3- Natural Language Processing (NLP)

By: eng. Esraa Madhi

What is Natural Language Processing (NLP)?

It is a subfield of artificial intelligence (AI) that f**ocuses on the interaction between computers and human (natural) languages.** NLP involves applying **algorithms to identify and extract the rules such that the unstructured language data is converted into a form that computers can understand.**

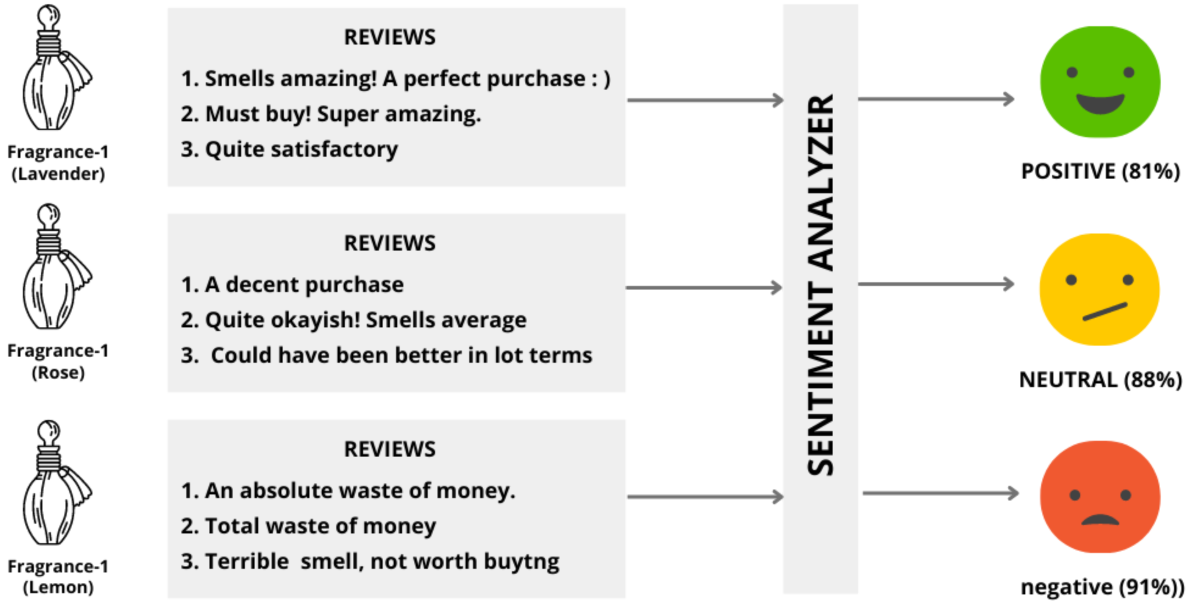
NLP Applications:

NLP tasks can be broadly categorized into two types → Understanding & Generation.

Understanding

This includes tasks where the computer system tries **to understand the input given by the user.**

1. **Sentiment Analysis:** Determining the emotional tone behind a series of words, used to gain an understanding of the attitudes, opinions and emotions expressed.



