Natural Language Processing (NLP)

By

Dr. Ravirajsinh S. Vaghela

Assistant Professor, SCSFD.

NFSU,Gandhinagar

Introduction

- Natural Language Processing (NLP) is a subfield of machine learning which leverages analysis, generation, and understanding of human languages to derive meaningful insights.

- NLP is becoming popular as Large Language Models (LLMs) are growing and used widely in the market. Having foundational knowledge of NLP concepts and techniques can help you become an NLP data scientist, NLP engineer, or distinguished ML engineer to stand out in the job market.

NLP topic to understand

❑Text pre-processing

❑Text feature extraction

❑Text sort

❑Named entity recognition

❑Parts-of-speech tagging

❑Text generation

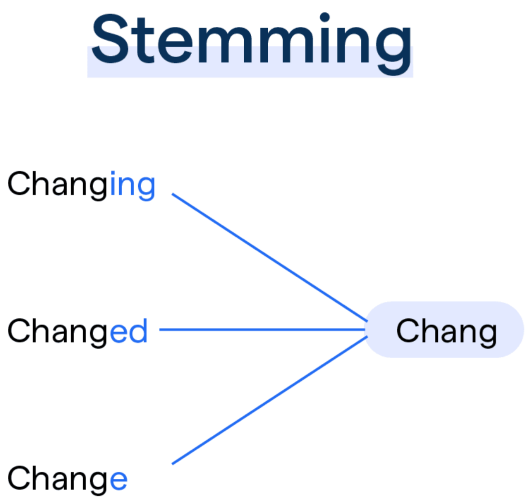
❑Text-to-speech and speech-to-text techniques

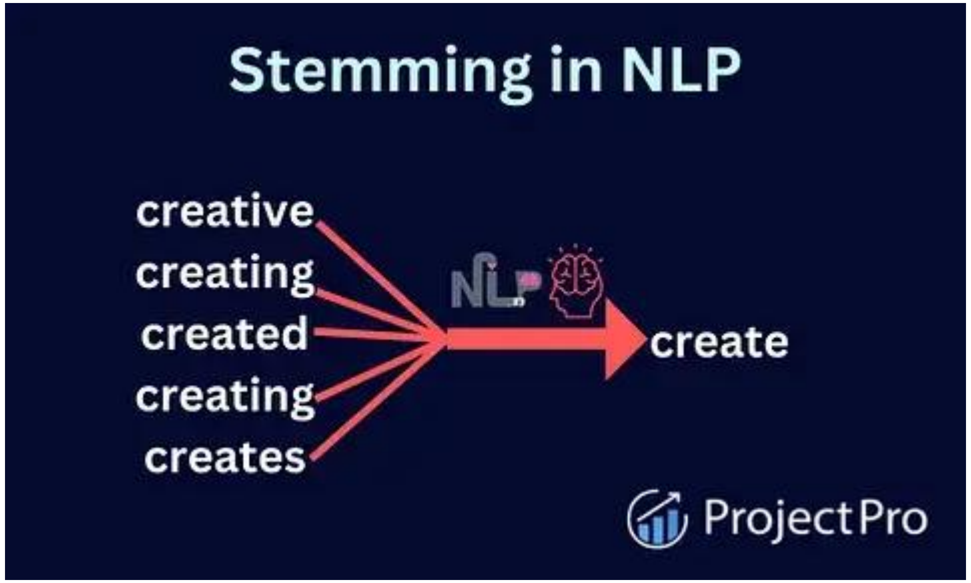
Text pre-processing

- Pre-processing techniques such as tokenization, stemming, and lemmatization, help convert raw text into a format that can be easily analyzed.
- Tokenization
- The fundamental concept in NLP is tokenization.
- It is the process of breaking down a complex piece of text into smaller units called tokens.

