Natural Language Processing

Agenda

01 What is NLP?

02 Text Analytics

03 Sentiment Analysis

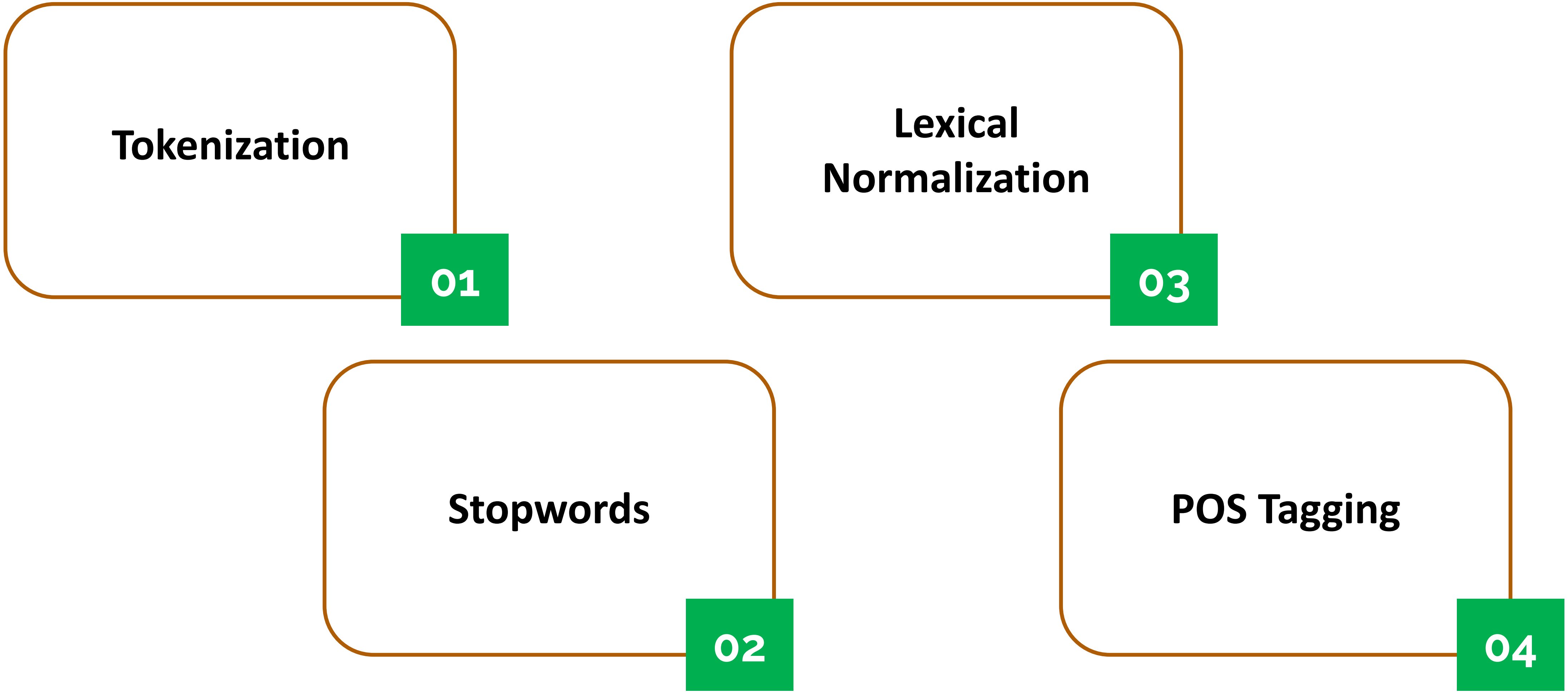
What is Natural Language Processing?

What is Natural Language Processing?

Natural language is how humans communicate, and with the generation of tons and tons of textual data. To analyze and process the data has become necessary for business outcomes.

