NLTK vs SpaCy

By: Codie Munos

Introduction

This presentation is focused on the differences and similarities in two open source NLP libraries. These libraries include the NLTK and SpaCy platforms. Both of these work well, but they have specific areas that they are better used in.

Background of NLTK

Natural Language Processing Toolkit is a python based module and is one of the most popular platforms to be used in the handling of human language data. It is an open-source library and has a great set of tools and resources. It originated at the University of Pennsylvania in the early 2000's. The leaders of the team were Steven Bird and Edward Loper. Their goal was to have this large extensive library that would allow access for everyone to use NLP tools.

Sources: Pant, Anmol. "NLP 101-Data Preprocessing & Representation Using NLTK." *Medium*, CodeChef-VIT, 2 Aug. 2020, medium.com/codechef-vit/nlp-101-data-preprocessing-representation-using-nltk-def9b01194ce.

Nandgaonkar, Pushkar. "Getting Started With NLTK." *Codersarts*, 28 Sept. 2023, www.codersarts.com/post/nltk-library.

Background of SpaCy

SpaCy is another free open-source library that is used in NLP. The goal of the library is to allow room for a fast and accurate NLP pipeline. It was developed by Matthew Honnibal.

Source:

Desosa. *spaCy: For All Things NLP - DESOSA 2020*. 9 Mar. 2020,

desosa.nl/projects/spacy/2020/03/09/the-product-vision.html.

Key Features of NLTK

- NTLK has been around for decades compared to other libraries.
- It is customizable for more NLP projects in the sense of the library has pre-existing models along with the ability for the user to create their own if needed.
- There are more languages that it is compatible with.
- More advanced models can be built due to it being integrated with other libraries.

Source: Srivastava, Prabhu. “SpaCy Vs. NLTK: A Comprehensive Comparison of Two Popular NLP Libraries in Python”.” *Medium*, 29 Apr. 2023, medium.com/@prabhuss73/spacy-vs-nltk-a-comprehensive-comparison-of-two-popular-nlp-libraries-in-python-b66dc477a689.

Key Features of SpaCy

- SpaCy is a newer library, but is does work faster than other libraries such as NLTK.
- It is able to achieve faster processing of large text models because of the use of Cython.
- It is very user-friendly and also comes with pre-trained models to kickstart a project.
- The tokenization used is able to process more complex sentence structure.
- It has the ability to classify text, recognize context, and create more advanced NLP models.

Source: Srivastava, Prabhu. “SpaCy Vs. NLTK: A Comprehensive Comparison of Two Popular NLP Libraries in Python”.” *Medium*, 29 Apr. 2023, medium.com/@prabhuss73/spacy-vs-nltk-a-comprehensive-comparison-of-two-popular-nlp-libraries-in-python-b66dc477a689.

Real world applications of NLTK

Spam filtering is done by using NLTK, an example being emails that get directed to different folders. It does this by recognizing similarities/patterns along with extracting keywords that can be a sign of which folder it needs to go in. It is also favorable for customer reviews on websites. The sentiment analysis allows for it to filter specific feedback such as bad or good reviews.

Source: Codezup. “Unlock Real-World NLP Applications With spaCy & NLTK.” *Codez Up*, 22 Nov. 2024, codezup.com/real-world-nlp-applications-spacy-nltk.

Real World Applications of SpaCy

SpaCy is helpful for extracting trends and insights on social media. This side of the media is accessible on business accounts. An example being Instagram or TikTok.

Source: Mayo, Matthew. “Applications of SpaCy in Real-World Text Analytics Projects.” *Statology*, 17 Feb. 2025, www.statology.org/applications-spacy-real-world-text-analytics-projects.

Comparison: SpaCy vs NLTK

- NLTK is an older library while SpaCy is much newer.
- SpaCy is more user-friendly which leads newer people in the field to lean towards using it.
- Compared to all NLP libraries, the tokenization used in SpaCy is top tier. It is designed for more complex wording.
- SpaCy is not as flexible as its older friend, NLTK, which is the better option for a customizable experience.
- NLTK offers more compatibility with more languages around the world allowing a wider range of users accessibility.
- SpaCy is better when handling large amounts of text data.
- NLTK is better in scholarly settings such as research.

Summary:

SpaCy and NLTK are both great Natural Language Processing libraries. They both have pro’s and con’s depending on the intended purpose of the model. For a better experience, it is best to consider what type of data will be used and which library is best suited to reach the goal.