Natural Language Processing

An Introduction and Preprocessing using NLTK

Nitin Mahajan Feb 18 · 6 min read ★

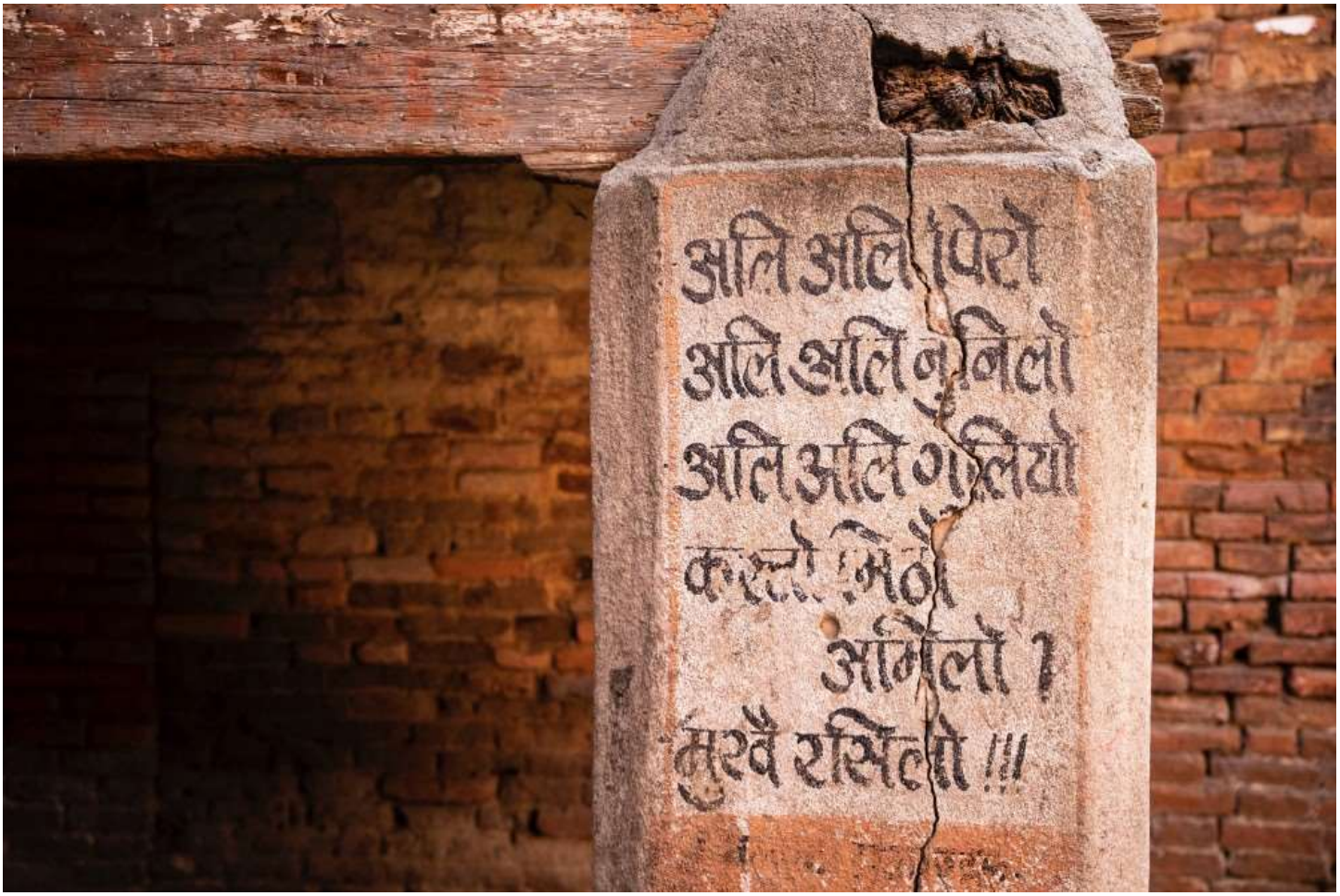


Photo by Kerensa Pickett on Unsplash

What is NLP?

