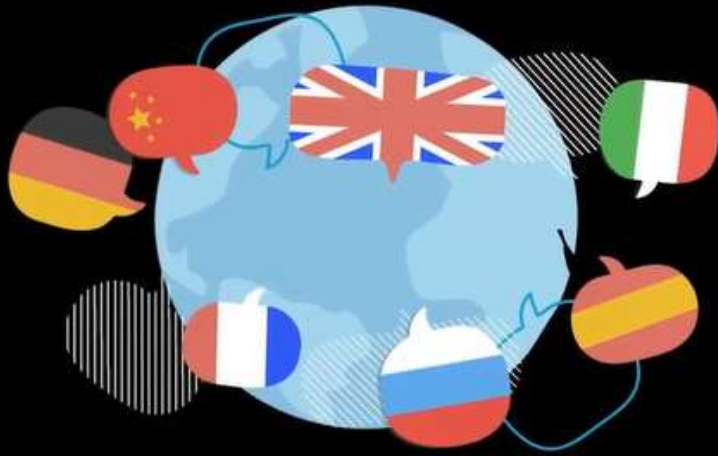
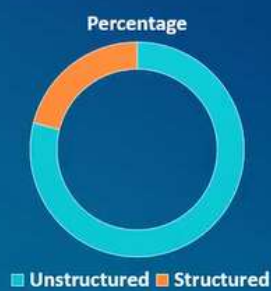
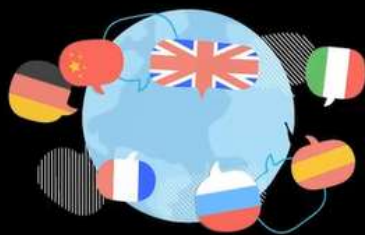


The Human Language



**6500
LANGUAGES**

The 21st Century



Text Mining and NLP

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

A Venn diagram consisting of three overlapping circles. The top circle is light blue and labeled "COMPUTER SCIENCE". The bottom-left circle is light pink and labeled "ARTIFICIAL INTELLIGENCE". The bottom-right circle is light green and labeled "HUMAN LANGUAGE". The central area where all three circles overlap is white and contains the text "NLP".

