NLP Introduction and Basics of Text Pre-processing (& extraction)

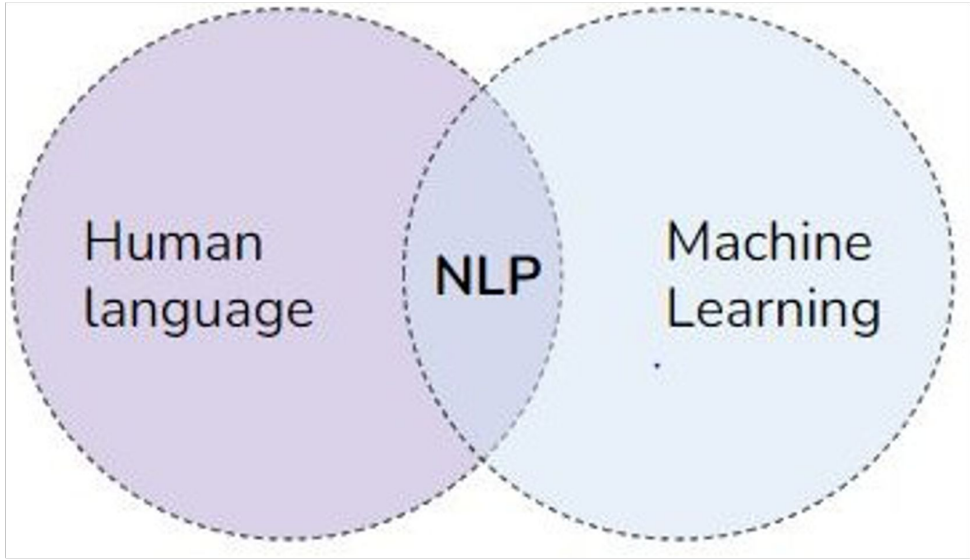
Content

- Introduction to Natural Language Processing
  - Need for NLP
  - Applications of NLP
- Introduction to Text Pre-processing
  - Challenges With Text Data
  - Importance of Text Processing
- Common Text Pre-processing /Cleaning Methods
- Brief Introduction to Web Scraping (Beautiful Soup)
- Name Entity Recognition
- Word Cloud
- NLP Essential Libraries : NLTK/Spacy

What is NLP?

- **Natural Language Processing**, abbreviated as NLP, is a branch of **artificial intelligence (A.I.)** that deals with the interaction between **machines and human languages.**

- The ultimate objective of NLP is to **automate the reading, interpretation and understanding** of **human languages**, also called **natural language**.



The Need for NLP

- Natural language processing holds the potential to **improve how we live and work**. It can help bring **progress** to areas that have been slow or burdensome to change without the partnership between **humans and technology**.

- The potential applications of NLP are incredibly diverse, and could help with nearly any situation involving the need to **rapidly and vigorously analyse unstructured text data.**

- **Text data is different**, because it cannot directly be input into **machine learning** and **deep learning** models like other numerical forms of data. Text data requires a series of pre-processing steps before it can be analysed and mined for insights.

The Need for NLP

- It is worth noting however, that even though text data is unstructured, it is sequential in nature. This is obvious from the fact that changing or reversing the order of characters or words in a sentence, changes its meaning or renders it meaningless. For example, consider the sentence:

"The cat and the dog sat on the wall."

Reversing the sentence to "wall the on sat dog the and cat the" or just swapping two words to form "The cat and the wall sat on the dog" can entirely change the meaning of the sentence.

