

Overview of NLP

What is NLP?

Natural Language Processing (NLP) is a branch of AI that deals with analyzing and understanding written and spoken language. It is used in a wide array of application areas including:

- Search engines
- Translation apps
- Spam filters
- Social media
- Speech engines (like Siri)
- Chat bots

NLP Vocabulary

Corpus - a corpus is a body of text. It is usually (but not always) purposefully collected and structured.

They can be used for doing NLP analysis.

Note that the plural form of corpus is corpora.

Example:

http://www.nltk.org/nltk_data/


NLP Process

The NLP process can be broadly broken into the following steps:

1. Data Collection
2. Data Cleaning/Manipulation
 - a. Remove special characters, symbols, and punctuation
 - b. Make text lower case
 - c. Tokenization
 - d. Stop words removal
 - e. Stemming/lemmatization
 - f. Parts of Speech (POS) Tagging
3. Sentiment Analysis
4. Text Representation
5. Modeling and/or Pattern Mining

Data Cleaning/Manipulation

Tokenization - tokenization is the process of splitting long strings of text into small pieces (tokens). Paragraphs can be tokenized into sentences, and sentences can be further tokenized into words.


Ex. A big dog ate the bacon  "A", "big", "dog", "ate", "the", "bacon"

Stop words - stop words are small filler words that are filtered out prior to processing text since they contribute little to the overall meaning of the text. Examples of stop words are "a", "the", "and" and "of".

Ex. A big dog ate the bacon  A big dog ate the bacon

Data Cleaning/Manipulation

Stemming - stemming is the process of deleting prefixes and suffixes from a word, leaving on the word “stem”.

Ex. hunting,  h~~un~~ted, hunter hunt

Lemmatization - lemmatization is similar to stemming, but lemmatization is able to capture the underlying meaning of the word.

 Ex. caring car (stemming)

 Ex. caring care (lemmatization)

Sentiment Analysis

Sentiment analysis can be used to understand the feeling or emotion tied to the text. Sentiment is defined using two metrics:

1. The **polarity score** (a float between -1.0 and 1.0). -1 is negative, 1 is positive.
2. The **subjectivity** (a float between 0.0 and 1.0). 0 is very objective, while 1 is very subjective.

Syntax vs Semantics vs Sentiment

Syntax - syntactic analysis involves ensuring text conforms to an established set of grammatical rules.

Semantics - semantic analysis involves trying to understand the meaning of text. Semantic analysis wants to know what the words actually mean.

Sentiment - sentiment analysis involves identifying the feeling or emotion tied to text. Sentiment can be negative, neutral or positive.

Text Representation

Text Data Vectorization - the process of mapping words to numbers. It is important to numerically represent text so that it can be used by computer algorithms.

We will discuss methods for text data vectorization in another lecture.

Modeling

Once text data has been cleaned and represented numerically, it can be used in a machine learning algorithm. NLP modeling can be used for the following cases:

- **Unsupervised learning** (clustering) to group together similar content (e.g., trying to group together social media posts on the same topic)
- **Supervised learning** to classify certain content (e.g., spam) or make a prediction based on content (e.g., trying to predict a viewer's star rating of a movie based on their comments)

We will discuss specific modeling techniques in another lecture.