

Professional Training Program in Large Language Models

Day 3



Hamza Farooq
Founder & CEO
Traversaal.ai



SDAIA
الهيئة السعودية للبيانات
والذكاء الاصطناعي
Saudi Data & AI Authority



Agenda



The world of NLP



Re-Introduction to LLMs



What is an RNN



What is ANN (later to come)



What is their role in AI and GenAI



Transformer & General Architecture



Lemmatization and stemming



Day 3:
Getting deeper into
LLMs and ML
System Design





Learning outcomes

- The value of Context
- NLP Recap
- How to Machine Understand Text

The value of context in NLP Models

The goal of natural language processing (NLP) is to find answers to four questions:

- Who is talking?
- What are they talking about?
- How do they feel?
- Why do they feel that way?

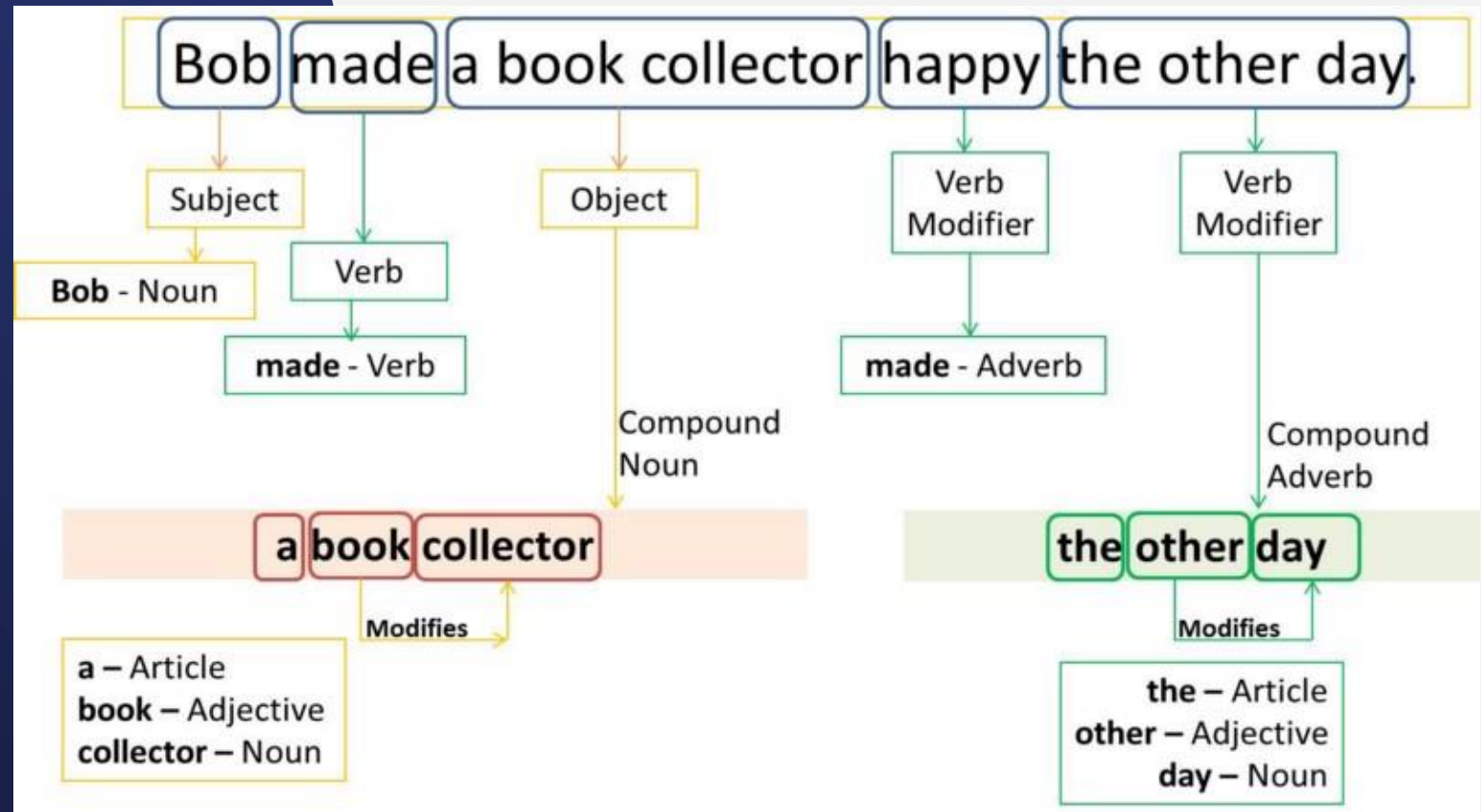
This last question is a question of context



Content vs Context

Content is the material/matter/medium contained within the work that's available for audience.

Context is the positioning of the content, storyline or purpose that provides value to the audience.