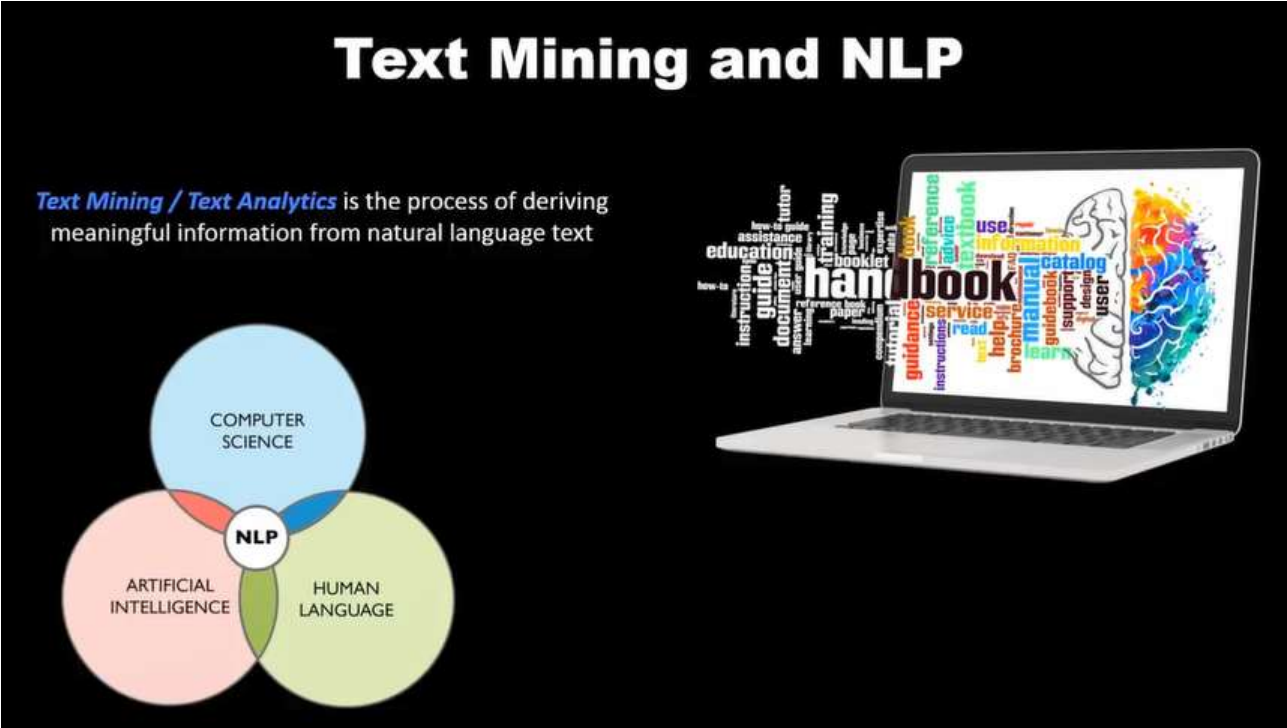
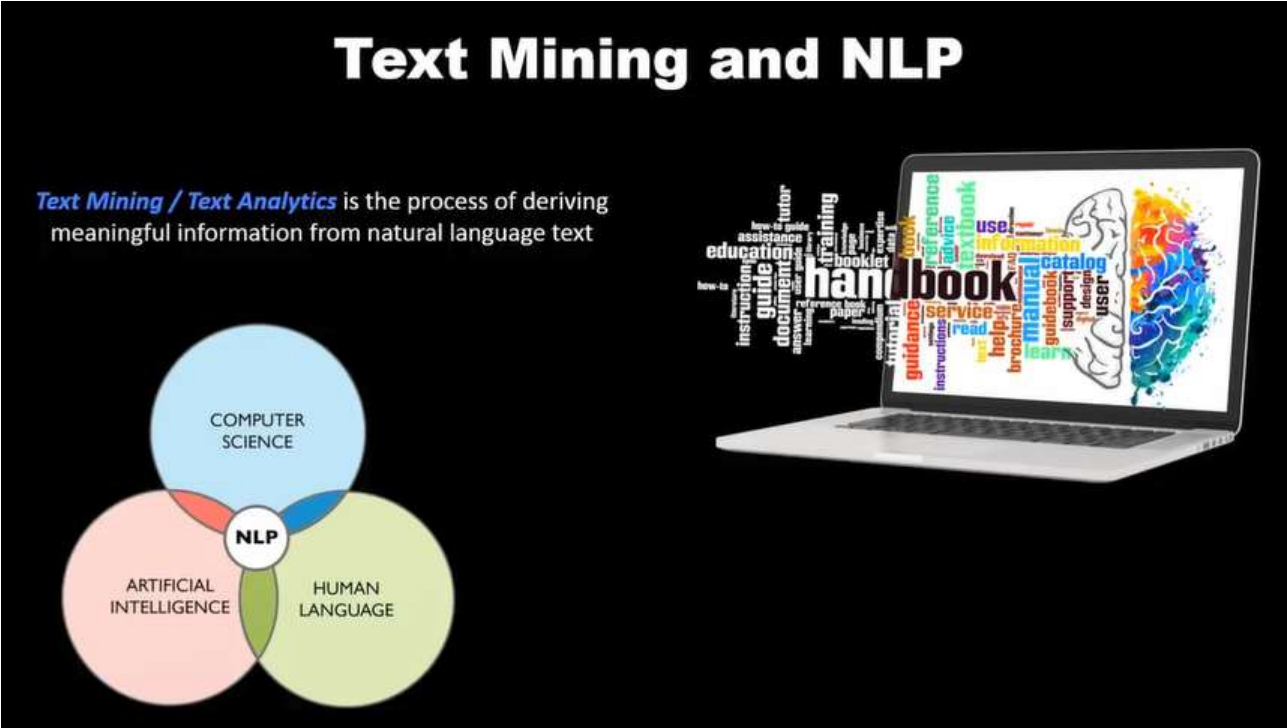
A silver laptop is shown from a slightly elevated angle. On its screen is a vibrant word cloud. The most prominent words are "handbook", "guide", "manual", "catalog", "service", "help", "instructions", "document", "tutor", "education", "reference", "book", "use", "design", "support", "user", "read", "brochure", "guidance", "instructions", "manual", "ears", "brain". To the right of the word cloud is a stylized illustration of a human brain with colorful, splattered paint-like textures.

