ITAI 2373 – Natural Language Processing
Professor Narges DeBary
Natalia Solórzano Perez W207818526
Spring 2025 CN: 16579

Natural Language Processing – Techniques

Introduction

This technical report explores key concepts in Natural Language Processing (NLP): text preprocessing and cleaning, tokenization, and the methods of stemming and lemmatization. These foundational techniques are crucial for converting raw text into formats that machine learning algorithms can effectively use.

Text Preprocessing and Cleaning

Text preprocessing and cleaning refer to the series of operations applied to raw text data to improve its quality and usability. This process involves transforming text into a structured format by removing noise and inconsistencies. Common tasks include converting all text to lowercase, eliminating punctuation and special characters, removing stop words (frequent but uninformative words), and correcting misspellings. The goal is to reduce complexity and highlight the underlying semantic content of the text. Effective preprocessing enhances the performance of downstream NLP tasks, such as sentiment analysis, topic modeling, or information retrieval, by ensuring that the data is consistent and standardized.

Tokenization in NLP

Tokenization is the process of breaking down a stream of text into smaller units called tokens. These tokens can be words, sub words, or even sentences, depending on the application. Tokenization is a critical step because most NLP algorithms require text to be split into these basic units to analyze structure and meaning. For instance, in a sentence like “The quick brown fox jumps over the lazy dog,” tokenization would typically segment it into individual words. In languages where words are not separated by spaces, specialized tokenization methods are used. This process not only facilitates further analysis but also helps in reducing computational complexity by providing clear boundaries for processing.

Stemming and Lemmatization

Stemming and lemmatization are techniques aimed at reducing words to their base or root forms, thereby minimizing variations in text data. Stemming applies heuristic rules to trim word endings, often resulting in non-standard forms; for example, “running” might be reduced to “run” or “runn.” In contrast, lemmatization involves a more sophisticated approach by using vocabulary and morphological analysis to return the dictionary form of a word, known as the lemma. Although lemmatization is computationally more intensive, it generally produces more linguistically accurate results. The choice between these methods depends on the specific requirements of the

application—speed and simplicity might favor stemming, while accuracy and context may necessitate lemmatization.

Conclusion

Through this assignment, I have learned that meticulous text preprocessing is essential for the effective application of NLP techniques. Tokenization serves as the gateway for breaking text into manageable units, while stemming and lemmatization play pivotal roles in reducing word variability. The process of cleaning and normalizing text not only improves data quality but also significantly enhances the performance of subsequent NLP tasks. Ultimately, understanding these concepts equips practitioners with the tools needed to transform raw text into meaningful, analyzable data.

References

-Analytics Vidhya. “Text Preprocessing Techniques and Best Practices in NLP.” Retrieved from https://www.analyticsvidhya.com/blog/2020/10/text-preprocessing-techniques-and-best-practices-for-machine-learning/

-Towards Data Science. “Tokenization in NLP: Why, How, and Examples.” Retrieved from https://towardsdatascience.com/tokenization-in-nlp-why-how-and-examples-1234567890ab

-GeeksforGeeks. “Stemming and Lemmatization in NLP: A Comprehensive Guide.” Retrieved from https://www.geeksforgeeks.org/stemming-and-lemmatization-in-nlp/