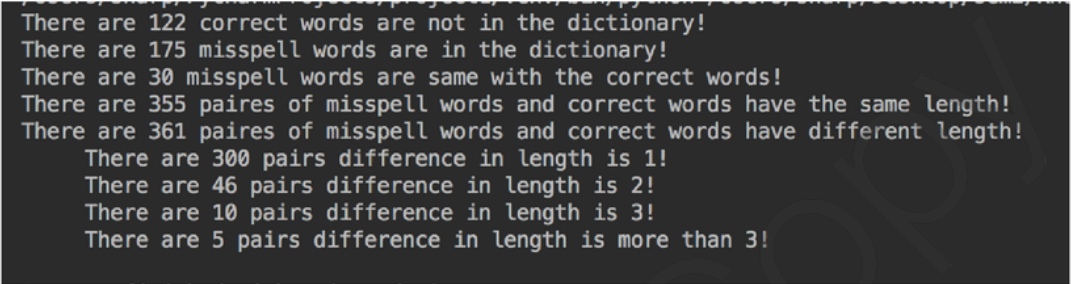
### 1. Introduction

In this project, I will choose two Spelling correcting methods- Global Edit Distance(GED) method and 'Soundex & Global Edit Distance' (S&GED) method to find the intended spelling words. By adopting the two methods above to the given dataset, we can evaluate the performance and efficiencies of these methods.

The dataset used here is curated by Naomi Saphra and Adam Lopez (2016) [1]. And, there are three files in this dataset: misspell, correct and dictionary. The misspell contains 716 words which are identified as misspelled; The correct contains 716 intended spelling words; And the dictionary with 400,000 words is used to retrieval intended spelling words.



However, the dataset is flawed, the following picture 1 shows some details about the dataset:



```
There are 122 correct words are not in the dictionary!  
There are 175 misspell words are in the dictionary!  
There are 30 misspell words are same with the correct words!  
There are 355 paires of misspell words and correct words have the same length!  
There are 361 paires of misspell words and correct words have different length!  
There are 300 pairs difference in length is 1!  
There are 46 pairs difference in length is 2!  
There are 10 pairs difference in length is 3!  
There are 5 pairs difference in length is more than 3!
```

**Picture 1: Some statistical results of the dataset**

Noting that there are 122 correct words are not in the dictionary. In other words, we cannot retrieval these words by implementing the original dataset. This is a much more serious issue than low precision, because it is a fatal weakness for a spelling correcting method to cannot find the intended words of users by all means. Beyond that, 175 misspell words exist in the dictionary, this inherit problems of the dataset will affect the precision of spelling correcting methods. Moreover, there are 361 pairs of misspell words and correct words have different length. This problem will have impact on the performance of some edit distance methods.

## **2. Methodology**

Two evaluation indicators will be used in this project: precision and recall.

### **(1) Precision**

The precision reflects the efficiency of Spelling correcting method. The higher the precision is, the more effective the retrieval is. In this project, precision will be used to compare the two methods above.



## **(2) Recall**

The recall reflects the ability of Spelling correcting method to retrieval the intended spelling words. The higher the recall is, the more intended spelling words are retrieved.

### **2.1 Global Edit Distance (GED)**

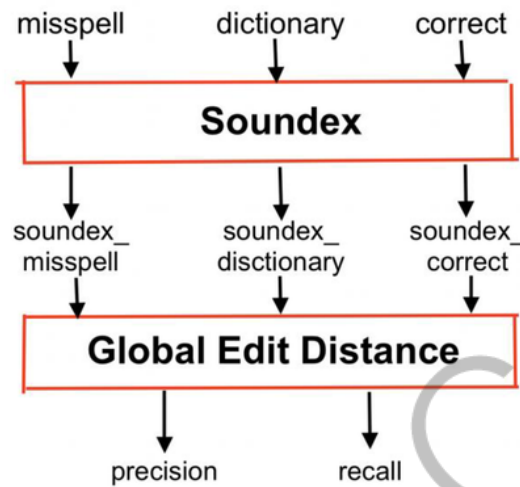
The global edit distance between two given words reflects the minimum steps needed to transform a word to another word. Levenshtein Distance (LD) is the GED when the parameters of GED are (match, insertion, deletion, replace) = (0, 1, 1, 1) [2]. And the Damerau-Levenshtein distance (DLD) has the parameter of (match, insertion, deletion, replace, transpose) = (0, 1, 1, 1, 1). The program package used here “weighted-levenshtein” is written by David Su (2018) [3].

### **2.2 Soundex & Global Edit Distance (S&GED)**

The main idea of Soundex algorithm is to generate, for each word, a “phonetic hash”, In this way, similar-sounding terms hash to the same value [4]. By implementing Soundex to the original dataset, we can improve the recall. The program package used here “Soundex” is written by SILPA (2017) [5]. Here, the max length of “Soundexed” words are set to 8, because some long words in the dataset have long “Soundexed” values. For example, the word “blackheartedness” in the dictionary with Soundex “b426352”, and the word “abiogenetically” in the dictionary with Soundex “a125324”. And “Soundexed” values can relatively improve the precision. To implementing the S&GED method, we adopted Soundex to the given dataset, and generate their corresponding “Soundexed” documents first. After that, we use GED to find the intended spelling words. The following picture 2 shows how to implement S&GED on the given dataset.







**Picture 2: The implementation of S&GED**

In this project, 122 correct words cannot be founded in dictionary. And in theory, the max recall should be 83%. However, several of them sound like some other words in the dictionary. In this case, we can use Soundex algorithm to match these words with some words in the dictionary. For example:

Correct words not in the dictionary	Soundex	Words in the dictionary	Soundex
ballsack	b42	ballcock	b42
beezie	b2	bessi	b2
bhenchod	b523	banqueted	b523
fagoot	f23	fagott	f23
thingee	t52	tingi	t52
skellie	s24	seggiola	s24
joggins	j252	joggings	j252
haterade	h363	heathered	h363

In this way, the recall can be improved.