Text Mining and NLP

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





A Venn diagram consisting of three overlapping circles. The top circle is light blue and labeled "COMPUTER SCIENCE". The bottom-left circle is light pink and labeled "ARTIFICIAL INTELLIGENCE". The bottom-right circle is light green and labeled "HUMAN LANGUAGE". The central area where all three circles overlap is white and contains the text "NLP".

A silver laptop is shown from a slightly elevated angle. On its screen is a vibrant word cloud. The most prominent words are "handbook", "guide", "manual", "catalog", "service", "help", "instructions", "document", "tutor", "education", "reference", "book", "use", "design", "support", "user", "read", "brochure", "guidance", "instructions", "manual", "ears", "brain", "catalyst". The words are arranged in various sizes and colors (red, yellow, green, blue, purple) against a dark background.







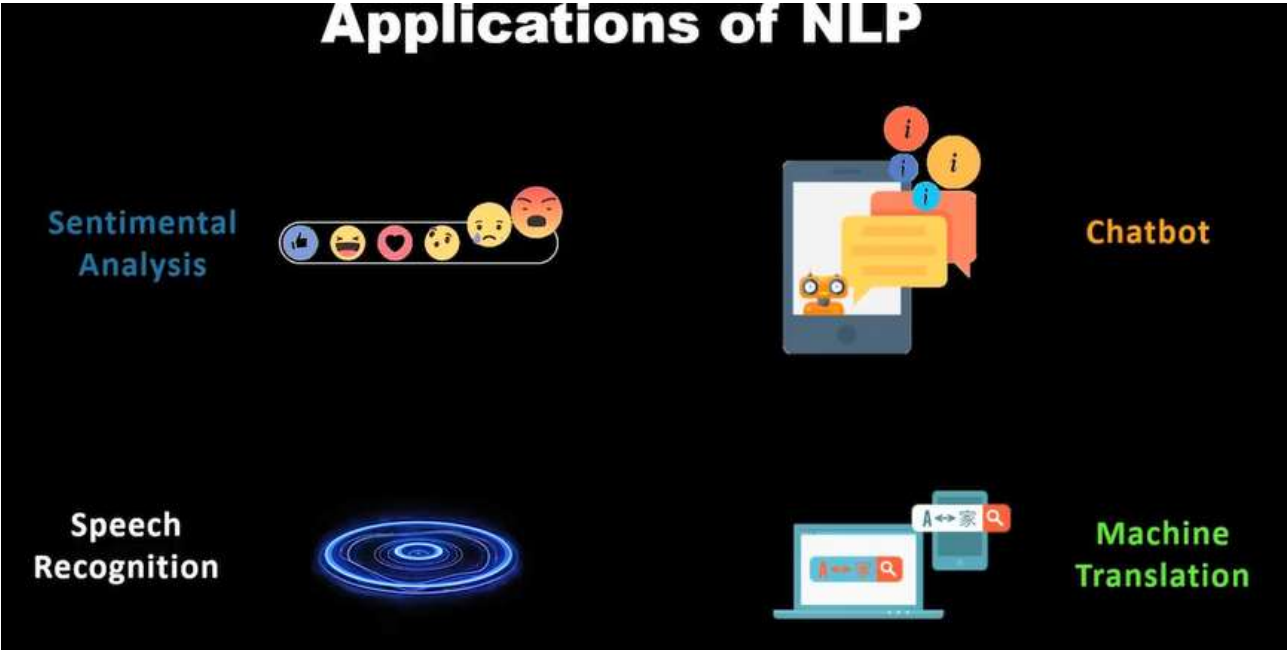
NLP: Natural Language Processing is a part of computer science and artificial intelligence which deals with human languages.