- Stemming

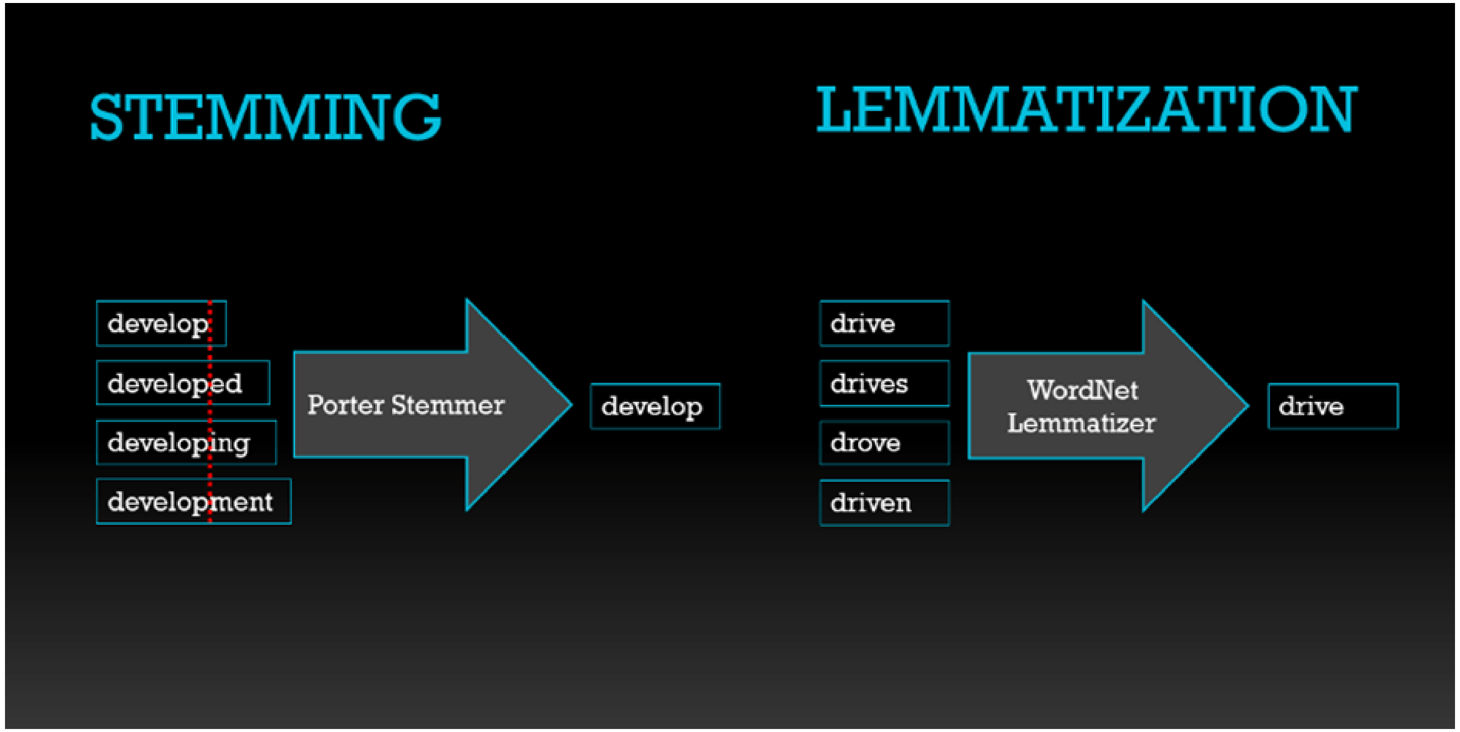
- Stemming is the process of reducing words to their base or root form. This can be useful in classification or information retrieval tasks.



