Natural Language Processing: NLTK vs spaCy

NLTK and spaCy are two of the most popular Natural Language Processing (NLP) tools available in Python. You can build chatbots, automatic summarizers, and entity extraction engines with either of these libraries. While both can theoretically accomplish any NLP task, each one excels in certain scenarios.

NLTK and spaCy are better suited for different types of developers. For scholars and researchers who want to build something from the ground up or provide a functioning model of their thesis, NLTK is the way to go. Its modules are easy to build on and it doesn't really abstract away any functionality. After all, NLTK was created to support education and help students explore ideas.

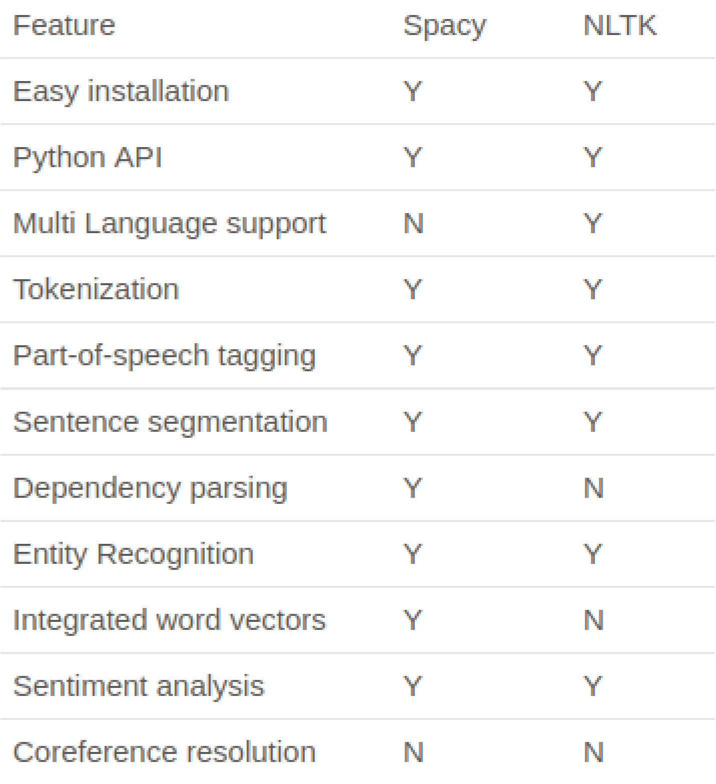
SpaCy, on the other hand, is the way to go for app developers. While NLTK provides access to many algorithms to get something done, spaCy provides the best way to do it. It provides the fastest and most accurate syntactic analysis of any NLP Library released to date. It also offers access to larger word vectors that are easier to customize. For an app builder mindset that prioritizes getting features done, spaCy would be the better choice.

Approach And Performance

A core difference between NLTK and spaCy stems from the way in which these libraries were built. NLTK is essentially a string processing library, where each function takes strings as input and returns a processed string. Though this seems like a simple way to use the library, in practice, you'll often find yourself going back to the documentation to discover new functions.

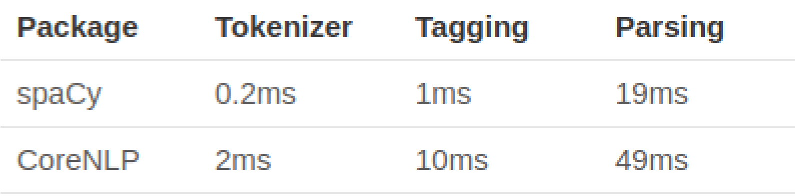
In contrast, spaCy takes an object-oriented approach. Each function returns objects instead of strings or arrays. This allows for easy exploration of the tool. Developers don't need to constantly check with documentation to understand context because the object itself provides it.

FEATURE AVAILABILITY:



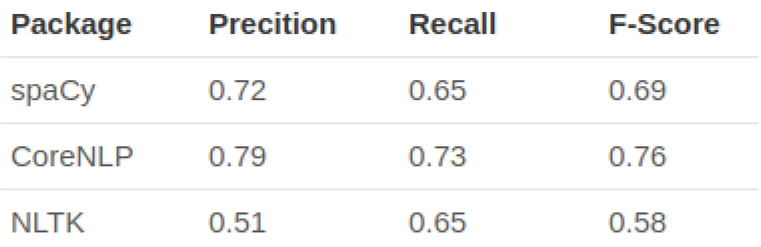
Feature	Spacy	NLTK
Easy installation	Y	Y
Python API	Y	Y
Multi Language support	N	Y
Tokenization	Y	Y
Part-of-speech tagging	Y	Y
Sentence segmentation	Y	Y
Dependency parsing	Y	N
Entity Recognition	Y	Y
Integrated word vectors	Y	N
Sentiment analysis	Y	Y
Coreference resolution	N	N

SPEED: KEY FUNCTIONALITIES — TOKENIZER, TAGGING, PARSING:



Package	Tokenizer	Tagging	Parsing
spaCy	0.2ms	1ms	19ms
CoreNLP	2ms	10ms	49ms

ACCURACY: ENTITY EXTRACTION:



Package	Precition	Recall	F-Score
spaCy	0.72	0.65	0.69
CoreNLP	0.79	0.73	0.76
NLTK	0.51	0.65	0.58

INSTALL:

1. IMPORT

SPACY:

```
sudo pip install spacy
```

NLTK:

```
pip install nltk
```

2. WORD TOKENIZE

text = """Most of the outlay will be at home. No surprise there, either. While Samsung has expanded overseas, South Korea is still host to most of its factories and research engineers. """

[SPACY OUTPUT]:

```
['Most', 'of', 'the', 'outlay', 'will', 'be', 'at', 'home', '.', 'No', 'surprise', 'there', ',', 'either', '.', 'While', 'Samsung', 'has', 'expanded', 'overseas', ',', 'South', 'Korea', 'is', 'still', 'host', 'to', 'most', 'of', 'its', 'factories', 'and', 'research', 'engineers', '.']
```

[NLTK OUTPUT]:

```
['Most', 'of', 'the', 'outlay', 'will', 'be', 'at', 'home', '.', 'No', 'surprise', 'there', ',', 'either', '.', 'While', 'Samsung', 'has', 'expanded', 'overseas', ',', 'South', 'Korea', 'is', 'still', 'host', 'to', 'most', 'of', 'its', 'factories', 'and', 'research', 'engineers', '.']
```

3. SENTENCE TOKENIZE:

```
text = """Natural Language Toolkit, or more commonly NLTK, is a suite of libraries and programs for
symbolic and statistical natural language processing (NLP) for English written in the Python
programming language. It was developed by Steven Bird and Edward Loper in the Department of
Computer and Information Science at the University of Pennsylvania."""
```

[SPACY OUTPUT]:

```
[Natural Language Toolkit, or more commonly NLTK, is a suite of libraries and programs for symbolic and
statistical natural language processing (NLP) for English written in the Python programming language.,

 It was developed by Steven Bird and Edward Loper in the Department of Computer and Information
Science at the University of Pennsylvania.]
```

[NLTK OUTPUT]:

```
['Natural Language Toolkit, or more commonly NLTK, is a suite of libraries and programs for symbolic
and statistical natural language processing (NLP) fXor English written in the Python programming
language.',

'It was developed by Steven Bird and Edward Loper in the Department of Computer and Information
Science at the University of Pennsylvania.']
```

4. STOP WORDS REMOVAL

```
text = """Most of the outlay will be at home. No surprise there, either. While Samsung has expanded
overseas, South Korea is still host to most of its factories and research engineers. """
```

[SPACY OUTPUT]:

```
['Most', 'outlay', 'home', 'No', 'surprise', 'While', 'Samsung', 'expanded', 'overseas', 'South', 'Korea',
'host', 'factories', 'research', 'engineers']
```

[NLTK OUTPUT]:

```
['Most', 'outlay', 'home', '.', 'No', 'surprise', ',', 'either', '.', 'While', 'Samsung', 'expanded', 'overseas', ',',
'South', 'Korea', 'still', 'host', 'factories', 'research', 'engineers', '.']
```

5. Lemma

text = """While Samsung has expanded overseas, South Korea is still host to most of its factories and research engineers. """

[SPACY OUTPUT]:

```
While while

Samsung samsung

has have

expanded expand

overseas overseas

South south

Korea korea

is be

still still

host host

to to

most most

of of

its -PRON-

factories factory

and and

research research

engineers engineer
```

[NLTK OUTPUT]

```
['While', 'Samsung', 'ha', 'expanded', 'overseas', ',', 'South', 'Korea', 'is', 'still', 'host', 'to', 'most', 'of', 'it',
'factory', 'and', 'research', 'engineer', '.']
```

6. get word frequency

text = """Most of the outlay will be at home. No surprise there, either. While Samsung has expanded overseas, South Korea is still host to most of its factories and research engineers. """

[SPACY OUTPUT]:

```
[('factories', 1), ('engineers', 1), ('No', 1), ('Most', 1), ('research', 1)]
```

[NLTK OUTPUT]:

```
[('factories', 1), ('still', 1), ('engineers', 1)]
```

7. pos tags

text = """Natural Language Toolkit, or more commonly NLTK."""

[SPACY OUTPUT]:

```
Natural PROPN

Language PROPN

Toolkit PROPN

, PUNCT

or CCONJ

more ADJ

commonly ADV

NLTK NOUN

. PUNCT
```

[NLTK OUTPUT]:

```
[('Natural', 'JJ'),
 ('Language', 'NNP'),
 ('Toolkit', 'NNP'),
 (',', ','),
 ('or', 'CC'),
 ('more', 'JJR'),
 ('commonly', 'RB'),
 ('NLTK', 'NNP'),
 ('.', '.')]
```

8. NER

text = """Most of the outlay will be at home. No surprise there, either. While Samsung has expanded overseas, South Korea is still host to most of its factories and research engineers. """

[SPACY OUTPUT]:

```
ORG ['Samsung ']
GPE ['South Korea ']
```

[NLTK OUTPUT]:

```
['Samsung', 'South Korea']
```