Applications of NLP





Sentimental Analysis		Chatbot	
Speech Recognition		Machine Translation	

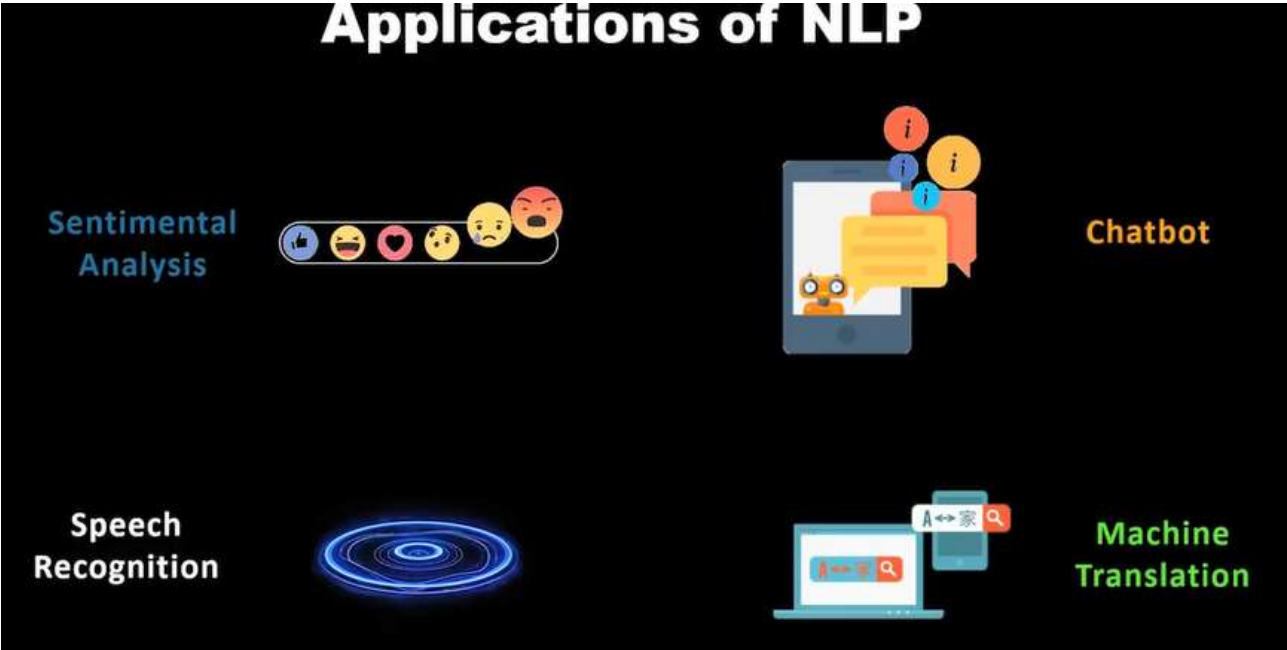
Applications of NLP

Sentimental Analysis		Chatbot	
Speech Recognition		Machine Translation	







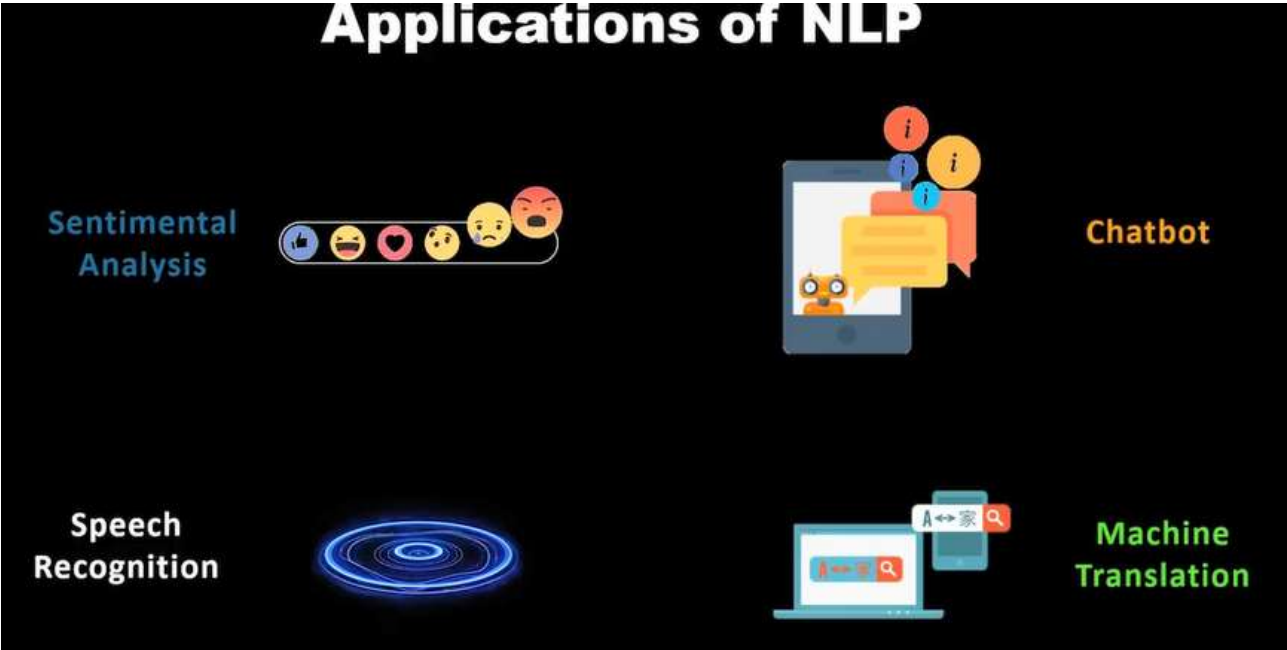
Applications of NLP

Sentimental Analysis		Chatbot	
Speech Recognition		Machine Translation	







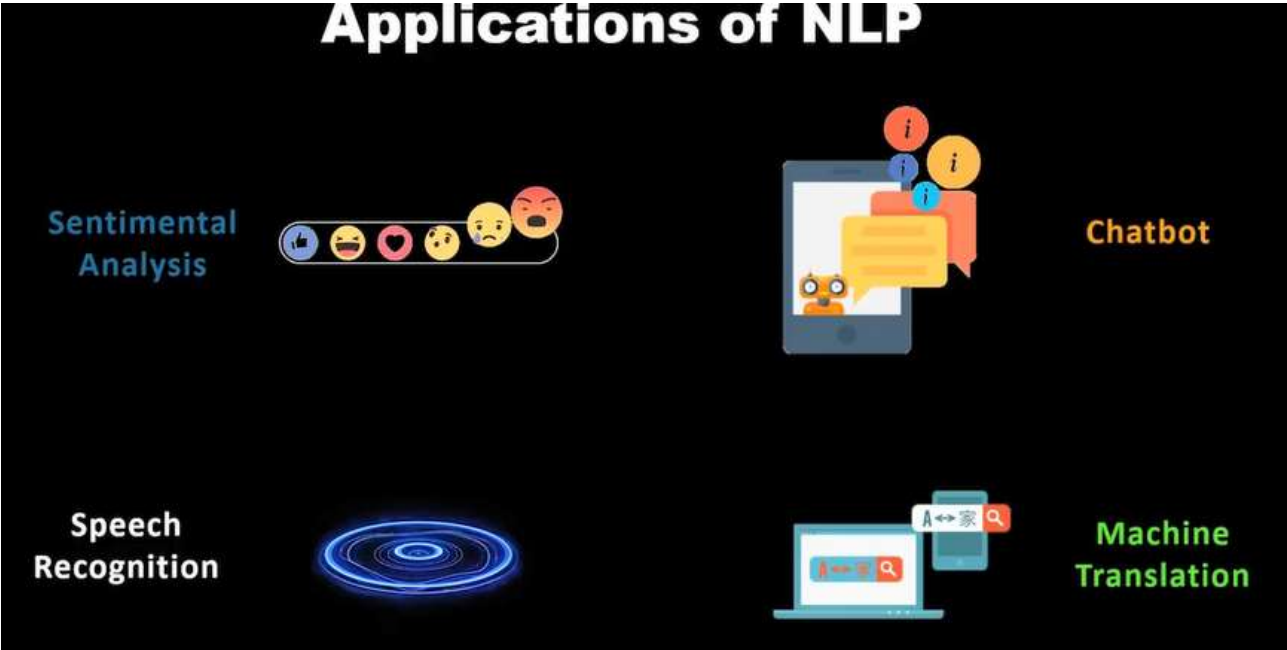
Applications of NLP

Sentimental Analysis		Chatbot	
Speech Recognition		Machine Translation	




Applications of NLP

Sentimental Analysis		Chatbot	
Speech Recognition		Machine Translation	




Applications of NLP


Spell Checking




Keyword Searching



Information Extraction



Advertisement Matching





Keyword Searching

Components of NLP

The diagram illustrates the components of NLP. At the top, the text "Components of NLP" is displayed. Below it, a central graphic shows the letters "NLP" in a large, white, bold font, surrounded by a network of white dots connected by lines, with two red arrows pointing outwards. Below this central graphic are two circular icons. The left icon is blue and features a lightbulb with a brain inside, surrounded by various icons representing language and thought. Below this icon is the text "Natural Language Understanding". The right icon is white and features a tree with a black trunk and branches, where the leaves are replaced by various words related to family and home. Below this icon is the text "Natural Language Generation".

NLP

Natural Language Understanding

Natural Language Generation





Tokenization



Stemming



Lemmatization

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



Affect

Lemmatization

L

Lemmatization

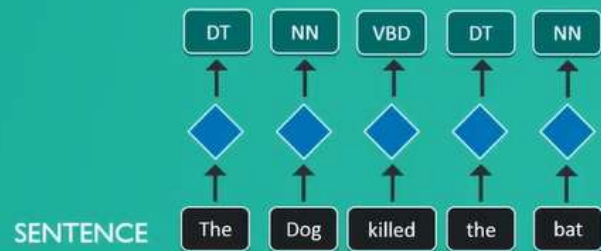
Groups together different inflected forms of a word, called Lemma

