Introduction to NLP

Penyusun Modul : Dennis Laorens Bawole

Editor : Citra Chairunnisa

The Human Language

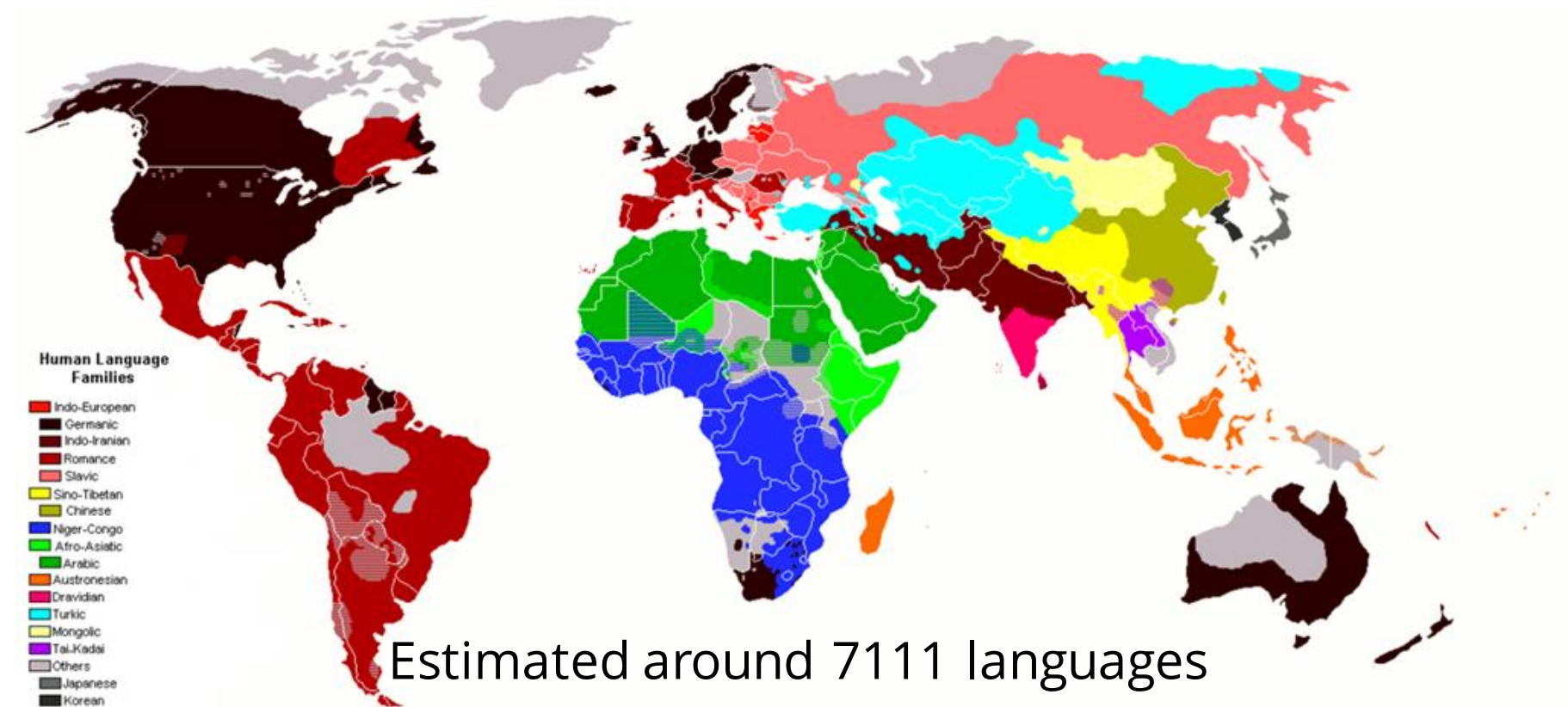A way to communicate with other people
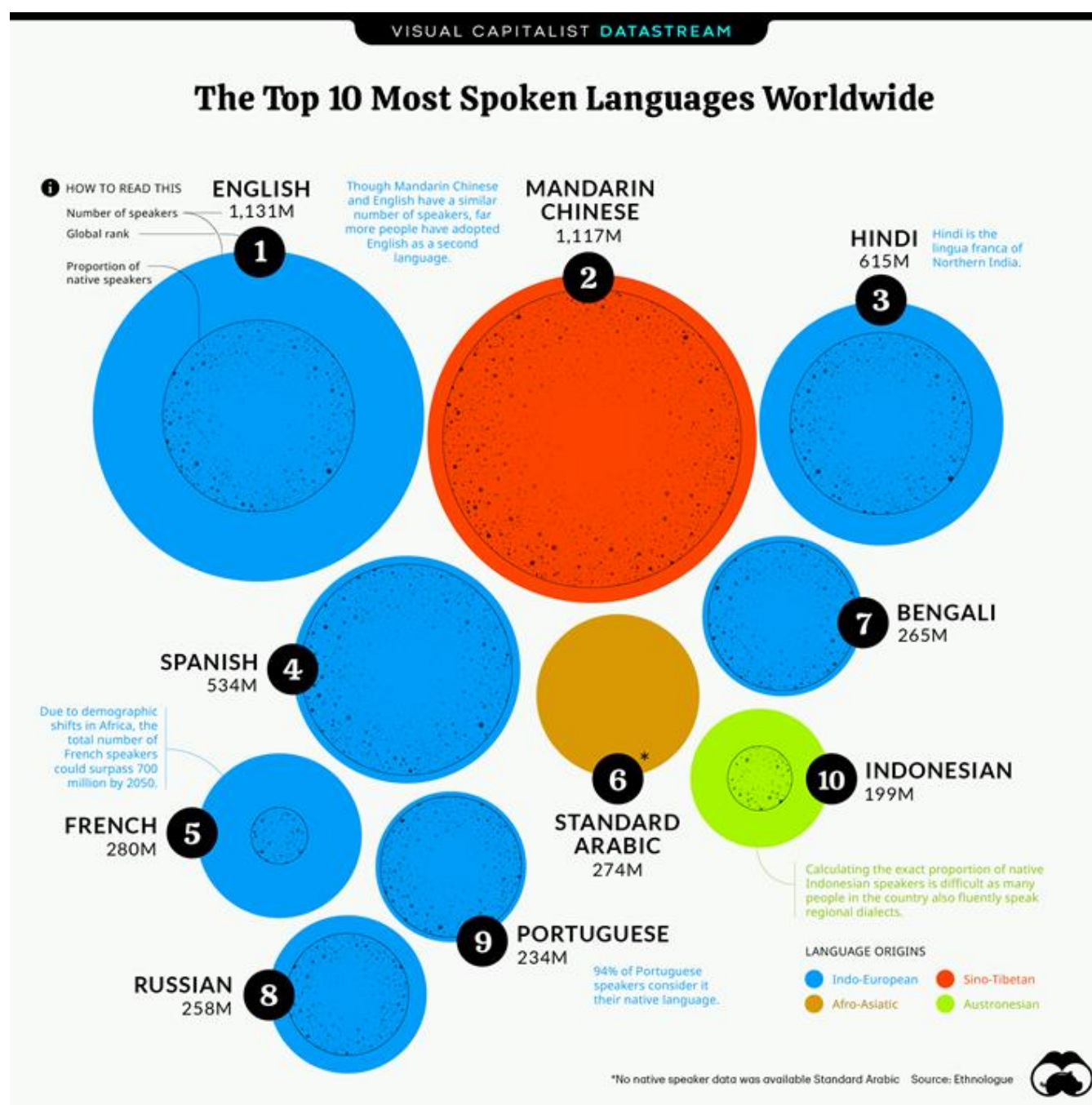
718

