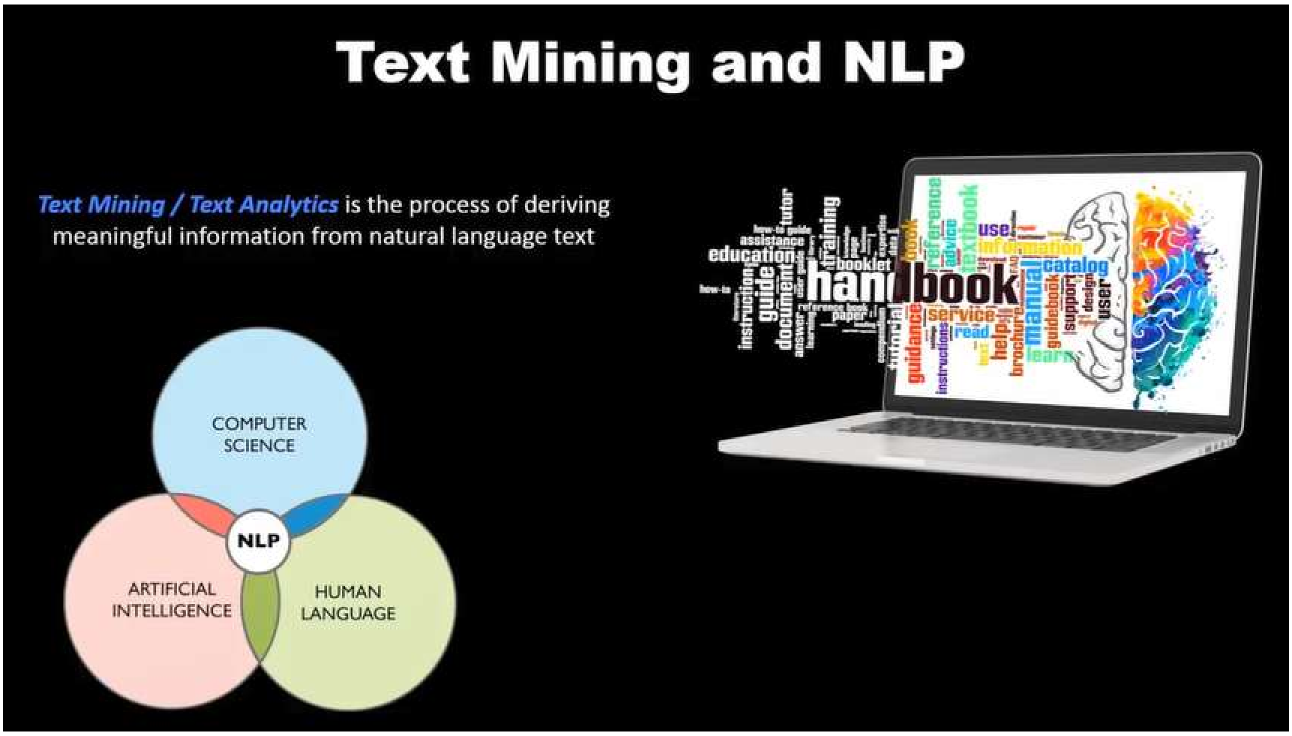
The Human Language

6500 LANGUAGES

The 21^{st} Century

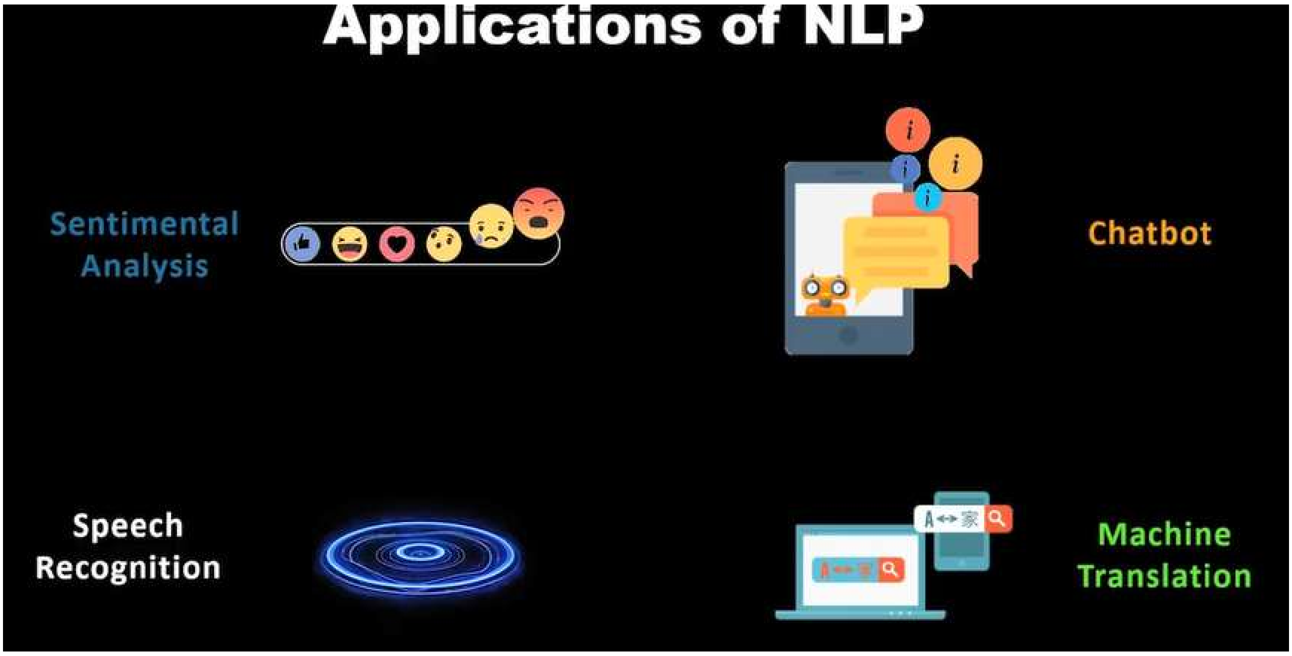
Percentage
🟦 Unstructured 🟧 Structured
BIG DATA



