# INTRODUCTION

Natural language processing (NLP) is a subfield of computer science and artificial intelligence concerned with the interactions between computers and human languages, in particular how to program computers to process and analyze large amounts of natural language data.

Natural Language Processing (NLP) uses tools, techniques and algorithms to process and understand natural language-based data, which is usually unstructured like text & speech.

Challenges in natural language processing frequently involve speech recognition, natural language understanding, and natural language generation.

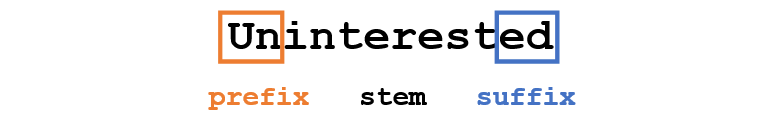
Natural language processing (NLP) is all about creating systems that process or “understand” language in order to perform certain tasks. These tasks may include:

* Question Answering (What Siri, Alexa, and Cortana do)
* Sentiment Analysis (Determining whether a sentence has a positive or negative connotation)
* Image to Text Mappings (Generating a caption for an input image)
* Machine Translation (Translating a paragraph of text to another language)
* Speech Recognition
* Part of Speech Tagging
* Name Entity Recognition

The traditional approach to NLP involved a lot of domain knowledge of linguistics itself. Understanding terms such as phonemes and morphemes were pretty standard as there are whole linguistic classes dedicated to their study. Let’s look at how traditional NLP would try to understand the following word.



Let’s say our goal is to gather some information about this word (characterize its sentiment, find its definition, etc). Using our domain knowledge of language, we can break up this word into 3 parts.



We understand that the prefix “un” indicates an opposing or opposite idea and we know that “ed” can specify the time period (past tense) of the word. By recognizing the meaning of the stem word “interest”, we can easily deduce the definition and sentiment of the whole word.

However, when you consider all the different prefixes and suffixes in the English language, it would take a very skilled linguist to understand all the possible combinations and meanings.

## RULE-BASED VS. STATISTICAL NLP

In the early days, many language-processing systems were designed by hand-coding a set of rules. However, this is rarely robust to natural language variation.

Since the so-called "statistical revolution in the late 1980s and mid-1990s, much natural language processing research has relied heavily on machine learning.

The machine-learning model uses statistical inference to automatically learn such rules through the analysis of large corpora of typical real-world examples.

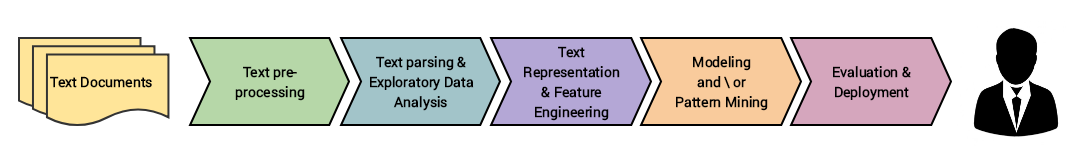
Many different classes of machine-learning algorithms have been applied to natural-language-processing tasks. These algorithms take as input a large set of "features" that are generated from the input data. Some of the earliest-used algorithms, such as decision trees, produced systems of hard if-then rules similar to the systems of hand-written rules that were then common. Increasingly, however, research has focused on statistical models, which make soft, probabilistic decisions based on attaching real-valued weights to each input feature.

Systems based on machine-learning algorithms have many advantages over hand-produced rules:

* The learning procedures used during machine learning automatically focus on the most common cases, whereas when writing rules by hand it is often not at all obvious where the effort should be directed.
* Automatic learning procedures can make use of statistical-inference algorithms to produce models that are robust to unfamiliar input (e.g. containing words or structures that have not been seen before) and to erroneous input (e.g. with misspelled words or words accidentally omitted). Generally, handling such input gracefully with hand-written rules—or, more generally, creating systems of hand-written rules that make soft decisions—is extremely difficult, error-prone and time-consuming.
* Systems based on automatically learning the rules can be made more accurate simply by supplying more input data. However, systems based on hand-written rules can only be made more accurate by increasing the complexity of the rules, which is a much more difficult task.
* In particular, there is a limit to the complexity of systems based on hand-crafted rules, beyond which the systems become more and more unmanageable. However, creating more data to input to machine-learning systems simply requires a corresponding increase in the number of man-hours worked, generally without significant increases in the complexity of the annotation process.

