**NLP 5**

It is the part 5 article of NLP series, if you are new kindly refer to the previous articles here.

In this article, we will learn the following topics.

* Part of speech Tagging
  + Nouns
  + Noun Phrases
  + Dependency Parsing
* Topic Modelling
  + Latent Dirchlet Allocation Model
* Amazon reviews Case Study

Libraries used: spacy, genism

**NLP 1**

In this series of articles, we go through various techniques and model used in Natural Language Processing.

Series covers the following topics:

1. Text Preprocessing
2. Text Representation
3. Word Embedding
4. Language Modeling
5. Topic Modeling

Pre requisite for this article

Python basics

In this article, we learn the real time problem statements in NLP. The following topics are covered in this article.

* What is Natural Language Processing (NLP)
* Need of NLP
* Text Pre-processing
* Case Study: Tweet Sentiment Classification

**What is NLP?**

**An area of computer science that deals with methods to analyze, model, and understand human language. Every intelligent application involving human language has some NLP behind it.**

**Need of NLP?**

**Applications of NLP:**

**Language modelling:**

**It is the task of predicting what the next word in a sentence will be based on the history of previous words. The goal of this task is to learn the probability of a sequence of words appearing in a given language. Language modelling is useful for building solutions for a wide variety of problems, such as**

* **Spelling correction**
* **Speech recognition**
* **Machine translation.**

**Text classification:**

**This is the task of bucketing the text into a known set of categories based on its content**

* **Email spam identification**
* **Sentiment analysis.**

**Information extraction:**

**It is the task of extracting relevant information from text, such as calendar events from emails or the names of people mentioned in a social media post.**

* **Gmail extracts destination from tickets whenever we are traveling.**

**Information retrieval:**

**It is the task of finding documents relevant to a user query from a large collection.**

* **Semantic search engine**

**Question answering:**

**It is the task of building dialogue systems that can converse in human languages.**

* **Chat bots**
* **Voice Assistant**

**Text summarization:**

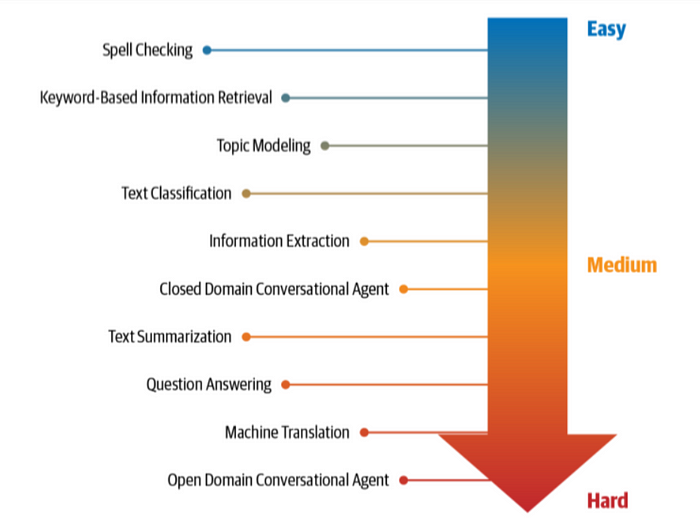
**It aims to create short summaries of longer documents while retaining the core content and preserving the overall meaning of the text.**

* **Inshorts**
* **Creating automated abstracts**

**Topic modeling:**

**This is the task of uncovering the topical structure of a large collection of documents. Topic modeling is a common text-mining tool and is used in a wide range of domains, from literature to bioinformatics.**

**Difficulty level of NLP Applications:**



**Why NLP is Challenging ?**

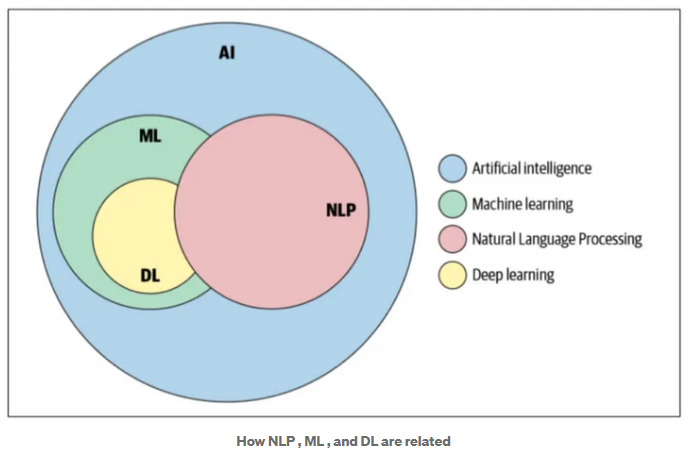
**What makes NLP a challenging problem domain ? The ambiguity and creativity of human language are just two of the characteristics that make NLP a demanding area to work in .**

