A02 A Comparative Analysis of Machine Learning and Deep Learning Tools and Frameworks

Team 5

Contributors: Martin Demel, Bradley Johnson, Jesus Ocampo, Ethan Phan

Department of Science, Technology, Engineering & Math, Houston Community College

ITAI 2376 Deep Learning in Artificial Intelligence

Patricia McManus

January 29th, 2025.

**Introduction**

NLTK and SpaCy are the most popular libraries for Natural Language Processing (NLP). Both have proven effective in various real-world applications, though they excel in different areas.

**Body**

NLTK is known for its comprehensive tools and flexibility, making it a popular choice for research and educational purposes. One of its most common applications is sentiment analysis, where the library’s VADER tool determines whether a piece of text expresses positive, negative, or neutral sentiments. This is especially useful for businesses that analyze customer feedback or social media conversations to understand public opinion about their products. Another key application of NLTK is in historical text analysis. Researchers and libraries use NLTK to process centuries-old documents, extracting meaningful information such as names, events, and dates. NLTK’s ability to handle complex linguistic features, like parsing and tokenization, makes it ideal for working with historical texts where language can be archaic or nuanced.

SpaCy, on the other hand, is known for its speed and efficiency, making it better suited for real-world, large-scale applications. For instance, financial firms use SpaCy’s Named Entity Recognition (NER) to automatically extract important information, such as company names, monetary values, and dates, from large volumes of financial documents. This significantly reduces the time needed for manual data extraction. In healthcare, SpaCy has been applied to process medical records, extracting details like patient diagnoses, medications, and procedures. SpaCy’s pre-trained models allow easy adaptation to different industries, making it an effective tool for the real-time processing of specialized documents.

During our research, we created a Jupyter notebook to test SpaCy and NLTK, focusing on how they processed the text by splitting it into sentences and tokens. We converted the text into tokens—words and punctuation. This demonstrated basic NLP tasks like sentence segmentation and tokenization. During our presentation, we will show in Visual Studio Code how SpaCy and NLTK can perform these tasks in real-time.

**Conclusion**

While SpaCy and NLTK are excellent choices for Natural Language Processing, they are meant for different use cases. NLTK is ideal for research and smaller-scale applications, where SpaCy excels in speed and scalability for industry-level projects. Both libraries offer valuable tools for various NLP tasks, making them essential in their respective domains. Deciding between Spacy and NLTK depends on performance requirements and project complexity.

**Resources:**

* GeeksforGeeks. (2022, September 16). Tokenization Using Spacy library. GeeksforGeeks. <https://www.geeksforgeeks.org/tokenization-using-spacy-library/> ​
* GeeksforGeeks. (2024, May 1). Introduction to NLTK: Tokenization, stemming, lemmatization, POS tagging. GeeksforGeeks. <https://www.geeksforgeeks.org/introduction-to-nltk-tokenization-stemming-lemmatization-pos-tagging/> ​
* 1. Language Processing and Python. (2024). <https://www.nltk.org/book/ch01.html> ​
* Nathan, P. (2024, June 28). Natural language processing in Python using spaCy: An introduction. <https://domino.ai/blog/natural-language-in-python-using-spacy>
* Extractive Automatic Text Summarization using SpaCy in Python & NLP. (2021, March 4). IEEE Conference Publication | IEEE Xplore. <https://ieeexplore.ieee.org/document/9404712>