Source:
Pixabay

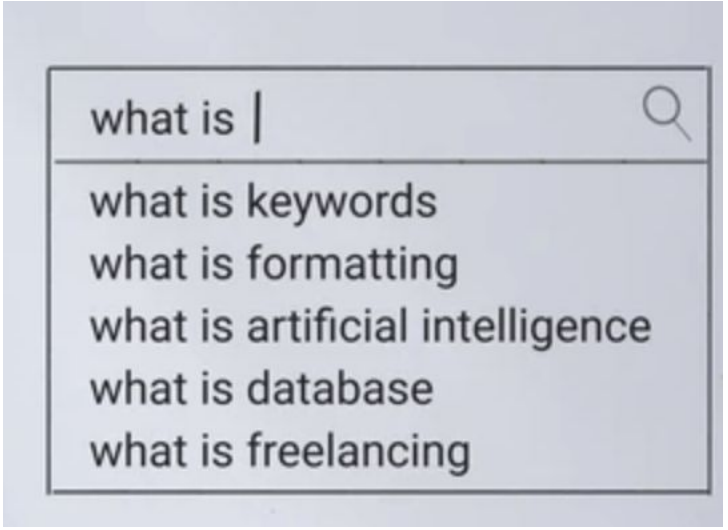
Applications of NLP

Applications of NLP : Auto-suggest/complete/ correct

- An important recent addition to search engine functionality has been that whenever one searches for anything on the search bar, it returns suggestions for the possible search keywords after you type a few characters. This is sometimes referred to as Autocomplete.

- If the search query has spelling mistakes, search engines also recommend a corrected version while still returning relevant results. This is also called Autocorrect, although that term is also used for automatic spelling corrections when typing on devices like computers and smartphones.

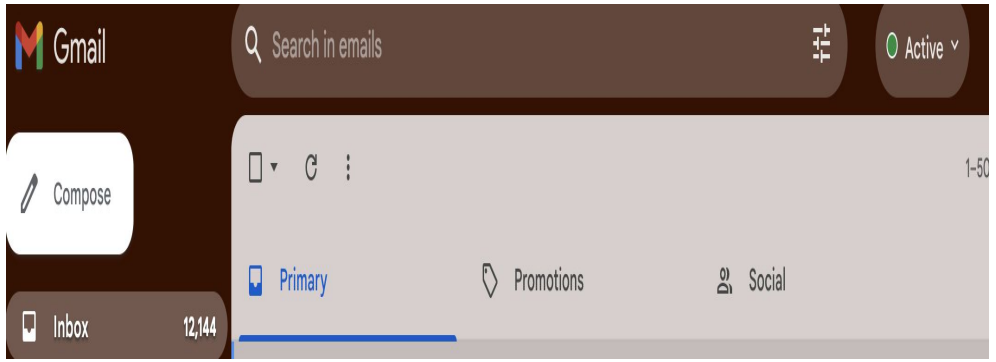
- All of the above functionalities related to search engines leverage the power of NLP in order to automate browsing tasks and make our lives easier.



Applications of NLP :Automatic Email Filtering

- You may have observed that on certain email service providers such as Gmail, when an email arrives, it is sometimes already assigned a category: primary, social, or promotional.

- This is possible due to the NLP technique of **Text Classification**. In this use case, the contents of the email are automatically analyzed through an NLP machine learning algorithm to identify patterns, which in turn allows for the email to be classified into the relevant label.



Applications of NLP :Language Translation

- Due to the nuances of word usage and sentence construction that widely vary among languages around the world, it is not feasible to manually translate text word-by-word from one language to another.

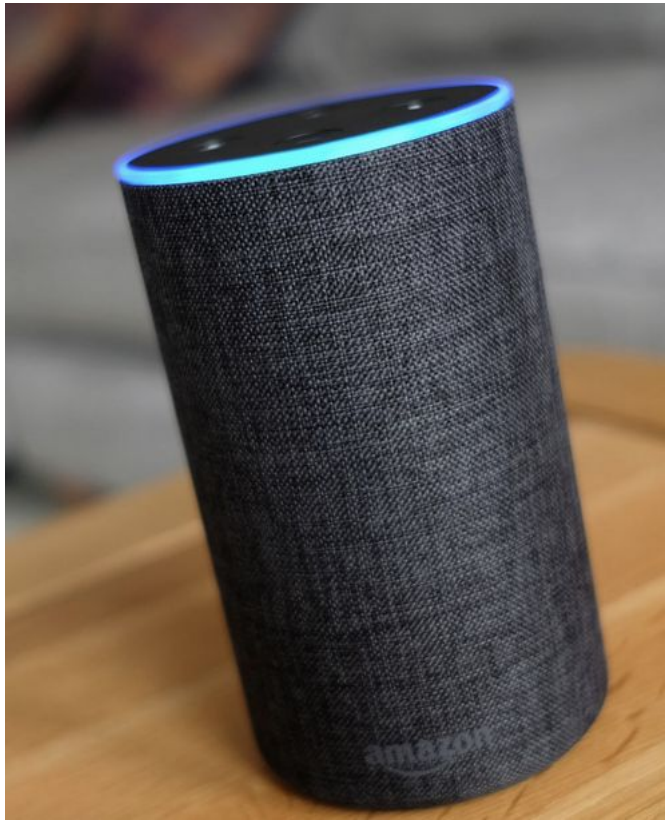
- However, the latest advances in NLP have been leveraged by products like **Google Translate**, in order to understand the intricacies of **meaningful translation** from one language to another.