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



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

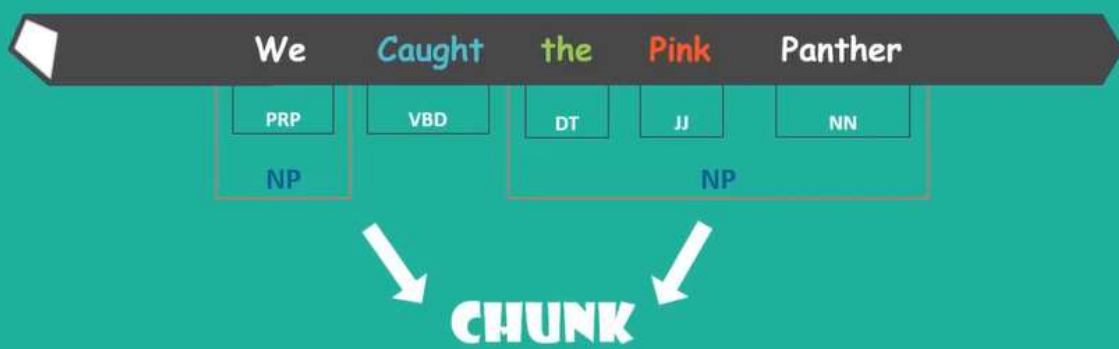
Organization

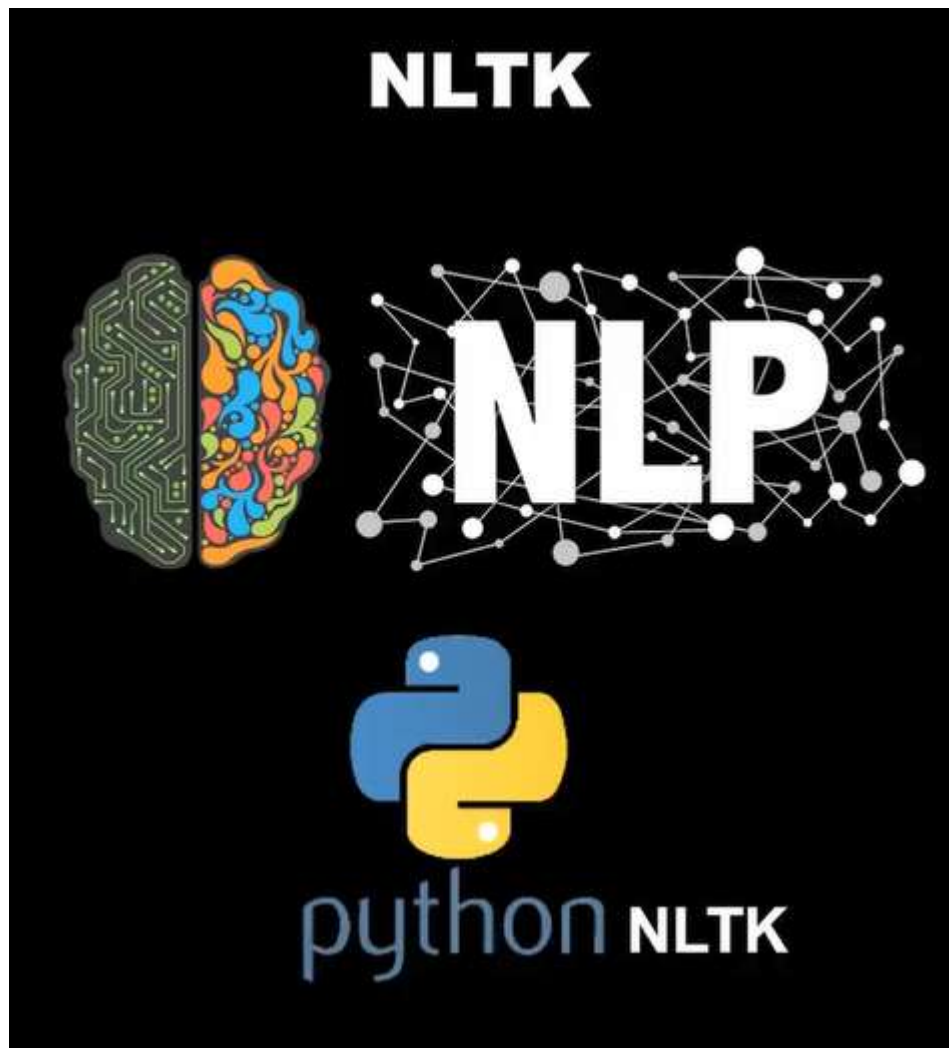
Chunking

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



Chunking





- How does spaCy compare to NLTK?

SPACY

- Over 400 times faster
- State-of-the-art accuracy
- Tokenizer maintains alignment
- Powerful, concise API
- Integrated word vectors
- English only (at present)

NLTK

- Slow
- Low accuracy
- Tokens do not align to original string
- Models return lists of strings
- No word vector support
- Multiple languages