Natural Language Processing (NLP) is the technology used to help machines to understand and learn text and language. With NLP data scientists aim to teach machines to understand what is said and written to make sense of the human language. It is used to apply **machine learning** algorithms to **text** and **speech**.

> *In this article we will walk through how NLP jobs are carried out for understanding human language using machine learning.*

IS IT IMPORTANT TO US?

We can use NLP to create systems like **speech recognition**, **machine translation**, **spam detection, text simplifications**, **question answering, autocomplete, predictive typing, sentiment analysis, document summarization** and **many more**. Few examples include:

Google Translate — Language translation applications

Grammarly — Employ NLP to check grammatical accuracy of texts

Personal assistance — Applications such as OK Google, Siri, Cortana, and Alexa.

Keyword search, spelling check, synonyms search.

Queries like product price, location, company names etc. on search engines.

IS THIS AN EASY TASK?

> **NLP, Emotions *and Ambiguity:*** *Teaching computers about exact sense and emotions of the language considered as a difficult problem in computer science.*

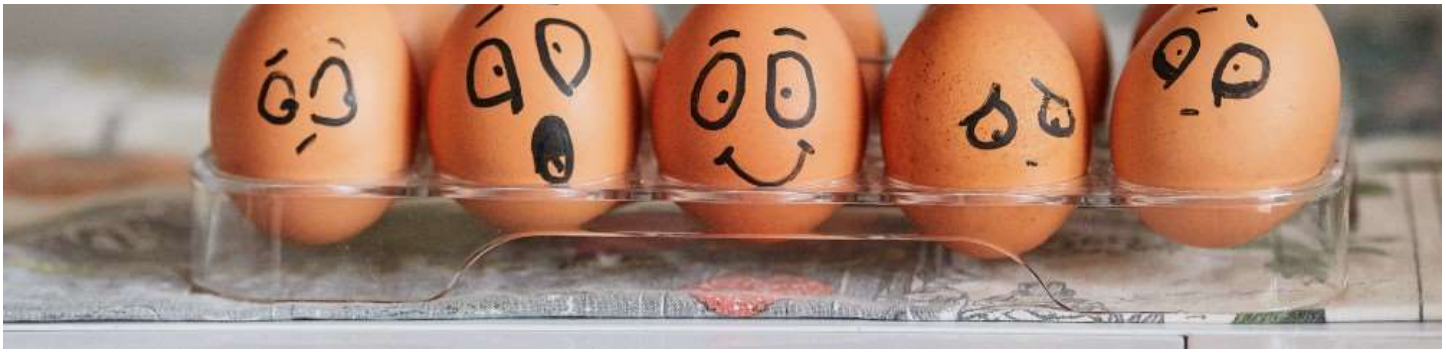


Photo by Tengyart on Unsplash

Fully understanding and representing the meaning of language is an extremely difficult goal. One of the major challenges to developing NLP applications is computers most likely need structured data, but as far as human speech is concerned it is unstructured and often ambiguous.

The rules that direct the information using natural languages are not easy for computers to understand and translate especially to sense the tone for example if someone uses a sarcastic remark to pass information. This means computers have to comprehensively understand the meaning of words as well as the intention or emotion behind the words. Unlike programming human languages are ambiguous which make them complex and hard to learn.

What are the techniques used in NLP?

NLP has primarily two aspects: natural language understanding (NLU) or natural language interpretation (NLI) (i.e. human to machine) and natural language generation (NLG) (i.e. machine to human). In simple words, one can say that NLG is inverse of NLU (broadly called as NLP). Natural language generation (NLG) is when software automatically transforms data into written narrative.

SYNTACTIC & SEMANTIC ANALYSIS

Natural Language Processing tasks are primarily achieved by syntactic analysis and semantic analysis.

