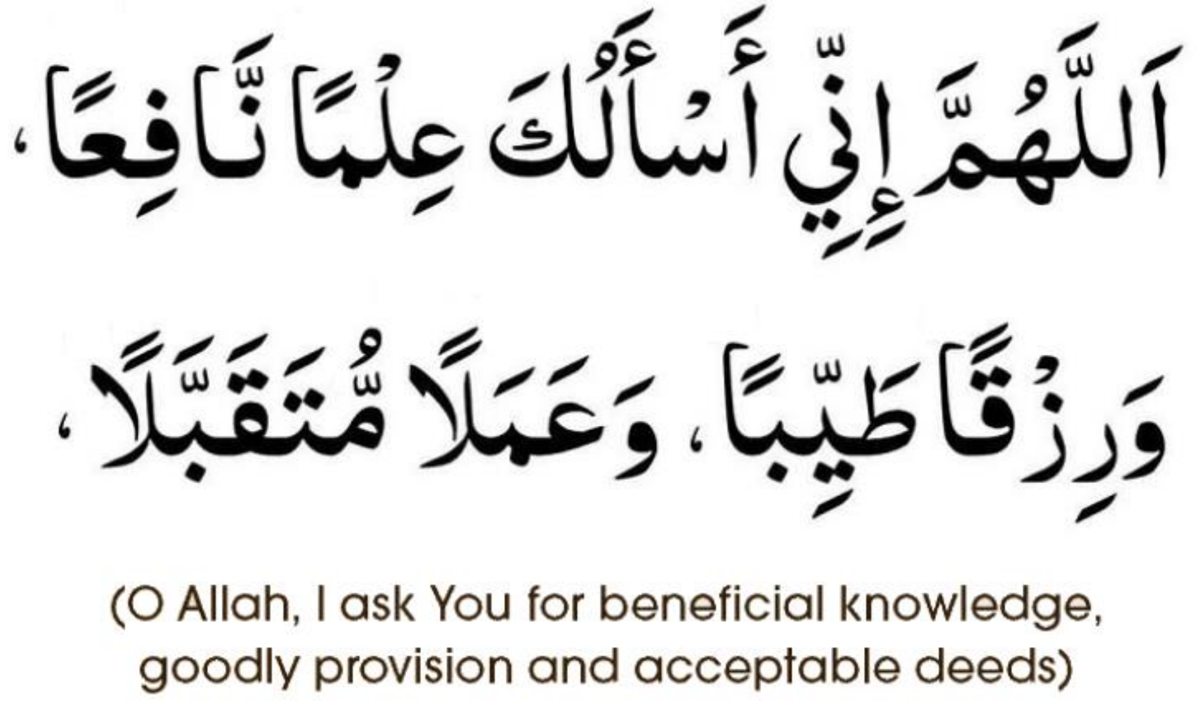
Natural Language Processing (NLP)

Unit 3
Basic Test Processing using NLTK and spaCy

By:
Syeda Saleha Raza

AL NAFI,
A company with a focus on education,
wellbeing and renewable energy.



اَللَّهُمَّ إِنِّي أَسْأَلُكَ عِلْمًا نَّافِعًا ،

وَرِزْقًا طَيِّبًا ، وَعَمَلًا مُّتَقَبَّلًا ،

(O Allah, I ask You for beneficial knowledge,
goodly provision and acceptable deeds)

اے اللہ ، میں آپ سے سوال کرتی ہوں نفع بخش علم کا، طیّب رزق کا، اور اس عمل کا

(Sunan Ibn Majah: 925)

Outline

- Text Pre-processing
  - Tokenization
  - Stemming
  - Lemmatization
  - Stop-words
  - Part-of-Speech(POS) Tagging
- Text Processing using NLTK
- Text Processing using Spacy