Applications of NLP :Optical Character Recognition (OCR)

- Optical Character Recognition (OCR) is the general name given to the technique of **converting images of handwritten, typed, or printed text into machine-encoded language.** It is one of the most prevalent methods for **digitizing printed documents** so that they may be saved, modified, and searched electronically.
- Although predominantly a Computer Vision use case, **OCR requires techniques from NLP to convert image encodings into natural language.** OCR provides a variety of benefits that could assist organizations with any document digitization activities such as reading / recognizing ID cards or passports, cheques, receipts and other financial documentation.

Applications of NLP :Voice Assistants

- A voice assistant is software that understands a user's spoken requests to perform actions based on **speech recognition, natural language comprehension, and natural language processing.**
- It can efficiently interpret spoken words, voice modulations, and accents in order to convert speech into text. In the form of a responsive two-way voice assistant, it would also convert text into speech, to provide replies to a user in a voice and intonation that tries to mimic human conversations for a personalized experience.
- Amazon’s **Alexa** and Apple’s **Siri** are two popular examples of voice assistants in the world today.



Applications of NLP : Customer analytics in E-commerce platforms

- Unstructured text data can be a wealth for **marketing intelligence**. With the ability to quickly scan such datasets and **find hidden patterns in customer behavio**r, decision-makers learn what products or services are most compelling for their specific market.
- This has important applications for **product development**, as well as digging out what marketing initiatives are most worthwhile.
- They also use **chatbots** to solve the basic queries of the customers. Now, chatbots have evolved and have become a virtual companions to the shoppers and have the ability to **recommend products** and also **get data** about the likes and dislikes of a customer.

Applications of NLP : Social media

- Social media uses NLP in various ways including understanding the likes and dislikes of a user for **targeted advertising**. The posts, comments, reviews and other data are analyzed to understand the personality and interests of the user.
- Social media is a free platform and thus, a major hub for the flow of fake news and hate speech. Though restrictions cannot be imposed, these platforms **flag posts and comments** which might be false or offensive.

Applications of NLP : Customer analytics

- One use case for unstructured text data is **customer analytics**. When organizations are able to integrate unstructured data from a variety of sources such as call center transcripts, online reviews of products, chatbot conversations with customers, social media mentions, and use **artificial intelligence** to extract patterns in the information from these sources, they have the information available to make swift decisions that can **improve customer relationships**.

Introduction to Text Pre-processing

What is Text Pre-processing ?

- The process of transforming the raw and uncleaned text data into a form that is analysable for the model is known as Text pre-processing.

- Genrally, raw text data coming from various sources is not entirely clean.

- For every NLP task, a lot of pre-processing goes on behind the scenes so that a model can understand and interpret text data. If we train our model on unstructured/uncleaned data that hasn't been pre-processed, the model can miss out on learning important information

Need for Text Pre-Processing : Toy Example

- Consider the below example of text data scraped from a website

<html> This is the first header line in the newspaper.
 It is not highlighted.</html>

- In order to display certain text on a website, they are written in a specific format like the above.

- In the above sentence the tags like ‘<html>’ , ‘
’ or the punctuators like ‘,’ are not required to understand the context of the sentence. Therefore, these tags or punctuators or special symbols can be considered as noise in the input data.

- Before such input can be passed to a model, it needs some clean-up so that the model can focus on important words instead of words like ‘<html>’ which adds minimal to no value.

