NLP

Introduction to NLP & NLTK

1) Natural language processing (NLP) is an area of computer science and artificial intelligence concerned with the interaction between computers and humans in natural language.

2) The ultimate goal of NLP is to help computers understand language as well as we do. It is the driving force behind things like virtual assistants, speech recognition, sentiment analysis, automatic text summarization, machine translation and much more.

3) Applications of NLP techniques include voice assistants like Amazon's Alexa and Apple's Siri,but also things like machine translation and text-filtering.

4) Syntactic analysis (syntax) and semantic analysis (semantic) are the two primary techniques that lead to the understanding of natural language. Language is a set of valid sentences, but what makes a sentence valid? Syntax and semantics.

5) Syntax is the grammatical structure of the text, whereas semantics is the meaning being conveyed. A sentence that is syntactically correct, however, is not always semantically correct. For example, **“cows flow supremely”** is grammatically valid (subject—verb—adverb) but it doesn't make any sense.

- Semantic analysis is the process of understanding the meaning and interpretation of words, signs and sentence structure. This lets computers partly understand natural language the way humans do. I say partly because semantic analysis is one of the toughest parts of NLP and it's not fully solved yet.
- Speech recognition, for example, has gotten very good and works almost flawlessly, but we still lack this kind of proficiency in natural language understanding. Your phone basically understands what you have said, but often can’t do anything with it because it doesn’t understand the meaning behind it. Also, some of the technologies out there only make you think they understand the meaning of a text.
- An approach based on keywords or statistics or even pure machine learning may be using a matching or frequency technique for clues as to what the text is “about.” These methods are limited because they are not looking at the real underlying meaning.

What is NLTK?

- NLTK is a standard python library with prebuilt functions and utilities for the ease of use and implementation. It is one of the most used libraries for natural language processing and computational linguistics.
- It provides us various text processing libraries with a lot of test datasets
- A variety of tasks can be performed using NLTK such as tokenizing, parse tree visualization, etc… In this article, we will go through how we can set up NLTK in our system and use them for performing various NLP tasks during the text processing step.

To install

```
pip install nltk
```

Sentence Segmentation

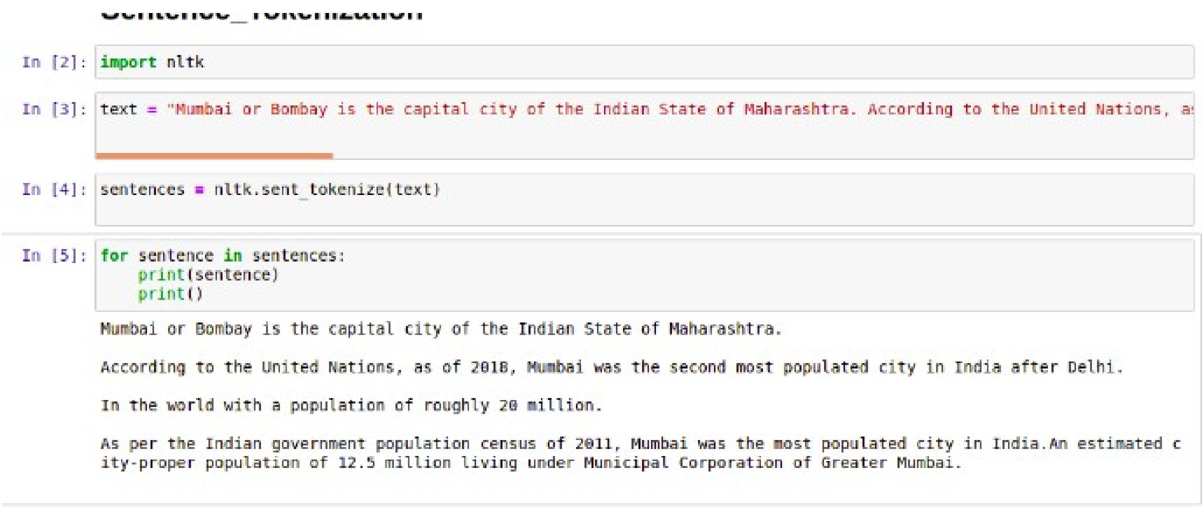
The first step in the pipeline is to break the text apart into separate sentences. That gives us this:

“Mumbai or Bombay is the capital city of the Indian state of Maharashtra.”