REVIEWS

1. Smells amazing! A perfect purchase :)
2. Must buy! Super amazing.
3. Quite satisfactory

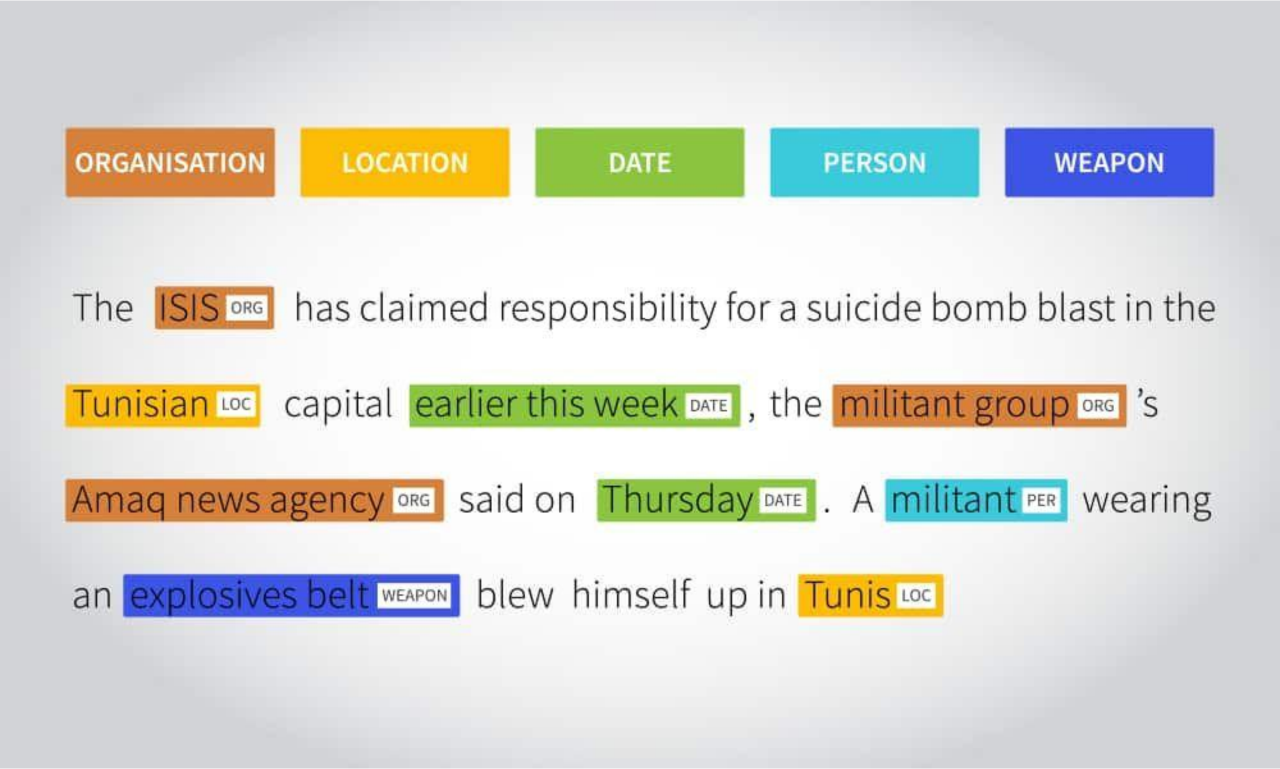
REVIEWS

1. A decent purchase
2. Quite okayish! Smells average
3. Could have been better in lot terms

REVIEWS

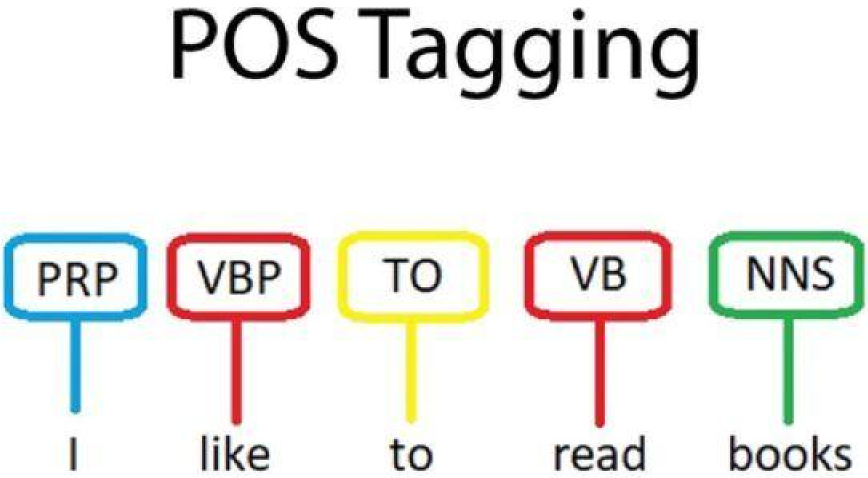
1. An absolute waste of money.
2. Total waste of money
3. Terrible smell, not worth buytng

2. **Named Entity Recognition (NER):** Identifying and classifying named entities mentioned in text into predefined categories such as the names of persons, organizations, locations, expressions of times, quantities, monetary values, percentages, etc.

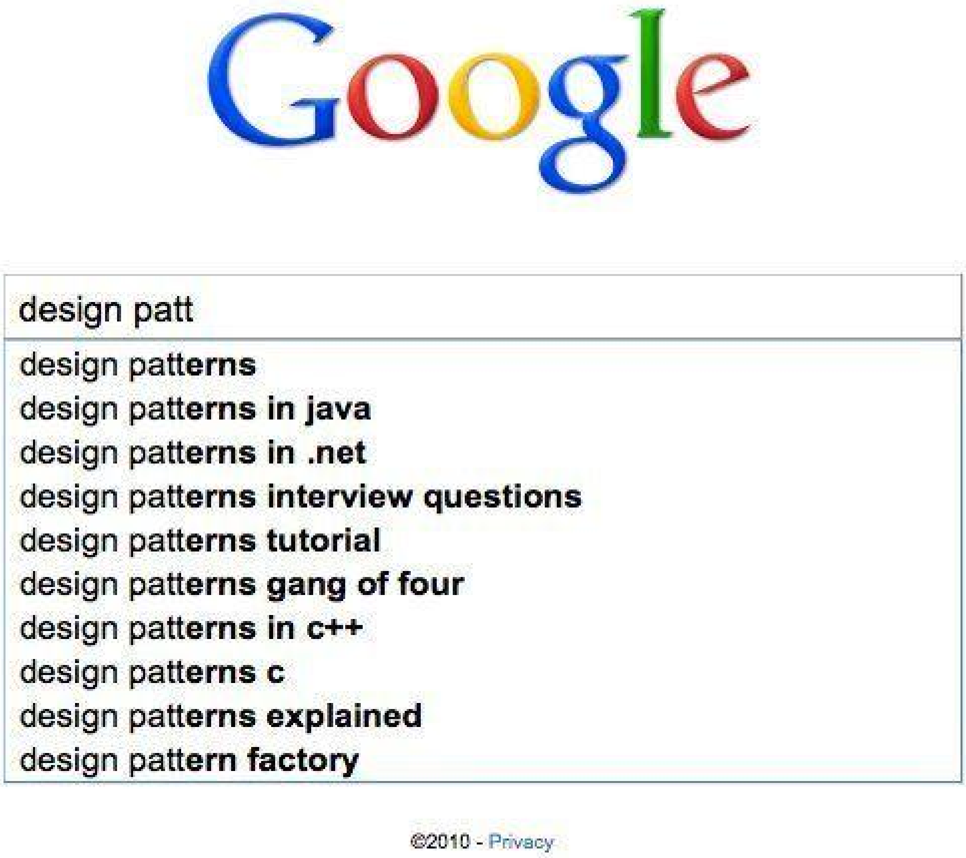


The ISIS ORG has claimed responsibility for a suicide bomb blast in the Tunisian LOC capital earlier this week DATE , the militant group ORG 's Amaq news agency ORG said on Thursday DATE . A militant PER wearing an explosives belt WEAPON blew himself up in Tunis LOC

3. **Part-of-Speech Tagging:** Assigning word types to each word in a sentence, such as noun, verb, adjective, etc.



4. **Language Modeling**: Predicting the probability of a sequence of words. This is used in applications like auto-complete or text generation.