Local Languages

Human Language Families

- Indo-European
- Germanic
- Indo-Iranian
- Romance
- Slavic
- Sino-Tibetan
- Chinese
- Niger-Congo
- Afro-Asiatic
- Arabic
- Austronesian
- Dravidian
- Turkic
- Mongolic
- Tai-Kadai
- Others
- Japanese
- Korean

Estimated around 7111 languages in the world

The Human Language

VISUAL CAPITALIST DATASTREAM

The Top 10 Most Spoken Languages Worldwide

HOW TO READ THIS
Number of speakers
Global rank
Proportion of native speakers

ENGLISH 1,131M
①

Though Mandarin Chinese and English have a similar number of speakers, far more people have adopted English as a second language.

MANDARIN CHINESE 1,117M
②

HINDI 615M
③
Hindi is the lingua franca of Northern India.

SPANISH 534M
④

BENGALI 265M
⑦

Due to demographic shifts in Africa, the total number of French speakers could surpass 700 million by 2050.

STANDARD ARABIC 274M *
⑥

INDONESIAN 199M
⑩

FRENCH 280M
⑤

Calculating the exact proportion of native Indonesian speakers is difficult as many people in the country also fluently speak regional dialects.

PORTUGUESE 234M
⑨

94% of Portuguese speakers consider it their native language.