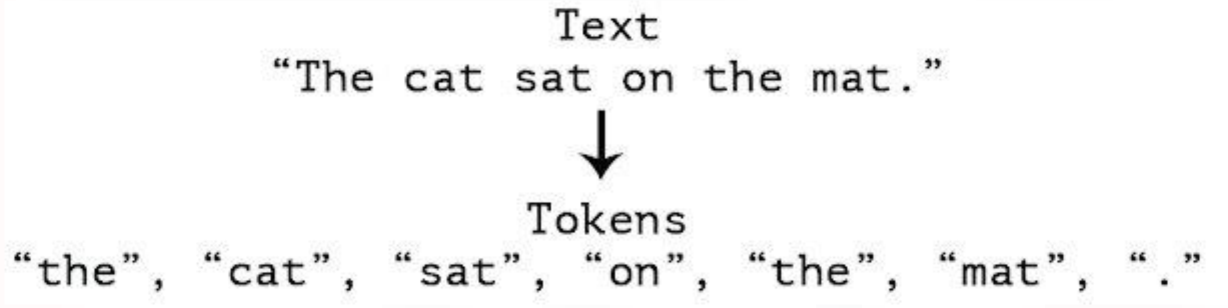
References

- https://medium.com/towards-artificial-intelligence/text-mining-in-python-steps-and-examples-78b3f8fd913b
- Comparison of Top 6 Python NLP Libraries | by Igor Bobriakov | ActiveWizards — AI & ML for startups | Medium

Text Pre-Processing

Tokenization

- Tokenization is the first step in NLP. It is the process of breaking strings into tokens which in turn are small structures or units.

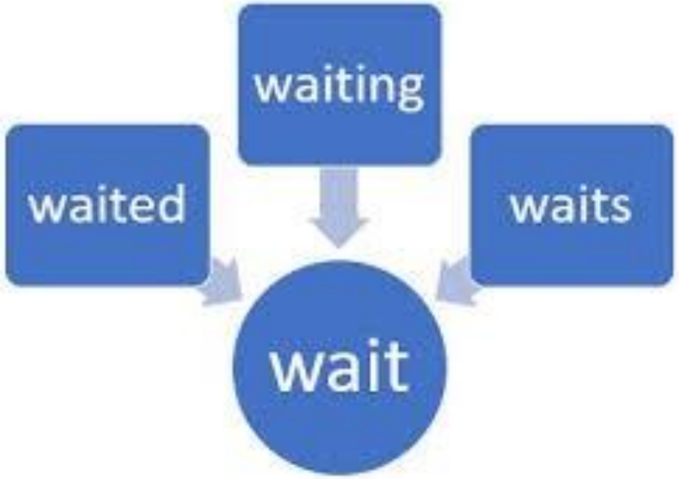


Text
“The cat sat on the mat.”
↓
Tokens
“the”, “cat”, “sat”, “on”, “the”, “mat”, “.”

https://medium.com/towards-artificial-intelligence/text-mining-in-python-steps-and-examples-78b3f8fd913b

Stemming

- *Stemming usually refers to normalizing words into its base form or root form.*

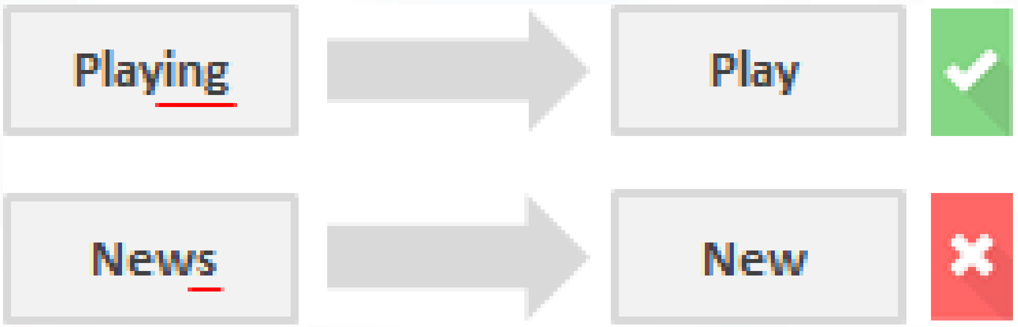


https://medium.com/towards-artificial-intelligence/text-mining-in-python-steps-and-examples-78b3f8fd913b

Stemming

- NLTK provides two methods in Stemming namely,
  - Porter Stemming (removes common morphological and inflectional endings from words) and
  - Lancaster Stemming (a more aggressive stemming algorithm).