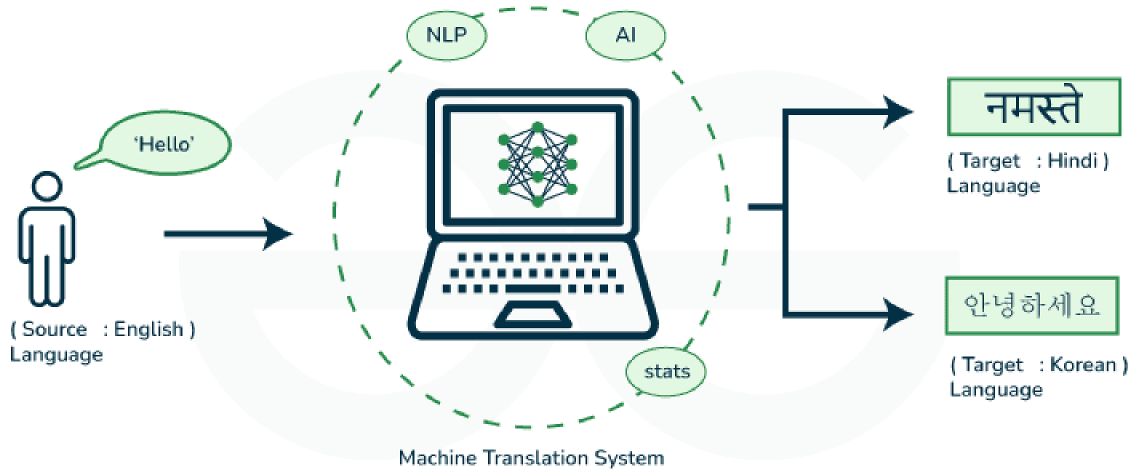


design patt

design patt**erns**
design patt**erns in java**
design patt**erns in .net**
design patt**erns interview questions**
design patt**erns tutorial**
design patt**erns gang of four**
design patt**erns in c++**
design patt**erns c**
design patt**erns explained**
design patt**ern factory**

©2010 - Privacy

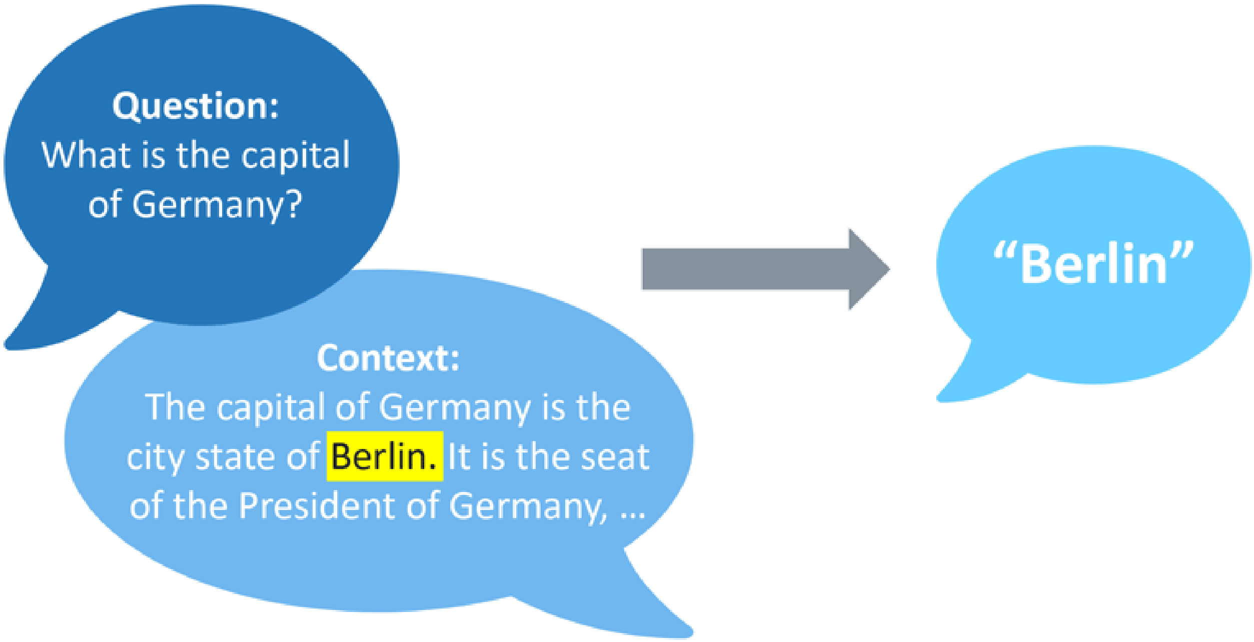
5. **Machine Translation**: Translating text or speech from one language to another.


'Hello'
(Source : English) Language
NLP
AI
stats
Machine Translation System
नमस्ते
(Target : Hindi) Language
안녕하세요
(Target : Korean) Language

Machine Translation Model

6. **Speech Recognition:** Translating spoken language into text.

7. **Question Answering:** Building systems that automatically answer questions posed by humans in a natural language.



Question: What is the capital of Germany?

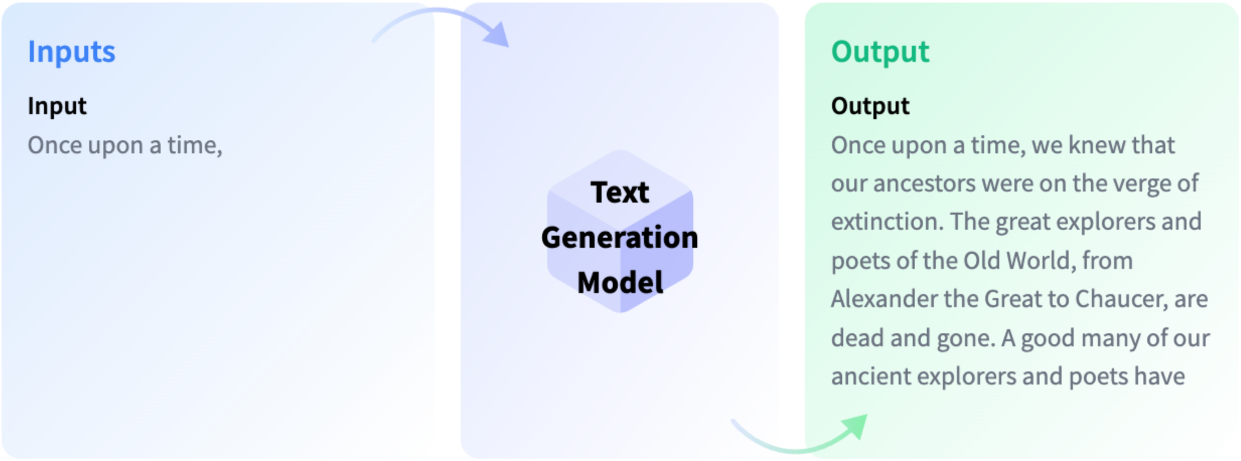
Context: The capital of Germany is the city state of Berlin. It is the seat of the President of Germany, ...