RUSSIAN 258M
⑧

LANGUAGE ORIGINS
● Indo-European ● Sino-Tibetan
● Afro-Asiatic ● Austronesian

*No native speaker data was available Standard Arabic   Source: Ethnologue

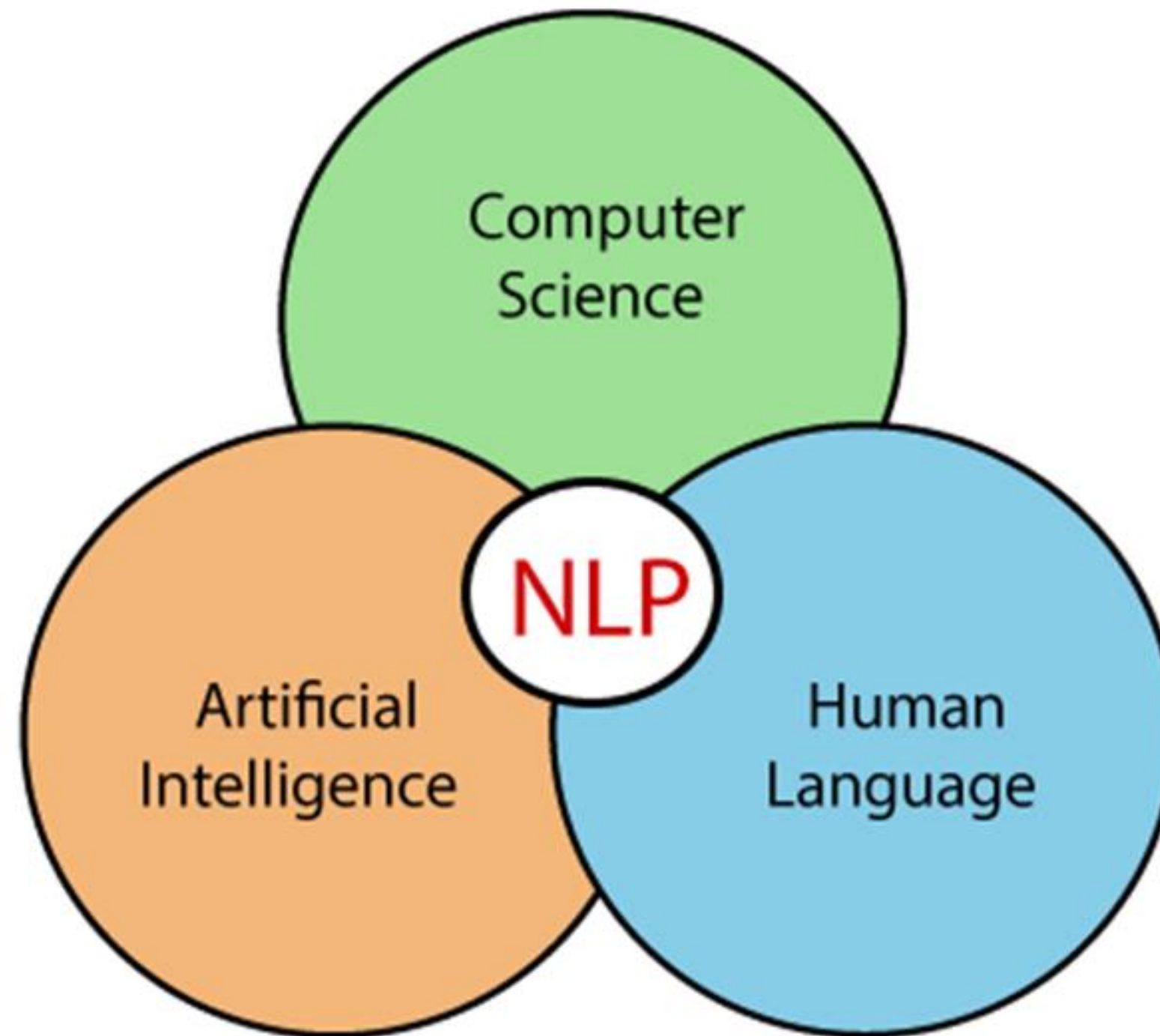# The Human Language

Understanding the Human Language = Understanding the basics of NLP
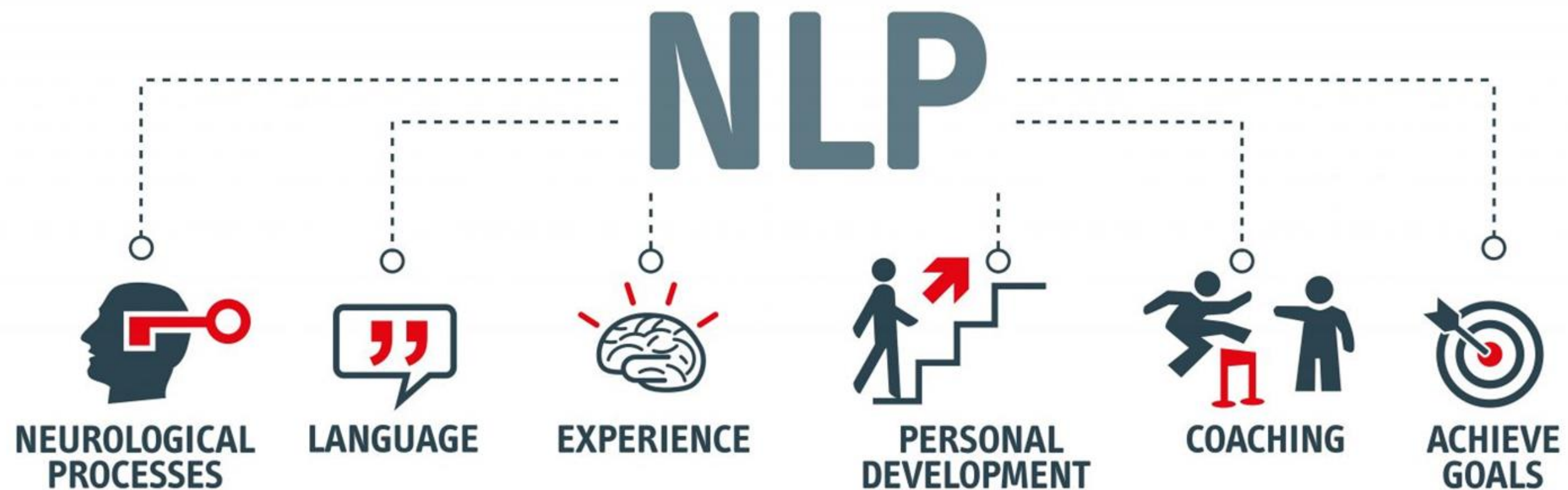
Why?
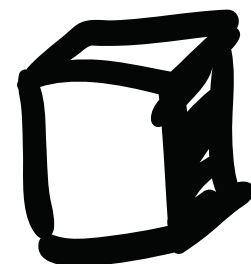
Uses Computer science as the programming basis

Uses AI to predict behaviors based on the dataset

Uses the Human Language as its dataset

So What is NLP?

NLP, also called Natural Language Processing is the manipulation of natural languages such as speech or text by software
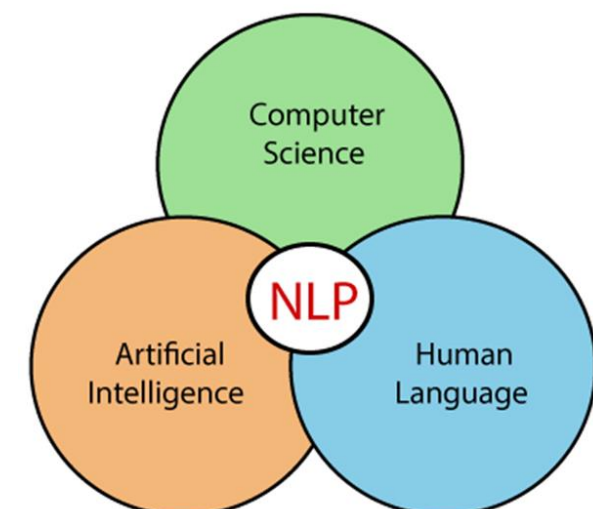
Where have you seen NLP?