Stemming

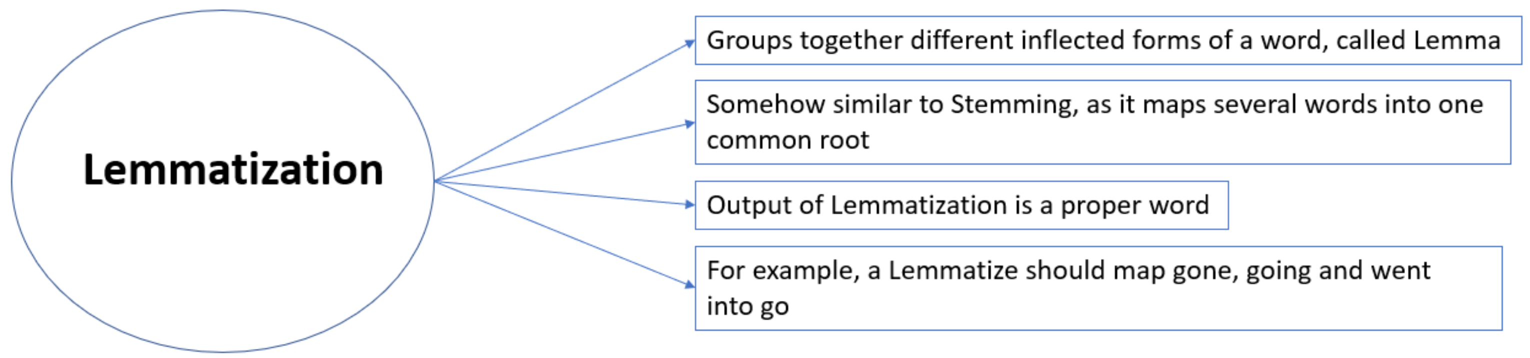


Playing → Play ✓

News → New ✗

Lemmatization

- Reducing a word to its base form and grouping together different forms of the same word



Lemmatization

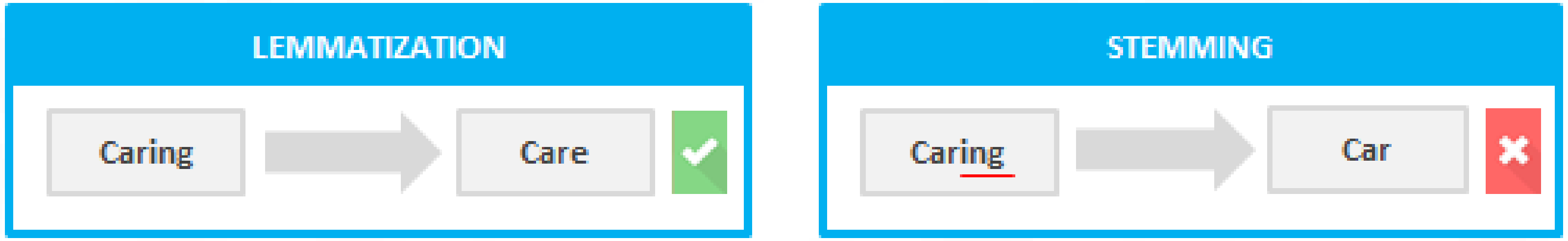
Groups together different inflected forms of a word, called Lemma

Somehow similar to Stemming, as it maps several words into one common root

Output of Lemmatization is a proper word

For example, a Lemmatize should map gone, going and went into go

https://medium.com/towards-artificial-intelligence/text-mining-in-python-steps-and-examples-78b3f8fd913b