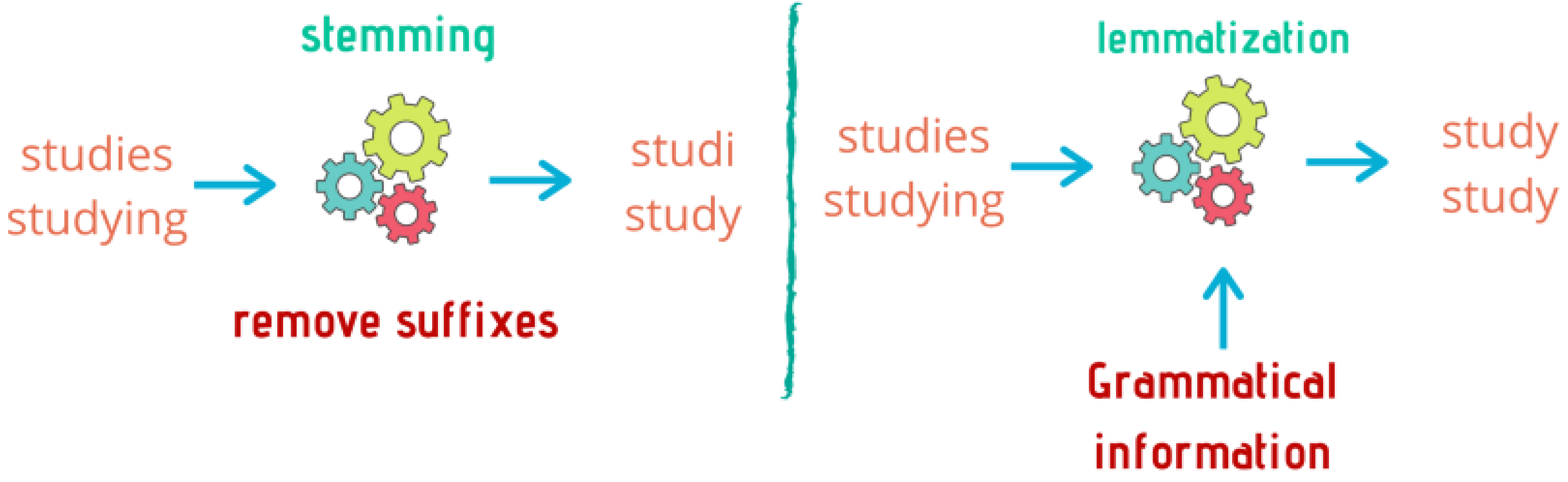


- Lemmatization

- Lemmatization is the process of reducing a word to its base or root form, which is known as the lemma. It is a more sophisticated version of stemming, as it takes into account the context and the part of speech of the word.



STEMMING VS. LEMMATIZATION



stemming

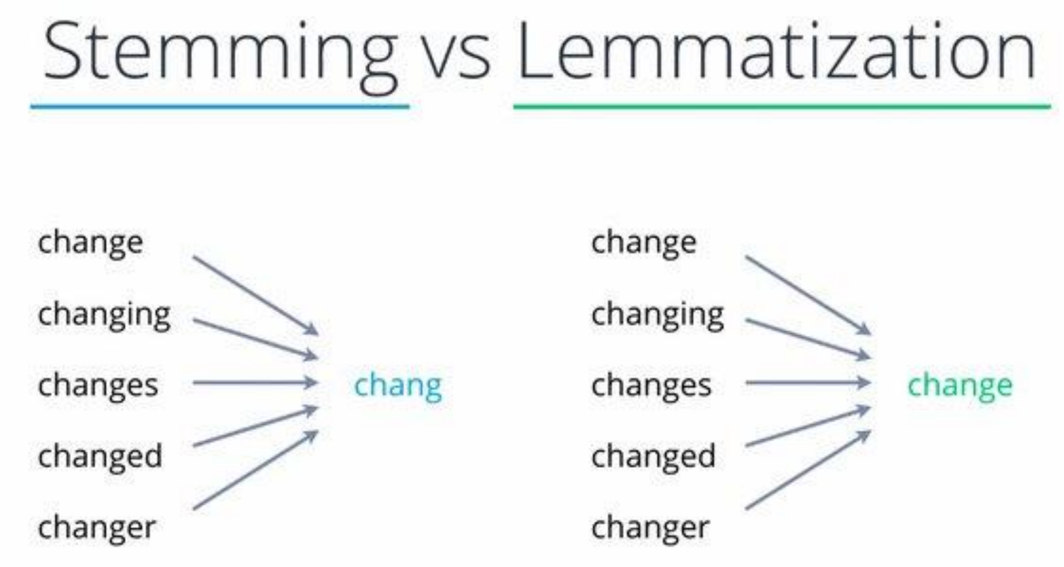
studies
studying → studi
study

remove suffixes

lemmatization

studies
studying → study
study

Grammatical information



Stemming vs Lemmatization

change
changing
changes → chang
changed
changer

change
changing
changes → change
changed
changer

Text feature extraction

Techniques such as

a) Vocabulary/bag-of-words,
b) n-grams,
c) count vectorization, and
d) word embeddings

are used to represent text as numerical features in machine learning models.