How is NLP applied?
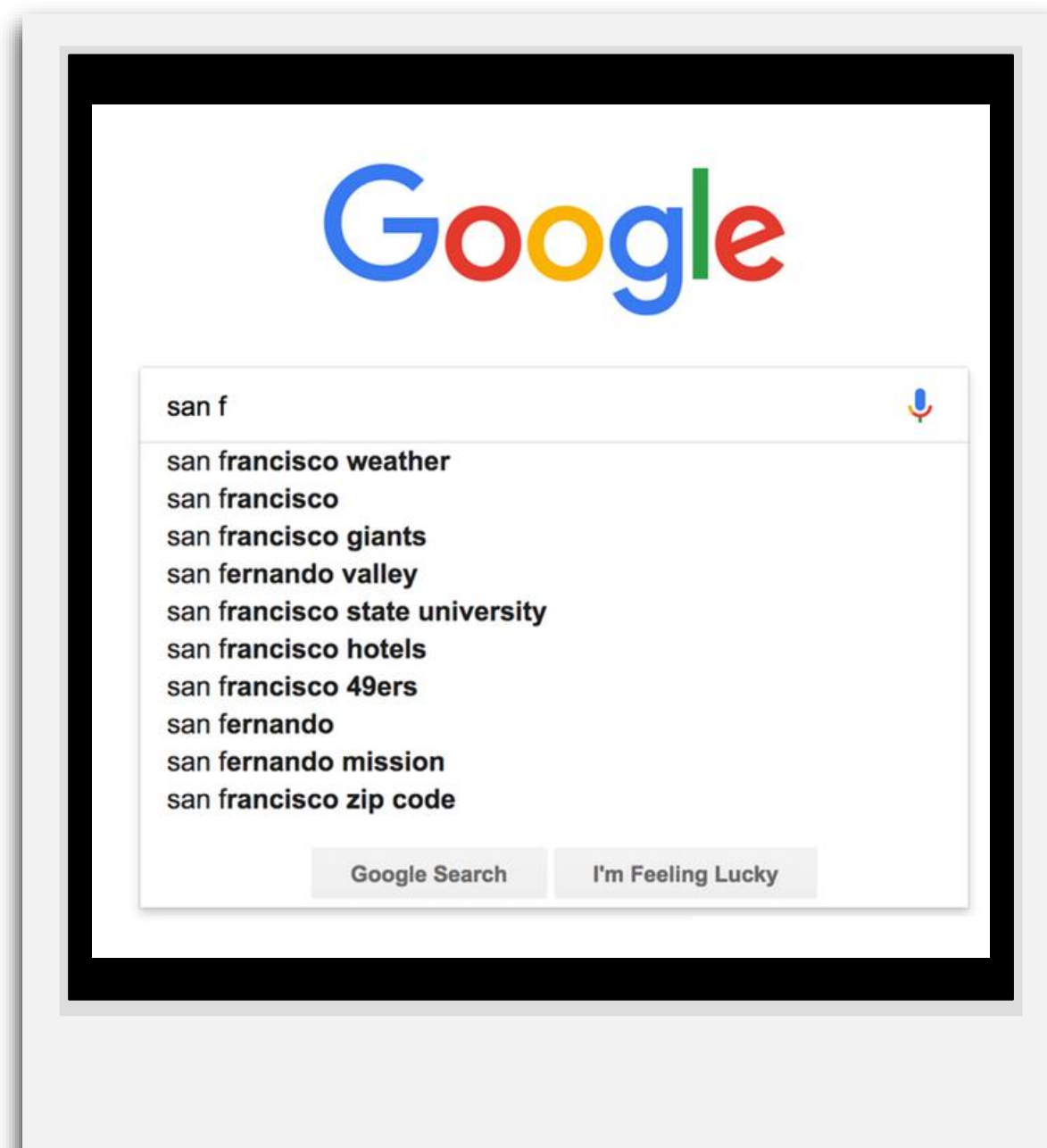
▶ The google search bar is designed using computer science and software techniques

▶ While the input and output itself are words based on the human language, in this case the words are in English

▶ The search recommendations are available with the help of AI where research is done to predict the chances of words that will appear after a certain word is typed
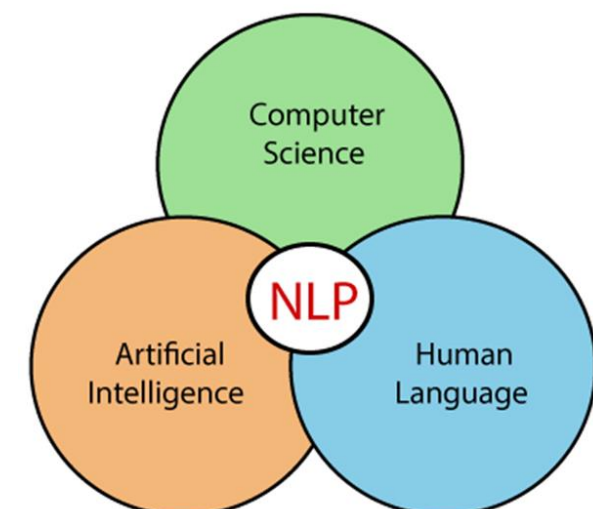
How is NLP applied?
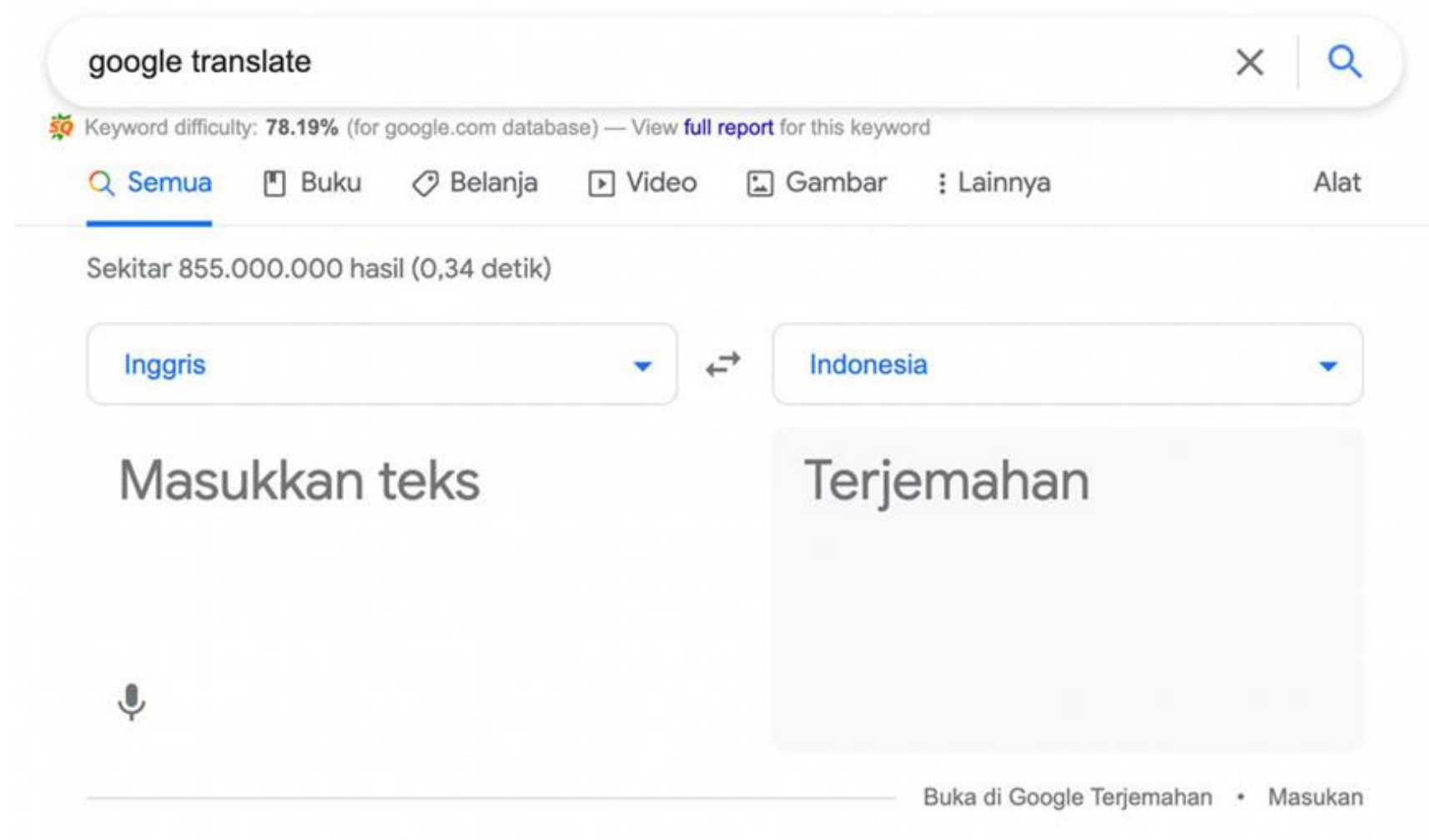
▶ For the google translate page, the impact of AI is much more diverse as the type of language used as the input may vary while the resulted output is in a different language