LEMMATIZATION

Caring → Care ✓

STEMMING

Caring → Car ✗

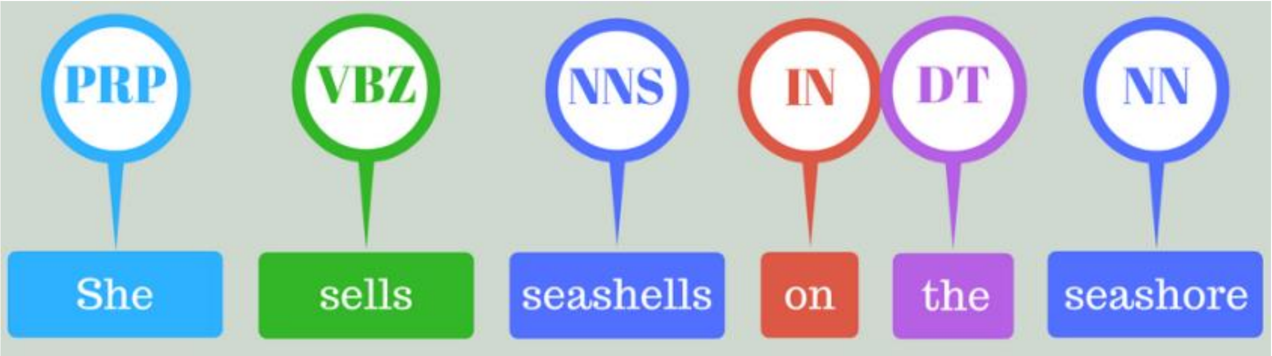
Stop Words Removal

- “Stop words” are the most common words in a language like “the”, “a”, “at”, “for”, “above”, “on”, “is”, “all”. These words do not provide any meaning and are usually removed from texts.

https://medium.com/towards-artificial-intelligence/text-mining-in-python-steps-and-examples-78b3f8fd913b

Part of speech tagging (POS)

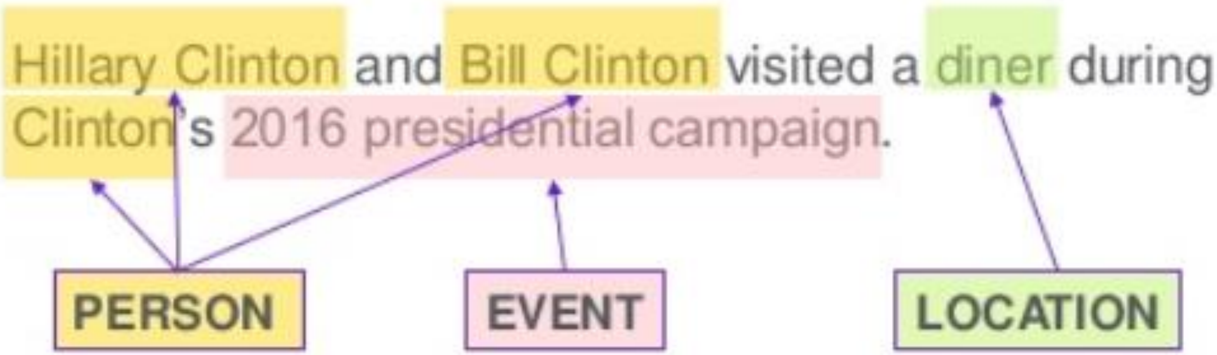
- Part-of-speech tagging is used to assign parts of speech to each word of a given text (such as nouns, verbs, pronouns, adverbs, conjunction, adjectives, interjection) based on its definition and its context.



https://medium.com/towards-artificial-intelligence/text-mining-in-python-steps-and-examples-78b3f8fd913b

Named Entity Recognition (NER)

- It is the process of detecting the named entities such as the person name, the location name, the company name, the quantities and the monetary value.