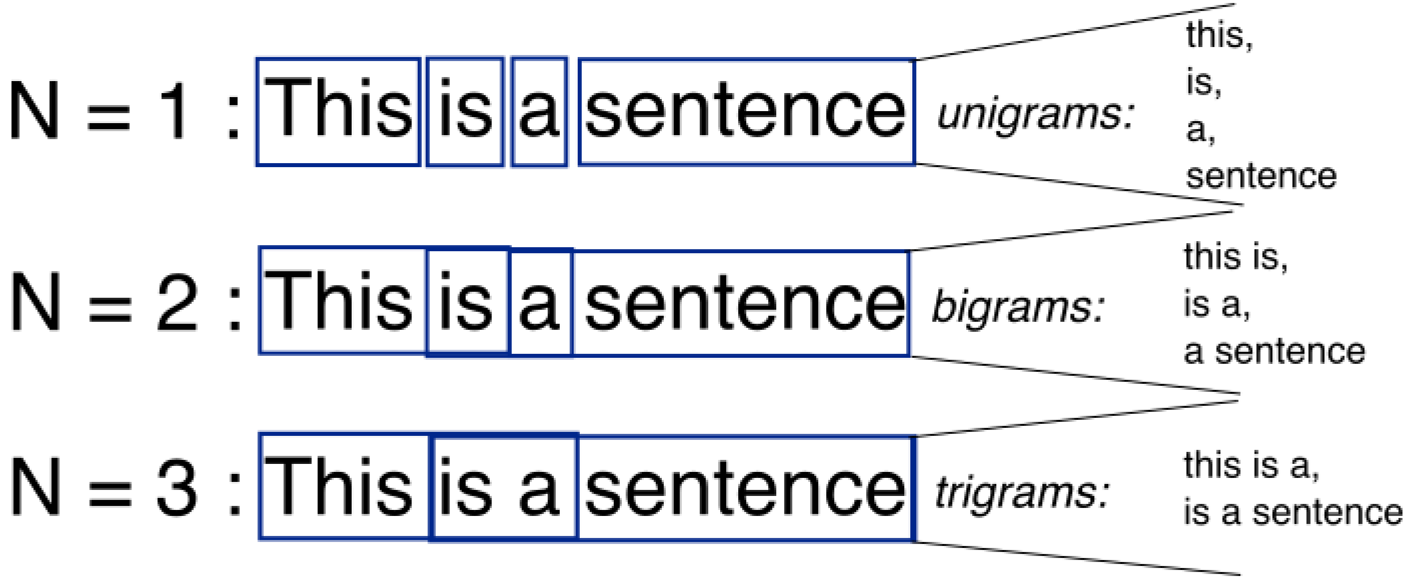
Vocabulary or Bag-of-Words

Vocabulary in NLP refers to the set of unique words or tokens in a given text or corpus.

N-grams

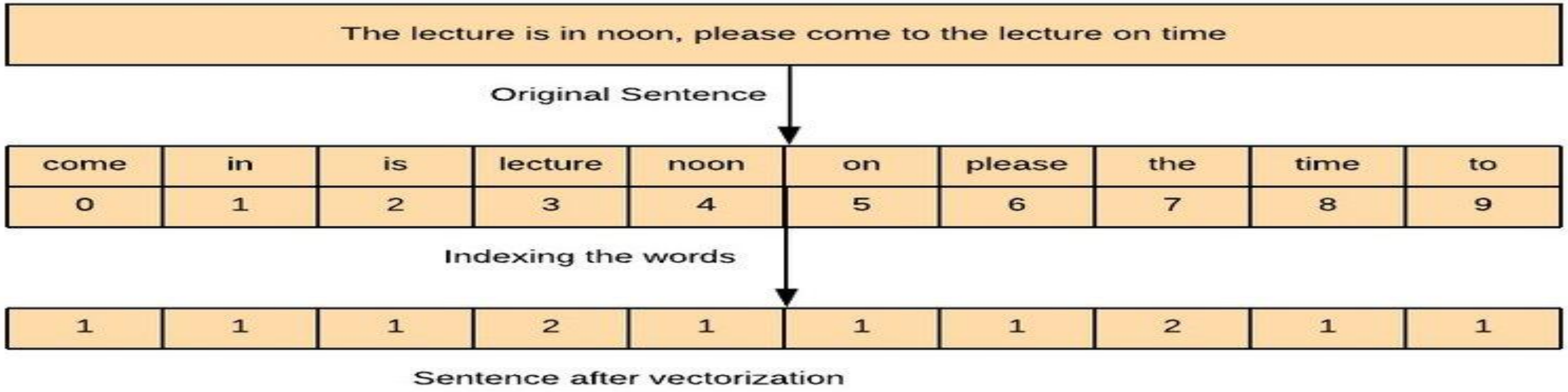
- An n-gram is a contiguous sequence of n items from a given sample of text or speech, where n can be any positive integer. In NLP, n-grams are often used to capture the context of words in a text.

N-grams are simply all combinations of adjacent words or letters of length n that you can find in your source text.