▶ This requires the AI too have a deep understanding towards grammar for every type of language used
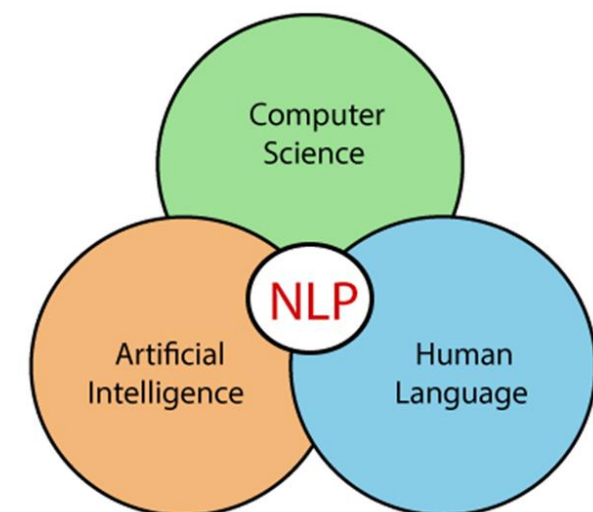
How is NLP applied?

Google Assistant

▶ The google assistant is a very complex product of the NLP field.
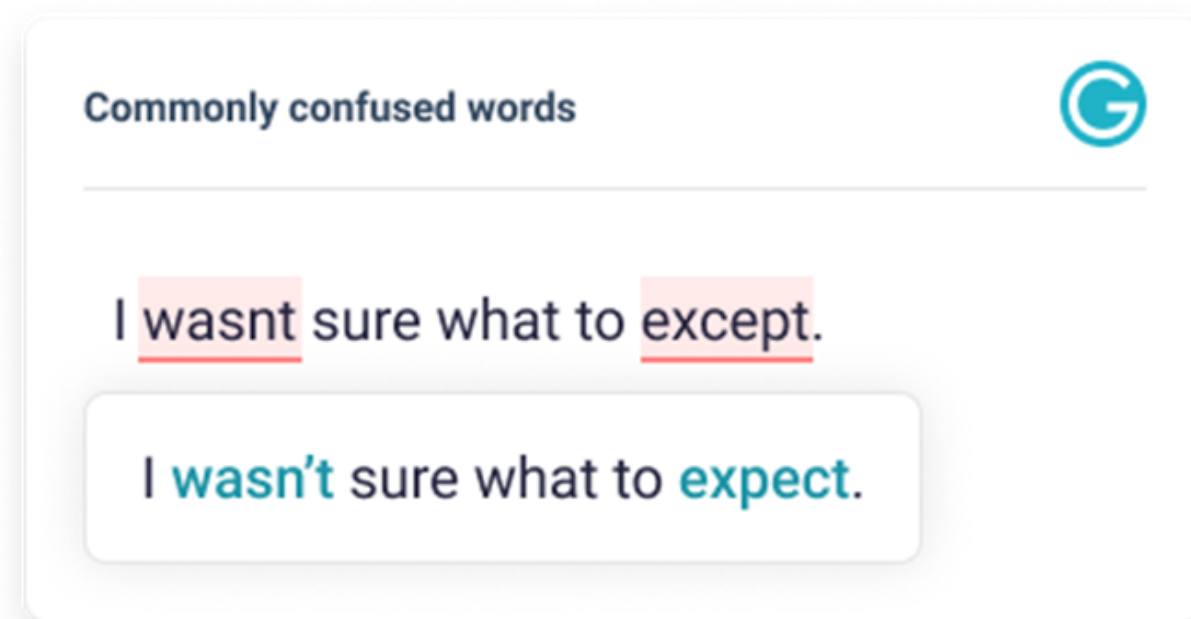
▶ It requires not only the understanding of basic grammar but also includes the use of STT (Speech To Text) technology which is another field of science that is highly complex.
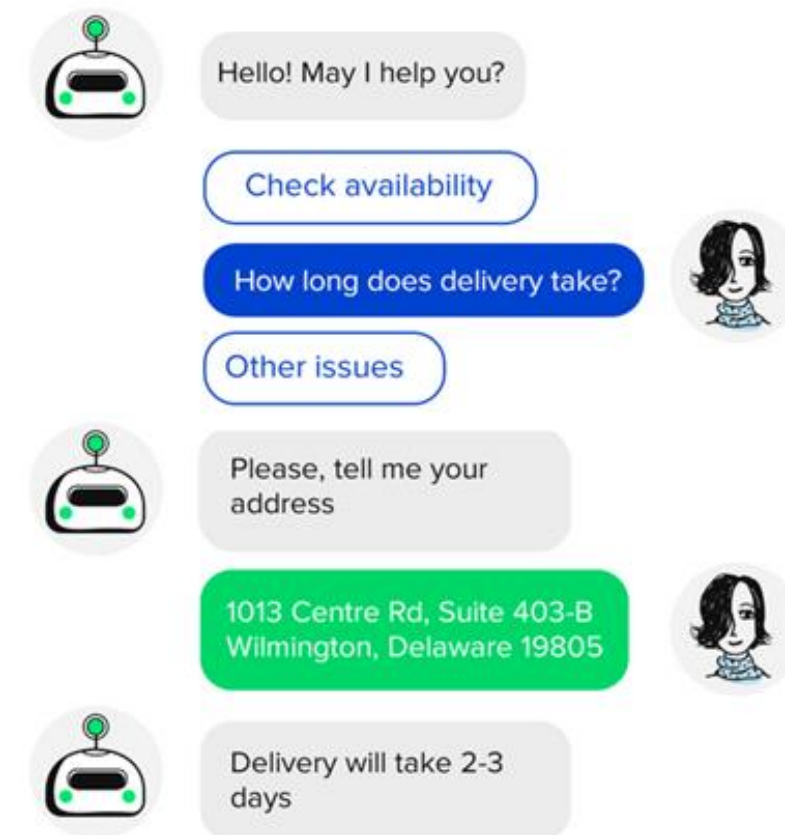
How is NLP applied?