Text Mining and NLP

Text Mining / Text Analytics is the process of deriving meaningful information from natural language text

NLP: Natural Language Processing is a part of computer science and artificial intelligence which deals with human languages.



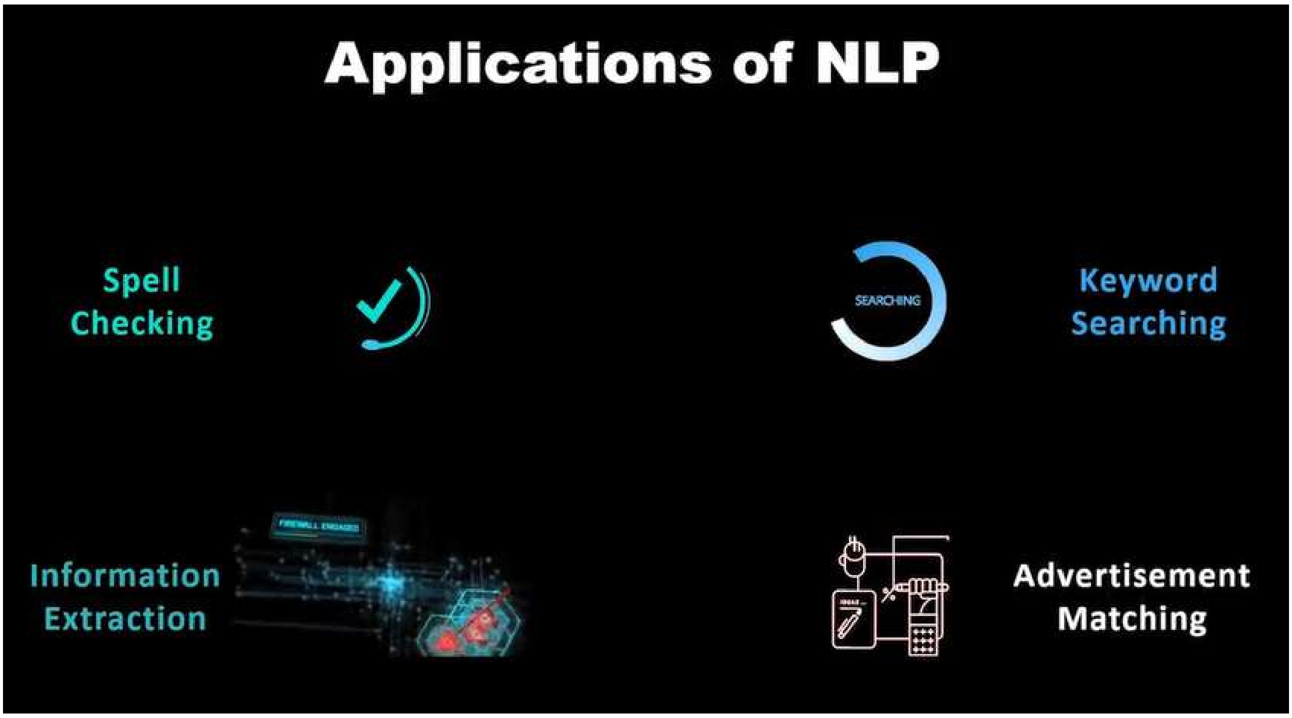
Applications of NLP

Sentimental Analysis