> *The term* **syntax** *refers the grammatical structure of the text, whereas* **semantics** *refers to the meaning of the sentence. A sentence that is syntactically correct does not mean to be always semantically correct.*

Syntactic analysis (syntax analysis or parsing), analyzes natural language using formal grammar. Grammatical rules are applied to categories and groups of words, but not to the individual words.

Examples:

- *Parsing* — Involves undertaking grammatical analysis for the provided sentence.
- *Lemmatization* — reduces various inflected forms of a word into a single form.
- *Stemming* — Cuts the inflected words to their root form.
- *Word segmentation* — divides a large piece of continuous text into distinct units.
- *Morphological segmentation* — divides words into individual units.
- *Part-of-speech tagging* — identify the part of speech for every word.
- *Sentence breaking* — Places sentence boundaries on a large piece of text

Semantic analysis is the process of understanding and interpreting the words, signs, tone and structure of the sentence. This task analyzes the meaning or logic behind the sentence. We understand what someone has said by primarily relying on our intuition, knowledge about language and tone.

Examples:

- *Named entity recognition (NER)* — determine the parts of a text that can be identified and categorized into preset groups, like names of people and objects.
- *Word sense disambiguation* — give meaning to words based on their context
- *Natural language generation (NLG):*— It involves using databases to derive semantic intentions and convert them into human language.

In recent years there has been a surge in unstructured data in the form of text, videos, audio and photos. NLU aids in extracting valuable information from text such as social media data, customer surveys, and complaints.

Reading text data & why do we need to clean the text?

Text data can be in a structured or unstructured format.

> *A **structured** format has a well-defined pattern whereas **unstructured** data has no proper structure. In between the 2 structures, we have a **semi-structured** format which is a comparably better structured than unstructured format.*

In python, we have several libraries to work with text.

- *Scikit-learn, Keras, TensorFlow* — has some text processing capabilities
- *NLTK* — Natural language toolkit.
- *SpaCy* — is an industrial strength NLP package with many practical tools in a nice API.
- *Other libraries* — TextBlob, gensim, Stanford CoreNLP, OpenNLP.

Here in this article, we will walk through various methods and techniques to preprocess the text data using the most commonly used python library i.e. NLTK.

Natural Language Toolkit (NLTK)

(Python Library for Text Processing)

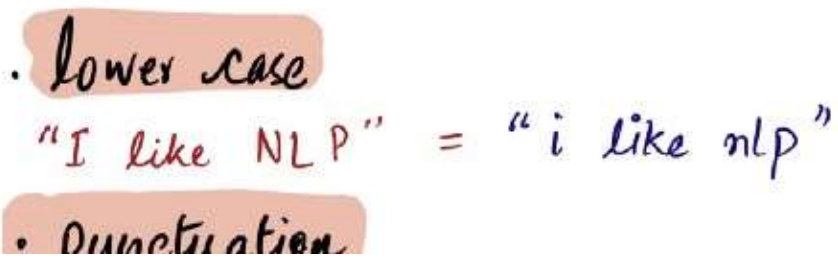
Natural Language Toolkit (NLTK) is a known open-source package in Python which allows us to run all common NLP tasks. It provides easy-to-use interfaces and a suite of **text processing libraries** valorous steps involved during preprocessing like classification, tokenization, stemming, tagging, parsing, and semantic reasoning.