There are much more examples on the impact of NLP in our daily life such as :

Spell Checking

Chat Bots

Goals

For this course, our target in learning NLP is to:

- Gain a basic understanding of how NLP is used

- Gain a basic understanding on the features that NLP has

- Applying basic functions from the most used NLP libraries

- Applying basic Exploratory Data Analysis on text data

Components of NLP

NLU (Natural Language Understanding)

- Enables human – computer interaction
- The field of NLP where computer tries to give meaning to words

NLG (Natural Language Generation)

- Generates words based on what has been understood by the computer
- As words may have multiple meanings, NLG generates words based on the context of the words prior
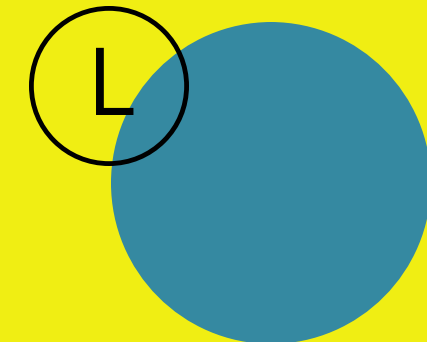
General processes for NLP

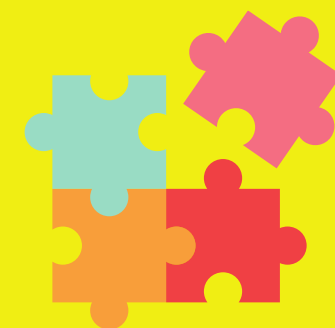Takenization

Stemming

Lemmatization

POS Tags

Named Entity Recognition

Chunking

Tokenization

Dividing a sentence based on each word
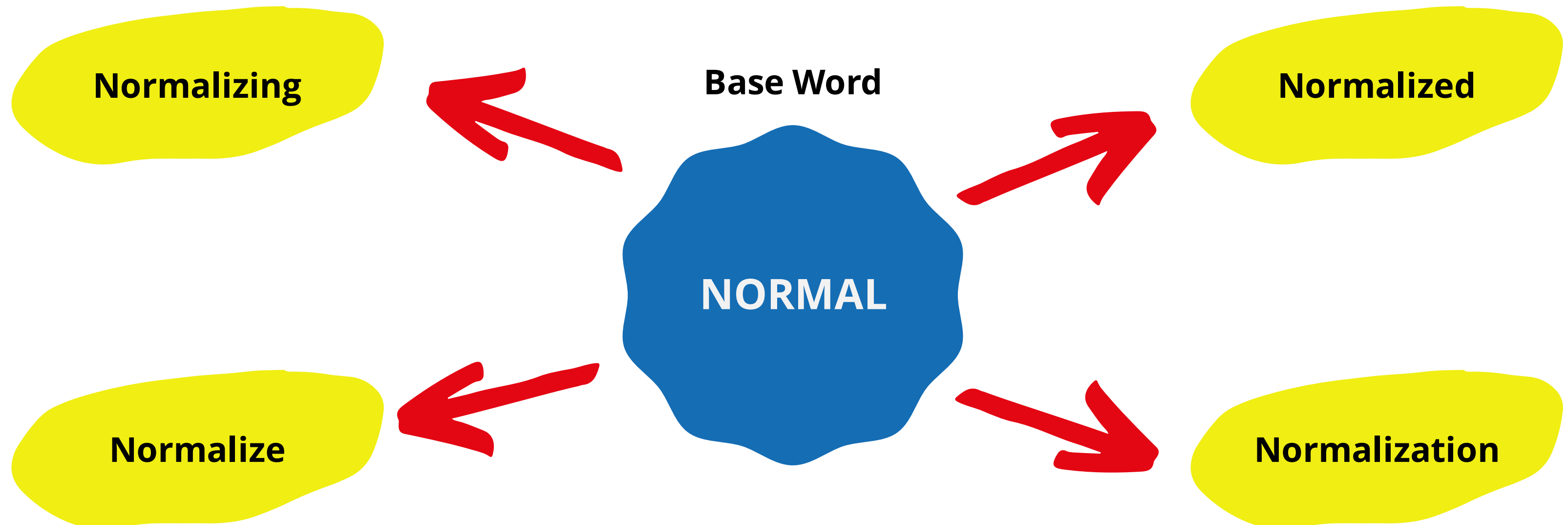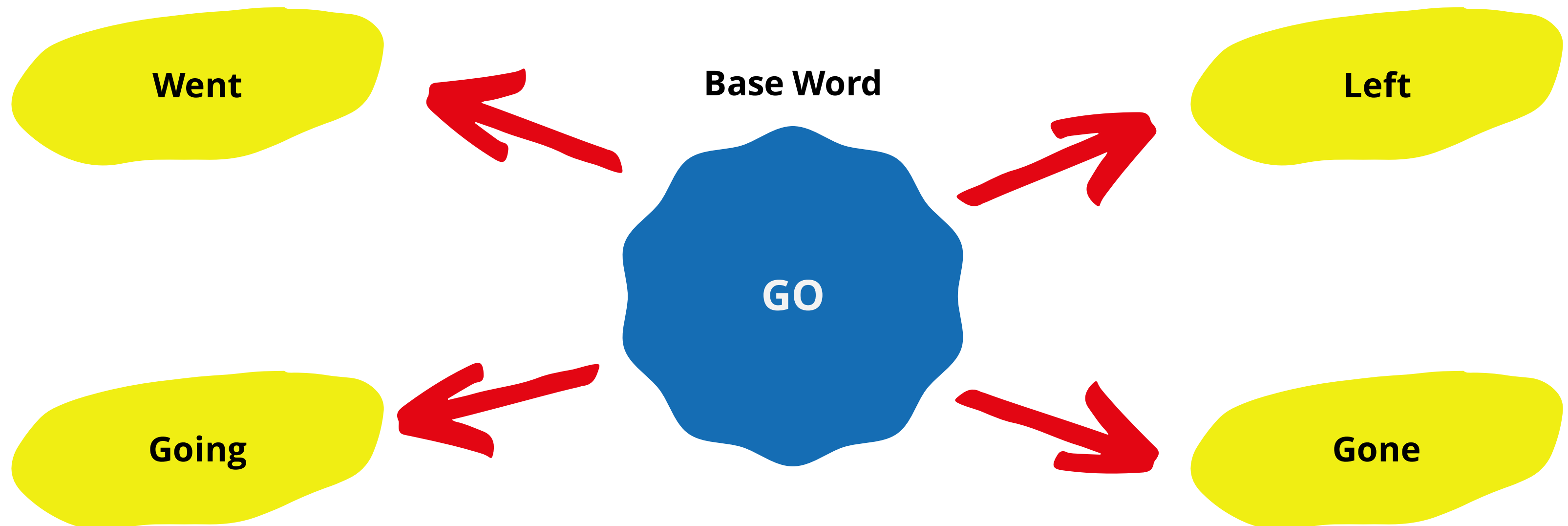
| Dividing | a | sentence | based | on | each | word |