Hillary Clinton and Bill Clinton visited a diner during Clinton's 2016 presidential campaign.

https://medium.com/towards-artificial-intelligence/text-mining-in-python-steps-and-examples-78b3f8fd913b

About Natural Language Toolkit (NLTK)

NLTK

NLTK (Natural Language Toolkit) is a powerful Python library for working with human language data. It provides tools and resources for text processing related tasks, making it a valuable resource for natural language processing and text analysis.

NLTK Functionality

- Tokenization
- Stopwords
- Stemming and Lemmatization
- Part-of-Speech Tagging
- Frequency Distribution
- WordNet Interface
- Named Entity Recognition (NER)
- Corpora and Resources
- Collocations
- Concordance
- Parsing
- Machine Learning with NLTK

NLTK Corpora

- NLTK has built-in support for dozens of corpora and trained models.
  - Words
  - Stopwords
  - Wordnet
  - Chat
  - News
- you use the NLTK corpus downloader, `>>> nltk.download()`
- https://www.nltk.org/nltk_data/

Setup

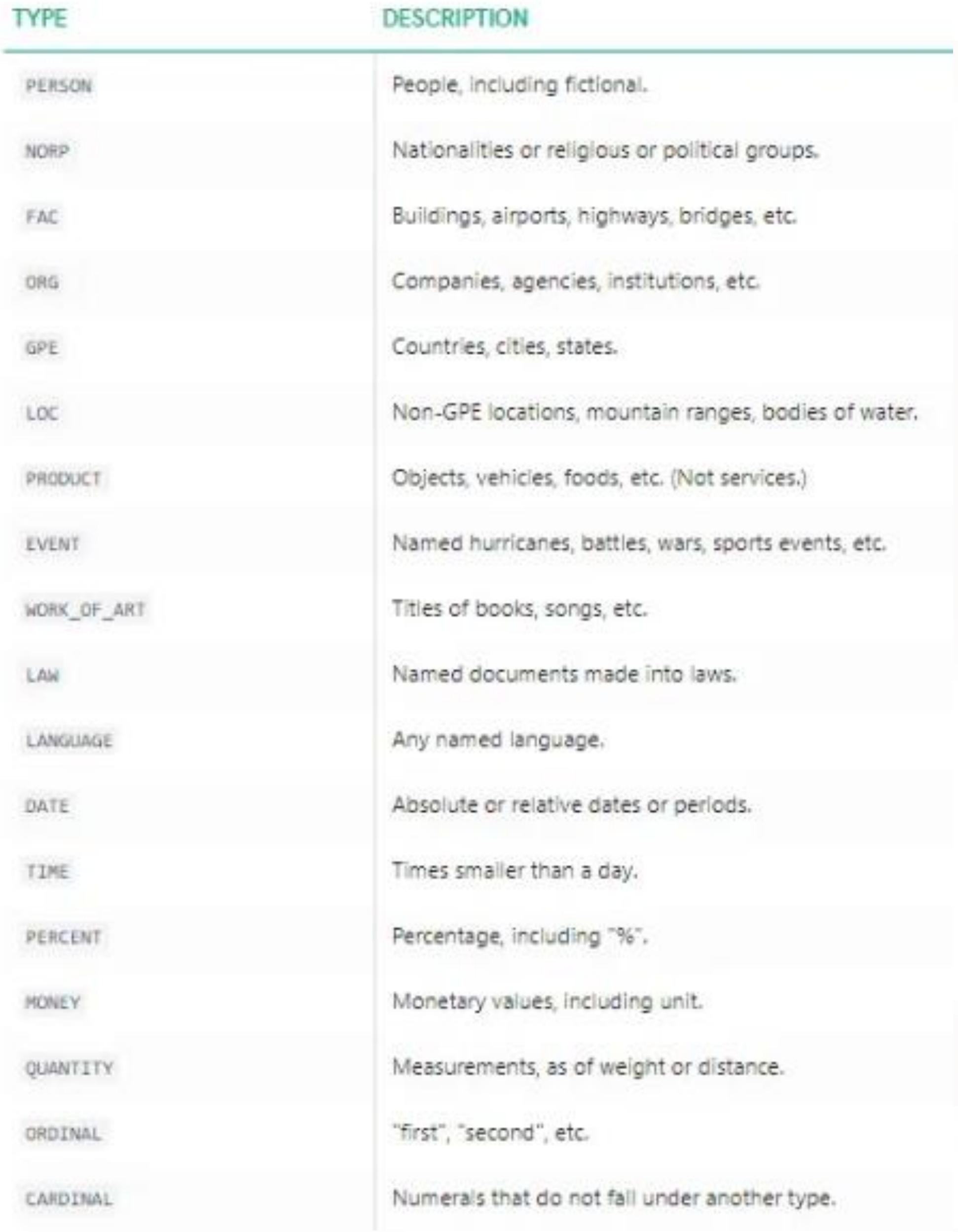
- NLTK :: Installing NLTK

```
pip install nltk
pip install matplotlib
```