Challenges with Text Data

- There are multiple challenges with text data such as **misspellings**, **foreign characters**, **out-of-vocabulary words**, **acronyms**, **punctuation**, words used for syntactical requirements that do not add any additional information to the text.
- For example, social media sites generate large amounts of sentiment-rich textual data that is inherently unstructured and contains irrelevant information like emoticons, over-usage of punctuations, etc. Given is as an example of text data from social media.
- The English used in social media channels differs significantly from conventional English.
- In social media, terms that are not part of the standard English language are frequently used, including acronyms, accented letters, hashtags, and misspelt words.

Importance of Text Processing

- Reduces noise of the data. Noisy data is one where there are unexplained variations in word usage under similar conditions. For example,
  - Hey! I like the movie a lot!.
  - I liked the movie a lot.
  - I like that movie a lot
- However, test processing tasks varies from text to text i.e use case

Text Pre-processing/Cleaning Methods

- Text Standardization/Normalization
  - Case Conversion
  - Spelling Correction*
  - Stemming
  - Parts Of Speech(POS) tagging
  - Lemmatization
- Eliminate Unessential Items from Text*
  - Stopwords
  - Accented Characters
  - Spaces, Digits, Punctuations, URLs
- Text Extraction

*Refer notebook

Text standardization/Normalization

- Process of converting text data into a consistent and standardized format
- Involves several techniques and tasks aimed at ensuring that text is uniform and can be processed, analyzed, and understood more easily by machines and humans

Case Conversion/Lowercasing

- **Lowercasing converts all the words into lowercase to ensure that repeated occurrences of the same word in different cases are still treated as the same word.**
- **The word "text" is written in two different cases in the example below, giving the model redundant information. This can be avoided by applying lowercasing.**

- **First letter is capitalized**

- **Using Natural Language Processing, we make use of the Text data available across the internet to generate insights for the business. To make this huge amount of data usable for a Natural Language Processing task,**
- **we use text preprocessing.**

- **No capitalization**

Base & Derived Words

- Words can be classified into Base word and Derived word.

- For example,
  - 'go' is a base word and 'going', 'gone' and 'went' are its derived words. During data cleaning phase derived word shall be converted to their base counterparts.
  - ‘Likes’ , ‘Liked’, ‘Liking’, ‘Likely’ are derived words of the word Like

Stemming

- Stemming helps us in standardizing the text by extracting the base word from the given words.

- Stemming algorithms are typically rule-based, which means that a word is analyzed and run through a series of conditions to determine how to cut it down to the base form or stem.
- Limitation : In Stemming Base word is identified by chopping the word at end. Forexample 'going' and 'gone' will get converted to 'go' but 'went' will not. Also If two or more words can be incorrectly reduced to a single stem for example university and universe to univers

Different types of Stemmers

- Stemming algorithms, also known as Stemmers, were proposed by various researchers and have been in use since the 1960s.

- The following Stemmers (algorithms) are among the commonly used ones. A comparative table is also shown.
  - Porter Stemmer
  - Snowball Stemmer
  - Lancaster Stemme

Word	Suffix	Lancaster	Snowball	Porter
Connecting	- ing	Connect	Connect	Connect
Connected	- ed	Connect	Connect	Connect
generous	- erous	Gen	Generous	Gener
generously	- erously	Gen	Generous	Gener

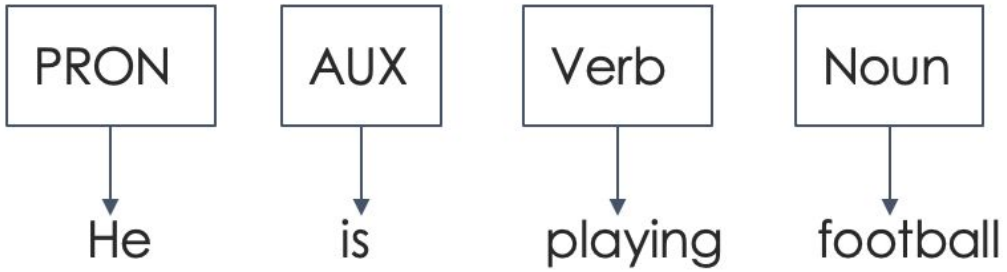
Lemmatization

- Lemmatization follows the same idea as Stemming, which is to reduce word inflections, in turn reducing the redundancy in the text to produce a more standardized result.
- Although the broad idea behind both techniques is the same, their implementation differs:
  - Lemmatization attempts to minimize inflections to their canonical or dictionary form , referred to as aLemma.
  - Before reducing a word to its base form, lemmatization also considers the context in which it is used.