**You shall know a word
by the company it keeps.**

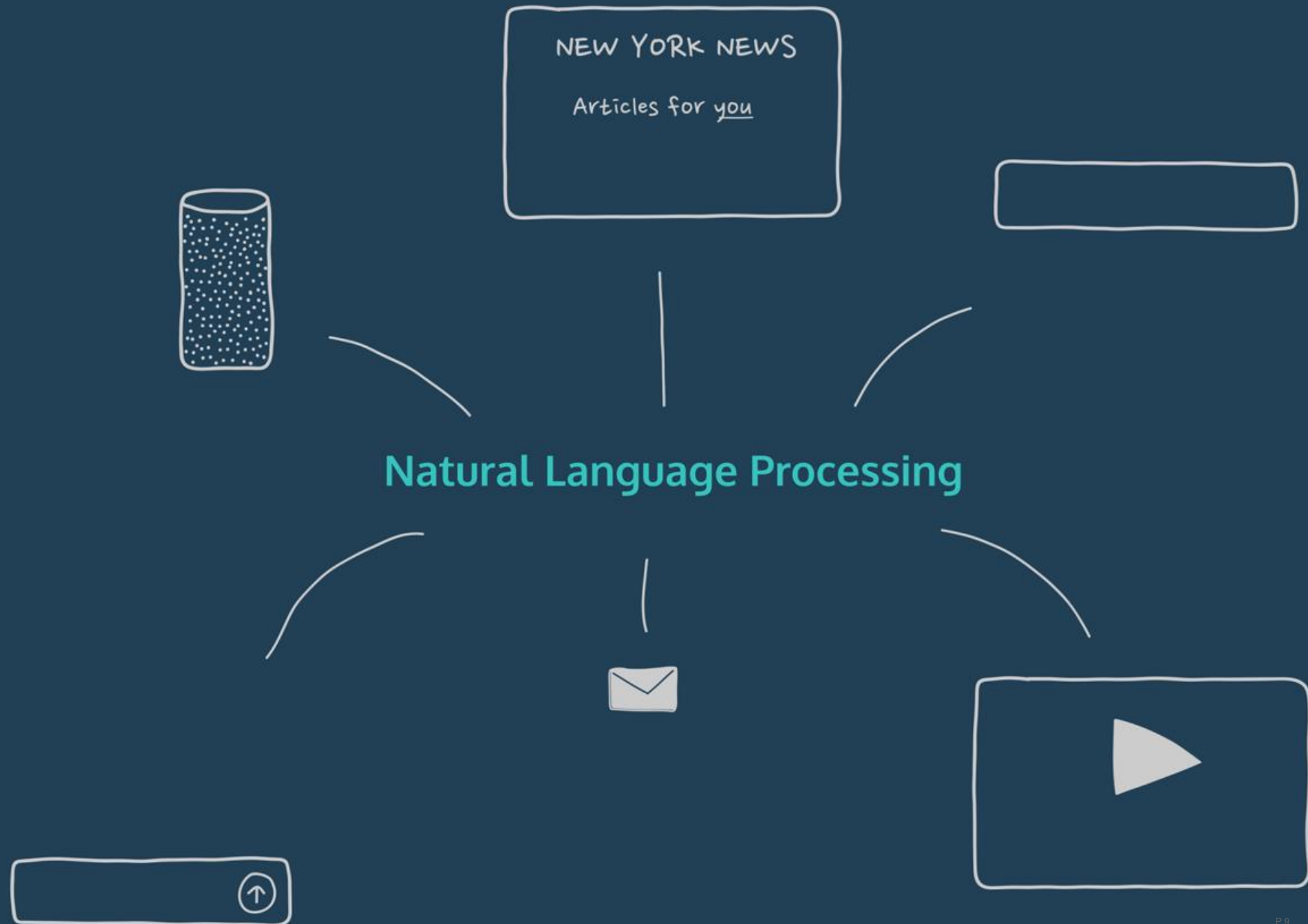
John Rupert Firth

British linguist specializing in contextual theories of meaning and prosodic analysis.

01

Let's go back in time

**We live in a
world of NLP**



What is NLP anyways?

[Natural Language Processing](#) (NLP) is defined as the branch of Artificial Intelligence that provides computers with the capability of understanding text and spoken words, in the same way a human being can.

It incorporates machine learning models, statistics, and deep learning models into computational linguistics i.e. rule-based modeling of human language to allow computers to understand text, spoken words and understands human language, intent, and sentiment.

How does NLP work?

Before you can ingest anything into your NLP model, you need to keep in mind that **computers only understand numbers**. Therefore, when you have text data, you will need to use text vectorization to transform the text into a format that the machine learning model can understand.

- Once the text data is vectorized in a format the machine can understand, the NLP machine learning algorithm is then fed training data.
- This training data helps the NLP model to understand the data, learn patterns, and make relationships about the input data.
- Once the model has gone through the training phase, it will then be put to the test through the testing phase to see how accurately the model can predict outcomes

I like apples and pears.
I know you like apples,
but what about pears?



2	I
2	like
2	apples
1	and
2	pears
1	know
1	you
1	but
1	what
1	about



Applications - 1


- Information retrieval
- Information extraction
- Question answering

Google

list of good sushi restaurant in nyc

Q All News Shopping Maps Images More Tools

About 505,000,000 results (1.29 seconds)



4.0+ rating Sushi Price Hours

Sushi Nakazawa
4.7 ★★★★★ (1,038) · \$\$\$\$ · Sushi
23 Commerce St
Closes soon · 11PM
Dine-in · No takeout · No delivery

Sushi Yasuda
4.4 ★★★★★ (1,119) · \$\$\$\$ · Japanese
204 E 43rd St
Closes soon · 11PM
"Good sushi, but over priced"

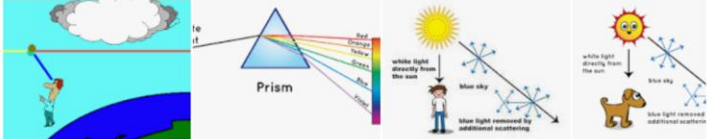
Blue Ribbon Sushi
4.5 ★★★★★ (1,193) · \$\$ · Sushi
119 Sullivan St
Closes soon · 11PM
"Good sushi, extensive menu."

→ View all

why is the sky blue

Q All Books Videos Images News More Tools

Q Child Q Bill Nye Q Adult Q Daddy



Thus, as sunlight of all colors passes through air, the **blue** part causes charged particles to oscillate faster than does the red part. ... More of the sunlight entering the atmosphere is **blue** than violet, however, and our eyes are somewhat more sensitive to **blue** light than to violet light, so the **sky** appears **blue**. Apr 7, 2003