Text Processing with NLTK - Demo

List of POS tags

- POS Tagging with NLTK and Chunking in NLP [EXAMPLES] (guru99.com)



TYPE	DESCRIPTION
PERSON	People, including fictional.
NORP	Nationalities or religious or political groups.
FAC	Buildings, airports, highways, bridges, etc.
ORG	Companies, agencies, institutions, etc.
GPE	Countries, cities, states.
LOC	Non-GPE locations, mountain ranges, bodies of water.
PRODUCT	Objects, vehicles, foods, etc. (Not services.)
EVENT	Named hurricanes, battles, wars, sports events, etc.
WORK_OF_ART	Titles of books, songs, etc.
LAW	Named documents made into laws.
LANGUAGE	Any named language.
DATE	Absolute or relative dates or periods.
TIME	Times smaller than a day.
PERCENT	Percentage, including "%".
MONEY	Monetary values, including unit.
QUANTITY	Measurements, as of weight or distance.
ORDINAL	"first", "second", etc.
CARDINAL	Numerals that do not fall under another type.

Text Processing with Spacy - Demo

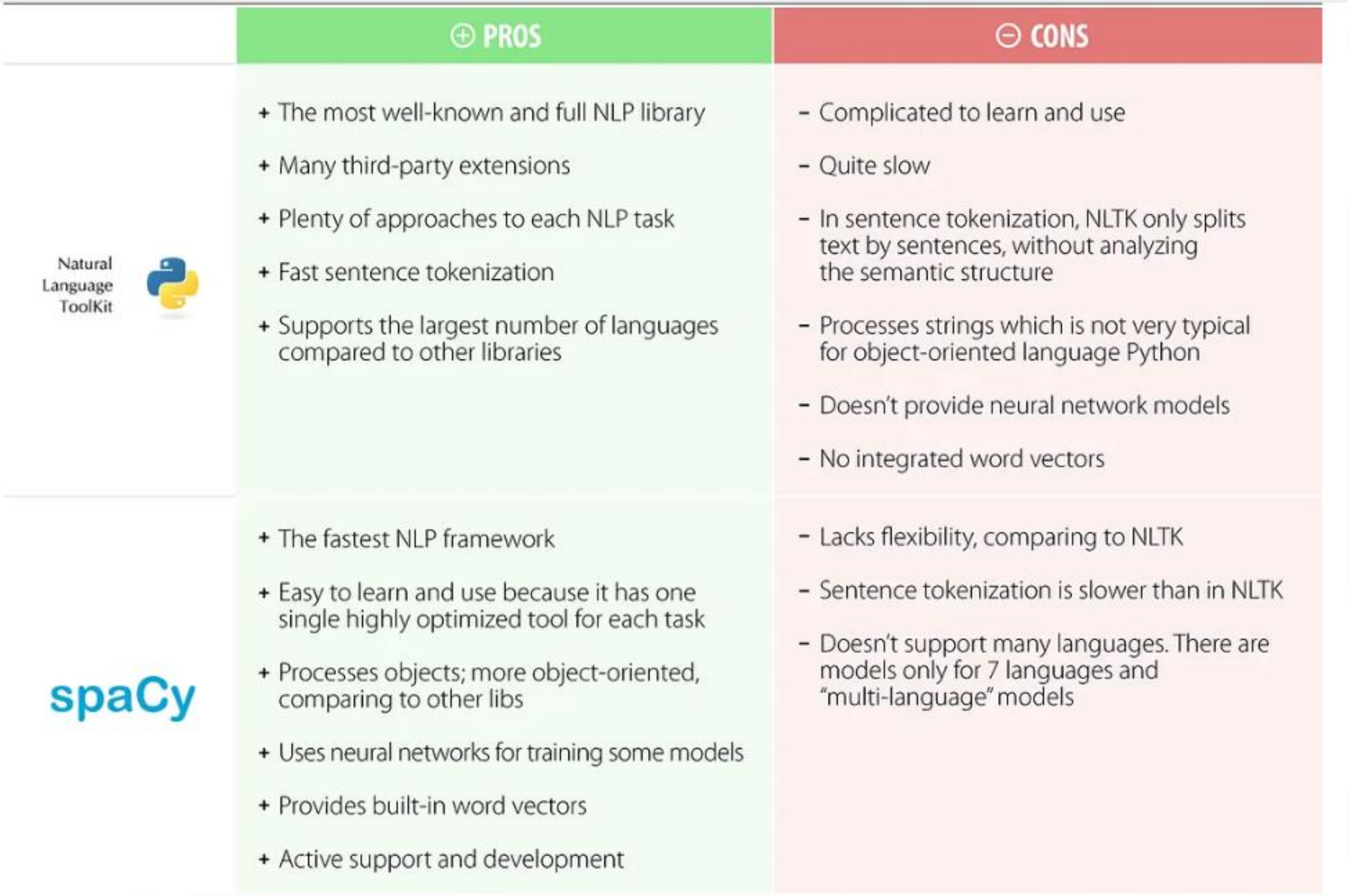
About spaCy

- spaCy is a free, open-source library for NLP in Python written in Cython. spaCy is designed to make it easy to build systems for information extraction or general-purpose natural language processing.

- pip install spacy
- spacy download en_core_web_sm

Spacy Functionality

- Tokenization
- Stopwords removal
- Stemming and Lemmatization
- Part-of-Speech Tagging
- Dependency Parsing
- Named Entity Recognition (NER)
- Word Vectors
- Visualisation
- Similarity Matching
- Custom Pipeline
- Large Pre-trained Models
- Text Classification
- Text Summarization



	⊕ PROS	⊖ CONS