Chatbot

Speech Recognition

Machine Translation



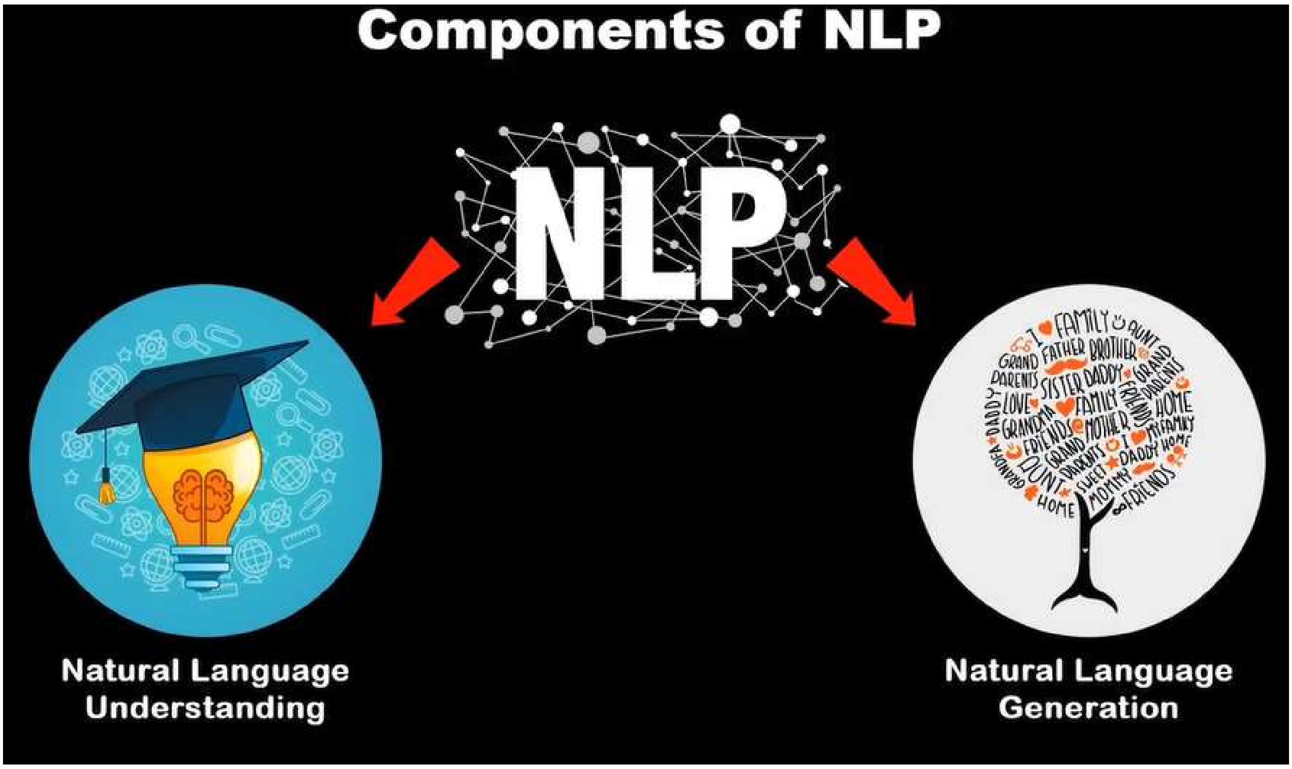
Applications of NLP

Spell Checking

Keyword Searching

Information Extraction

Advertisement Matching

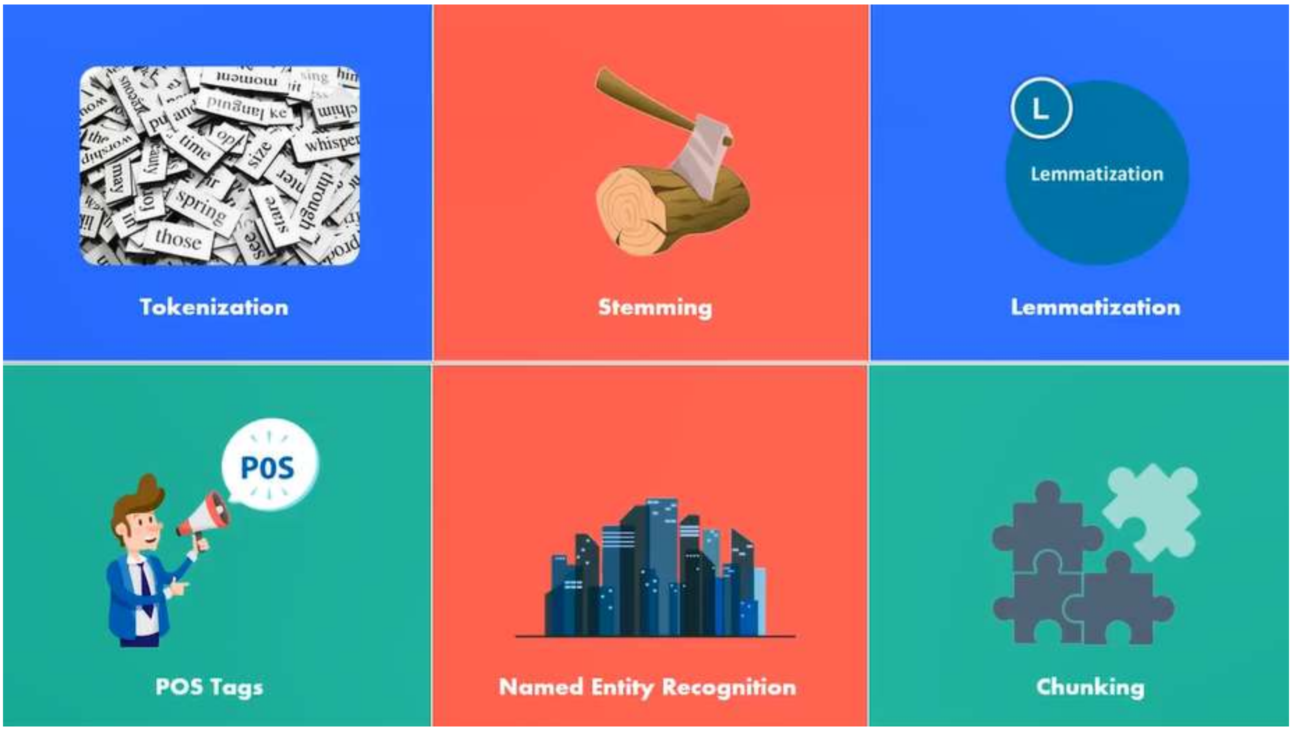


Components of NLP

NLP

Natural Language Understanding

Natural Language Generation