<https://www.scientificamerican.com/article/why-is-the-sky-blue/>

Why is the sky blue? - Scientific American

About featured snippets Feedback

People also ask

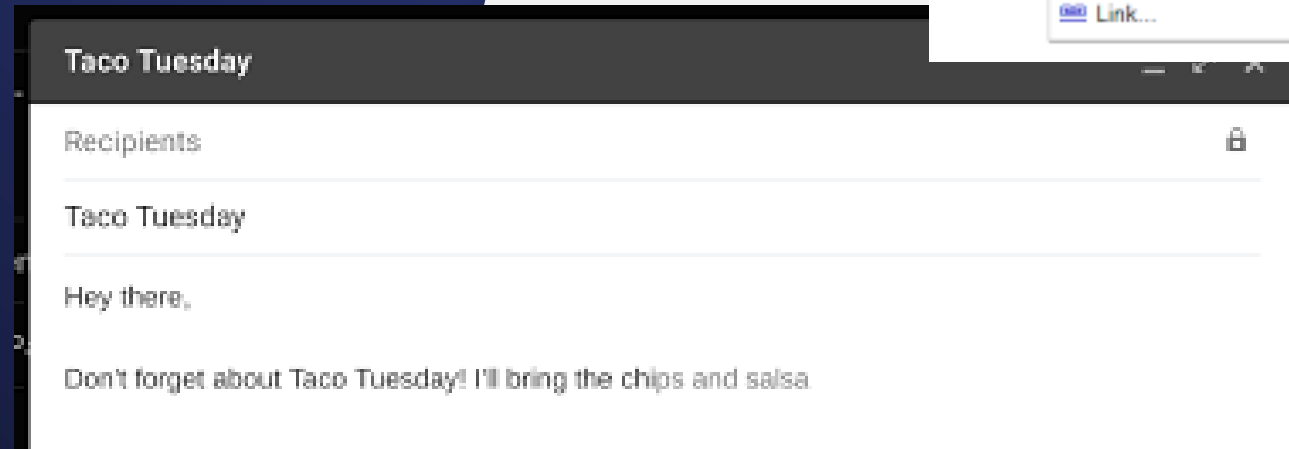
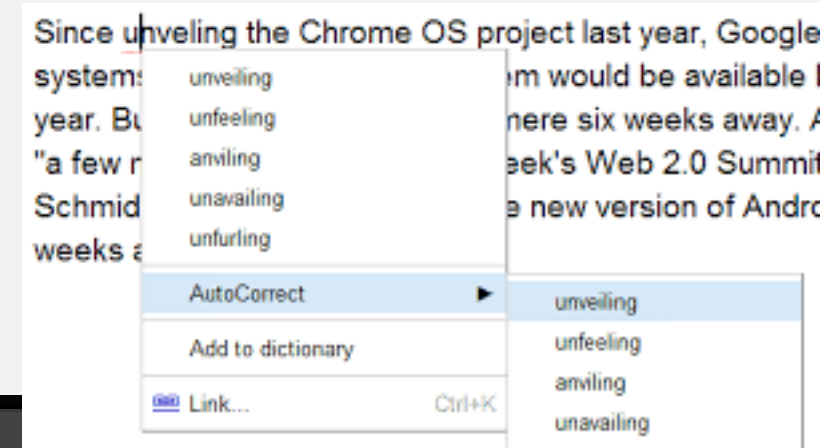
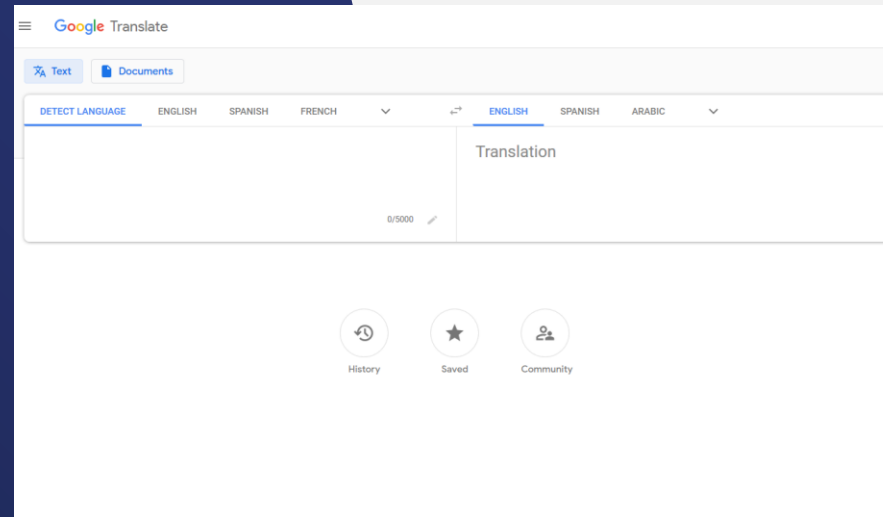
- Why is the sky blue short answer?
- Is the sky blue because of the ocean?
- Why is the sky blue explain to a child?
- What is the reason the sky looks blue?

Feedback



Applications -2

- Machine Translation
 - Summarization
 - Auto Completion
 - Spell Correction
- Many More...



NLP Ambiguities

There are different types of ambiguities present in natural language:

1. Lexical Ambiguity: It is defined as the ambiguity associated with the meaning of a single word. A single word can have different meanings. Also, a single word can be a noun, adjective, or verb. For example, The word “bank” can have different meanings. It can be a financial bank or a riverbank. Similarly, the word “clean” can be a noun, adverb, adjective, or verb.



NLP Ambiguities

2. Syntactic Ambiguity: It is defined as the ambiguity associated with the way the words are parsed. For example, The sentence “Visiting relatives can be boring.” This sentence can have two different meanings. One is that visiting a relative’s house can be boring. The second is that visiting relatives at your place can be boring.

NLP Ambiguities

3. Semantic Ambiguity: It is defined as ambiguity when the meaning of the words themselves can be ambiguous. For example, The sentence “Mary knows a little french.” In this sentence the word “little french” is ambiguous. As we don’t know whether it is about the language french or a person.

Common NLP tasks

Common NLP tasks

NLP systems

Natural language understanding

Natural language generation and
summarization

Natural language translation

Natural language understanding

- Extract information (e.g. about entities or events) from text
- Translate raw text into a meaning representation
- Reason about information given in text
- Execute NL instructions

Natural language generation and summarization

- Translate database entries or meaning representations to raw natural language text
- Produce (appropriate) utterances/responses in a dialog
- Summarize (newspaper or scientific) articles, describe images

Natural language translation

- Translate one natural language to another



Common NLP tasks

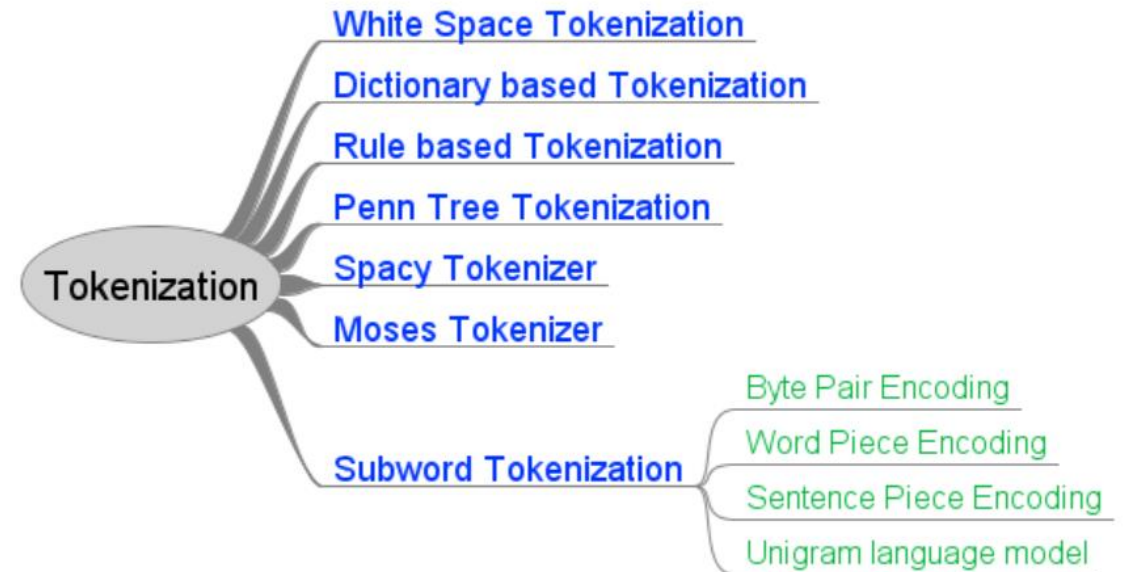
- **Tokenization**
- POS tagging
- Word sense disambiguation
- Dependency Parsing
- Syntactic parsing
- Semantic analysis
- Coreference resolution
- Named Entity Recognition (NER)
- Text representation
- Text classification
- Natural language generation
- Multimodal NLP

- Tokenization is the process of breaking down a text into individual units called tokens.
- Tokens are typically words, but can also be phrases or even individual characters, depending on the application.
- Tokenization is a crucial step in natural language processing tasks such as machine translation, sentiment analysis, and named entity recognition.