Text Analytics

Text Analytics



01 Tokenization

02 Stopwords

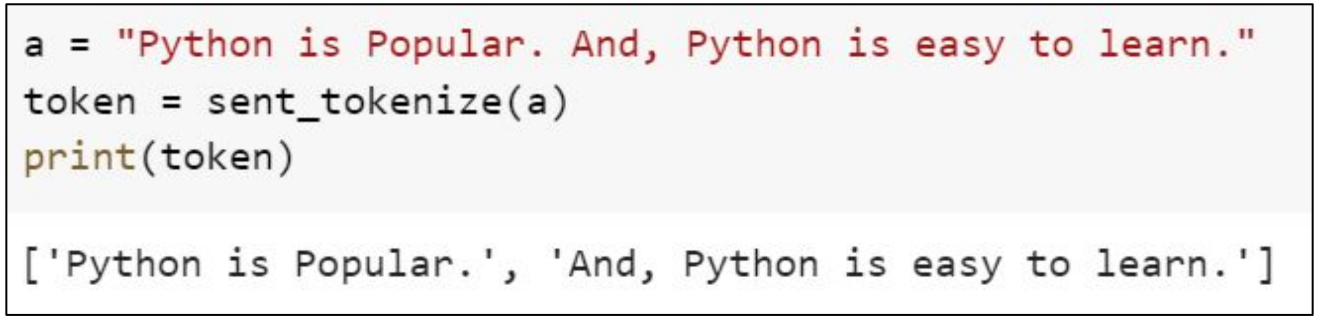
03 Lexical Normalization

04 POS Tagging

Text Analytics

Tokenization

01



```
a = "Python is Popular. And, Python is easy to learn."
token = sent_tokenize(a)
print(token)
```

```
['Python is Popular.', 'And, Python is easy to learn.']
```

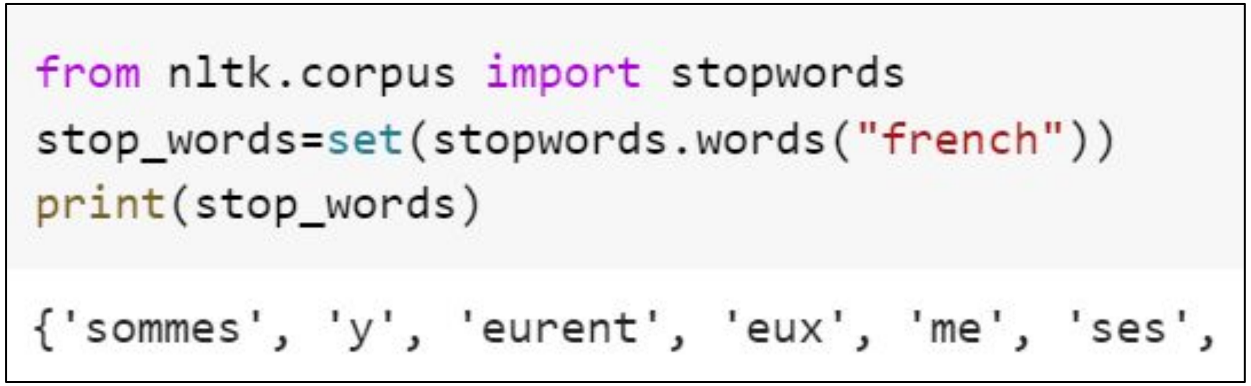
Tokenization is the process of breaking down the text into small tokens.

The processing time increases with tokenization and helps in faster computation of text analytics.

Text Analytics

Stopwords

02



```
from nltk.corpus import stopwords
stop_words=set(stopwords.words("french"))
print(stop_words)
```

```
{'sommes', 'y', 'eurent', 'eux', 'me', 'ses',
```

Stop words are the noise in the textual data, that must be removed from the data.

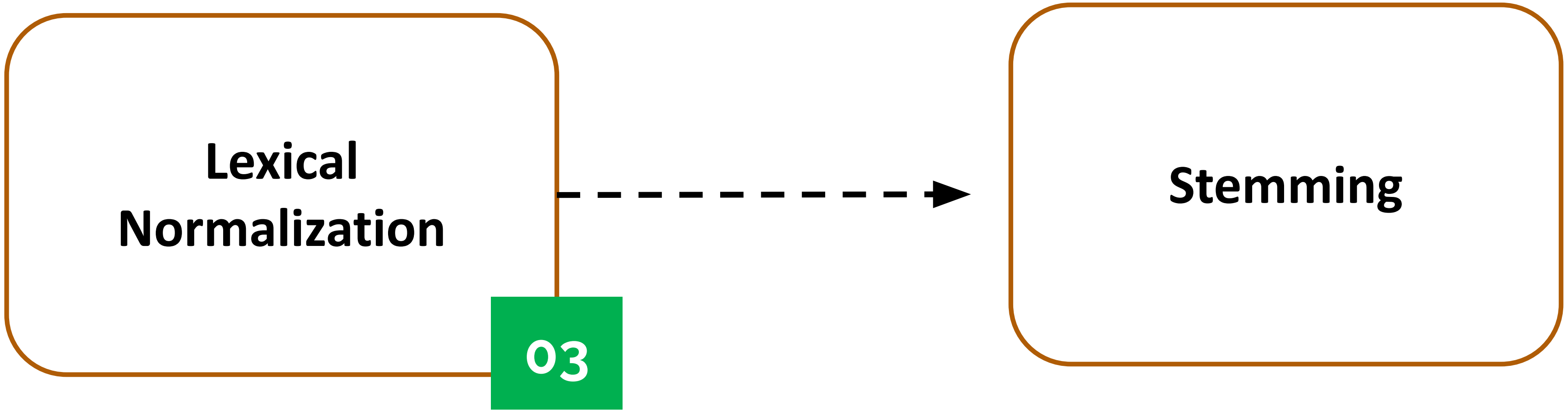
Text Analytics

Lexical Normalization

03

Lexical normalization is breaking down the tokens into their base forms, There are two types of lexical normalization – Stemming and Lemmatization.