## STANDARD NLP WORKFLOW

CRISP-DM model is typically an industry standard for executing any data science project. Typically, any NLP-based problem can be solved by a methodical workflow that has a sequence of steps. The major steps are depicted in the following figure



We usually start with a corpus of text documents and follow standard processes of text wrangling and pre-processing, parsing and basic exploratory data analysis. Based on the initial insights, we usually represent the text using relevant feature engineering techniques. Depending on the problem at hand, we either focus on building predictive supervised models or unsupervised models, which usually focus more on pattern mining and grouping. Finally, we evaluate the model and the overall success criteria with relevant stakeholders or customers, and deploy the final model for future usage.

# NLP KEY TERMS

Below are few important terms one should be familiar with while working for NLP model.

## TOKENIZATION

Tokenization is, generally, an early step in the NLP process, a step which splits longer strings of text into smaller pieces, or tokens. Larger chunks of text can be tokenized into sentences, sentences can be tokenized into words, etc. Further processing is generally performed after a piece of text has been appropriately tokenized.

## NORMALIZATION

Before further processing, text needs to be normalized. Normalization generally refers to a series of related tasks meant to put all text on a level playing field: converting all text to the same case (upper or lower), removing punctuation, expanding contractions, converting numbers to their word equivalents, and so on. Normalization puts all words on equal footing, and allows processing to proceed uniformly.

## STEMMING

Stemming is the process of eliminating affixes (suffixed, prefixes, infixes, circumfixes) from a word in order to obtain a word stem.



## LEMMATIZATION

Lemmatization is related to stemming, differing in that lemmatization is able to capture canonical forms based on a word's lemma.

For example, stemming the word "better" would fail to return its citation form (another word for lemma); however, lemmatization would result in the following:



## DIFFERENCE BETWEEN STEM AND LEMMA

The stem is the part of the word that never changes even when morphologically inflected; a lemma is the base form of the word. For example, from "produced", the lemma is "produce", but the stem is "produc-". This is because there are words such as production

In linguistic analysis, the stem is defined more generally as the analyzed base form from which all inflected forms can be formed

Some lexemes have several stems but one lemma. For instance, the verb "to go" (the lemma) has the stems "go" and "went" due to suppletion.

## CORPUS

In linguistics and NLP, corpus (literally Latin for body) refers to a collection of texts. Such collections may be formed of a single language of texts, or can span multiple languages -- there are numerous reasons for which multilingual corpora (the plural of corpus) may be useful. Corpora may also consist of themed texts (historical, Biblical, etc.). Corpora are generally solely used for statistical linguistic analysis and hypothesis testing.

## STOP WORDS

Stop words are those words which are filtered out before further processing of text, since these words contribute little to overall meaning, given that they are generally the most common words in a language. For instance, "the," "and," and "a," while all required words in a particular passage, don't generally contribute greatly to one's understanding of content.



## PARTS-OF-SPEECH (POS) TAGGING

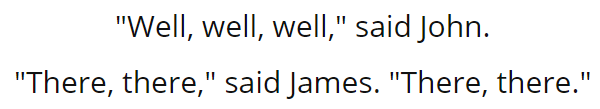
POS tagging consists of assigning a category tag to the tokenized parts of a sentence. The most popular POS tagging would be identifying words as nouns, verbs, adjectives, etc.