Common NLP tasks

- **Tokenization**
- POS tagging
- Word sense disambiguation
- Dependency Parsing
- Syntactic parsing
- Semantic analysis
- Coreference resolution
- Named Entity Recognition (NER)
- Text representation
- Text classification
- Natural language generation
- Multimodal NLP



Text Preprocessing Techniques: Tokenization

The text is split into smaller units. We can use either sentence tokenization or word tokenization based on our problem statement. You can easily tokenize the sentences and words of the text with the tokenize module of NLTK.

```
[17] import nltk
      nltk.download('punkt')
      from nltk.tokenize import word_tokenize, sent_tokenize
```

```
[18] ar_text= 'اهلا ومرحبا بكم في معسكر علم'
      text='The weather is warm today. It is a great day to go to the beach.'
```

```
[20] print(word_tokenize(ar_text))
      print(sent_tokenize(ar_text))
```

```
⇒ ['اهلا', 'ومرحبا', 'بكم', 'في', 'معسكر', 'علم']
   ['اهلا ومرحبا بكم في معسكر علم']
```

```
[21] print(word_tokenize(text))
      print(sent_tokenize(text))
```

```
⇒ ['The', 'weather', 'is', 'warm', 'today', '.', 'It', 'is', 'a', 'great', 'day', 'to', 'go', 'to', 'the', 'beach', '.']
   ['The weather is warm today.', 'It is a great day to go to the beach.']
```





Text Preprocessing Techniques: **Tokenization**

Types of Tokenization:

- **Word Tokenization:** Breaks text into individual words, suitable for languages with clear word boundaries like English.
 - Input: "Natural Language Processing is fascinating!"
 - Output: ["Natural", "Language", "Processing", "is", "fascinating!"]
- **Character Tokenization:** Segments text into individual characters, beneficial for languages without clear word boundaries or tasks like spelling correction.
 - Input: "Hello, world!"
 - Output: ['H', 'e', 'l', 'l', 'o', ',', ' ', 'w', 'o', 'r', 'l', 'd', '!']
- **Sub-word Tokenization:** Divides text into units larger than a single character but smaller than a full word, striking a balance between word and character tokenization. Useful for languages forming meaning by combining smaller units or handling out-of-vocabulary words in NLP tasks.
 - Input: "Chatbots are becoming increasingly popular."
 - Output: ["Chat", "bots", "are", "becoming", "increasingly", "popular", "."]



Text Preprocessing Techniques: Tokenization

Challenges and Limitations of the tokenization task:

- In general, this task is used for text corpus written in English or French where these languages separate words by using white spaces, or punctuation marks to define the boundary of the sentences.
- In the tokenization of Arabic texts since Arabic has a complicated morphology as a language.

For example, a single Arabic word may contain up to six different tokens like the word (“دفع”)

Necklace	عُقْدٌ
Decade	عِقْدٌ
Contract	عَقْدٌ
Held	عَقَدَ
Complicated	عَقَّ
Knots	عُقَّ





Common NLP tasks

- Tokenization
- **POS tagging**
 - POS stands for Part-of-Speech, which is a linguistic term used to describe the grammatical category of a word in a sentence.
 - POS tagging is the process of assigning each word in a text with its corresponding POS category, such as noun, verb, adjective, or adverb.
 - POS tagging is a critical component in various natural language processing tasks, including text-to-speech conversion, information retrieval, and machine translation.
- Word sense disambiguation
- Dependency Parsing
- Syntactic parsing
- Semantic analysis
- Coreference resolution
- Named Entity Recognition (NER)
- Text representation
- Text classification
- Natural language generation
- Multimodal NLP

Common NLP tasks

- Tokenization
- **POS tagging**
- Word sense disambiguation
- Dependency Parsing
- Syntactic parsing
- Semantic analysis
- Coreference resolution
- Named Entity Recognition (NER)
- Text representation
- Text classification
- Natural language generation
- Multimodal NLP

Open the pod door, Hal.



Verb Det Noun Noun , Name .
Open the pod door , Hal .

open:

verb, adjective, or noun?