Stemming

Normalizing a word into its base form

Lemmatization

Maps words that inflict the same meaning

Went

Left

Base Word

GO

Going

Gone

POS Tags

Tagging a word based on its grammatical part of speech

WORD
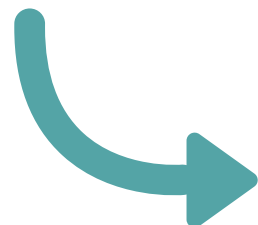
PARTS OF SPEECH
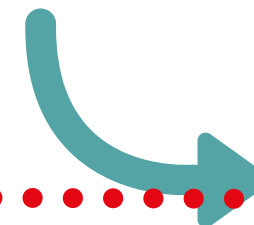
- Jakarta .... • Noun
- Dance .... • Verb
- Closely .... • Adverb
- Beautiful .... • Adjective

Named Entity Recognition

Process of detecting a named entity such as location or person

"President Jokowi is heading over to Bogor in a car"

WORD

ENTITY

- President .... • Job

- Jokowi .... • Person

- Bogor .... • Location

- Car .... • Vehicle

Chunking

Picking up individual information and combining them into bigger pieces

"Budi likes Football"

| WORD	BASE WORD	POS	ENTITY
Budi	Budi	Noun	Person
likes	likes	Verb	-
Football	Football	Noun	Activity

What do we do with the information?

- Computers are digital meaning information only has meaning when it is defined
- Computers cannot add 1+1 if the values behind the numbers are not previously defined
- How does the computer use this information to do things like the google search bar does?
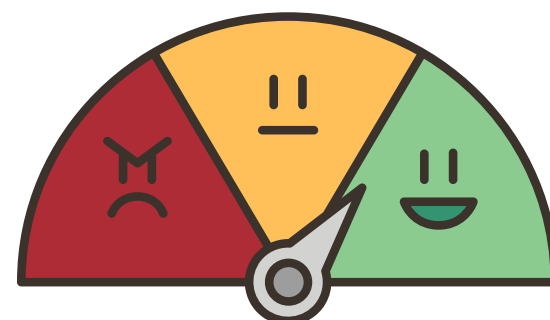
How does a computer understand words?
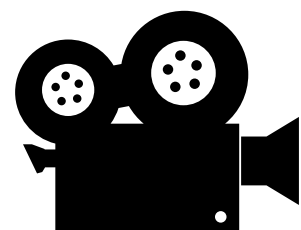
MoviE

The Dark Knight - 2008

R E V I E W

This may seem like faint praise, but about the highest compliment I can give Christopher Nolan's The Dark Knight right now is to say that there were many long stretches during which I didn't even realize it was a superhero movie

December 13, 2017 | Full Review...

Glenn Kenny
MUBI
★ TOP CRITIC

You will feel utterly numb after the screening of The Dark Knight. The film is bleak and brilliant.

May 31, 2017 | Full Review...

James Christopher
Times (UK)
★ TOP CRITIC

Among the great strengths of The Dark Knight is the way it combines hardboiled naturalism with the kind of stunts and setpieces you expect in summer blockbusters.

August 20, 2014 | Full Review...

Geoffrey Macnab
Independent (UK)
★ TOP CRITIC

Too much psychology and not enough pop. It's possible to be too serious, you know.

June 2, 2017 | Full Review...

Ryan Gilbey
New Statesman
★ TOP CRITIC

This is an impressive film in many ways, and Nolan directs with real confidence, yet the overall result feels cumbersome.

August 20, 2014 | Full Review...

Jonathan Romney
Independent on Sunday
★ TOP CRITIC

To talk of Heath Ledger's performance is difficult, merely because, as gigantic as it is, it is only one important feature of an intricately wrought canvas.

August 20, 2014 | Full Review...