- Count vectorization

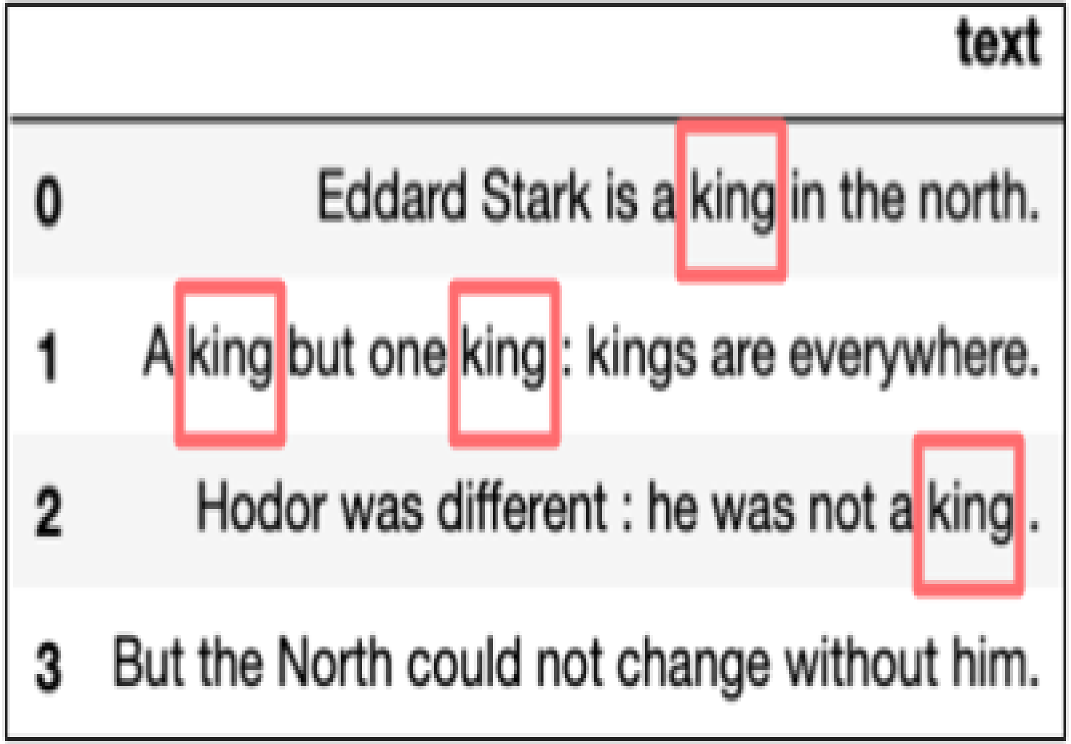
- Text vectorization is the process of converting text data into numerical vectors, which can be used as input for machine learning models.
- One of the most common techniques for text vectorization is bag-of-words, which represents text as a bag (or multiset) of its words, disregarding grammar and word order but keeping track of the number of occurrences of each word.



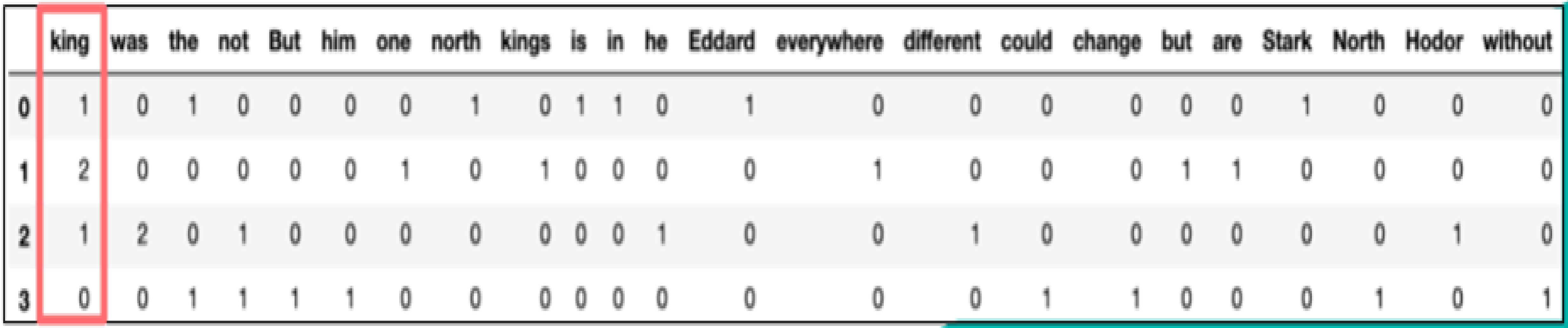
The lecture is in noon, please come to the lecture on time

come	in	is	lecture	noon	on	please	the	time	to
0	1	2	3	4	5	6	7	8	9

1	1	1	2	1	1	1	2	1	1



	text
0	Eddard Stark is a king in the north.
1	A king but one king : kings are everywhere.
2	Hodor was different : he was not a king .
3	But the North could not change without him.



