NLP with Python

Goals

1. Familiarising with the set of Python libraries for NLP
2. Appreciating the use cases of the individual Python libraries for NLP

Python has many libraries for NLP

TextBlob

scikit learn

CoreNLP

Stanza

NLTK

spaCy

AllenNLP

GENSIM

topic modelling for humans

tomotopy

Topic Modeling Tools for Python

PyTorch

Research tools for small projects

NLTK

Key features

- It is the most popular Python library for NLP.
- Exceptionally well-documented
- A Swiss-army knife can achieve a very diverse set of tasks (from tokenisation, through First order Logic, to sentiment analysis).
- It does not excel at any, though.

Domains/use cases

- Education/research.
- Low-memory intensive applications.
- Many text pre-processing tasks.
- Many text analysis tasks.

TextBlob

Key features

- It builds on NLTK.
- It provides high-level functionalities to get the sentiment scores of words and sentences.
- It implements also sentiment analysis via Naïve Bayes Classifier.

Domains/use cases

- Education/research.
- Low-memory intensive applications.
- Some text pre-processing
- ‘Quick & dirty’ sentiment score analysis and classification.

TextBlob

NLP pipelines

spaCy

Key features

- Production-oriented library with an emphasis on efficiency, reliability, and transparency.
- Known for wide inter-operability — it integrates with PyTorch, Tensorflow, Stanford’s GloVe embedding model, and CoreNLP/Stanza.
- It is the centre of a constellation of libraries, called spaCy Universe.
- It runs on GPUs too.
- Cutting-edge documentation.

Domains/use cases

- Industry application.
- Text pre-processing.
- Text manipulation.
- Front-end to train statistical models of language.
- Training/fine-tuning statistical models of language.

spaCy

CoreNLP / Stanza

Key features

- Overall, it offers a set of features comparable to spaCy.
- It is written in Java — it can be accessed from a variety of languages (e.g., Go, R, Ruby).
- Python's interface to CoreNLP comes as an independent library called Stanza.

Domains/use cases

- Industry application.
- Text pre-processing.
- Few text analysis options via additional tools (e.g., sentiment analysis).

CoreNLP

Stanza

Topic modeling

Gensim

Key features

- It offers several features to estimate the Latent Semantic Index and topic models.
- It also allows training word embeddings via word2vec and FastText.
- Efficient in reading/writing text corpora.
- Mature and robust library.

Domains/use cases

- For research and industry application.
- Text analysis: discovering hidden themes in text corpora (via LSI or LDA).
- Modelling training: training word and document embeddings.

GENSIM
topic modelling for humans

Tomotopy

Key features

- Specialistic software for topic modelling
- It integrates the latest research—developments in topic modelling.
- It implements many different flavours of topic modelling (e.g., LDA, dynamic topic modelling, correlated topics models, hierarchical topic models).
- Very efficient — the library is a Python wrapper around C++ code.

Domains/use cases

- For research and practice.
- Text analysis: Topic modelling estimation.

tomotopy

Topic Modeling Tools for Python

Generalist libraries

scikit-learn

Key features

- The industry standard ML library implements a few NLP tools, such as LDA.
- On top of that, it complements NLP libraries by offering numerous post-processing capabilities (e.g., scikit-learn may be used after Tomotopy to analyse doc2topic probabilities).

Domains / use cases

- For research and industry applications.
- Text analysis (e.g., LDA).
- Post-processing of text analysis outcome achieved with specialised NLP libraries.

PyTorch

Key features

- PyTorch, the acclaimed framework for deep learning, has an NLP library called TorchText
- TorchText offers data processing utilities and model training capabilities

Domains / use cases

- Both for research and industry application
- Text pre-processing
- Training statistical models of language

Wrap-up

Main points

Python has one of the richest NLP ecosystems

Here, I alluded to a fraction of established NLP libraries for Python

Altogether, these libraries cover the entire NLP spectrum, from data pre-processing, though model training, to dedicates text analysis tools

It is worth notice that some of the libraries I mentioned are substitutes of similar value (e.g., spaCy and CoreNLP)