**Natural Language Processing (NLP) is challenging for several reasons, including:**

* **Ambiguity: Natural language is highly ambiguous, with many words and phrases having multiple possible meanings, depending on context. For example, the word “bank” could refer to a financial institution or the edge of a river.**
* **Complexity: Language is complex and contains many subtle nuances that are difficult to capture in a set of rules or algorithms. For example, idioms, sarcasm, and metaphors can be challenging to interpret for both humans and machines.**
* **Data availability: NLP systems rely heavily on large amounts of data to learn patterns and make predictions. However, obtaining high-quality labeled data can be expensive and time-consuming.**
* **Ethical considerations: NLP technologies can have significant societal impacts, and ethical considerations such as bias, privacy, and accountability must be carefully considered and addressed.**

While humans use intuition to understand language, NLP models don't have intuition in the human sense. They learn by recognizing patterns in the data they are trained on.

NLP algorithms incorporates AI and its subsets to solve a problem.



The different approaches used to solve NLP problems commonly fall into following three categories

* **Heuristics-Based:** Creating a set of rules to analyze the text based on patterns, syntax, and grammar. Heuristics-Based NLP systems are designed to make decisions and solve problems using a set of pre-defined rules
  + Regular expressions (regex) is a great tool for text analysis and building rule-based systems. A regex is a set of characters or a pattern that is used to match and find substrings in text. Regexes are a great way to incorporate domain knowledge in your NLP system.
* **Machine Learning based:** Machine learning techniques are applied to textual data just as they’re used on other forms of data, such as images, speech, and structured data. Supervised ML techniques such as classification and regression methods are heavily used for various NLP tasks.
  + Supervised machine learning algorithms can be used for classification tasks that would classify news articles into a set of news topics like sports or politics.
  + Similarly, unsupervised clustering algorithms can be used to club together text documents.
* **Deep Learning based:** This approach involves training neural networks with multiple layers to process language. The advantage of this approach is that it can learn complex relationships between words and their meanings, but the downside is that it requires even more computing resources than traditional machine learning.
  + A neural network to generate text or to classify text.

**Text Pre-processing:**

**Stop words:**

Stop words are common words that frequently appear in a language but often carry little meaningful information in the context of natural language processing (NLP) tasks.

In English, stop words include articles ("the," "a," "an"), prepositions ("in," "on," "at"), conjunctions ("and," "or," "but"), pronouns ("he," "she," "it"), and auxiliary verbs ("is," "am," "are"). Despite their high frequency, these words usually do not contribute significantly to the understanding of the main content of the text.

The removal of stop words helps reduce the dimensionality of the text data, making it easier to process and analyse. However the decision to remove stop words can be context-dependent.

Overall, handling stopwords is a fundamental step in text pre-processing.

**Contractions:**

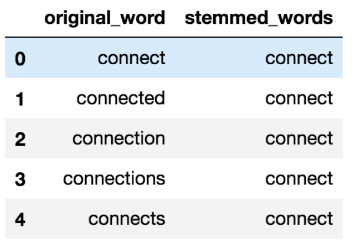
Contractions are shortened forms of words or phrases created by omitting certain letters and sounds, often replacing them with an apostrophe.

Examples:

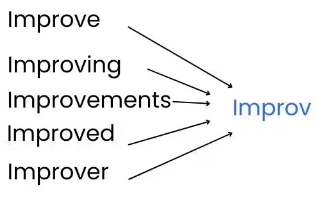
* "I'm" for "I am"
* "They’ve" for "They have"
* "don't" for "do not"
* "we'd" for "we would"

**Stemming:**

Stemming is applied to reduce words to their root or base form. This process helps in treating different morphological variants of a word as a single entity, thereby simplifying the text and reducing dimensionality.

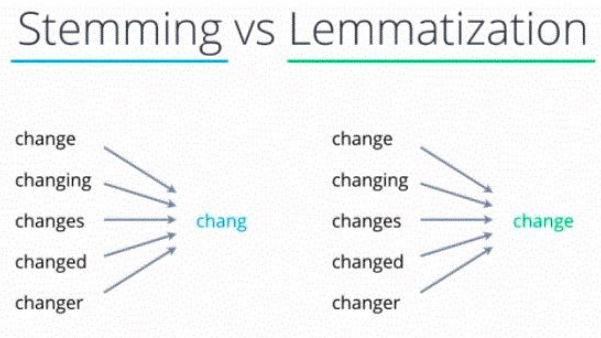


However, sometimes stemming can be too aggressive, reducing words to stems that are not meaningful.



**Lemmatization:**

**Lemmatization** reduces words to their base or dictionary form, known as a **lemma**. Unlike stemming, which often reduces words to their root forms through a set of rules, lemmatization considers the context and meaning of the word, ensuring that the base form (lemma) is a valid word in the language.