Statistical Language Modeling

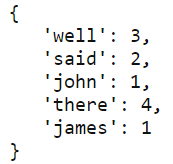
Statistical Language Modeling is the process of building a statistical language model which is meant to provide an estimate of a natural language. For a sequence of input words, the model would assign a probability to the entire sequence, which contributes to the estimated likelihood of various possible sequences. This can be especially useful for NLP applications which generate text.

## BAG OF WORDS

Bag of words is a particular representation model used to simplify the contents of a selection of text. The bag of words model omits grammar and word order, but is interested in the number of occurrences of words within the text.



The resulting bag of words representation as a dictionary:



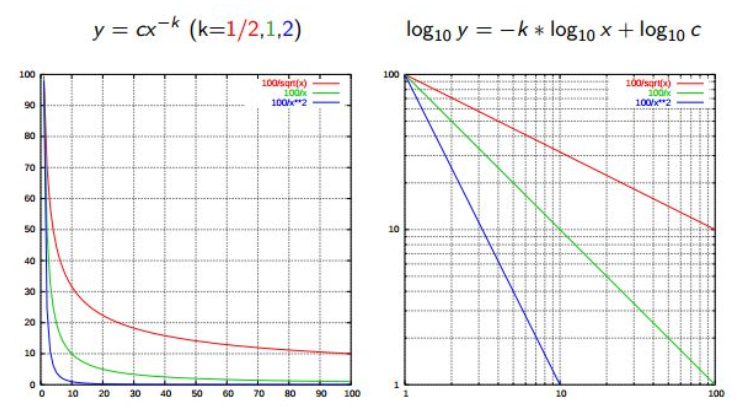
## N-GRAMS

n-grams is another representation model for simplifying text selection contents. As opposed to the order less representation of bag of words, n-grams modeling is interested in preserving contiguous sequences of N items from the text selection.

An example of trigram (3-gram) model of the second sentence of the above example ("There, there," said James. "There, there.") appears as a list representation below:

## ZIPF'S LAW

Zipf's law states that given some corpus of natural language utterances, the frequency of any word is inversely proportional to its rank in the frequency table. Thus, the most frequent word will occur approximately twice as often as the second most frequent word, three times as often as the third most frequent word, etc.



## SIMILARITY MEASURES

There are numerous similarity measures which can be applied to NLP. What are we measuring the similarity of? Generally, strings.

* **Levenshtein** - the number of characters that must be deleted, inserted, or substituted in order to make a pair of strings equal.
* **Jaccard** - the measure of overlap between 2 sets; in the case of NLP, generally, documents are sets of words.
* **Smith Waterman** - similar to Levenshtein, but with costs assigned to substitution, insertion, and deletion.

## SYNTACTIC ANALYSIS

Also referred to as parsing, syntactic analysis is the task of analyzing strings as symbols, and ensuring their conformance to an established set of grammatical rules. This step must, out of necessity, come before any further analysis which attempts to extract insight from text -- semantic, sentiment, etc. -- treating it as something beyond symbols.

## SEMANTIC ANALYSIS

Also known as meaning generation, semantic analysis is interested in determining the meaning of text selections (either character or word sequences). After an input selection of text is read and parsed (analyzed syntactically), the text selection can then be interpreted for meaning.

Simply put, syntactic analysis is concerned with what words a text selection was made up of, while semantic analysis wants to know what the collection of words actually means. The topic of semantic analysis is both broad and deep, with a wide variety of tools and techniques at the researcher's disposal.

## SENTIMENT ANALYSIS

Sentiment analysis is the process of evaluating and determining the sentiment captured in a selection of text, with sentiment defined as feeling or emotion. This sentiment can be simply positive (happy), negative (sad or angry), or neutral, or can be some more precise measurement along a scale, with neutral in the middle, and positive and negative increasing in either direction.

## INFORMATION RETRIEVAL

Information retrieval is the process of accessing and retrieving the most appropriate information from text based on a particular query, using context-based indexing or metadata. One of the most famous examples of information retrieval would be Google Search.