"Berlin"

Generation

This includes tasks where the computer system generates text on its own.

1. **Text Generation**: Producing text with similar characteristics to a human-written text. This can be in the form of narratives, responses to questions, or translations.



Inputs

Input

Once upon a time,

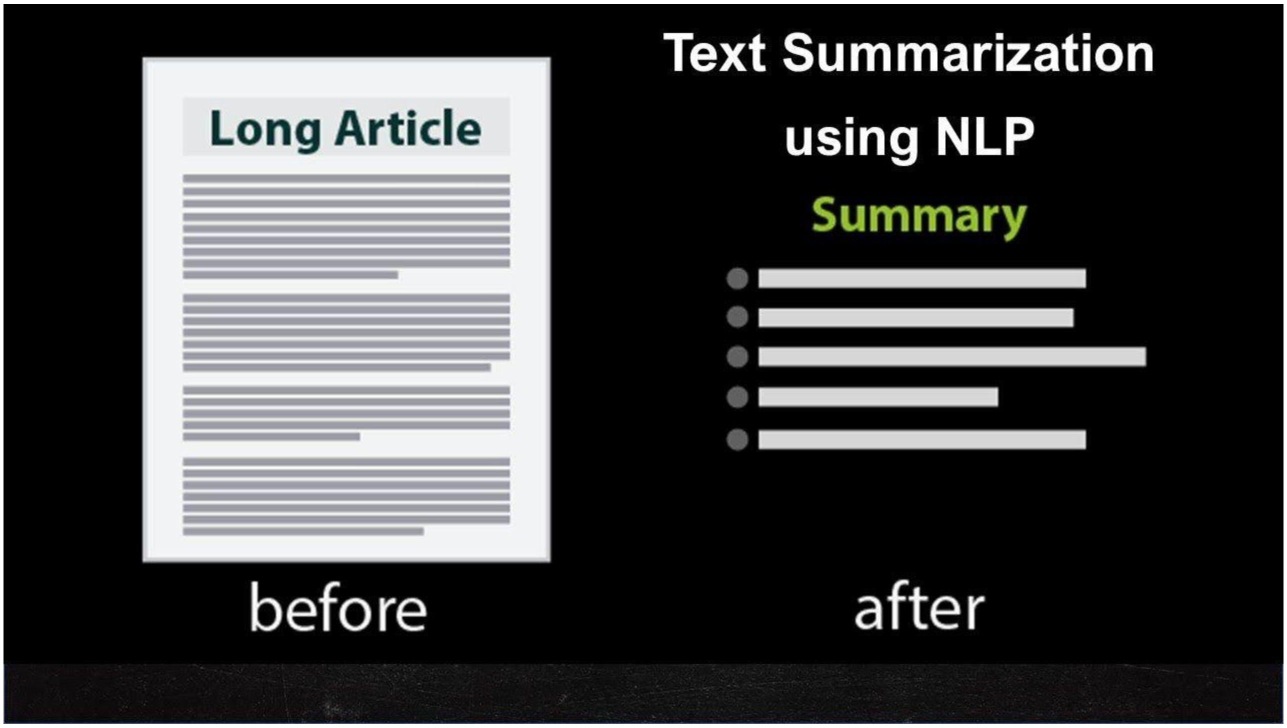
Text Generation Model

Output

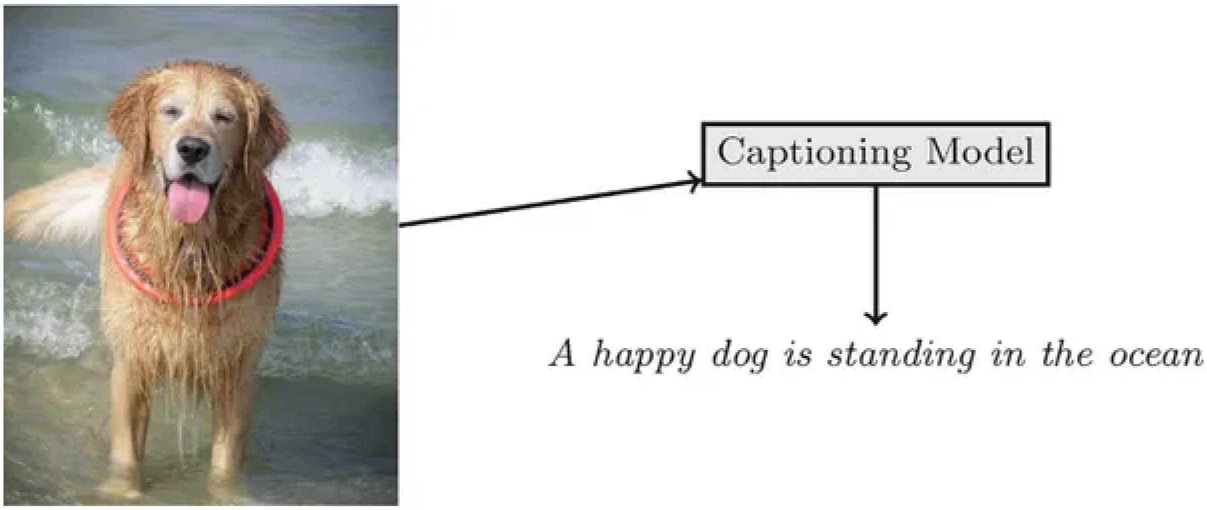
Output

Once upon a time, we knew that our ancestors were on the verge of extinction. The great explorers and poets of the Old World, from Alexander the Great to Chaucer, are dead and gone. A good many of our ancient explorers and poets have

2. **Summarization**: Creating a short and coherent version of a longer document while retaining the key information and overall meaning.