- Lemmatization analyzes a word and links it to its lemma using dictionaries. Because the lemma of a word returned always has a dictionary meaning, it is also known as a dictionary-based approach.
- Programmatically, , lemmatization requires Parts of speech tagging and has high compute requirement.

POS Tagging

- **Part-of-speech (POS) tagging** is a popular Natural Language Processing process which refers to categorizing words in a corpus in correspondence with a particular **part of speech**, depending on the definition of the word and its context.
- Allows understanding of language structure and syntax.
- These properties can used to extract information by using language rules.
- Multiple NLP libraries support POS tagging e.g. NLTK, spaCy



Examples of tags

- Nouns
  - Singular noun, NN (Cat)
  - Plural noun, NNS (Cats)
  - Proper noun, NNP (Garfield)
  - Personal pronouns, PRP (He)
- Verb
  - Base verb, VB (sleep)
  - Gerund, VBG (sleeping)
- Preposition, IN (over)
- Adjective
  - Basic, JJ (bad)
  - Comparative, JJR (worse)
- Adverb
  - Basic, RB (quickly)
- Determiner
  - Basic, DT (a, an, the)
  - WH, WDT (which, who)
- Coordinating conjunction, CC (and, but)

Some PoS Tagging Challenges

- Ambiguity that needs *context*
  - It is a quick **read** (NN)
  - I like to **read** (VB)
- Differences in numbers of tags
  - Brown has 87 tags
  - British National Corpus has 61 tags
  - Penn Treebank has 45 tags (several merged)

Approaches to PoS Tagging

- Learn from corpora
- Use regular expressions
  - Words ending with '*ed*' or '*ing*' are likely to be of a certain kind
- Use context
  - POS of preceding words and grammar structure
  - For example, n-gram approaches
- Map untagged words using an embedding
- Use recurrent neural networks

Stop words

- stop words are a set of commonly used intervening words in a language. Some examples of stop words in English are “a”, “the”, “is”,“are” etc.
- The idea behind removing stop words is that by eliminating low-information parts of the text, we can concentrate on the key words.
- Note however that the incorrect removal of stop words can even alter the meaning of our text. for example ‘The movie was **not** bad’

Accented and special characters

- Following is an example of text data with accented and special characters.

Using Nātùrăl Làñgûägè Processing, we make use of the text data available across the internet to generate insights for the business. To make this huge amount of #$ ^_^ $# data usable for a Natural Language Processing task, we use text preprocessing.

- we usually remove accented characters. If we didn't do so, our model will consider words with and without an accent symbol as separate words, even though they may be the same word

- special characters add to the noise in unstructured text, and add no value to the meaning of the text. Removing them is usually preferable.

Web Scrapping

- Web scraping is the process of collecting structured web data in an automated fashion. It’s also called web data extraction.
- Few Python Packages for Web Scraping
  - Beautiful Soup --
  - Scrapy (Less for Web Scrapping ---more for building web spider for web crawler)
  - XTML
  - Specific API packages : newsapi , tweepy

Beautiful Soup : Typical web scrapping Approach

i. Extract the web page HTML content & convert it to text to view the Elements of the HTML page

ii. Convert the HTML content into XML object

iii. `prettify` helps to have a look proper intendent look of the XML page

iv. `title` helps to extract the title of the web page

v. `string` helps to convert the tags into string

vi. We can fetch content from first mention of any tags by having the tag extension with the XML content

vii. `find_all` helps to fetch content from all the mentioned of any tags by having the tag extension with the XML content

viii. We can also fetch information for a tag with a specific class type.

ix. Fetch and arrange the information following the above steps.