Verb: ***open the door***

Adjective: ***the open door***

Noun: ***in the open***

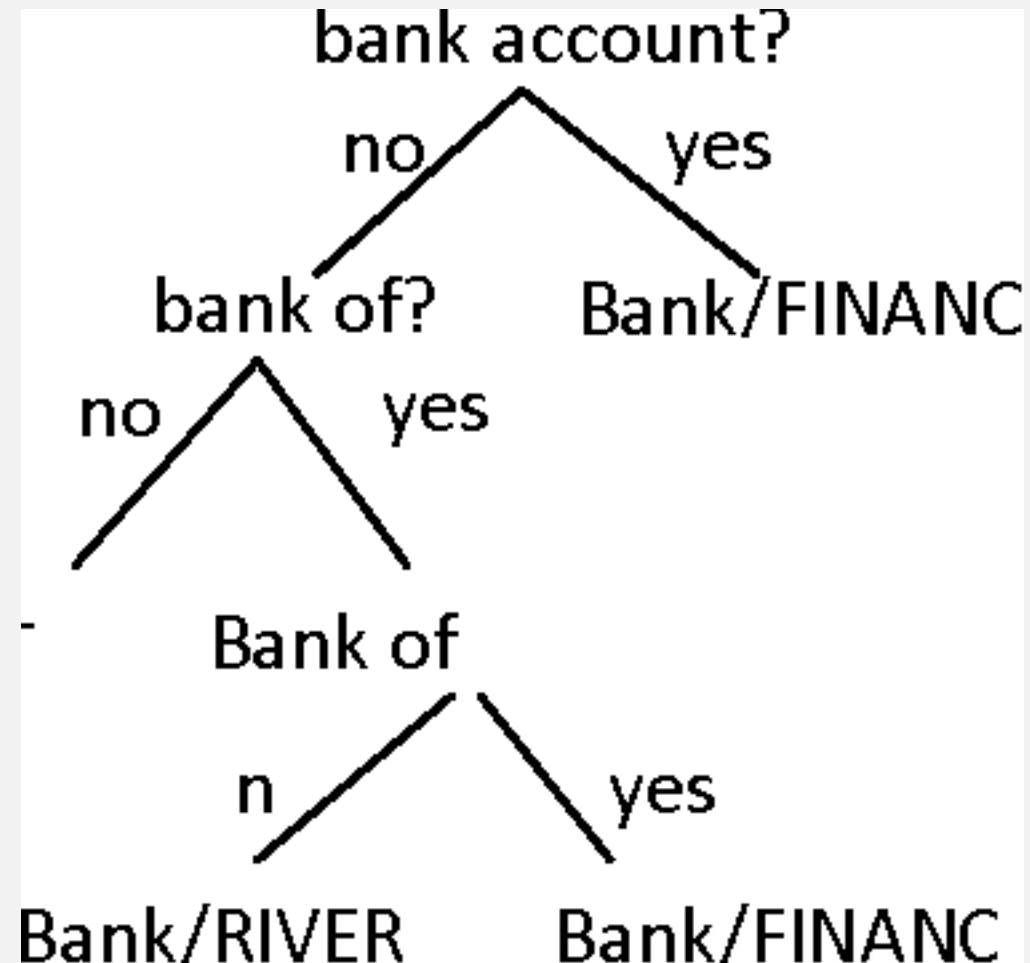


Common NLP tasks

- Tokenization
 - POS tagging
 - **Word sense disambiguation**
 - Dependency Parsing
 - Syntactic parsing
 - Semantic analysis
 - Coreference resolution
 - Named Entity Recognition (NER)
 - Text representation
 - Text classification
 - Natural language generation
 - Multimodal NLP
- Word sense disambiguation is the process of identifying the correct meaning of a word with multiple possible meanings based on the context in which it appears.
 - This is a crucial task in natural language processing because words often have different meanings depending on the context in which they are used.
 - Word sense disambiguation is used in various applications, including information retrieval, machine translation, and question answering systems.

Common NLP tasks

- Tokenization
- POS tagging
- **Word sense disambiguation**
- Dependency Parsing
- Syntactic parsing
- Semantic analysis
- Coreference resolution
- Named Entity Recognition (NER)
- Text representation
- Text classification
- Natural language generation
- Multimodal NLP





Common NLP tasks

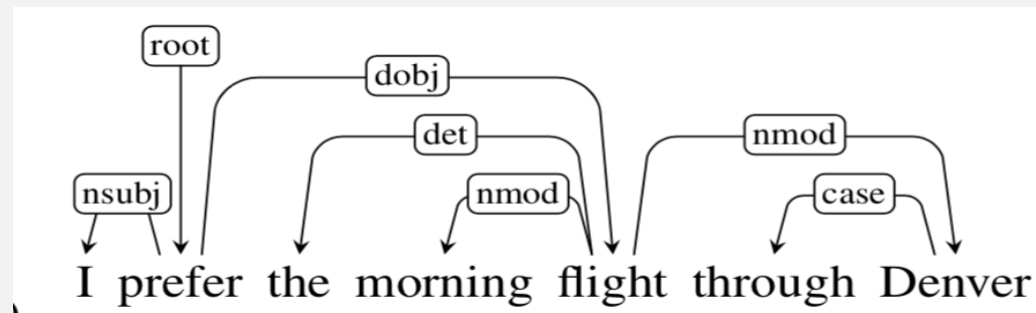
- Tokenization
 - POS tagging
 - Word sense disambiguation
 - **Dependency Parsing**
 - Syntactic parsing
 - Semantic analysis
 - Coreference resolution
 - Named Entity Recognition (NER)
 - Text representation
 - Text classification
 - Natural language generation
 - Multimodal NLP
- Dependency parsing is the process of analyzing the grammatical structure of a sentence by identifying the relationships between words in a sentence.
 - It involves identifying the subject, object, and other dependent clauses and phrases, and representing them as a tree-like structure known as a dependency tree.
 - Dependency parsing is used in various natural language processing applications, including sentiment analysis, named entity recognition, and machine translation.



Common NLP tasks

- Tokenization
- POS tagging
- Word sense disambiguation
- **Dependency Parsing**
- Syntactic parsing
- Semantic analysis
- Coreference resolution
- Named Entity Recognition (NER)
- Text representation
- Text classification
- Natural language generation
- Multimodal NLP

- **Head-Dependent:** In the arrows representing relationship, the origin word is the Head & the destination word is Dependent.
- **Root:** Word which is the root of our parse tree. It is 'prefer' in the above example.
- **Grammar Functions and Arcs:** Tags between each Head-Dependent pair is a grammar function determining the relation between the Head & Dependent. The arrowhead carrying the tag is called an Arc.



Clausal Argument Relations	Description
NSUBJ	Nominal subject
DOBJ	Direct object
IOBJ	Indirect object
CCOMP	Clausal complement
XCOMP	Open clausal complement
Nominal Modifier Relations	Description
NMOD	Nominal modifier
AMOD	Adjectival modifier
NUMMOD	Numeric modifier
APPOS	Appositional modifier
DET	Determiner
CASE	Prepositions, postpositions and other case markers
Other Notable Relations	Description
CONJ	Conjunct
CC	Coordinating conjunction



Common NLP tasks

- Tokenization
 - POS tagging
 - Word sense disambiguation
 - Dependency Parsing
 - **Syntactic parsing**
 - Semantic analysis
 - Coreference resolution
 - Named Entity Recognition (NER)
 - Text representation
 - Text classification
 - Natural language generation
 - Multimodal NLP
- Syntactic parsing is the process of analyzing the grammatical structure of a sentence to determine its syntactic components, such as nouns, verbs, adjectives, and adverbs.
 - It involves identifying the parts of speech of each word in the sentence and grouping them together into phrases and clauses based on their syntactic relationships.
 - Syntactic parsing is used in various natural language processing applications, including text-to-speech conversion, machine translation, and information retrieval.



Common NLP tasks

- Tokenization
 - POS tagging
 - Word sense disambiguation
 - Dependency Parsing
 - **Syntactic parsing**
 - Semantic analysis
 - Coreference resolution
 - Named Entity Recognition (NER)
 - Text representation
 - Text classification
 - Natural language generation
 - Multimodal NLP
- POS tagging is the process of labeling individual words in a sentence with their part of speech, such as noun, verb, adjective, or adverb, while syntactic parsing involves analyzing the relationships between the words to determine the overall grammatical structure of the sentence.
 - For example, consider the sentence "John eats pizza." POS tagging would label "John" as a proper noun and "eats" as a verb, while syntactic parsing would identify "John" as the subject of the verb "eats" and "pizza" as the object of the verb.
 - In short, POS tagging is concerned with the individual words, while syntactic parsing focuses on the overall sentence structure.