### 3. Evaluation

In the following evaluation, much more attention will be paid on recall rather than precision for the following two reasons: Firstly, the precisions of both Global Edit Distance method and 'Soundex and Global Edit Distance' method are low. In real world, several tools can help us to improve the precision of the spelling correcting methods, such as machine learn, analyzing the context and analyzing the users' input habits. The precision is only used as a relative reference parameter to compare the two methods in this report. The other reason is that low recall is regarded as a serious problem in spelling correcting methods. And the issue of intended words cannot be retrieved with the reason that they are not in the dictionary should be avoided.

#### 3.1 Global Edit Distance (GED)

The following two tables showed the precision and recall of Global Edit Distance method:

	<b>LD</b>	<b>DLD</b>
<b>Distance&lt;=1</b>	5.289%	4.381%
<b>Distance&lt;=2</b>	0.346%	0.353%
<b>Distance&lt;=3</b>	0.034%	0.035%

**Table1: The precision of LD and DLD**

	<b>LD</b>	<b>DLD</b>
<b>Distance&lt;=1</b>	40.642%	50.140%
<b>Distance&lt;=2</b>	70.810%	71.090%
<b>Distance&lt;=3</b>	78.492%	78.771%

**Table2: The recall of LD and DLD**



There is no obvious difference between the precisions and recalls of LD and DLD. When the max edit distance was set to 1, the precisions of LD and DLD reach 5.289% and 4.381%. However, the recall of both two methods are relatively low. When the max edit distance increased, the precision of GED declined quickly. However, the recall increased relatively slow. When the max edit distance is less or equal to 3, the recall reaches 78% which is an unsatisfactory level. At the same time, precision is only about 0.035%. As I mentioned before, there are 122 correct words are not in the dictionary, so the max recall is 83%. The GED cannot break through this limit.

### 3.2 Soundex & Global Edit Distance (S&GED)

The following two tables showed the precision and recall of Global Edit Distance method and 'Soundex & Global Edit Distance' method:

	<b>LD</b>	<b>Soundex &amp; LD</b>
<b>Distance=0</b>	3.429%	0.478%
<b>Distance&lt;=1</b>	4.381%	0.021%

**Table3: The precision of LD and S&LD**

	<b>LD</b>	<b>Soundex &amp; LD</b>
<b>Distance=0</b>	0.837%	70.810%
<b>Distance&lt;=1</b>	40.642%	92.178%

**Table4: The recall of LD and S&LD**

Noting that when the max edit distance was set to 0, the precision of S&LD method is about 0.478% and the precision of LD method is about 3.429%. It is clear that the precisions of these two methods are on the same order of



magnitude. But as for the recall, the recall of S&LD method reaches 70.810% while the recall of LD method is about only 0.8%. The recall of two methods vary greatly. When the max edit distance came to 1, the recall of S&LD method became 92.178%, which had broken the limits of original dataset. The reason is that some correct words not appeared in the dictionary come to exist in the “soundexed” dictionary after the process of Soundex. Because the correct words not in the dictionary have the same Soundex values with some similar-sounding words in the dictionary. However, everything has two side, by implementing Soundex, lots of words will get similar Soundex values, the precision will decline.

By comparison, the S&LD method has stronger retrieval ability than GED method. However, the precision of S&LD method is lower than GED method. The S&LD method improves the recall at the price of reducing the precision.

## 4. Improvement

### (1)

The precision of the two methods above are unsatisfactory. While, some other methods can improve the precision. For example, when adopting the N-gram method with parameters of N is equal to 2, and similarity is higher or equal to 0.8, the precision can reach 13.768%; when adopting Custom Edit Distance method with parameter of (match, insertion, deletion, replace) = (-1, 3, 3, 15) to this dataset, the precision can reach 16.594%. However, the recall of the Custom Edit Distance method and N-gram method are low. Some integrated methods may have better performance over when handling this dataset.





## (2)

In real life, we can improve the precision of spelling correcting methods and guarantee the recall stay on a satisfactory level. For example, analyzing the context and the users' spelling habits can improve the precision quite a lot.

## 5. Conclusion

In this report, I analyzed precision and recall of two spelling correcting methods: Global Edit Distance (GED) method and 'Soundex and Global Edit Distance' (S&GED) method. The precision of GED method is higher than S&GED method. As for the performance on the recall, S&GED method is much better than GED method. Moreover, S&GED method is able to break through the limit of original dataset on recall.

When handling the spelling of UrbanDictionary headwords, the performance of approximate matching methods is unsatisfactory for following reasons: Firstly, without the help of linguistic context, the approximate matching methods cannot improve the precision of their methods. Secondly, the principles of approximate matching methods are too strict and rigid to handle these flexible and diversified headwords in UrbanDictionary.



## References

- [1] Naomi Saphra and Adam Lopez (2016) Evaluating Informal-Domain Word Representations with UrbanDictionary. In Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP, Berlin, Germany. pp. 94–98.
- [2] V. I. Levenshtein, Binary codes capable of correcting deletions, insertions, and reversals," in Soviet physics doklady, vol. 10, pp. 707-710, 1966.
- [3] David Su. (2018). weighted-levenshtein.  
<https://github.com/infoscout/weighted-levenshtein>.
- [4] Christian, P. (1998) Soundex - can it be improved? COMPUTERS IN GENEALOGY. 6(5):215-221.
- [5] SILPA. (2017). Soundex. <https://github.com/libindic/soundex>.