Name Entity Recognition

- Named Entity Recognition (NER) is the process of identifying and categorizing named entities in given text. Examples of categories could be Name, Organization, Location, Money, Time etc.
- There are three major steps to Named Entity Recognition:
  a) Text is read by an NER System
  b) Named Entities are identified from a corpus of unstructured text
  c) Classification and category creation is being done - ex: person, organization etc.

Approaches to NER

- Match to an NE in a tagged corpus
  - Fast, but cannot deal with ambiguities
- Rule based
  - E.g. capitalization of first letter
  - Does not always work, especially between different types of proper nouns
- Recurrent neural network based
  - Learn from a NE tagged corpus

Some challenges with NER

- Different entities sharing the same name
  - Manish *Jindal* , Person
  - *Jindal* Steel ,Thing (company)
- Common words that are also names
  - Do you want it with *curry* or dry
  - Tyler *Curry*
- Ambiguity in the order, abbreviation, style
  - Jindal, Manish
  - Dept. of Electrical Engineering
  - De Marzo, DeMarzo

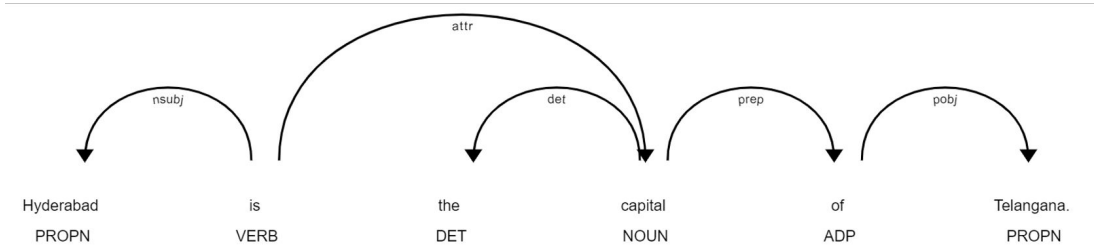
Some challenges with NER

- Different entities sharing the same name
  - Manish *Jindal* □ Person
  - *Jindal* Steel □ Thing (company)
- Common words that are also names
  - Do you want it with *curry* or dry
  - Tyler *Curry*
- Ambiguity in the order, abbreviation, style
  - Jindal, Manish
  - Dept. of Electrical Engineering
  - De Marzo, DeMarzo

Dependency Parsing

1. Shows how words in a sentence relate to each other.

2. Allows further understanding of language structure and syntax.



NLTK (Natural language Toolkit)

- Natural Language Toolkit (NLTK) was developed in 2001 by Steven Bird and Edward Loper at the University of Pennsylvania’s Department of Computer and Information Science.
- The developers' goal was to create long-lasting software that would facilitate easy learning of NLP, and as a result, they created one of the most widely used NLP packages in the world today. As NLTK allows for a wide range of experiments to be performed on text data, it also receives contributions from researchers all over the world. However, many of the functionalities provided by NLTK are currently limited to English only.
- NLTK is most commonly used for educational purposes and supports a wide range of tasks such as:
    - Tokenization
    - Stemming
    - Stopword removal
    - Part of Speech Tagging, etc.

spaCy

- spaCy enables you to create applications that handle and "understand" massive amounts of text because it is made primarily for usage in production environments. It can be used to create systems for information extraction, NLU, or text preprocessing.

- Modern NLP practitioners prefer spaCy because of its industry strength and production ready packages.

- It can perform similar tasks to the NLTK library such as text classification and tokenization, but it also provides pre-trained language models and pipelines that users can customize. These features are also not limited to the English language, increasing spaCy's popularity worldwide.

NLTK vs spaCy

NLTK	Spacy
NLTK is suitable for research and educational purposes, **and is hence preferred by researchers.**	spaCy is preferred by NLP practitioners **because of its ability to work with large-scale data and production-ready packages.**
Many operations/features of NLTK are limited only to the English language.	**spaCy provides many pre-trained models and pipelines that can be customized by users** for multiple languages and not just English.
NLTK even provides support for Natural Language Generation **tasks like Translation and Chatbots.**	spaCy doesn't support **tasks in the domain of** Natural Language Generation.
The NLTK package requires more storage space, but is less memory intensive.	**spaCy offers different package sizes (small, medium, and large)** and is more memory-intensive in comparison to NLTK.