3. **Image Captioning**: Generating a textual description of an image.



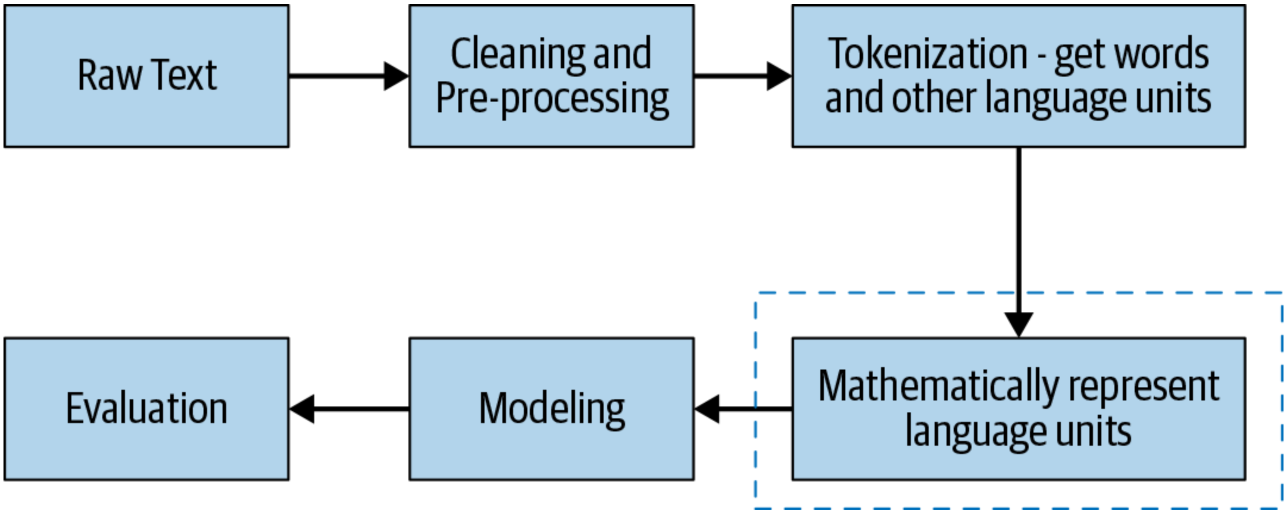
Challenges in NLP:

NLP is a difficult field due to the complexity of human languages, including:

- **Ambiguity:** Words and sentences can have multiple meanings.
  - Ex: الساعة
- **Contextual Meaning:** The meaning of a sentence can change depending on the context in which it is used.
  - Ex: عطوني عين
- **Sarcasm:** These can be very difficult for a computer system to detect and interpret.
  - Ex: اليوم الجو رائع (وهو بالحقيقة حر قوايل)

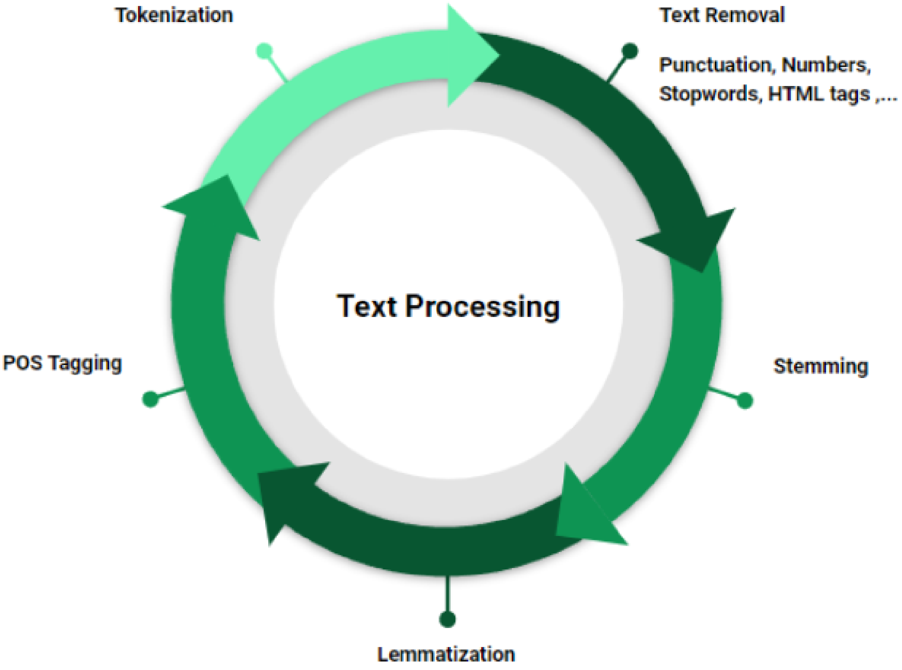
Main steps in any NLP task (Handle Text Problem):

https://youtu.be/CMrHM8a3hqw


Raw Text
Cleaning and Pre-processing
Tokenization - get words and other language units
Mathematically represent language units
Modeling
Evaluation

1. **Text Preprocessing**

Text Processing is an essential task in NLP as it helps to clean and transform raw data into a suitable format used for analysis or modeling.