Tokenization

Stemming

Lemmatization

POS Tags

Named Entity Recognition

Chunking

Tokenization

Tokenization is the first step in NLP

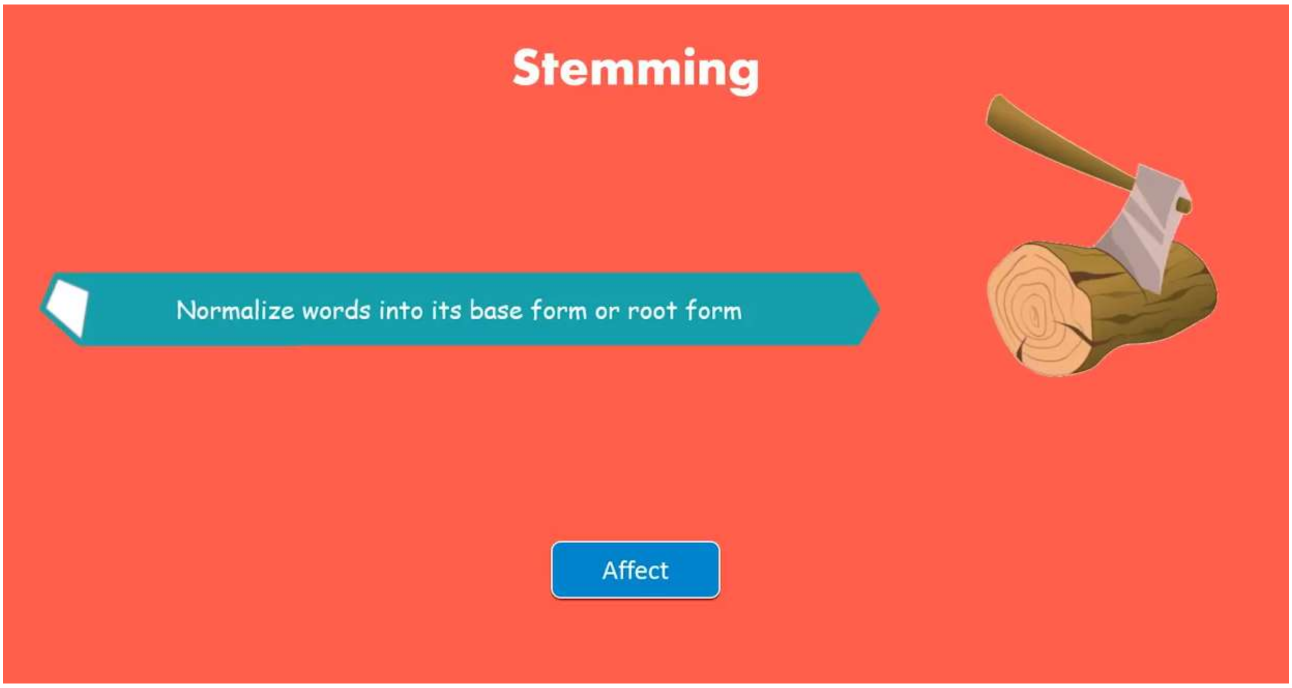
Tokenization

Tokenization | is | the | first | step | in | NLP

Stemming

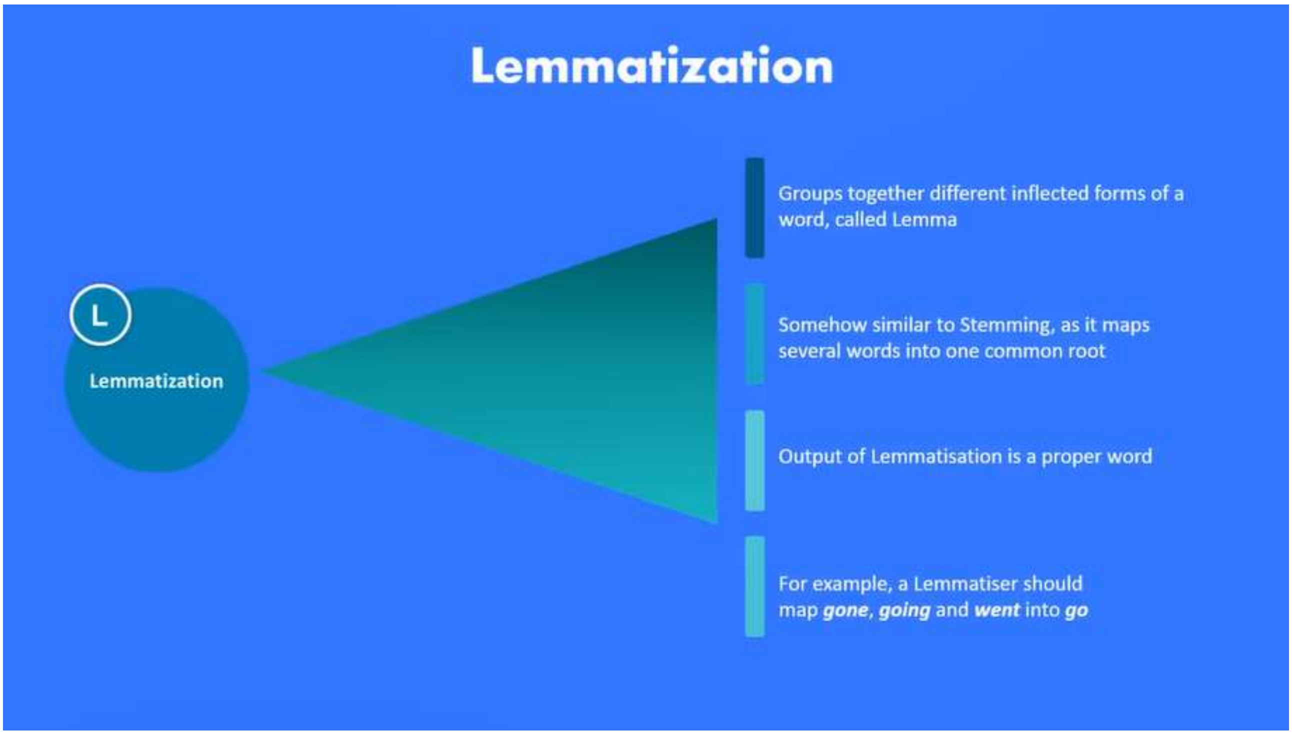
Normalize words into its base form or root form