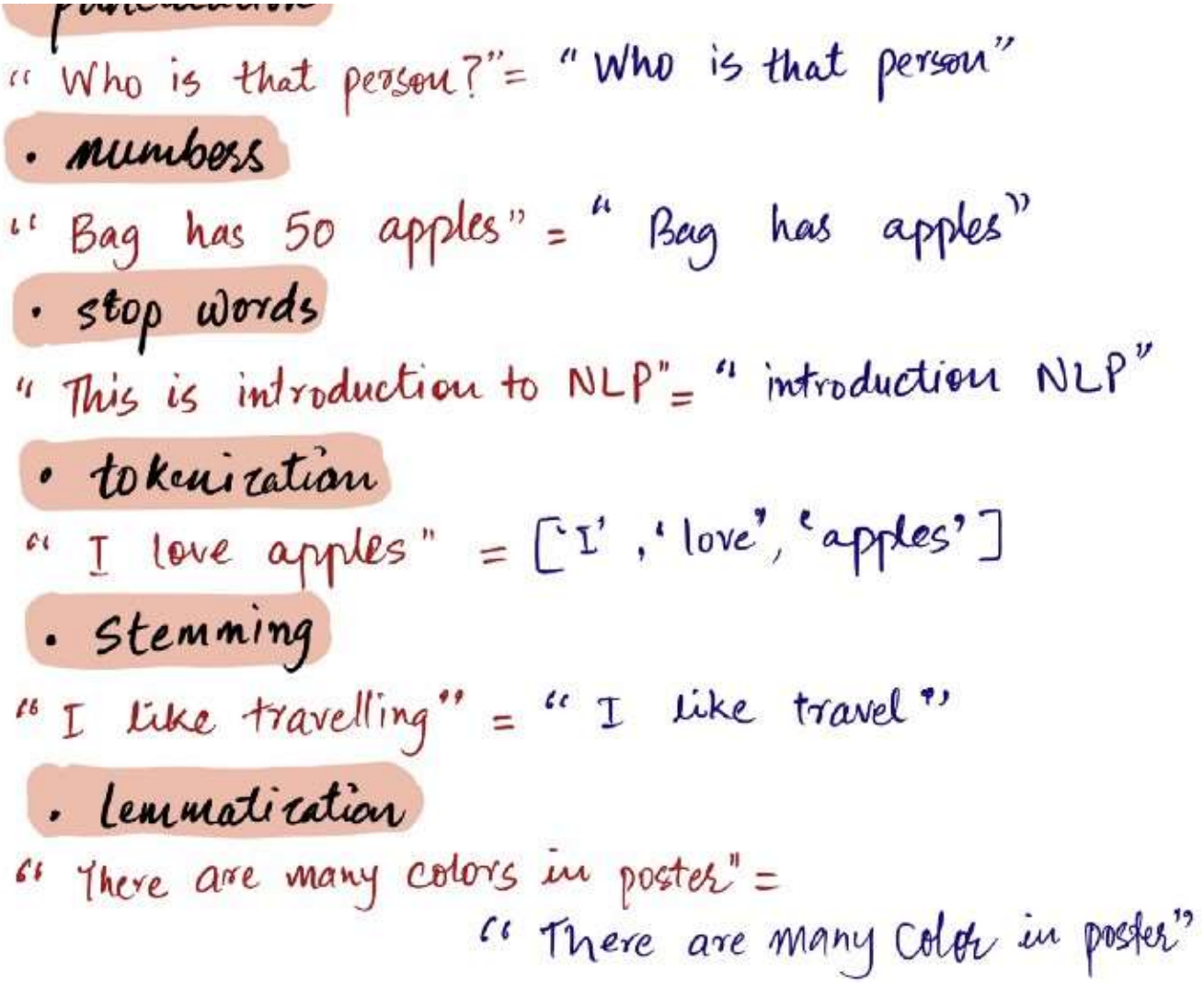
Need for preprocessing the text data

- *Real data* — often incomplete, inconsistent, and filled with a lot of noise and errors
- *Data Structure* — Approx. 90% of the data is unstructured
- *Variety* — Text can come from a list of individual words, sentences, multiple paragraphs (with and without correct spellings and punctuation)
- *Noise* — And this data is never clean and consists of a lot of noise.

> *It is a necessary to preprocess the data before building models and analysis. In simple words preprocessing transforms raw text data into an understandable format for the computer.*

PREPROCESSING OF TEXT





Let's create a **data frame (df)** to practice the preprocessing of text data.

```
text = [
    'This article was published in 2002..',
    'NLP stands for Natural Language Processing',
    'NLP is an emerging field in the Data Science!!!',
    'We can preprocess text data using python library - NLTK',
    'It is likely to be useful, to people',
    'here we will discuss 5-7 steps for preprocessing'
]
```

```
import pandas as pd
df = pd.DataFrame({'Message': text})
df
```

	Message
0	This article was published in 2002..
1	NLP stands for Natural Language Processing
2	NLP is an emerging field in the Data Science!!!

