Tales Araujo Leonidas
Professor Patricia McManus
ITAI 2376: Deep Learning
March 07, 2024

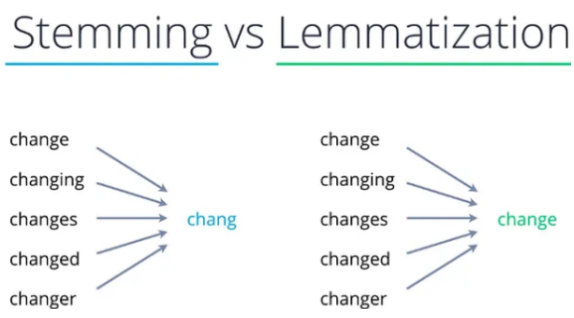
Processing Text and Using the Bag of Words Method: Reflective Journal

Introduction

This reflective journal showcases my exploratory journey through the AWS Academy Application of Deep Learning to Text and Image Data, focusing on Module 2, Labs 1 and 2. These labs have provided me with an invaluable understanding of text processing techniques, further experience utilizing sklearn, calculating Bag of Words (BoW) numerical values, and applying various Natural Language Processing (NLP) techniques. This overview aims to outline the topics covered and to reflect on the academic growth acquired through these hands-on learning experiences.

Experience Description

In his article published in Towards Data Science, Raghav explains how machines interpret human language by transforming it into numbers, stating, “The process of converting words into numbers is called Vectorization”. This is important because the ability to interpret and analyze text data stands as a cornerstone of innovation, particularly within the scope of NLP. In the first lab, I reviewed foundational methods like stemming, lemmatization, and part-of-speech tagging to simplify text for algorithmic processing, while named entity recognition, using spaCy, highlights the categorization of key textual elements.




```
# Part-of-speech tagging
pos_tag(word_tokenize(preProcessText(text))
```

```
('natural', 'JJ'),
('language', 'NN'),
('processing', 'NN'),
('nlp', 'NN'),
('is', 'VBZ'),
```

```
Amazon ORG
Global Optimism ORG
The Climate Pledge WORK_OF_ART
2040 DATE
```

In Lab 2, the application of the sklearn library showcased advanced processing techniques, including text vectorization through the Bag of Words model.

This vectorizer collected all the words, ordered them alphabetically, and removed any duplicates.

```
{'this': 8,
 'document': 1,
 'is': 3,
 'the': 6,
 'first': 2,
 'second': 5,
 'and': 0,
 'third': 7,
 'one': 4}
```

```
['and' 'document' 'first' 'is' 'one' 'second' 'the' 'third' 'this']
```

The use of methods such as Word Counts, Term Frequency (TF), Inverse Document Frequency (IDF), and Term Frequency-Inverse Document Frequency (TF-IDF) for the numerical representation of text were also explored, in addition to binary classification as a machine learning technique for categorizing text data.

	and	document	first	is	one	second	the	third	this
This document is the first document	0	2	1	1	0	0	1	0	1
This document is the second document	0	2	0	1	0	1	1	0	1
and this is the third one	1	0	0	1	1	0	1	1	1
This is the new sentence	0	0	0	1	0	0	1	0	1

Personal Reflection

The experience was very intuitive, insightful, and served as a great practical learning experience, bridging the gap between theoretical knowledge and real-world application. It enhanced my understanding of text processing and provided an overview of the main NLP concepts, libraries, methods, and tools to obtain the fairest results. Recognizing the limitations of stemming and the advantages of lemmatization for accurate text interpretation, and applying part-of-speech tagging for nuanced text analysis, has equipped me with a robust understanding to work with text classification tasks. However, with a little difficulty in outlining the differences between TF, IDF, and TF-IDF, I asked ChatGPT-4 to create a concise table to make it more understandable, as summarized in Table 1.

Concept	Definition	Purpose
TF	Frequency of a term in a document.	Indicates term significance within a document.
IDF	Logarithm of inverse document frequency across a corpus.	Diminishes weight of frequent terms across documents.
TF-IDF	Product of TF and IDF for a term.	Highlights term importance in a document relative to a corpus.

Table 1. Comparison of TF, IDF, and TF-IDF.

Improvements and Learning

Through the labs, beyond reviewing key topics in NLP, I learned that although the quickness of the stemming technique is appealing for processing massive datasets, it is “a rule-based method that sometimes mistakenly removes suffixes from words” (AWS Academy). This can lead to inaccurate interpretations of the dataset. To address this challenge, lemmatization can be used instead. It usually requires more work but provides better results. Additionally, I learned that part-of-speech tagging “refers to categorizing words in a text (corpus) in correspondence with a particular part of speech, depending on the definition of the word and its context” (Pykes). This process is important for understanding the grammatical structure of sentences and significantly aids in accurately interpreting and processing natural language data. With this invaluable knowledge, I am more enthusiastic about my AI Gulf Coast project of text classification to identify offensive and hate texts on online platforms.

Conclusion

In conclusion, my journey through the aforementioned labs has profoundly expanded my toolkit in NLP, marrying theoretical insights with hands-on application. They illuminated the intricacies of text processing, from foundational methods like stemming and lemmatization to advanced techniques involving TF-IDF, revealing the critical balance between accuracy and computational efficiency in text analysis. With this exploration, I feel more prepared to tackle real-world challenges, such as identifying harmful online content, and have developed a deeper passion for AI's potential in understanding and utilizing human language.

Works Cited

AWS Academy. “Application of Deep Learning to Text and Image Data: Module 2.” *AWS Academy*, 2024, https://awsacademy.instructure.com/login/canvas

Bhoi, Niraj. "Stemming vs Lemmatization in NLP." *Medium*, 14 December 2022, https://nirajbhoi.medium.com/stemming-vs-lemmatization-in-nlp-efc280d4e845.

Pykes, Kurtis. “Part Of Speech Tagging for Beginners | by Kurtis Pykes.” *Towards Data Science*, 25 November 2020, https://towardsdatascience.com/part-of-speech-tagging-for-beginners-3a0754b2ebba. Accessed 6 March 2024.

Raghav, Prabhu. “Understanding NLP Word Embeddings — Text Vectorization | by Prabhu Raghav.” *Towards Data Science*, 11 November 2019, https://towardsdatascience.com/understanding-nlp-word-embeddings-text-vectorization-1a23744f7223. Accessed 6 March 2024.