Affectation | Affects | Affections | Affected | Affection | Affecting



Stemming

Normalize words into its base form or root form



Lemmatization

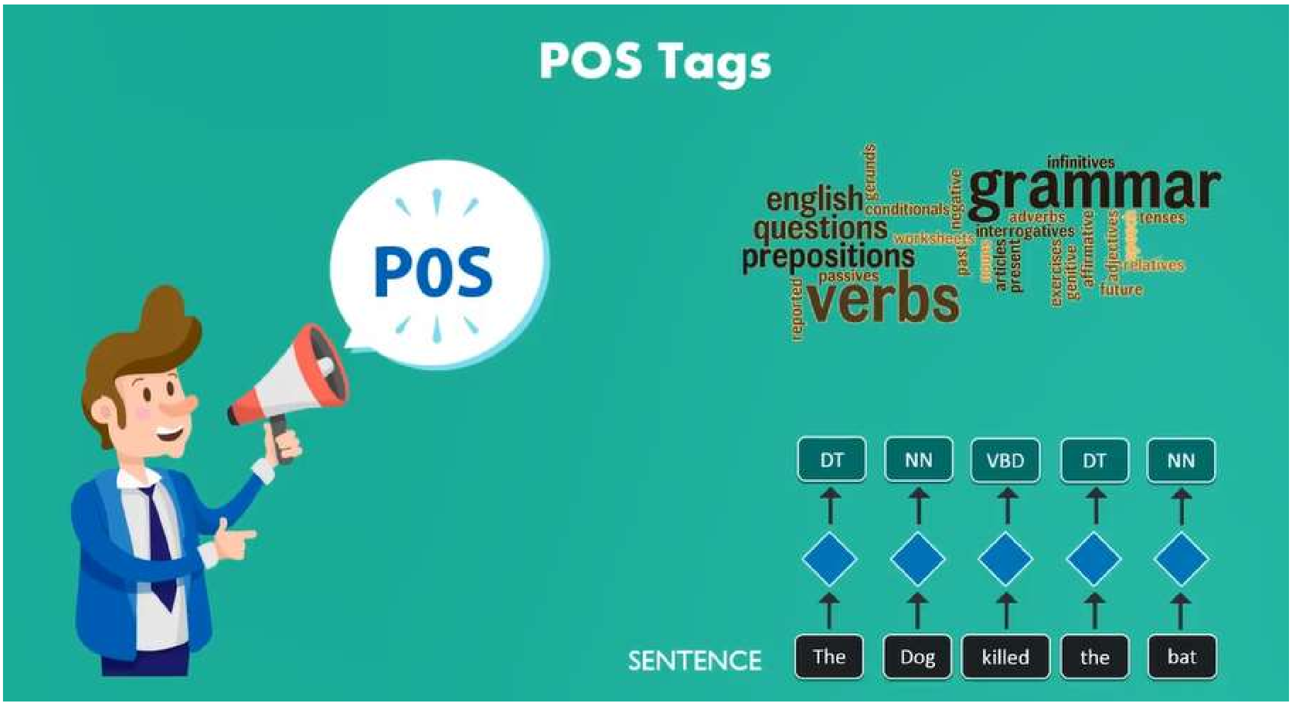
L
Lemmatization

Groups together different inflected forms of a word, called Lemma

Somehow similar to Stemming, as it maps several words into one common root

Output of Lemmatisation is a proper word

For example, a Lemmatiser should map ***gone***, ***going*** and ***went*** into ***go***