3	We can preprocess text data using python libra...
4	It is likely to be useful, to people
5	here we will discuss 5-7 steps for preprocessing

- Covert text to **lowercase** to make all the data in uniform format

```
# lowercase #method-1
df = df['Message'].str.lower()
print(df)

# lowercase #method-2
df['Message'] = df['Message'].apply(
  lambda x: " ".join(x.lower() for x in x.split()))
df['Message']
```

```
0                 this article was published in 2002..
1          nlp stands for natural language processing
2      nlp is an emerging field in the data science!!!
3    we can preprocess text data using python libra...
4                 it is likely to be useful, to people
5     here we will discuss 5-7 steps for preprocessing
Name: Message, dtype: object
```

- **Punctuation —** punctuation doesn't add any extra information. This step reduces the size of the data and therefore increase computational efficiency

```
df['Message'] = df['Message'].str.replace('[^\w\s]','')
df['Message']
```

```
0                 this article was published in 2002
1          nlp stands for natural language processing
2        nlp is an emerging field in the data science
3    we can preprocess text data using python libra...
4                 it is likely to be useful to people
5     here we will discuss 57 steps for preprocessing
Name: Message, dtype: object
```

- **Numbers —** converting numbers into words or removing numbers. Remove numbers if they are not relevant or change to words.

```
#removing numbers
df['Message'] = df['Message'].str.replace('\d+', '')
df['Message']
```

```
0                          this article was published in
1            nlp stands for natural language processing
2          nlp is an emerging field in the data science
3     we can preprocess text data using python libra...
4                    it is likely to be useful to people
5         here we will discuss  steps for preprocessing
Name: Message, dtype: object
```

- **Non-significant words "Stop words" —** are very common words that carry no meaning or less meaning compared to other keywords. If we remove the words that are less commonly used, we can focus on the important keywords instead.

```
#Stop words
import nltk
from nltk.corpus import stopwords
stop = stopwords.words('english')
df['Message'] = df['Message'].apply(
    lambda x: " ".join(x for x in x.split() if x not in stop))
df['Message']
```

```
0                                  article published
1            nlp stands natural language processing
2                  nlp emerging field data science
3    preprocess text data using python library nltk
4                              likely useful people
5                       discuss steps preprocessing
Name: Message, dtype: object
```

- **Tokenizing text —** A mandatory step in text preprocessing where text is split into **minimal meaningful units**. It can be for words (word_tokenize) or sentences (sent_tokenize).

```
text = [
    'This article was published in 2002..',
    'NLP stands for Natural Language Processing',
    'NLP is an emerging field in the Data Science!!!',
    'We can preprocess text data using python library - NLTK',
    'It is likely to be useful, to people',
    'here we will discuss 5-7 steps for preprocessing'
]

import pandas as pd
```

```
import pandas as pd
df = pd.DataFrame({'Message': text})


from nltk.tokenize import word_tokenize
df['tokenized_text_w']= df['Message'].apply(word_tokenize)
df['tokenized_text_w']
```

```
0          [This, article, was, published, in, 2002..]
1    [NLP, stands, for, Natural, Language, Processing]
2    [NLP, is, an, emerging, field, in, the, Data, ...
3    [We, can, preprocess, text, data, using, pytho...
4      [It, is, likely, to, be, useful, ,, to, people]
5    [here, we, will, discuss, 5-7, steps, for, pre...
Name: tokenized_text_w, dtype: object
```

- **Stemming —** Stemming is a process of extracting a root word by removing the suffix from a word. In this article, we will be focusing on 3 stemming techniques: *Porter Stemmer, Snowball Stemmer and Lancaster Stemmer*

```
from nltk.stem import PorterStemmer
PorterStemmer = PorterStemmer()
df['Message'][:6].apply(
    lambda x: " ".join([PorterStemmer.stem(word) for word in x.split()]))
```

