CS410 – Fall 2020
Technology Review

Thiago Seuaciuc-Osorio
thiagos2@illinois.edu

Natural Language Processing Toolkits: NLTK and spaCy

Introduction

Natural Language Processing

Natural Language Processing (NLP) focuses on automatically processing and analyzing textual data. Despite being faced with formidable challenges, such as the unstructured nature of textual data, assumed reader prior knowledge and ambiguities commonly found in human languages, there has been increased interest in the field. This can be attribute, among other reasons, to the massive collection of such data readily available (in the internet, for instance), continuing advances in computation power and the many fields where it finds applications, from speech recognition to areas of commercial value such as virtual assistants and product review sentiment analysis.

Efforts in NLP started in as early as the 1950s [1], with efforts based on increasingly complex sets of rules. The introduction of machine learning approaches (facilitated by the increase in available computation power) led to statistical approaches in the 1990s; most of the robust and widely used NLP techniques follow this approach. More recently, neural networks for NLP are becoming more common, although deep networks are not yet robust enough for wide-spread applications, being typically restricted to problems with narrow and well-defined scope.

Common Tasks

Natural Language Processing deals with a variety of tasks. Some examples include:

- *Pre-Processing*: examples include tokenization (dividing text in logical units—typically words—for processing) and stemming & lemmatization (reducing words to their most basic root or form)
- *Lexical analysis*: determining part-of-speech
- *Syntactic analysis*: determining structure of sentences and phrases
- *Semantic analysis*: interpreting meaning from text

The list above is in order of complexity. State-of-the-art NLP cannot perform error-free lexical analysis because of word-level ambiguities, and syntactic analysis is less accurate, especially in view of many possible syntactic ambiguities. Semantic analysis is only achieved with reasonable accuracy in certain scenarios, such as sentiment analysis for some domains.

NLP tasks are language-specific; approaches that work in one language may not work in another. Even tokenization, which may be the simplest pre-processing steps in NLP, is different in English and Chinese, for example, being considerably more complex for the latter. Additionally, as with many other machine learning fields, NLP solutions will be more successful and robust with domain-specific development.

These considerations indicate each NLP problem or application will likely require at least some level of specific, custom development for better results. A number of tools are available to help interested practitioners, novice or experienced, to develop their own solutions using NLP. This document introduces two such NLP toolkits: *NLTK* and *spaCy*. Brief comparative summaries can be found in Table 1 and Table 2.

NLTK: a Great Place to Learn

Natural Language Toolkit, or NLTK [2], is a Python open source library for working with human language originally developed by the Department of Computer and Information Science at the University of Pennsylvania in 2001. Including a large collection of corpora, trained models and grammars in different languages [3], it is widely used by researchers for prototyping and developing research systems and as a teaching tool by educators as well as a self-study tool by interested learners. The relatively early release year has contributed to its current broad use and adoption.

In terms of technical capabilities, NLTK supports various NLP tasks such as classification, tokenization, stemming, tagging, parsing, and semantic reasoning. It provides tools for lexical analysis, n-grams and collocations, part-of-speech tagging, generation of tree models and named-entity recognition, among others. It does not contain neural network models. Multiple language support is available for certain tasks. Stemming and sentence tokenization is supported in multiple languages, for instance, while most NLTK taggers only support English (NLTK does provide a module to interact with the Stanford tagger [4], which does provide multilingual support—provided the Stanford tool is available on the machine).

Technical capabilities aside, the popularity of this tool among the scientific and learning community can in part be attributed to its companion book, *Natural Language Processing with Python* [5] (also available online for free [6]), which provides an accessible, practical introduction to NLP using the NLTK, and an active discussion forum [7]. The NLTK website [2] provides additional support through API documentation [8], a Wiki page [9] and a series of guiding “How To’s” [10].

spaCy: a Production Tool

spaCy [11] is another, fairly recent (2015) open source library for natural language processing, developed by Explosion AI. In contrast to NLTK, spaCy is more focused on providing a platform for production usage, rather than research and development.

spaCy’s technical capabilities extend a little beyond that of NLTK (see Table 2), especially by adding neural network models and linkage to popular machine learning libraries such as TensorFlow and PyTorch. It also provides support for different languages for different tasks (see full list of available models for each language at [12]), and interesting visualization tools.

Its website [11] provides a number of short guides on different NLP tasks, providing an introduction not only to how to perform them with the tool, but what the task means in the NLP context and how to interpret the results as well. It also provides a number of in-depth examples with full-length sample code to aid users.

Table 1
NLTK and spaCy: General Information

	NLTK	spaCy
Website	https://www.nltk.org/	https://spacy.io/
Programming Language	Python (3.5, 3.6, 3.7 or 3.8)	Python, Cython (2.7 or 3.5+)
Supported OS	macOS, Unix, Windows	
License	Apache 2.0 [13]	MIT [14]
Original Release Year	2001	2015
Latest Stable Release	3.5 (April 2020)	2.3.2 (July 2020)
Supported Languages	Multiple (task dependent)	
Repository	https://github.com/nltk/nltk	https://github.com/explosion/spaCy

Table 2
NLTK and spaCy: Capabilities (source: [11])

	NLTK	spaCy
Neural network models	✗	✓
Integrated word vectors	✗	✓
Multi-language support	✓	✓
Tokenization	✓	✓
Part-of-speech tagging	✓	✓
Sentence segmentation	✓	✓
Dependency parsing	✗	✓
Entity recognition	✓	✓
Entity linking	✗	✓

Conclusion

Expression through human language in the form of text is ubiquitous, and therefore the ability to automatically process it finds applications in virtually any industry or domain area. However, the nuances and ambiguities commonly present in human expression make this a difficult task, and while great advances have been made, it continues to be the target of research as many challenges still remain.

This document provided a brief introduction to two popular software libraries commonly used by the community. Because of its long-standing and extensive support material, including a free book and active discussion forum, the *Natural Language Toolkit*, or *NLTK*, is the recommended tool for beginners, hobbyists and researchers in an academic environment. *spaCy*, a newer library, is recommended for those interested in developing production usage software and applications, as spaCy provides much better tools for that purpose, or for those interested in exploring more advanced approaches such as neural network models for NLP.

References

[1] https://en.wikipedia.org/wiki/Natural_language_processing, November 6, 2020

[2] https://www.nltk.org/, November 7, 2020

[3] http://www.nltk.org/nltk_data/, November 7, 2020

[4] https://nlp.stanford.edu/software/tagger.shtml, November 7, 2020

[5] Steven Bird, Ewan Klein, and Edward Loper (2009). *Natural Language Processing with Python*. O'Reilly Media Inc

[6] http://www.nltk.org/book/, November 8, 2020

[7] https://groups.google.com/g/nltk-users, November 7, 2020

[8] https://www.nltk.org/api/nltk.html#, November 7, 2020

[9] https://github.com/nltk/nltk/wiki, November 7, 2020

[10] http://www.nltk.org/howto/, November 7, 2020

[11] https://spacy.io/, November 8, 2020

[12] https://spacy.io/usage/models, November 9, 2020

[13] https://www.apache.org/licenses/LICENSE-2.0, November 7, 2020

[14] https://en.wikipedia.org/wiki/MIT_License, November 7, 2020