University of Macau

澳門大學
UNIVERSIDADE DE MACAU
UNIVERSITY OF MACAU

CISC3025 Natural Language Processing Project #1

by

Your Name

Student No: DC026157

Project Supervisor

Prof Derek, Fai Wong

01 February 2024

Table of Contents

1) Introduction

For this project, it aim to tackle one of the core issues & challenges in Natural Language Processing (NLP) — figuring out the edit distance between words and sentences. This is like the backbone for many tasks we see in NLP, such as correcting typos in texts, finding how similar two pieces of text are, and even translating languages. It's pretty crucial because it helps improve how systems understand and process languages, making them more efficient and accurate in understanding human language.

The project aims to implement a sequence comparison algorithm, specifically the Levenshtein Distance, to measure the similarity between two sequences by quantifying the minimum number of operations required to transform one sequence into the other, including edit distance between words, between sentences, and the algorithm also can read input files.

2) Background

- **Edit Distance (Levenshtein Distance):**

  Edit Distance, also known as Levenshtein Distance, measures the minimum number of operations required to transform one string into another. These operations typically include insertions, deletions, and substitutions of characters. The concept is widely used in NLP for tasks such as spell checking and measuring the similarity between texts.

- **Dynamic Programming:**

  Dynamic programming is a method for solving complex problems by breaking them down into simpler subproblems. It stores the results of these subproblems to avoid computing the same results multiple times, making the algorithm more efficient. This approach is especially useful in optimization problems and is widely used in fields like computer science, economics, and bioinformatics.

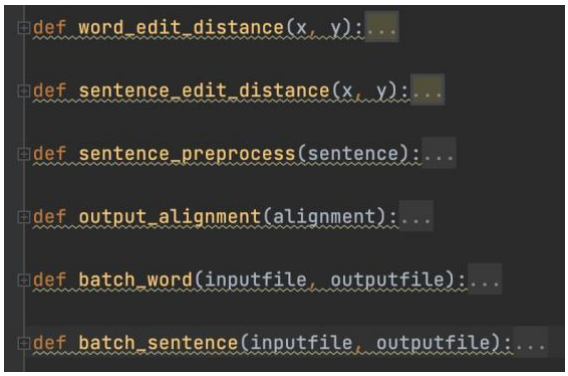
- **Tokenization:**

  Tokenization in Natural Language Processing (NLP) is the process of breaking down text into smaller units called tokens, which can be words, phrases, or symbols. This is a fundamental step in preparing text for further analysis or processing, as it helps in understanding the structure and meaning of the text by analyzing its individual components.

3) Approach and Challenges

- **Approach:**

  With the starting code that Prof provided, I further complete the algorithm by implementing dynamic programming ideas into the edit distance function. Total there are 6 functions. The code is also in OOP manner, so that future changes is possible. For example, the pre-processing part is separated from edit distance of sentences function, so that even with different pre-processing method, the code can be run with minimum changes.



- **Challenges:**

  The challenges is met is how alignment algorithm should be written after I tackle with the word distance algorithm, given its complexity and need for precision in tracking changes between string (like substitution is considered before insertion & deletion in my program). Additionally, working with command-line arguments was new territory, requiring research and learning to effectively parse and utilize inputs from the terminal for the program's various functionalities. These challenges were overcome through diligent study and practical application, enhancing both the project's success and personal skill development in software development and NLP.

4) Results

The results were promising, with my algorithms performing well on given test case, showing good potential for practical NLP applications. Below are some examples :

- word_edit_distance function

```
Cost between CAT & CARIO is 4.
An possible alignment is:
C A - - T
| | | | |
C A R I O
```

- sentence_edit distance function

```
Cost between I LOVE NATURAL LANGUAGE PROCESSING. & I ENJOY NATURAL LANGUAGE PROCESSING ALSO. is 3.
An possible alignment is:
        I       LOVE    NATURAL   LANGUAGE PROCESSING         -         .
        |         |         |         |         |         |
        I     ENJOY    NATURAL   LANGUAGE PROCESSING      ALSO         .
```

- batch_word function

```
R raining
H rainning 1
H raning 1
R writings
H writtings 1
R disparagingly
H disparingly 2
R yellow
H yello 1
```

- batch_sentence function

```
R 28-year-old chef found dead at san francisco mall
H the 28-year-old cook was found dead in a san francisco mall 7
H the 28 cook was found dead in a san francisco mall 11
H 28-year-old chef found dead in a shopping mall in san francisco 7
R a 28-year-old chef who had recently moved to san francisco was found dead in the stairwell of a local mall this week .
H a 28-year-old chef who recently moved to san francisco was found dead this week at a local shopping mall . 11
H a 28 chef , who has just moved to san francisco , was found dead on the stairs of a local mall this week . 14
H a 28-year-old chef who recently moved to san francisco was found dead in the stairwell of a local mall this week . 1
```

- **Main Finding :**
  The main findings of the project include the successful implementation of edit distance algorithms for both words and sentences, demonstrating their applicability in various NLP tasks. The development of a robust tokenization process to handle sentence comparisons effectively, and the integration of command-line arguments for dynamic input handling, were significant achievements.

- **Future Improvement :**
  Can focus on the efficiency, since the algorithms I coded didn't consider about time and space algorithm. And in NLP, often time, there will be a long string of text, resulting in many tokens, therefore, efficient algorithm is a very important to do.

5) Conclusion

Through this project, I gained significant insights into NLP, particularly in implementing edit distance algorithms. It was an enriching experience to delve into dynamic programming and tokenization, enhancing understanding of these core concepts. This project contributes to the field by providing a practical implementation of edit distance in NLP, demonstrating its applicability in text analysis and similarity assessment. It underscores the importance of foundational algorithms in advancing NLP tools and techniques. I'm excited for next advanced topics that I can learn.