Natural language processing (NLP) refers to the branch of computer science—and more specifically, the branch of artificial intelligence or AI—concerned with giving computers the ability to understand text and spoken words in much the same way human beings can.

**Spacy**:

* spaCy is designed specifically for **production use and helps you build applications that process and “understand” large volumes of text**. It can be used to build information extraction or natural language understanding systems, or to pre-process text for deep learning.
* spaCy **provides a variety of linguistic annotations to give you insights into a text's grammatical structure**. This includes the word types, like the parts of speech, and how the words are related to each other.

**En\_core\_web:**

"en" refers to the English language model in spaCy, which is a popular natural language processing library used for various tasks like text classification, named entity recognition, and more.

On the other hand, "en\_core\_web" is a specific version of the English language model that includes a larger set of features like part-of-speech tagging, dependency parsing, and named entity recognition. It is also pre-trained on a larger corpus of text data compared to the basic "en" model.

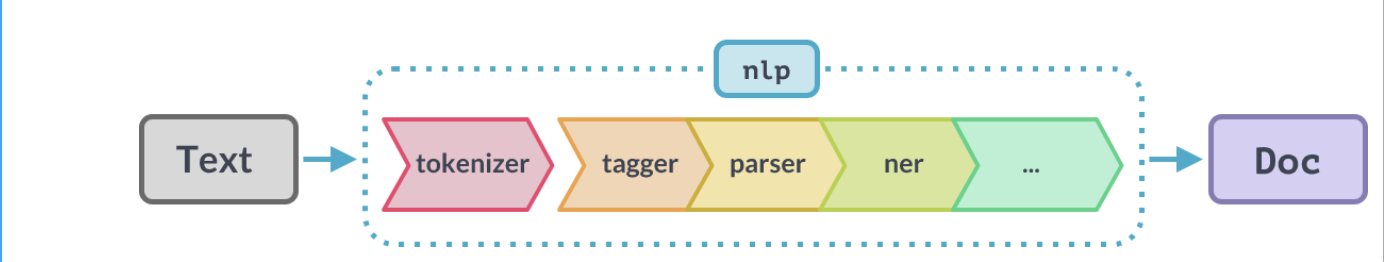
**Displacy**:

A modern syntactic dependency visualizer. Visualize spaCy's guess at the syntactic structure of a sentence. Arrows point from children to heads, and are labeled by their relation type.

**BeautifulSoup:**

Beautiful Soup is a python package and as the name suggests, **parses the unwanted data and helps to organize and format the messy web data by fixing bad HTML and present to us in easily-traversable XML structures**. In short, Beautiful Soup is a python package which allows us to pull data out of HTML and XML documents.

* So, the main difference between a tagger and POS is that a tagger is a more general term that can refer to any software component or algorithm that assigns labels to words in a text, while POS specifically refers to the grammatical category of a word.



**Tokenizer:**

* Tokenization is an essential task in natural language processing used to break up a string of words into semantically useful units called *tokens*.
* Sentence tokenization splits sentences within a text, and word tokenization splits words within a sentence. Generally, word tokens are separated by blank spaces, and sentence tokens by stops. However, you can perform high-level tokenization for more complex structures, like words that often go together, otherwise known as collocations (e.g., *New York*).

**Tagger**:

* Tagging is an essential feature of text processing where we tag the words into grammatical categorization. We take help of tokenization and pos\_tag function to create the tags for each word.
* A machine learning model that's been trained to classify natural language text at the word level.

**Paser**:

* Parsers are used when there is a need to represent input data from source code abstractly as a data structure so that it can be checked for the correct syntax.
* The processing of a piece of python program and converting these codes into machine language. In general, we can say parse is a command for dividing the given program code into a small piece of code for analyzing the correct syntax.
* Tries to find the dependence between the tokens.

**NER**:

**Vectorization and Similarity:**

Vectorization of tokens refers to the process of representing textual data (such as documents or sentences) as numerical vectors. This is necessary when applying machine learning algorithms to textual data because most algorithms operate on numerical data.

In the context of natural language processing (NLP), there are several ways to vectorize tokens. One common method is called bag-of-words (BOW) representation, where each document is represented as a vector of word counts. Another method is called term frequency-inverse document frequency (TF-IDF) representation, where each document is represented as a vector of TF-IDF scores, which measure how important a word is to a document relative to its importance in a corpus of documents.

**Similarity of documents** refers to the degree of similarity between two or more documents based on their vector representations. The closer the cosine similarity is to 1, the more similar the documents are, and the closer it is to 0, the less similar they are.

**Term frequency-inverse document frequency (TF-IDF)** is a numerical representation of a corpus of text documents, which takes into account both the importance of a term within a document and the rarity of the term in the corpus.

TF-IDF is calculated as the product of two factors: term frequency (TF) and inverse document frequency (IDF).

Term frequency (TF) measures how frequently a term appears in a document. It is calculated as the number of times a term appears in a document divided by the total number of terms in the document. A high TF value indicates that a term is important to the document.

Inverse document frequency (IDF) measures how rare a term is in the corpus. It is calculated as the logarithm of the total number of documents in the corpus divided by the number of documents

In natural language processing (NLP), a **corpus** refers to a large and structured collection of texts or documents that are used for linguistic analysis, language modeling, and machine learning.

**Latent Dirichlet Allocation (LDA)** is a popular topic modeling algorithm in natural language processing (NLP) that allows for the identification of latent topics or themes in a collection of text documents. There are several reasons why LDA is commonly used for topic modeling.

The **output** shows the top 10 words that are most strongly associated with one of the identified topics in the corpus, along with their corresponding weights or probabilities.

Polarity - -1 to 1

Subjectivity - Subjective or Objective

**Subjectivity scores** typically range from 0 to 1, with higher values indicating a greater degree of subjectivity. In this case, a score of 0.75 suggests that the text contains a significant amount of language that is indicative of personal opinions, emotions, or beliefs.