Sittichai Chaikamol

Introduction

This reflection is about my journey and growth in learning how text analysis plays a key part in getting data ready for Natural Language Processing (NLP) tasks. During this process, I focused on understanding text processing techniques and the Bag-of-Words (BoW) method, both of which are essential for turning plain text into data that machine learning can use.

The goal of this lab was to learn key methods for analyzing text data for NLP projects and to get hands-on experience in changing text into numerical formats that work with machine learning algorithms. By thinking over the challenges I faced, linking theory with its real-world use, and looking at the skills I improved on during the lab, I plan to solidify what I've learned and spot areas for more in-depth study.

Description of Experience or Topic:

In these labs I learned and gained more techniques to understand and manipulate textual data. Recognizing the importance of preparing text data for various NLP tasks, including:

- Stop word removal: Identifying and removing common, uninformative words (e.g., "the," "a," "is") that do not contribute significantly to the meaning of the text.
- Part-of-speech (POS) tagging: Identifying the grammatical function of each word (e.g., noun, verb, adjective).
- Stemming/Lemmatization: Reducing words to their base forms, improving consistency for analysis.

These techniques provided the foundation for understanding the structure and content of textual data. Additionally, I explored the BoW method, which focuses on word occurrence and disregards word order and grammatical structure. Different variations of BoW exist, including:

- Binary classification: Records whether a word exists in a sentence.
- Word counts: Captures the frequency of each word within a document.
- Term Frequency-Inverse Document Frequency (TF-IDF): Considers both word frequency within a document and its overall document frequency, penalizing frequently used words across documents.

Learning about BoW showed me how important text processing is, because it uses the words that are picked out and prepared by these methods.

Sittichai Chaikamol

Personal Reflection:

- Thoughts and Feelings: At first, it was tough to get the hang of some methods, like deciding which stemming or lemmatization technique to use. But making word clouds, which show how often words appear, was helpful and interesting. As I went on, I began to see how these ideas all fit together.
- Analysis and Interpretation: Getting deeper into BoW made it clearer how text that's been worked on turns into data machines can be used. Seeing that BoW builds on words picked out and prepped by earlier methods made me feel like I was really getting somewhere and showed me why each step matters.
- Connections to Theoretical Knowledge: I kept linking what I was learning to what we've covered. Things like POS tagging fit right in with what I know about grammar, and the idea of making words basic with stemming or lemmatization clicked with the whole concept of making words simple. Thinking about how different ways of using BoW, like the basic kind versus the more detailed TF-IDF, have their pros and cons made me think harder.
- Critical Thinking: I noticed that simple ways like binary classification or just counting words can fall short. TF-IDF, even though it takes more work, gives a better picture by weighing how common words are in one document against all documents. This shows that choosing the right BoW method is about finding a good middle ground between being simple and detailed.

Sittichai Chaikamol

Discussion of Improvements and Learning:

- Personal Growth: This journey has really pushed me forward in understanding NLP. Working hands-on with key methods and seeing how text analysis works in real life taught me a lot, including:
    i. How to use text processing methods like POS tagging, stemming/lemmatization, and NER.
    ii. How to change and work with text data.
    iii. How to think critically about information and compare different methods.
- Skills Developed:
    i. Working with text data
    ii. Handling data
    iii. Understanding NLP (like how to show text data, work on features)
    iv. Thinking critically
- Future Application:
    i. Sentiment Analysis: Figuring out if text is positive, negative, or neutral to see what people think or how happy customers are.
    ii. Information Extraction: Pulling out specific bits of information from texts to make gathering and adding data easier.
    iii. Chatbot Development: Creating chatbots that can understand and answer back

Looking back at this journey through text processing and the BoW method has been really rewarding. I've come to fully understand how these methods connect and why they're so important for getting text ready for NLP tasks.

Getting hands-on with things like identifying parts of speech, making words simpler through stemming/lemmatization, and spotting key information in text gave me great tools for digging into what text is all about. Also, trying out different ways to use BoW (like binary classification, counting words, and TF-IDF) helped me see how we turn processed text into numbers that machine learning can use.

This experience made me see the challenges of turning plain text into something useful. It showed me how choosing the right method for the job is key and that there's always more to learn and discover.

I'm excited to keep learning, especially about more complex topics like topic modeling or how computers can learn from text using neural networks. It's also important to think about how we use these NLP models responsibly. This whole reflection has not just made what I've learned stronger; it's also sparked a real interest in seeing what else NLP can do in different areas.

Sittichai Chaikamol

References:

- AWS Module2 Lab 1 and 2 Applications of Deep Learning to Text and Image Data
- https://www.youtube.com/watch?v=hhjn4HVEdy0&t=424s
- https://www.youtube.com/playlist?list=PLZLuc8eJafeEOqK22w_OfqIuzVBTQOE3K
- https://www.youtube.com/watch?v=kLMhePA3BiY
- https://www.youtube.com/watch?v=ATK6fm3cYfI