Natural Language ToolKit	+ The most well-known and full NLP library + Many third-party extensions + Plenty of approaches to each NLP task + Fast sentence tokenization + Supports the largest number of languages compared to other libraries	- Complicated to learn and use - Quite slow - In sentence tokenization, NLTK only splits text by sentences, without analyzing the semantic structure - Processes strings which is not very typical for object-oriented language Python - Doesn't provide neural network models - No integrated word vectors
spaCy	+ The fastest NLP framework + Easy to learn and use because it has one single highly optimized tool for each task + Processes objects; more object-oriented, comparing to other libs + Uses neural networks for training some models + Provides built-in word vectors + Active support and development	- Lacks flexibility, comparing to NLTK - Sentence tokenization is slower than in NLTK - Doesn't support many languages. There are models only for 7 languages and "multi-language" models

Comparison of Top 6 Python NLP Libraries | by Igor Bobriakov | ActiveWizards — AI & ML for startups | Medium

References

- **Natural Language Processing with Python, Steven Bird, Ewan Klein, and Edward Loper (**NLTK Book**)**
- https://medium.com/towards-artificial-intelligence/text-mining-in-python-steps-and-examples-78b3f8fd913b
- Comparison of Top 6 Python NLP Libraries | by Igor Bobriakov | ActiveWizards — AI & ML for startups | Medium
- NLP using NLTK Library | NLTK Library for Natural Language Processing (analyticsvidhya.com)

Thanks