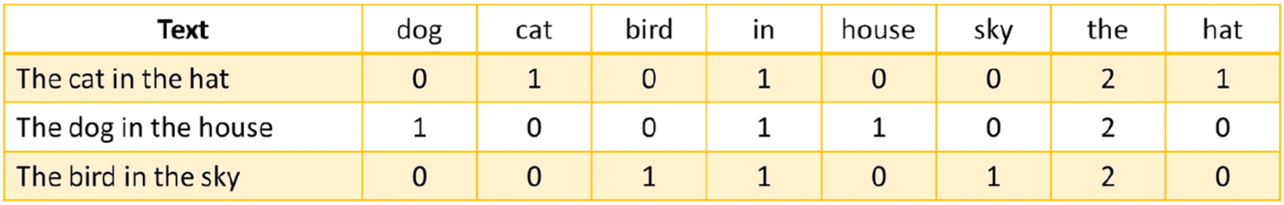
- Resources:

    - https://www.codecademy.com/learn/dsnlp-text-preprocessing/modules/nlp-text-preprocessing/cheatsheet
    - https://www.analyticsvidhya.com/blog/2021/06/text-preprocessing-in-nlp-with-python-codes/
    - https://medium.com/@maleeshadesilva21/preprocessing-steps-for-natural-language-processing-nlp-a-beginners-guide-d6d9bf7689c9
    - https://www.geeksforgeeks.org/text-preprocessing-in-python-set-1/
    - https://docs.cohere.com/docs/text-pre-processing-in-nlp
    - https://towardsdatascience.com/machine-learning-text-processing-1d5a2d638958

2. **Feature extraction form text (text representation)**
   Feature extraction where textual data is transformed into a numerical or symbolic format that can be used by machine learning algorithms. It is away to convert unstructured text data into a structured numeric format.

- **Bag of words:**

  Each word in the text is considered a feature, and the number of times a particular word appears in the text is used to represent the importance of that word in the text. Disregarding grammar and word order but keeping track of the frequency of each word.



Text	dog	cat	bird	in	house	sky	the	hat
The cat in the hat	0	1	0	1	0	0	2	1
The dog in the house	1	0	0	1	1	0	2	0
The bird in the sky	0	0	1	1	0	1	2	0

Is it a good representation?
Cons:

a. **Semantic Ignorance**: Ignores word order and context, losing important semantic information.
b. **Sparsity and High Dimensionality:** Produces sparse and high-dimensional feature vectors, which can be computationally inefficient.

c. **No Phrase Recognition:** Fails to capture phrases and multi-word expressions, treating every word independently.

- **N-gram:**
  An N-gram is a conventional method for representing text, where the text is divided into continuous sequences of n words. A uni-gram consists of individual words from a sentence. Bi-grams are created from pairs of consecutive words, while tri-grams are formed from sequences of three consecutive words, and the pattern continues similarly for higher values of n.

This is Big Data AI Book

Uni-Gram	This	Is	Big	Data	AI	Book
Bi-Gram	This is	Is Big	Big Data	Data AI	AI Book	
Tri-Gram	This is Big	Is Big Data	Big Data AI	Data AI Book		

Is it a good representation?
Cons:

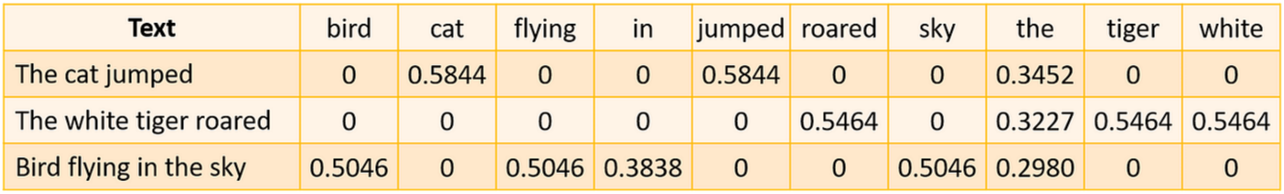
a. **Dimensionality**: Increases the feature space exponentially with the size of N, leading to higher computational costs.

b. **Sparsity and High Dimensionality:** Produces sparse and high-dimensional feature vectors, which can be computationally inefficient.

c. **Limited Context:** Despite capturing more context than single words, N-grams still offer a limited view, missing broader sentence-level or paragraph-level contexts.

- **TF-IDF**
  TF-IDF stands for Term Frequency-Inverse Document Frequency. The idea behind TF-IDF is to weight words based on how often they appear in a document (the term frequency) and how common they are across all documents (the inverse document frequency).

$$\textbf{TF-IDF} = \text{Term frequency in document} \;\; X \;\; \log\left(\frac{\text{Total number of documents}}{\text{Number of documents containing the term}} \right)$$



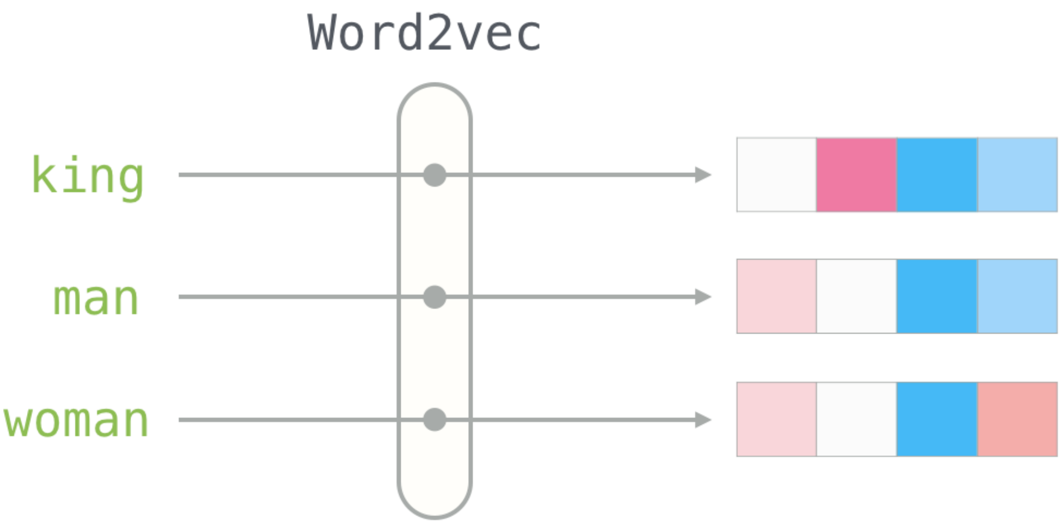
Text	bird	cat	flying	in	jumped	roared	sky	the	tiger	white
The cat jumped	0	0.5844	0	0	0.5844	0	0	0.3452	0	0
The white tiger roared	0	0	0	0	0	0.5464	0	0.3227	0.5464	0.5464
Bird flying in the sky	0.5046	0	0.5046	0.3838	0	0	0.5046	0.2980	0	0