Text Analytics



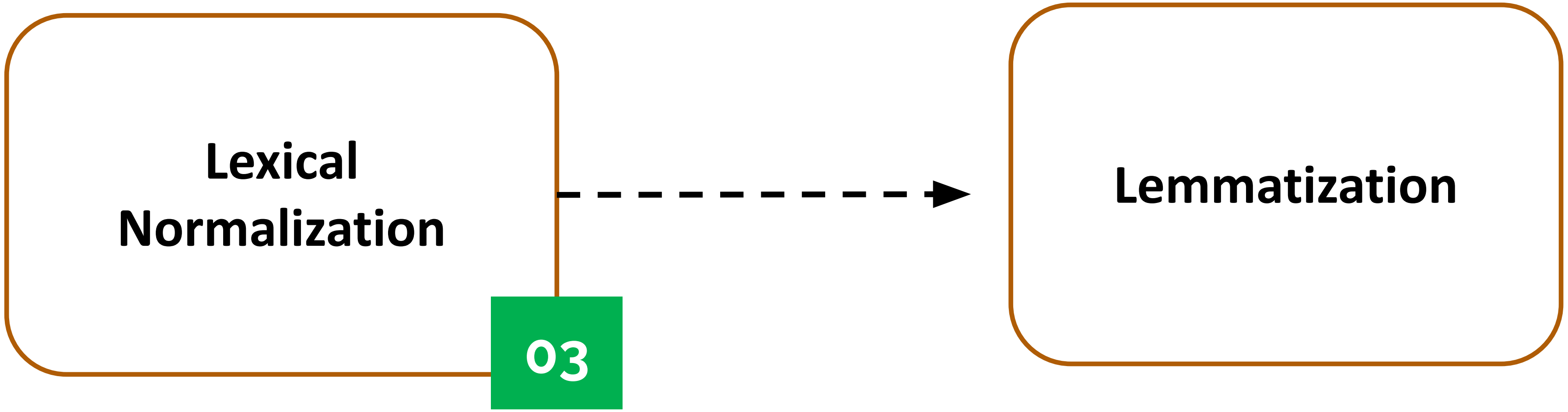
Lexical Normalization

03

Stemming

Stemming is a linguistic normalization procedure in which the word or the token is reduced to its root word and removes the derivational affixes.

Text Analytics



Lexical Normalization

03

Lemmatization

Lemmatization reduces words to their base word, which is linguistically correct lemmas. It transforms root word with the use of vocabulary and morphological analysis. For Example – Flying becomes fly.

Text Analytics

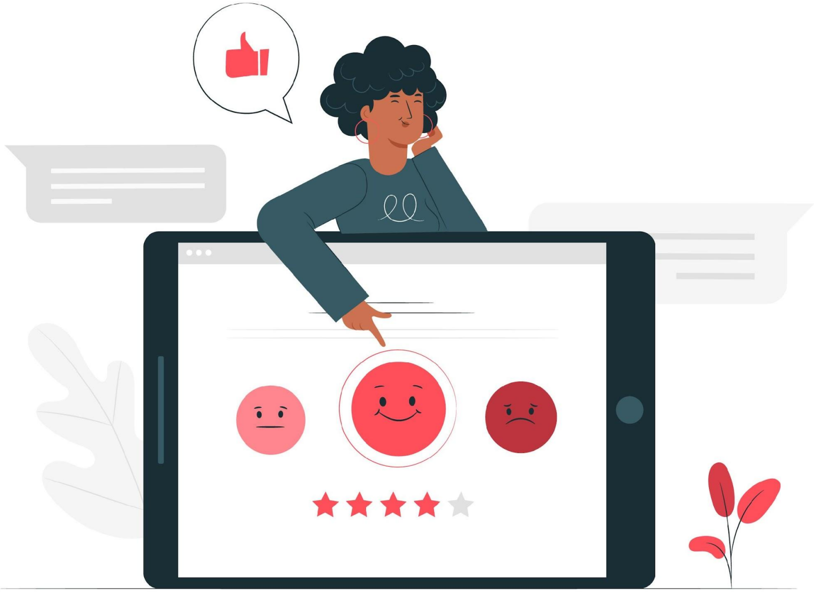
POS Tagging

04

POS tagging or the Part of Speech tagging is identifying the grammatical group for the given word. Like a noun, pronoun, etc.