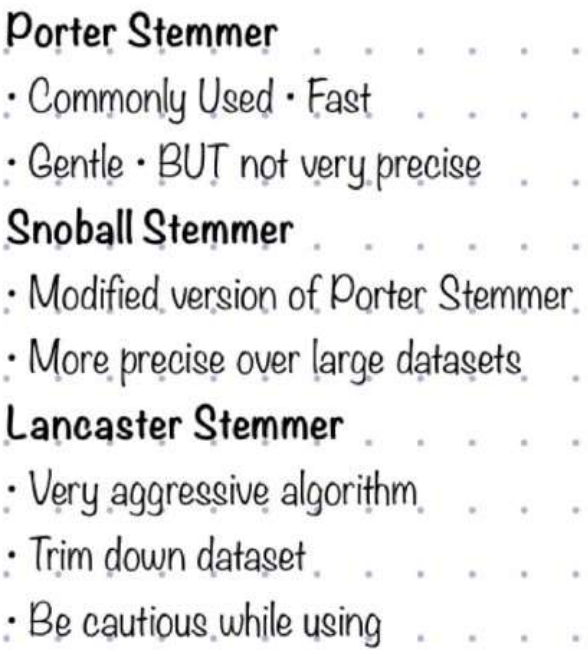
```
0                   thi articl wa publish in 2002..
1                nlp stand for natur languag process
2        nlp is an emerg field in the data science!!!
3    We can preprocess text data use python librari...
4                   It is like to be useful, to peopl
5        here we will discuss 5-7 step for preprocess
Name: Message, dtype: object
```

```
from nltk.stem.snowball import SnowballStemmer
SnowballStemmer = SnowballStemmer("english")
df['Message'][:6].apply(
    lambda x: " ".join([SnowballStemmer.stem(word) for word in x.split()]))
```

```
0                   this articl was publish in 2002..
1                nlp stand for natur languag process
2        nlp is an emerg field in the data science!!!
3    we can preprocess text data use python librari...
4                   it is like to be useful, to peopl
5        here we will discuss 5-7 step for preprocess
Name: Message, dtype: object
```

```
from nltk.stem.lancaster import *
lancaster = LancasterStemmer()
df['Message'][:6].apply(
    lambda x: " ".join([lancaster.stem(word) for word in x.split()]))
```

```
0                    thi artic was publ in 2002..
1                 nlp stand for nat langu process
2        nlp is an emerg field in the dat science!!!
3    we can preprocess text dat us python libr - nltk
4                   it is lik to be useful, to peopl
5         her we wil discuss 5-7 step for preprocess
Name: Message, dtype: object
```



- **Lemmatizing —** Lemmatization is a process of extracting a root word by considering the vocabulary. Lemmatization is a more powerful operation, and takes into consideration morphological analysis of the words.

```
text = ['leaves and leaf', 'fishes and fish', 'there are many colors in pack']

import pandas as pd
```

```
df = pd.DataFrame({'Message': text})

from nltk.stem import WordNetLemmatizer
lemmatizer = WordNetLemmatizer()
df['Message'][:3].apply(
    lambda x: " ".join([lemmatizer.lemmatize(word) for word in x.split()]))
```

```
0                  leaf and leaf
1                  fish and fish
2    there are many color in pack
Name: Message, dtype: object
```

> *"**Stemming** is a general operation while **lemmatization** is an intelligent operation where the proper form will be searched in the dictionary; as a result thee later makes better machine learning features."*

SUMMARY

In this post, we discussed the basic concepts and applications of natural language processing and preprocessing steps using python library NLTK. Once preprocessing is completed one can use the data for more complicated NLP tasks.

Enjoy NLP and Thanks for reading 😄.

Nltk Python NLP Naturallanguageprocessing Data Preprocessing