Example: $\text{tf-idf(cat)} = \frac{1}{3} \log\left(\frac{3}{1}\right)$

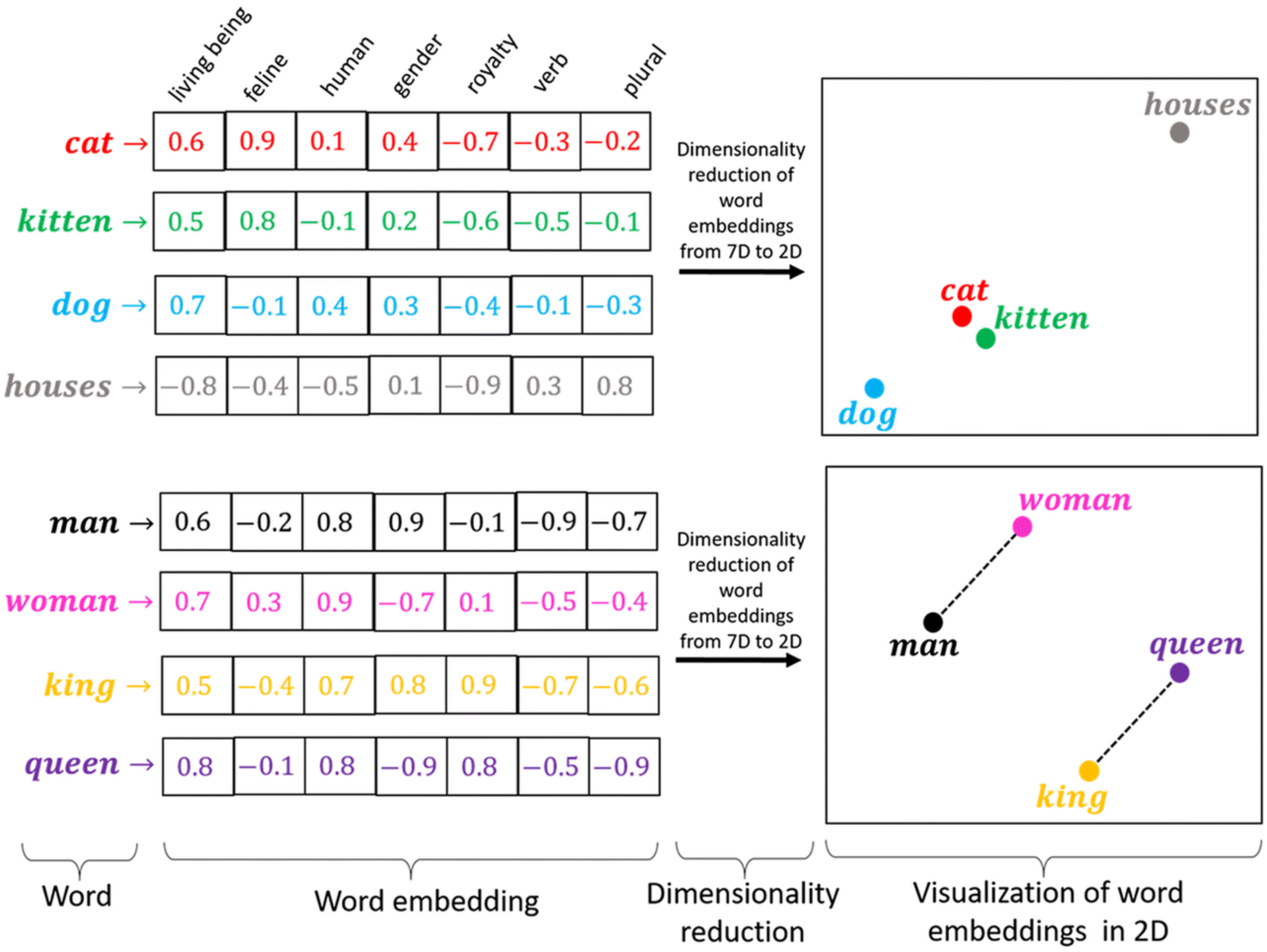
Is it a good representation?
Cons:

a. **Semantic Ignorance**: Ignores word order and context, losing important semantic information.
b. **Sparsity and High Dimensionality:** Produces sparse and high-dimensional feature vectors, which can be computationally inefficient.