	king	was	the	not	But	him	one	north	kings	is	in	he	Eddard	everywhere	different	could	change	but	are	Stark	North	Hodor	without
0	1	0	1	0	0	0	0	1	0	1	1	0	1	0	0	0	0	0	0	1	0	0	0
1	2	0	0	0	0	0	1	0	1	0	0	0	0	1	0	0	0	1	1	0	0	0	0
2	1	2	0	1	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	1	0
3	0	0	1	1	1	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0	1	0	1

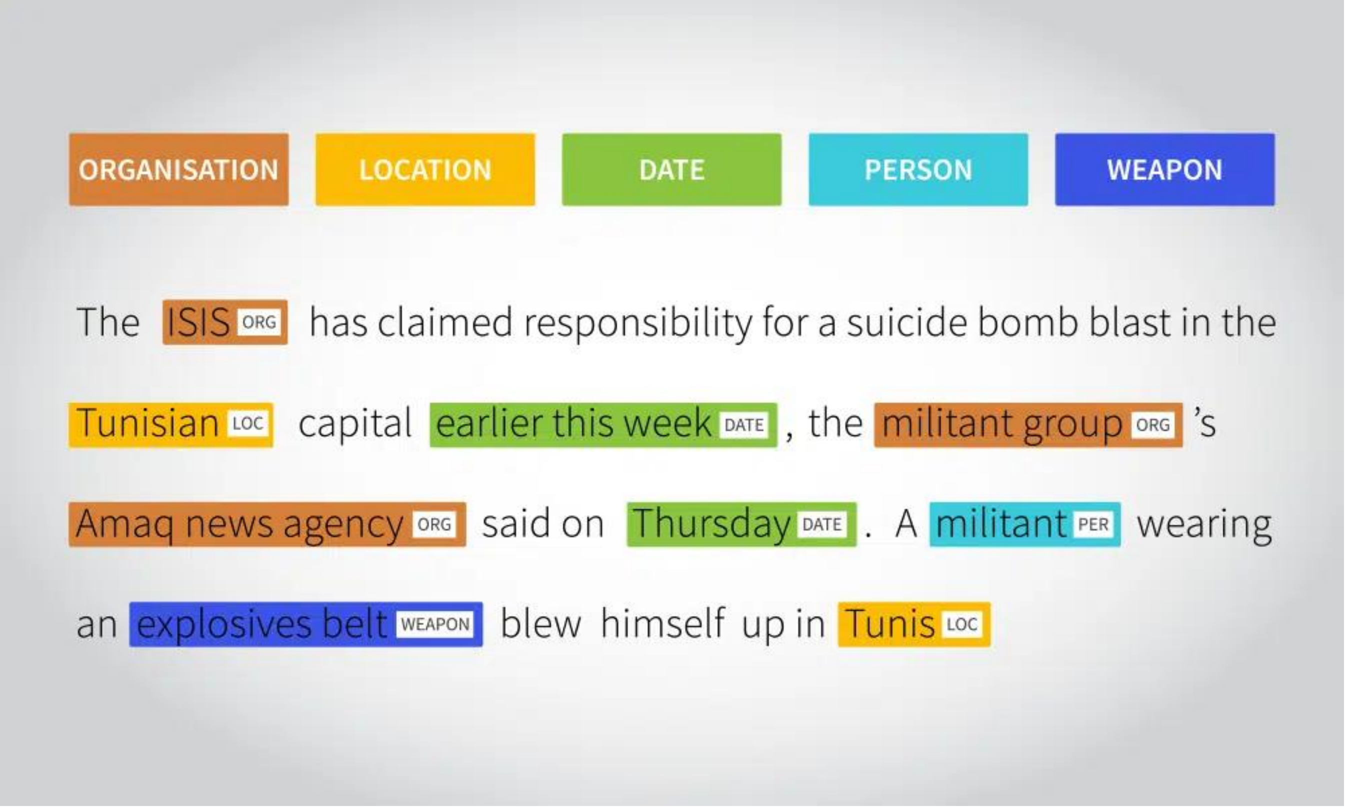
Text sort

- Techniques for classifying text into predefined categories, such as sentiment analysis and spam detection.
- Text sorting is one of the most important NLP tasks that involves assigning predefined categories or labels to a given piece of text.