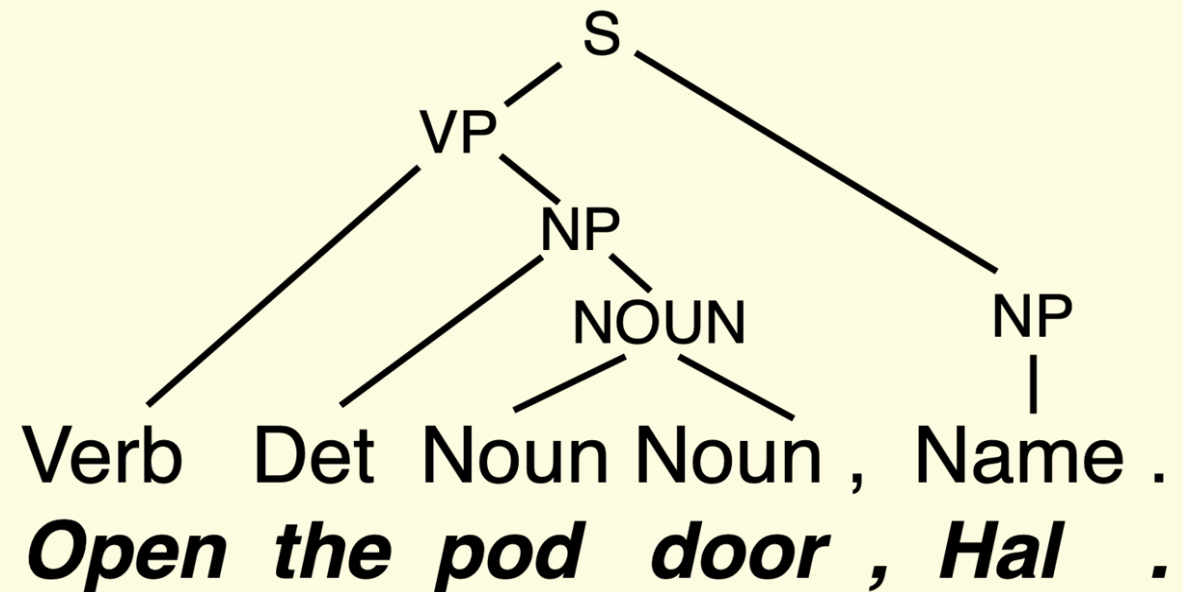
Common NLP tasks

- Tokenization
 - POS tagging
 - Word sense disambiguation
 - Dependency Parsing
 - Syntactic parsing
 - **Semantic analysis**
 - Coreference resolution
 - Named Entity Recognition (NER)
 - Text representation
 - Text classification
 - Natural language generation
 - Multimodal NLP
- Semantic analysis is the process of extracting the meaning of a text by analyzing the relationships between words and phrases in a sentence.
 - It involves identifying the underlying concepts and ideas conveyed by the text and representing them in a structured form, such as a knowledge graph or ontology.
 - Semantic analysis is used in various natural language processing applications, including question answering, information retrieval, and chatbots, to enable more accurate and intelligent responses.

Common NLP tasks

- Tokenization
- POS tagging
- Word sense disambiguation
- Dependency Parsing
- Syntactic parsing
- **Semantic analysis**
- Coreference resolution
- Named Entity Recognition (NER)
- Text representation
- Text classification
- Natural language generation
- Multimodal NLP

$\exists x \exists y (\text{pod_door}(x) \ \& \ \text{Hal}(y)$
 $\& \ \text{request}(\text{open}(x, y)))$





Common NLP tasks

- Tokenization
- POS tagging
- Word sense disambiguation
- Dependency Parsing
- Syntactic parsing
- **Semantic analysis**
- Coreference resolution
- Named Entity Recognition (NER)
- Text representation
- Text classification
- Natural language generation
- Multimodal NLP

We need a **meaning representation language**.

“Shallow” semantic analysis: **Template-filling**
(Information Extraction)

Named-Entity Extraction: Organizations, Locations, Dates,...
Event Extraction

“Deep” semantic analysis: (Variants of) **formal logic**
 $\exists x \exists y (\text{pod_door}(x) \& \text{Hal}(y) \& \text{request}(\text{open}(x, y)))$

We also distinguish between

Lexical semantics (the meaning of words) and
Compositional semantics (the meaning of sentences)



Common NLP tasks

- Tokenization
- POS tagging
- Word sense disambiguation
- Dependency Parsing
- Syntactic parsing
- Semantic analysis
- **Coreference resolution**
- Named Entity Recognition (NER)
- Text representation
- Text classification
- Natural language generation
- Multimodal NLP

More than a decade ago, **Carl Lewis** stood on the threshold of what was to become the greatest athletics career in history. **He** had just broken two of the legendary Jesse Owens' college records, but never believed **he** would become a corporate icon, the focus of hundreds of millions of dollars in advertising. **His** sport was still nominally amateur. Eighteen Olympic and World Championship gold medals and **21 world records later, Lewis has** become the richest man in the history of track and field -- a multi-millionaire.

Who is Carl Lewis?

Did Carl Lewis break any world records?
(and how do you know that?)



Common NLP tasks

- Tokenization
- POS tagging
- Word sense disambiguation
- Dependency Parsing
- Syntactic parsing
- Semantic analysis
- **Coreference resolution**
 - Coreference resolution is the task of identifying all the expressions (e.g., pronouns, names) in a text that refer to the same entity, and linking them together.
 - It is a crucial task in natural language processing as it enables a system to maintain a consistent representation of entities throughout a document, enabling more accurate information extraction and text understanding.
- Named Entity Recognition (NER)
- Text representation
- Text classification
- Natural language generation
- Multimodal NLP



Common NLP tasks

- Tokenization
- POS tagging
- Word sense disambiguation
- Dependency Parsing
- Syntactic parsing
- Semantic analysis
- Coreference resolution
- **Named Entity Recognition (NER)**
- Text representation
- Text classification
- Natural language generation
- Multimodal NLP

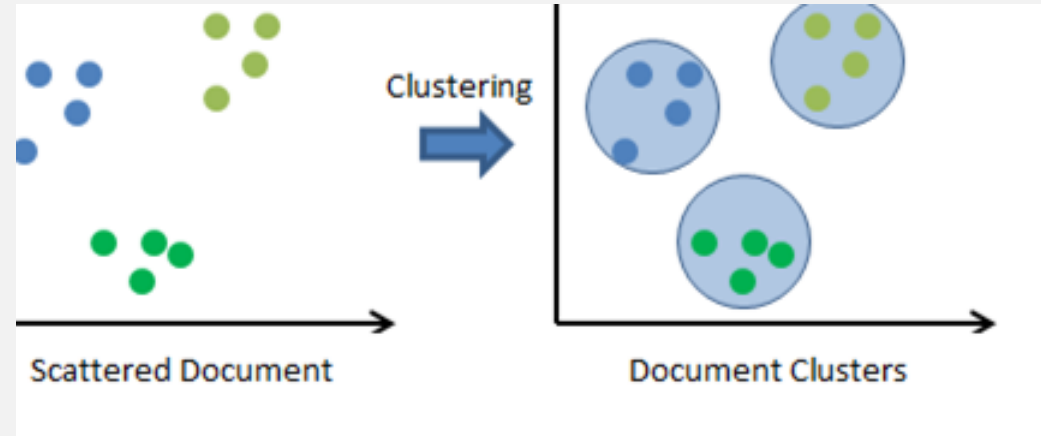
Named entity recognition (NER) is the process of identifying and categorizing named entities in a text, such as people, organizations, locations, and dates.