POS Tags

	DT	NN	VBD	DT	NN
SENTENCE	The	Dog	killed	the	bat

POS Tags

"Google" something on the internet

Named Entity Recognition

MOVIE

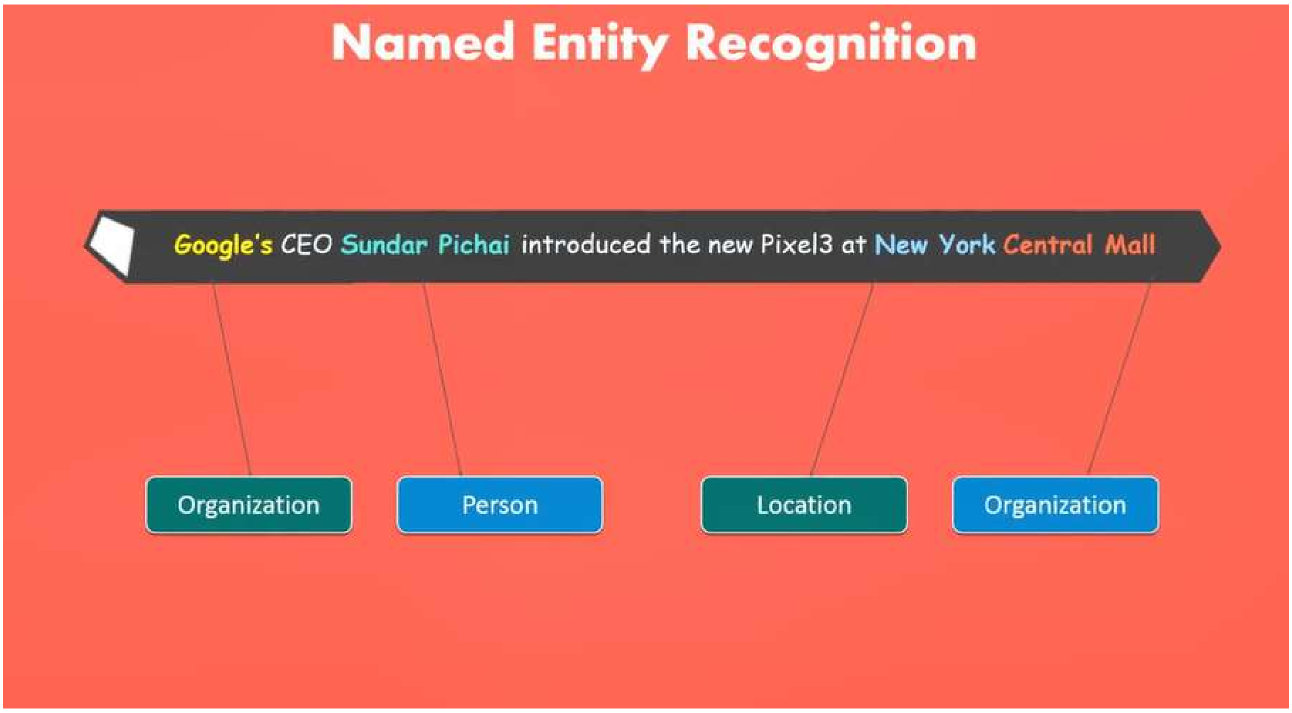
MONETARY VALUE

ORGANIZATION

LOCATION

QUANTITIES

PERSON



Named Entity Recognition

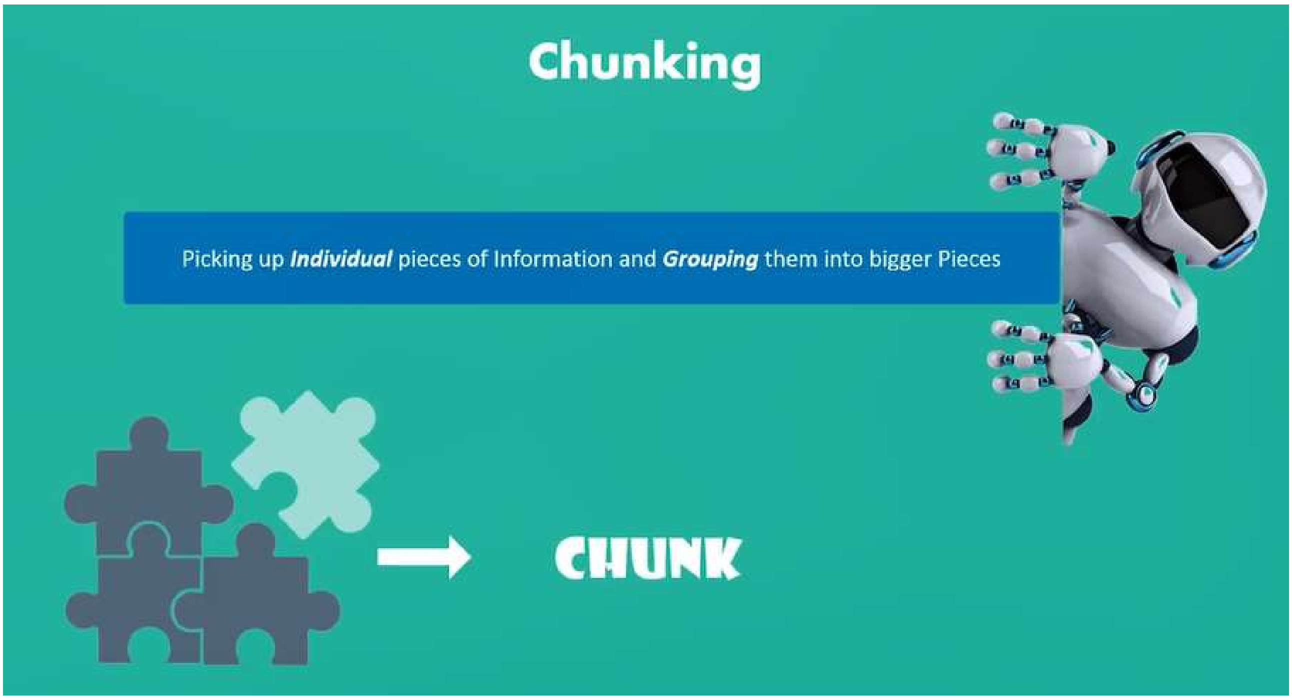
Google's CEO Sundar Pichai introduced the new Pixel3 at New York Central Mall