Tiru Thiruvilangam
Little White Lies
★ TOP CRITIC

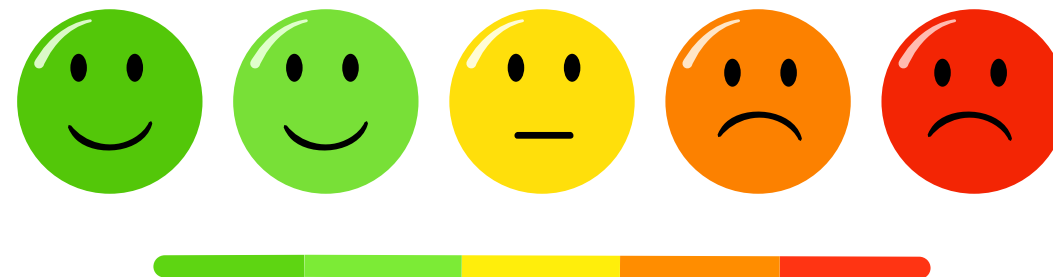# How does a computer understand words?
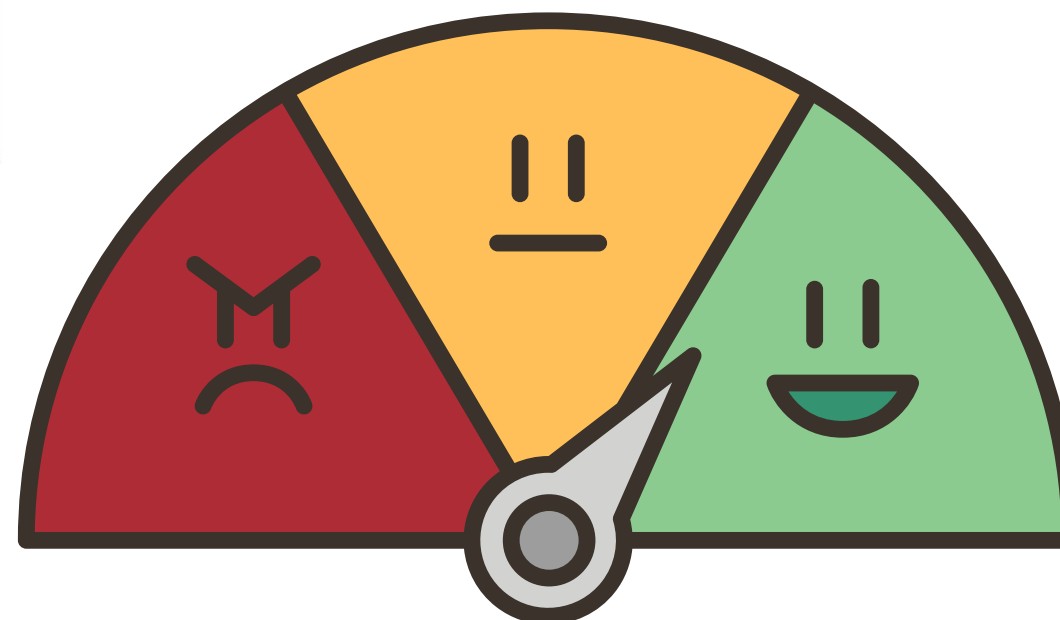
Sentiment Analysis

A technique in the field of NLP that tries to identify and extract opinions within a given text.

SENTIMENT ANALYSIS

Positive

Great service for an affordable price. We will definitely booking again.

Neutral

Just booked two nights at this hotel.

Negative

Horrible services. The room was dirty and unpleasant. Not worth the money.
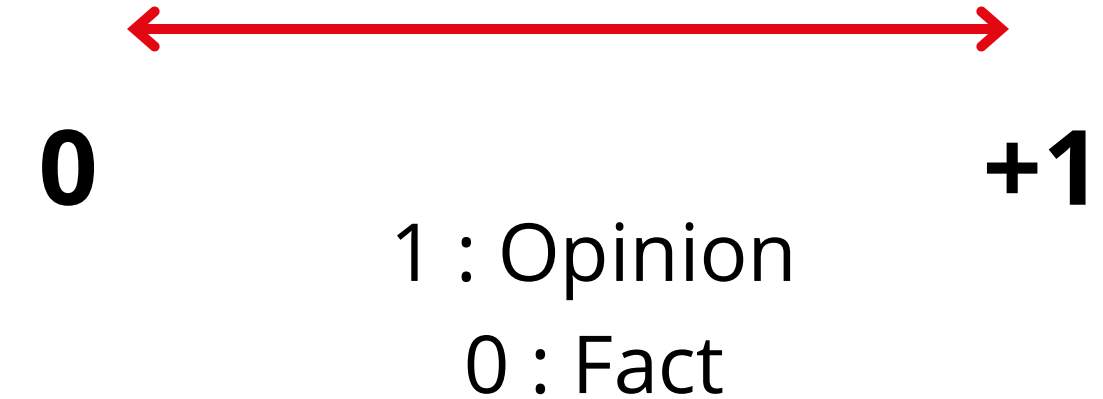
Sentiment Analysis

A combination of words can be reviewed as "Positive" or "Negative" based on the words that has been structured in the sentence. In general, every word has 2 parameters determining its positive or negative value which is

Polarity

Subjectivity

-1   **+1**

-1 : Negative

0 : Neutral

1 : Positive

0   **+1**

1 : Opinion

0 : Fact
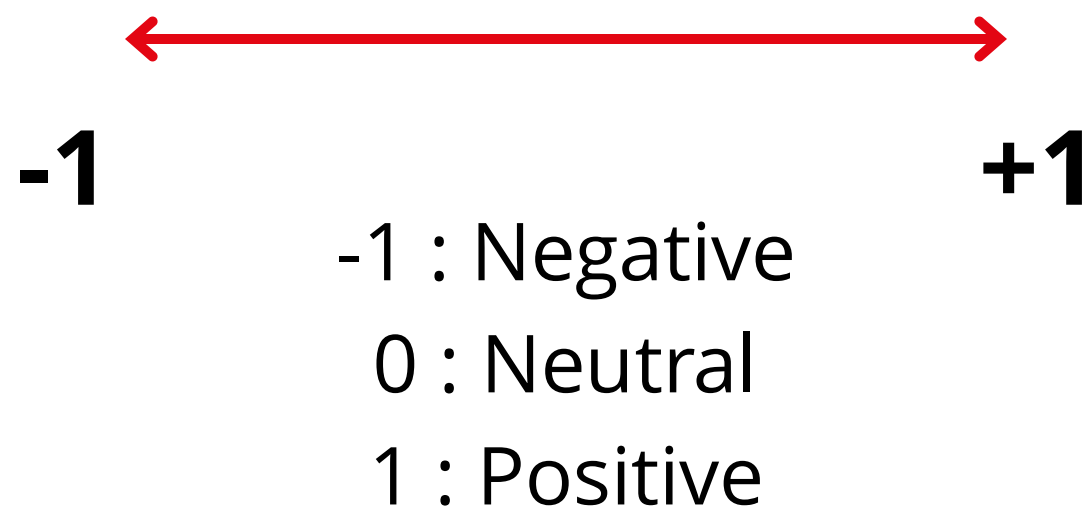
Sentiment Analysis

But some words can have multiple meanings such as the word great

Word Form	Sense	Polarity	Subjectivity
great	very good	1	1
great	of major significance or importance	1	1
great	relatively large in size or number or extent	0,4	0,2
great	remarkable or out of the ordinary in degree or magnitude of effect	0,8	0,8

But in the end, the computer averages the value from all variations of the word and gives it its final value

Sentiment Analysis

When words are combined, a final value of polarity and subjectivity is earned

Word Form	Polarity	Subjectivity
great	0,8	0,75
not great	0,4	0,75
great	1	0,75
I am great	0,8	0,75

"Not" divides polarity by 2

"Very" multiplies polarity by 1.3