**Context specific pre-processing:**

**In NLP algorithms, most of the context specific text pre-processing is handled with regular expressions**

* **Handling Hyperlinks**
* **Handling Punctuations**
* **Handling alpha numeric words**

**How can a machine understand a language?**

* Yes by recognizing the words that constitute a string of characters before processing a natural language.
* Meaning of the text could be interpreted by analyzing the words.

**This is where tokenization is essential to proceed with NLP (text data), which is used to break down the text into tokens (words).**

**Tokenization:**

In many alphabetic writing systems, words are separated by whitespace. However, even in a well-formed corpus, tokenization presents several challenges. Ambiguities often arise from the use of punctuation marks, such as periods, commas, quotation marks, apostrophes, and hyphens, since a single punctuation mark can serve multiple functions within a sentence.

**Example:**

**Clairson International Corp. said it expects to report a net loss for its second quarter ended March 26 and doesn’t expect to meet analysts’ profit estimates of $3.9 to $4 million, or 76 cents a share to 79 cents a share, for its year ending Sept. 24.**

This sentence has several items of interest that are common for alphabetic and space-delimited languages.

* First, it uses periods (.) in three different ways :
  + Within numbers as a decimal point ($3.9)
  + To mark abbreviations (Corp. and Sept.)
  + To mark the end of the sentence, in which case the period following the number 24 is not a decimal point.
* The sentence uses apostrophes in two ways:
  + To mark the genitive case (analysts’ )
  + To show contractions (doesn’t)

The tokenizer must thus be aware of the uses of punctuation marks and be able to determine when a punctuation mark is part of another token and when it is a separate token.

**word\_tokenize('''Clairson International Corp. said it expects to report a net loss for its second quarter ended March 26 and doesn’t expect to meet analysts’ profit estimates of $3.9 to $4 million, or 76 cents a share to 79 cents a share, for its year ending Sept. 24.''')**

**Sentence Tokenization:**

**Word Tokenization:**

**Case Study: Sentiment Analysis of Tweets**

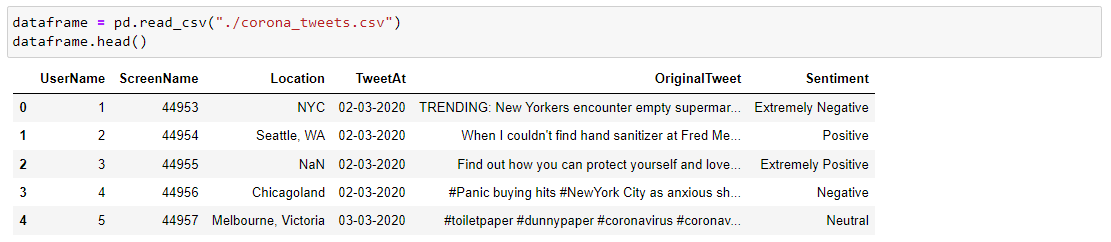
The tweets have been pulled from Twitter during Covid-19 breakout and the tweets are manually labelled. The government wants to learn the public sentiment on the new strain of COVID-19.

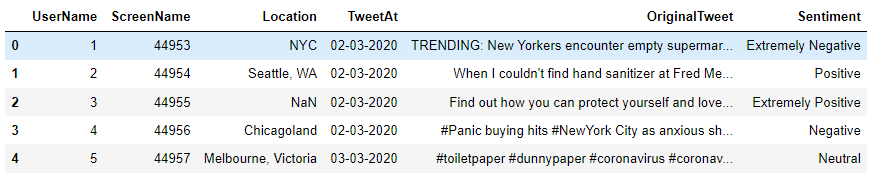
The task is to classify the sentiment of a tweet to either Positive or Negative.

**Understanding the Dataset:**

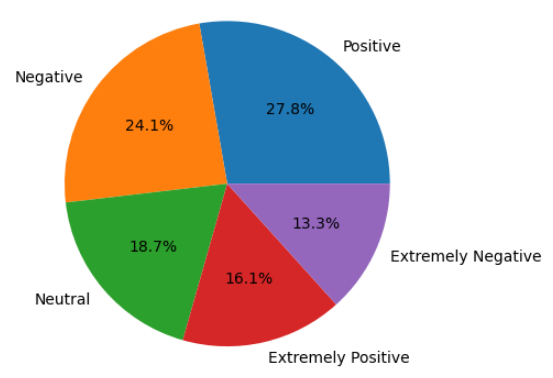
1. Dataset source: kaggle covid tweets [link](https://www.kaggle.com/datatattle/covid-19-nlp-text-classification)
2. Total 11,663 tweets with positive and negative labels

**Data Exploration:**





**Distribution of Labels**



**To simplify the analysis, we extract only “Positive” and “Negative” labelled Reviews**