When **Sebastian Thrun** PERSON started at **Google** ORG in **2007** DATE, few people outside of the company took him seriously. "I can tell you very senior CEOs of major **American** NORP car companies would shake my hand and turn away because I wasn't worth talking to," said **Thrun** PERSON, now the co-founder and CEO of online higher education startup Udacity, in an interview with **Recode** ORG **earlier this week** DATE.

A little **less than a decade later** DATE, dozens of self-driving startups have cropped up while automakers around the world clamor, wallet in hand, to secure their place in the fast-moving world of fully automated transportation.

Common NLP tasks

- Tokenization
- POS tagging
- Word sense disambiguation
- Dependency Parsing
- Syntactic parsing
- Semantic analysis
- Coreference resolution
- Named Entity Recognition (NER)
- **Text representation**
- Text classification
- Natural language generation
- Multimodal NLP

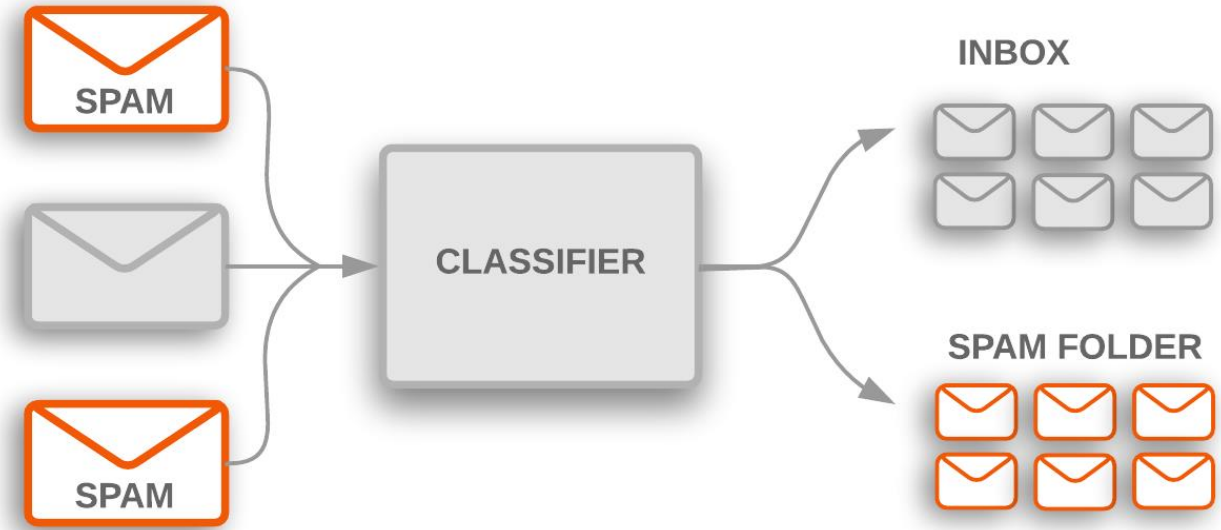


- Text representation is the process of converting unstructured text data into a structured format that can be used for natural language processing tasks.
- It involves selecting a suitable representation scheme, such as bag-of-words, word embeddings, or topic models, to capture the key features and characteristics of the text data in a numerical form that can be processed by machine learning algorithms.

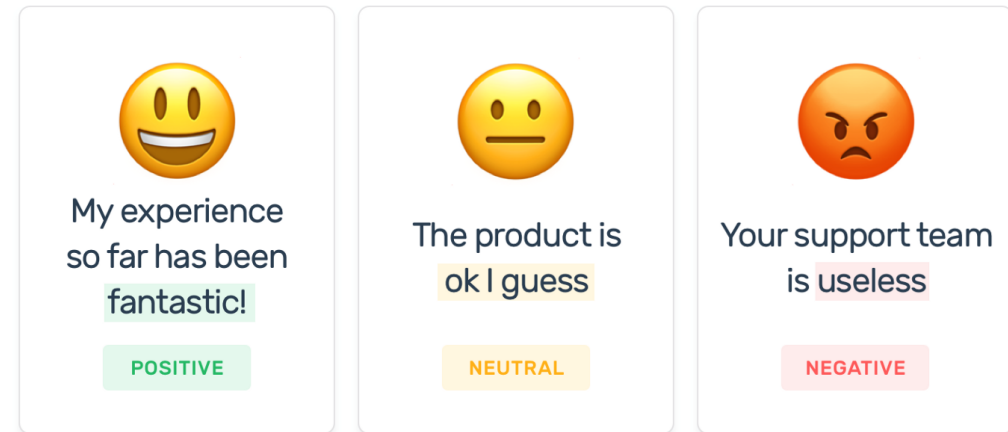


Common NLP tasks

- Tokenization
- POS tagging
- Word sense disambiguation
- Dependency Parsing
- Syntactic parsing
- Semantic analysis
- Coreference resolution
- Named Entity Recognition (NER)
- Text representation
- **Text classification**
- Natural language generation
- Multimodal NLP