“According to the United Nations, as of 2018, Mumbai was the second most populated city in India after Delhi.”

“In the world with a population of roughly 20 million.”

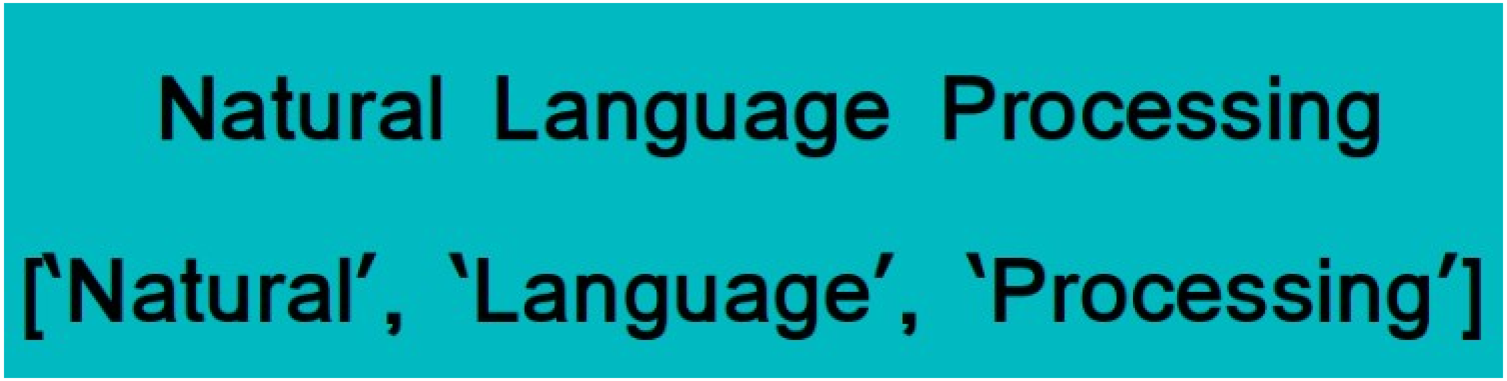
We can assume that each sentence in English is a separate thought or idea. It will be a lot easier to write a program to understand a single sentence than to understand a whole paragraph.



Sentence_Tokenization

```
In [2]: import nltk

In [3]: text = "Mumbai or Bombay is the capital city of the Indian State of Maharashtra. According to the United Nations, a

In [4]: sentences = nltk.sent_tokenize(text)

In [5]: for sentence in sentences:
            print(sentence)
            print()

        Mumbai or Bombay is the capital city of the Indian State of Maharashtra.

        According to the United Nations, as of 2018, Mumbai was the second most populated city in India after Delhi.

        In the world with a population of roughly 20 million.

        As per the Indian government population census of 2011, Mumbai was the most populated city in India.An estimated c
        ity-proper population of 12.5 million living under Municipal Corporation of Greater Mumbai.
```

Tokenization is essentially splitting a phrase, sentence, paragraph, or an entire text document into smaller units, such as individual words or terms. Each of these smaller units are called tokens.



Natural Language Processing

[‘Natural’, ‘Language’, ‘Processing’]

Why is Tokenization required in NLP?

Think about the English language here. Pick up any sentence you can think of and hold that in your mind as you read this section. This will help you understand the importance of tokenization in a much easier manner.

Before processing a natural language, we need to identify the words that constitute a string of characters. That's why tokenization is the most basic step to proceed with NLP (text data). This is important because the meaning of the text could easily be interpreted by analyzing the words present in the text.

Let's take an example. Consider the below string:

"This is a cat."

What do you think will happen after we perform tokenization on this string? We get ['This', 'is', 'a', cat'].

There are numerous uses of doing this. We can use this tokenized form to:

Count the number of words in the text

Count the frequency of the word, that is, the number of times a particular word is present

Tokenization using Python’s split() function

Let’s start with the split() method as it is the most basic one. It returns a list of strings after breaking the given string by the specified separator. By default, split() breaks a string at each space. We can change the separator to anything. Let’s check it out in the programming example.