Named entity recognition

- Named entity recognition are techniques for identifying and extracting named entities from text, such as people, organizations, and locations.
- An entity can be thought of as a category type present in a given text. For example, the name of a certain personality, the name of an organization, location, etc.

- https://spacy.io/usage



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

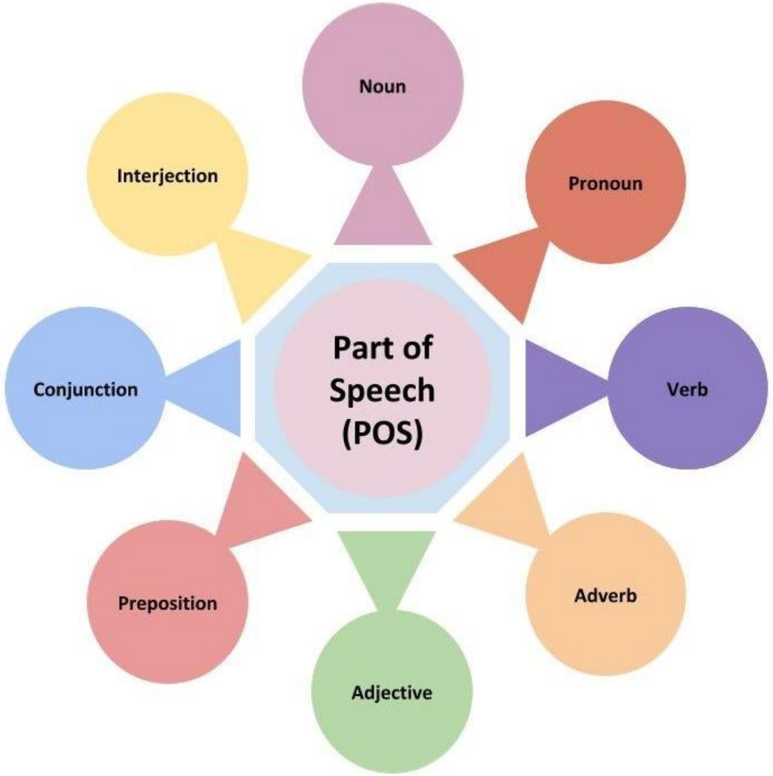
NER

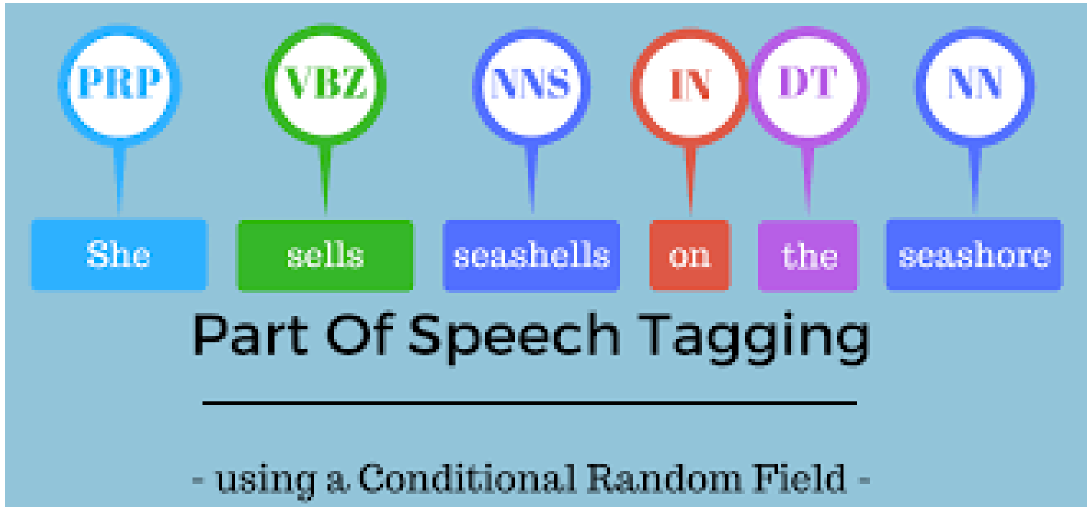
How it works: The intricacies of NER can be broken down into several steps:

1. **Tokenization**. Before identifying entities, the text is split into tokens, which can be words, phrases, or even sentences. For instance, "Steve Jobs co-founded Apple" would be split into tokens like "Steve", "Jobs", "co-founded", "Apple".
2. **Entity identification**. Using various linguistic rules or statistical methods, potential named entities are detected. This involves recognizing patterns, such as capitalization in names ("Steve Jobs") or specific formats (like dates).
3. **Entity classification**. Once entities are identified, they are categorized into predefined classes such as "Person", "Organization", or "Location". This is often achieved using machine learning models trained on labeled datasets. For our example, "Steve Jobs" would be classified as a "Person" and "Apple" as an "Organization".
4. **Contextual analysis**. NER systems often consider the surrounding context to improve accuracy. For instance, in the sentence "Apple released a new iPhone", the context helps the system recognize "Apple" as an organization rather than a fruit.
5. **Post-processing**. After initial recognition and classification, post-processing might be applied to refine results. This could involve resolving ambiguities, merging multi-token entities, or using knowledge bases to enhance entity data.

Part-of-speech tagging (P-o-S)

- Part-of-speech tagging are approaches for identifying the parts of speech of words in a sentence, such as nouns, verbs, and adjectives.
- NLTK library of python has a method called ‘pos_tag’ which allows tagging parts of speech with just one line of code.





The Different Parts of Speech and Their Tags

- There are nine main parts of speech: noun, pronoun, verb, adjective, adverb, conjunction, preposition, interjection, and article.

- Part-of-speech (POS) tags are labels that are assigned to words in a text, indicating their grammatical role in a sentence. The most common types of POS tags include:

The Different Parts of Speech and Their Tags

- Noun (NN): A person, place, thing, or idea
- Verb (VB): An action or occurrence
- Adjective (JJ): A word that describes a noun or pronoun
- Adverb (RB): A word that describes a verb, adjective, or other adverb
- Pronoun (PRP): A word that takes the place of a noun
- Conjunction (CC): A word that connects words, phrases, or clauses
- Preposition (IN): A word that shows a relationship between a noun or pronoun and other elements in a sentence
- Interjection (UH): A word or phrase used to express strong emotion

POS and categorise.

- This is just a sample of the most common POS tags, different libraries and models may have different sets of tags, but the purpose remains the same — to categorise words based on their grammatical function.

- Parts of speech can also be categorised by their grammatical function in a sentence. There are three primary categories: subjects (which perform the action), objects (which receive the action), and modifiers (which describe or modify the subject or object). Each primary category can be further divided into subcategories. For example, subjects can be further classified as simple (one word), compound (two or more words), or complex (sentences containing subordinate clauses).

Text generation

- Techniques for generating new text based on a given input, such as machine translation and text summarization.
- Text generation is the task of automatically producing new text based on a given input or model. It is a popular area of research in Natural Language Processing (NLP) and has numerous applications such as chatbots, content creation, and language translation.

Techniques for text generation.

- **Markov Chain:** A statistical model that predicts the next word based on the probability distribution of the previous words in the text. It generates new text by starting with an initial state, and then repeatedly sampling the next word based on the probability distribution learned from the input text.
- **Sequence-to-Sequence (Seq2Seq) Model:** A deep learning model that consists of two recurrent neural networks (RNNs), an encoder, and a decoder. The encoder takes the input text and produces a fixed-length vector representation, while the decoder generates the output text based on the vector representation.
- **Generative Adversarial Network (GAN):** A deep learning model that consists of two neural networks, a generator, and a discriminator. The generator produces new text samples, while the discriminator tries to distinguish between the generated text and the real text. The two networks are trained in an adversarial manner, where the generator tries to produce more realistic text, while the discriminator tries to become better at recognizing fake text.
- **Transformer-based Models**: Transformer models are a type of neural network architecture designed for NLP tasks, such as text classification and machine translation. They have been shown to perform well on text-generation tasks as well.

Text-to-Speech and Speech-to-Text

- Techniques for converting speech to text and text to speech.
- Text-to-Speech (TTS) and Speech-to-Text (STT) are two important applications of Natural Language Processing (NLP).
- **Text-to-Speech (TTS):** TTS is a technology that allows computers to generate human-like speech from written text. The goal of TTS is to produce speech that is natural, expressive, and matches the intonation, rhythm, and prosody of human speech as closely as possible. TTS systems typically consist of two components: a text analysis module that analyzes the input text, and a speech synthesis module that converts the analyzed text into speech.

- **Speech-to-Text (STT):**

- STT is a technology that allows computers to transcribe spoken words into written text. STT systems are used in a wide range of applications, including voice-controlled virtual assistants, dictation software, and automatic speech recognition (ASR) systems. STT systems typically use acoustic models and language models to transcribe speech into text.