- **Word embedding:**
  Word embedding represents each word as a dense vector of real numbers



The similar or closely related words are nearer to each other in the vector space.



	living being	feline	human	gender	royalty	verb	plural
cat	0.6	0.9	0.1	0.4	−0.7	−0.3	−0.2
kitten	0.5	0.8	−0.1	0.2	−0.6	−0.5	−0.1
dog	0.7	−0.1	0.4	0.3	−0.4	−0.1	−0.3
houses	−0.8	−0.4	−0.5	0.1	−0.9	0.3	0.8
man	0.6	−0.2	0.8	0.9	−0.1	−0.9	−0.7
woman	0.7	0.3	0.9	−0.7	0.1	−0.5	−0.4
king	0.5	−0.4	0.7	0.8	0.9	−0.7	−0.6
queen	0.8	−0.1	0.8	−0.9	0.8	−0.5	−0.9

Dimensionality reduction of word embeddings from 7D to 2D
houses
cat
kitten
dog
Dimensionality reduction of word embeddings from 7D to 2D
woman
man
queen
king
Word
Word embedding
Dimensionality reduction
Visualization of word embeddings in 2D

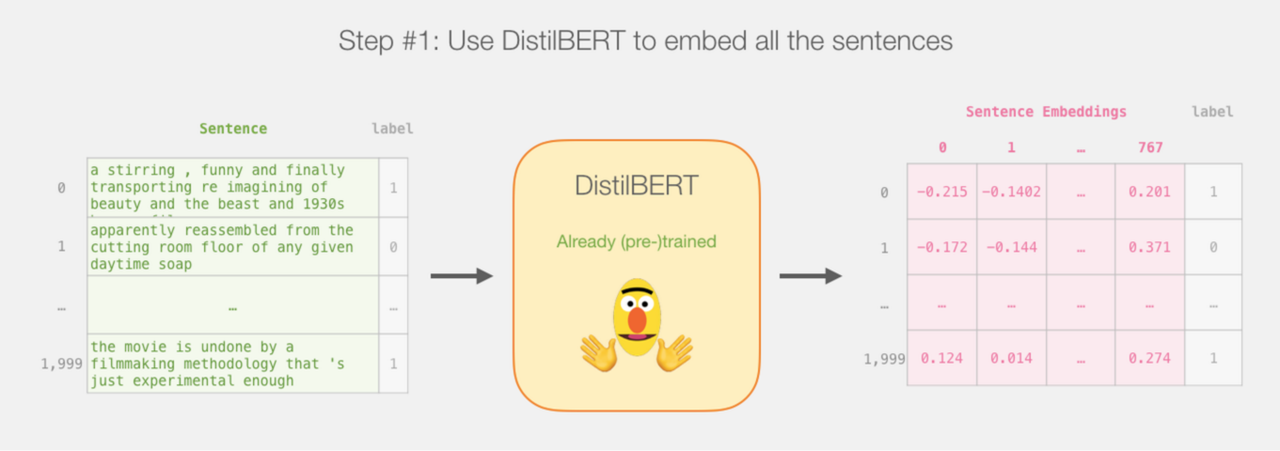
Demo: https://jalammar.github.io/illustrated-word2vec/

Resources for text representation:

- https://www.scaler.com/topics/nlp/text-representation-in-nlp/
- https://deysusovan93.medium.com/from-traditional-to-modern-a-comprehensive-guide-to-text-representation-techniques-in-nlp-369946f67497
- https://towardsdatascience.com/introduction-to-text-representations-for-language-processing-part-1-dc6e8068b8a4
- https://www.analyticsvidhya.com/blog/2022/02/machine-learning-techniques-for-text-representation-in-nlp/

- https://python.plainenglish.io/text-representation-in-natural-language-processing-nlp-23b44c9ca31f
- https://towardsdatascience.com/an-overview-for-text-representations-in-nlp-311253730af1
- https://medium.com/@mervebdurna/text-representation-techniques-d40741eb0916

- **Sentence embedding: BERT and Transformer Models:**
  It is similar to that of word embedding, the only difference is in place of a word, a sentence is represented as a numerical vector in a high-dimensional space. The goal of sentence embedding is to capture the meaning and semantic relationships between words in a sentence, as well as the context in which the sentence is used.



Step #1: Use DistilBERT to embed all the sentences

	Sentence	label
0	a stirring , funny and finally transporting re imagining of beauty and the beast and 1930s	1
1	apparently reassembled from the cutting room floor of any given daytime soap	0
…	…	…
1,999	the movie is undone by a filmmaking methodology that 's just experimental enough	1

	Sentence Embeddings				label
	0	1	…	767	
0	-0.215	-0.1402	…	0.201	1
1	-0.172	-0.144	…	0.371	0
…	…	…	…	…	…
1,999	0.124	0.014	…	0.274	1

Pros:

- It capture both the meaning of the individual tokens and their contextual relationships within the sentence.

3. **Machine learning models**

- Language Model:
  - https://jalammar.github.io/illustrated-word2vec/
- Text Classification:
  - https://www.datacamp.com/tutorial/text-classification-python

    - https://jalammar.github.io/a-visual-guide-to-using-bert-for-the-first-time/
    - https://github.com/EsraaMadi/Fewshot-text-classification-pipeline

---

Resources:

- https://www.linkedin.com/advice/3/how-do-you-evaluate-performance-machine-learning-model-gso2f
- https://www.linkedin.com/pulse/text-preprocessing-natural-language-processing-nlp-germec-phd/
- https://www.youtube.com/watch?v=6I-Alfkr5K4
- https://www.analyticsvidhya.com/blog/2022/02/machine-learning-techniques-for-text-representation-in-nlp/
- https://towardsdatascience.com/introduction-to-text-representations-for-language-processing-part-1-dc6e8068b8a4
- https://www.scaler.com/topics/nlp/text-representation-in-nlp/
- https://youtu.be/RO4Ip6pcBCk?si=nTOg6ICWO4HEr8Ck
- https://youtu.be/zLMEnNbdh4Q?si=TxB9A6JG0bJlv5bC
- https://deysusovan93.medium.com/from-traditional-to-modern-a-comprehensive-guide-to-text-representation-techniques-in-nlp-369946f67497
- https://medium.com/@mervebdurna/text-representation-techniques-d40741eb0916
- https://towardsdatascience.com/an-overview-for-text-representations-in-nlp-311253730af1