Organization

Person

Location

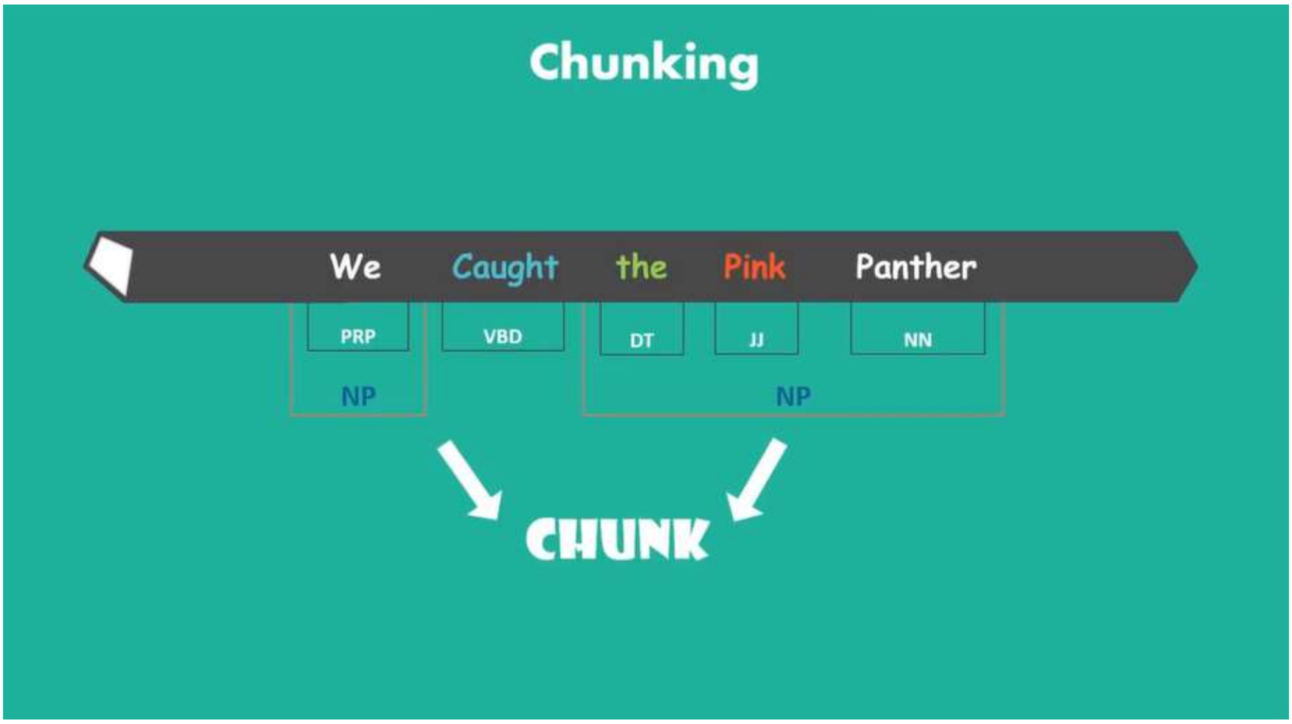
Organization



Chunking

Picking up ***Individual*** pieces of Information and ***Grouping*** them into bigger Pieces

