[NATURAL LANGUAGE PROCESSING](https://towardsai.net/p/category/nlp)

Five Must-learn Natural Language Processing Technologies

Natural language processing (NLP) is one of the areas in artificial intelligence that deals with the interaction between humans and machines through natural language [1]. According to the “State of AI Report 2020", NLP is one of the most trendy AI technologies areas unlocking many new use cases in the past few years by the emergence of sophisticated language models such as transformers [2]. NLP is all about how machines can understand human languages and replicate the human language abilities to automate the time-consuming human tasks or extract knowledge from the text data to generated actionable insights for business users.

To build high-performance NLP & text analytics software, data scientists need to use a wide range of methods such as exploratory data analysis & feature engineering on text data, knowledge engineering, language models, and machine learning algorithms. There are numerous open-source, ready-made NLP libraries available to data scientists. These libraries are powered by mathematical and statistical methods, Machin learning, and linguistics to deal with various NLP tasks. Among the popular NLP tasks are parts-of-speech (POS) tagging, name-entity recognition (NER), dependency parsing, sentiment analysis, document classification, topic modeling.

Here, we will present the top 5 must-learn libraries for data scientists who would like to work on NLP use cases.

**NLTK**

Natural Language Toolkit (NLTK) is one of NLP’s most famous libraries, a suite of open-source python modules used offering a wide range of functions for language operations. NLTK was Written in Python and developed by Steven Bird and Edward Loper at the University of Pennsylvania to support research and teaching activities in NLP. NLTK provides all types of functions components required in NLP tasks such as tokenization, tagging, classification, stemming, semantic reasoning, and parsing. Although NLTK is very flexible, it processes the text data in the string format, causing some text processing pipelines.

**SpaCy**

SpaCy is an open-source library for conducting advanced NLP tasks. It was written in Python and Cython by Matthew Honnibal under the MIT license. Mostly works on providing software applications to build modules for further production. Compared to other libraries, SpaCy is a powerful library when it comes to effectiveness in text processing pipelines and generating insightful visualizations based on large-scale text data. SpaCy is an object-oriented library, i.e., the text data is defined as objects instead of string format. The other advantage of SpaCy is that it is easily integrated with other data science frameworks, libraries, or software.

**CoreNLP**

CoreNLP or Stanford NLP is a Java-based library for basic NLP tasks like POS tagging, NER, dependency parsing, and sentiment analysis. Initially developed by Stanford University, CoreNLP currently supports six languages: Arabic, Chinese, English, French, German, and Spanish. CoreNLP performs well-known in named entity recognizer and grammatical functions. Also, CoreNLP offers an easy-to-use web interface by which everybody calls its NLP functions through web-based API.

**Gensim**

Gensim was initially implemented in Python and Cython to solve unsupervised topic modeling, a famous NLP task. Gensim has advantages compared to other libraries in real-time processing, the large-scale text datasets using data streaming and aggregation pipelines. Also, Gensim can effectively convert (de-)vectorize the text data and be used for text summarization.

**Textblob**

Textblob is an open-source Python library recommended for beginners because of its simple user interface. It offers many features like non-phrase extraction, POS tagging, sentiment analysis, and machine translation. It is also worth noting that Textblob provides its features with high processing speed and memory usage optimization.

**Conclusion**

As data science use cases evolve in the industry and more sophisticated technologies & methodologies are released, data scientists need to upgrade their toolbox to solve more business problems at deeper levels. Natural language processing is, in fact, one of those areas that has been very much in-demand in the industry within the past few years. Therefore, mastering the technologies mentioned above can create massive return-on-investment for data scientists as it can enable many new opportunities for young data science talents.