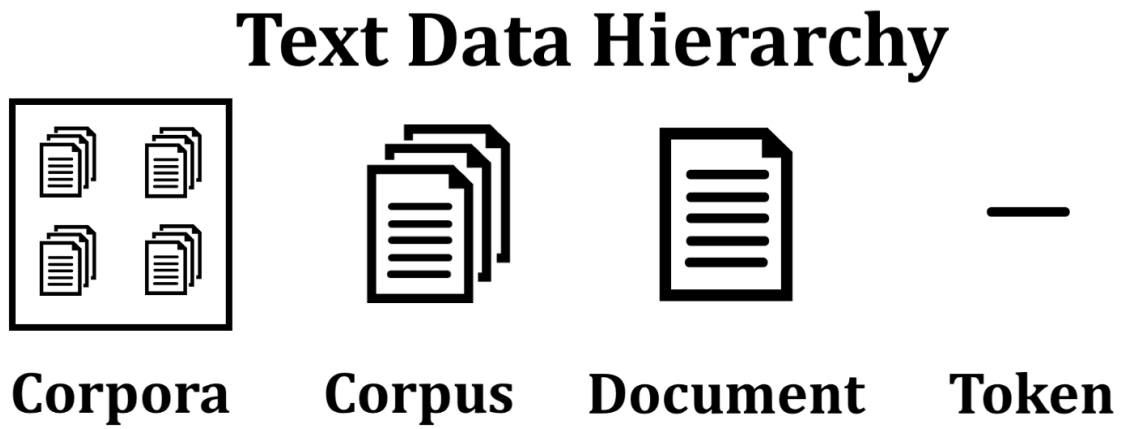
An Overview of NLP

Natural Language Processing (NLP) is the process of how computers interact with human languages. With the aid of modern machine learning capabilities, NLP allows Data Scientists to draw valuable actionable insights from text based data that has previously been overlooked.

NLP techniques range in complexity, from simple frequency based models to more complex deep learning models that take the order of the words into account. Computers extract meaning from text based data by evaluating words as numerical vectors.

Deconstructing Text

NLP analysis works by comparing numerical vectorizations of text. For instance, an article about a restaurant review will have many sentences that pertain to food. This article's sentences will be more similar to other articles of restaurant reviews than the sentences of an article about NASA. NLP uses a hierarchy to determine which groups of words and sentences belong to each other. The smallest level of text is a **token** which can be a sentence or an individual word. A group of tokens is called a **document**, *for instance a paragraph or chapter*. A group of documents is called a **corpus**, *for instance a book or article*. Finally, a group of corpus is called a **corpora**, which can be several books or articles that data scientists wish to compare and evaluate.



A visualization of the text data hierarchy

The NLP Process

In order to tokenize a document the text data needs to be cleaned. This means removing punctuation and capitalization to ensure identical words are not hindered by stray characters. For instance *Apple*, *apple*., and *apple* would all be read in as different words because of punctuation and capitalization. Once identical words appear identical the text is ready for the process of tokenization.

Tokenization is the process of extracting the root meaning from tokens utilizing an NLP tool kit. For instance *driven*, *drove*, and *driving* all have the same root word, *drive*. These words represent the same action, albeit in different tenses. We need to extract the root to find the meaning of the token. There are two main processes for extracting the root of a word: stemming and lemmatization.

Stemming is the process of using a set of rules to extract the root of a word. For instance, some of the rules can be to extract

an *-n* when a word ends in *-en* and to replace *-ing* with *-e*. Driven becomes *drive* and driving also becomes *drive*. This creates a problem when we have the word *drove*. Drove and drive will be read as two different words despite having the same root meaning. This is why stemming is considered rudimentary.

Lemmatization is the process of extracting the basic form of a word or **lemming**. for instance the stem of *driving* is *driv* and the lemma of *driving* is *drive*. There are numerous tokenization resources so you don't have to write a rule for every word in every human language. The **Natural Language Toolkit** (NLTK) has numerous sentence tokenizers that are free to use such as the Punkt toolkit.

Once the text data is cleaned and tokenized it is ready for NLP analysis. Vectorization of the tokens allows data scientist to mathematically represent text as vectors. There are numerous ways to create these vectors.

Count Vectorization is the process of assigning a count to each token in a document. This is a simple process that follows the principle that words that appear more frequently in a document are more important. Each element in a vector is a numerical count of one individual word. You can count only words in the document or all words in the english language, where words that do not appear are assigned a zero value. If you want to compare numerous documents you would have each column represent a word count and each row represent a document's vector.

TF-IDF (Term Frequency-Inverse Document Frequency) follows a more complex approach. Words that are more unique are considered more important to the meaning of a document. For instance, commonly used articles like *a* or *of* are not as important as extracting the meaning from an article than other words like *elected* or *profit*. By placing more importance on these more unique words we can determine the meaning of the article.

Both of these processes have one fatal flaw, the sequence of the words is not taken into account. There are more advanced methods involving deep learning that allows data scientist to evaluate the theme of a document, and even the sentiment of text to quickly and efficiently extract meaning from large quantities of text data.

Ref: https://medium.com/@amitrani/an-overview-of-nlp-fe597ed7e8b6