CHUNK



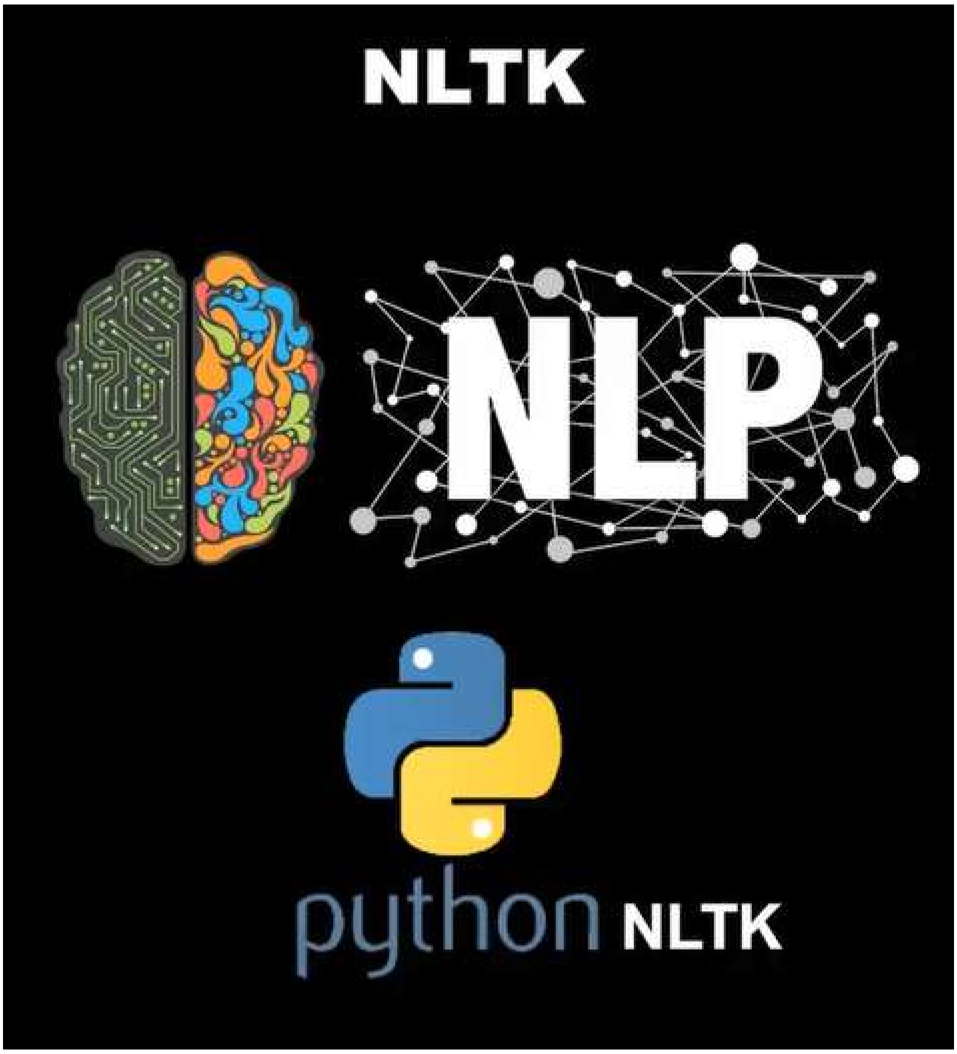
Chunking

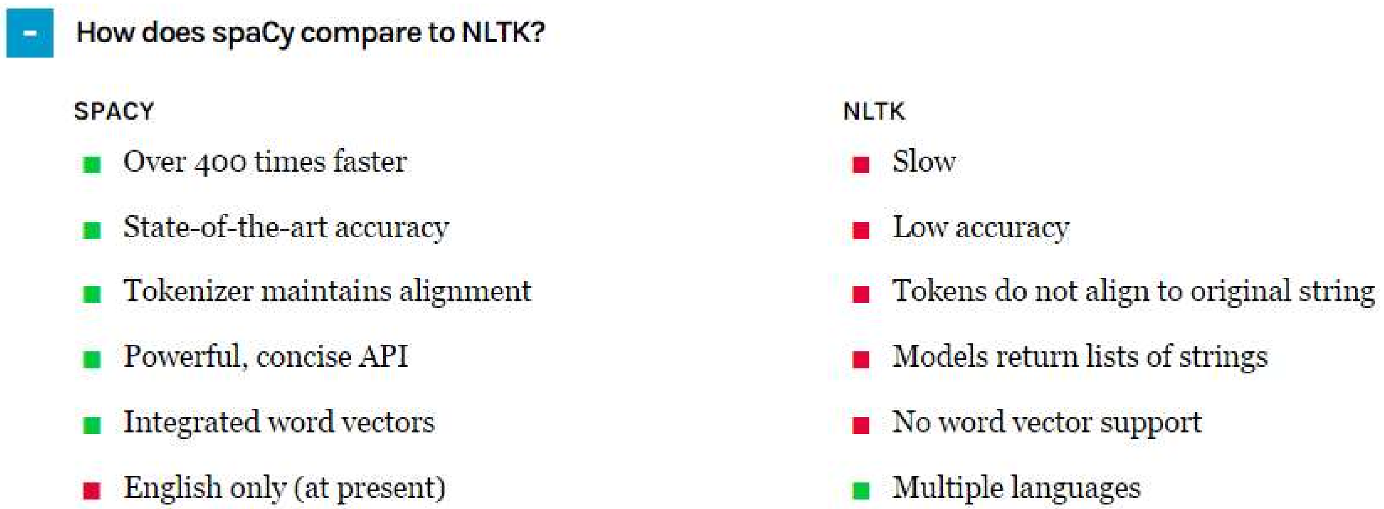
We Caught the Pink Panther

PRP VBD DT JJ NN

NP NP

CHUNK





- How does spaCy compare to NLTK?

SPACY	NLTK
🟩 Over 400 times faster	🟥 Slow
🟩 State-of-the-art accuracy	🟥 Low accuracy
🟩 Tokenizer maintains alignment	🟥 Tokens do not align to original string
🟩 Powerful, concise API	🟥 Models return lists of strings
🟩 Integrated word vectors	🟥 No word vector support
🟥 English only (at present)	🟩 Multiple languages