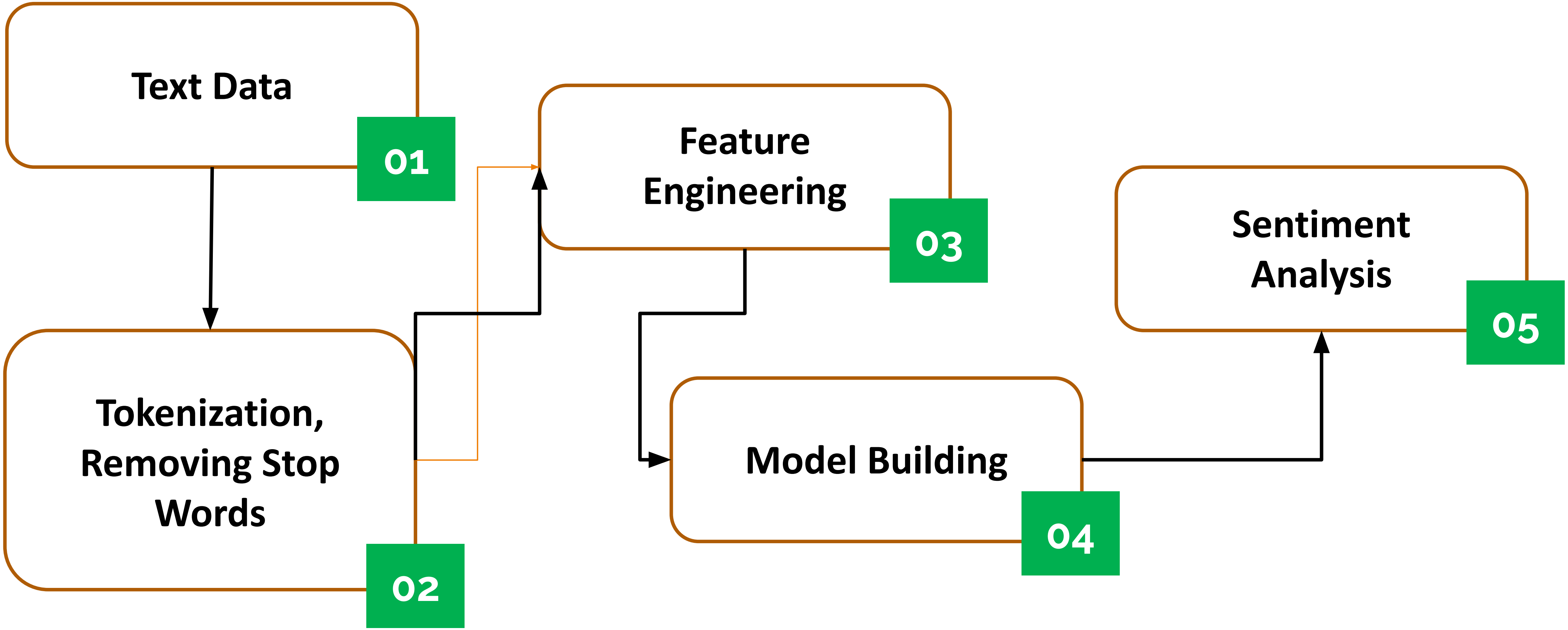
Sentiment Analysis

Sentiment Analysis



Sentiment Analysis is the predictive modeling technique in which a model is trained on sentiment labels, often generated by NLTK functions and then are used to predict the sentiments associated with the texts.

Sentiment Analysis



Sentiment Analysis

Text Data

01

The sentiment analysis lifecycle starts from gathering the textual data and transforming the data removing the outliers and null values before the further computation.

Sentiment Analysis

Tokenization, Removing Stop Words

02

Tokenization is initially done to simplify the computation, and stop words are removed to remove the noise from the data.

Sentiment Analysis

Feature Engineering

03

Feature engineering is done on the textual data by Bag-of-words and TF-IFD, that can be used to train the model.

Sentiment Analysis

Model Building

04

After the feature engineering has been finished, the model is trained on the transformed data, and we can use them to predict the sentiments on the test sets/validation sets, and unseen data as well.

Sentiment Analysis

Sentiment Analysis

05

For data, that may not have labels, we generate the sentiments associated with the texts based on the intensity using NLTK, and train the model with the exact same approach as we train any other model.