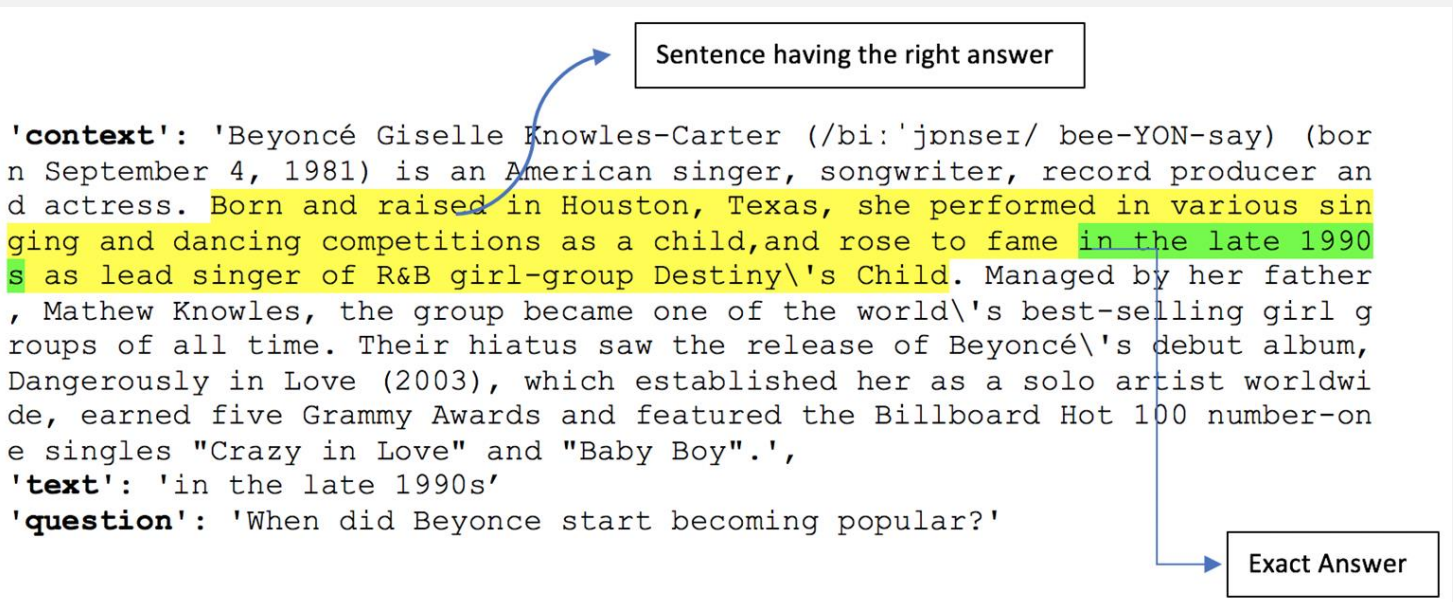
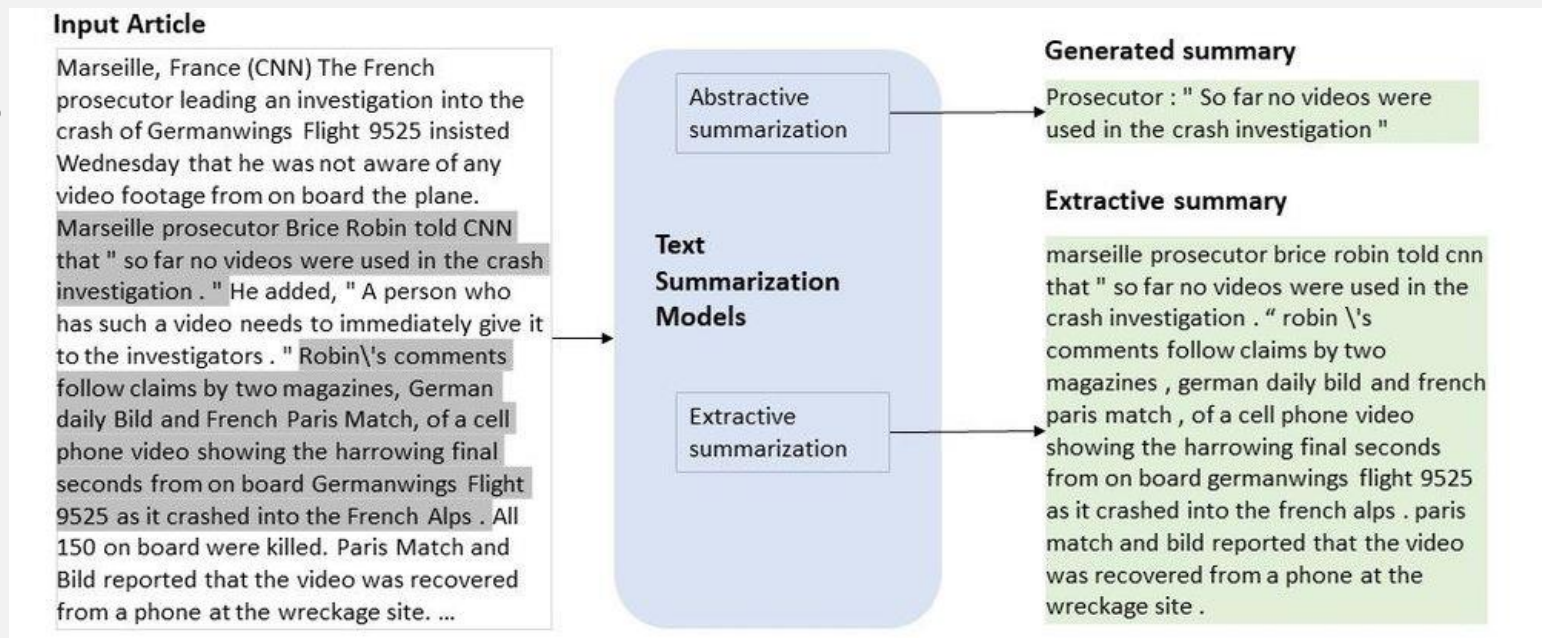


Sentiment Analysis



Common NLP tasks

- Tokenization
- POS tagging
- Word sense disambiguation
- Dependency Parsing
- Syntactic parsing
- Semantic analysis
- Coreference resolution
- Named Entity Recognition (NER)
- Text representation
- Text classification
- **Natural language generation**
- Multimodal NLP



Common NLP tasks

- Tokenization
- POS tagging
- Word sense disambiguation
- Dependency Parsing
- Syntactic parsing
- Semantic analysis
- Coreference resolution
- Named Entity Recognition (NER)
- Text representation
- Text classification
- Natural language generation
- **Multimodal NLP**

Multimodal NLP: mapping from language to the world

$\exists x \exists y (\text{pod_door}(x) \ \& \ \text{Hal}(y) \ \& \ \text{request}(\text{open}(x, y)))$



`request(open(door2, Sys))`



spaCy Package

spaCy is an open-source library used for natural language processing in python. It is extremely popular for processing a large amount of unstructured data generated at a vast scale in the industry and generate useful and meaningful insights from the data.

04

How do Machine Understand Text?



Lemmatization

Query: buy

John **bought** some candies.

I will **buy** a new computer.

I didn't consider **buying** a new car.

Lemmatization

John **buy** some candy.

I will **buy** a new computer.

I do not consider **buy** a new car.

Match

No Match



What is lemmatization

Lemmatization is a text normalization technique used in natural language processing (NLP) to reduce words to their base or dictionary form, known as the lemma. It is used to process words such that different grammatical variants of a word are treated as the same term.



What is the purpose of lemmatization

Its primary purpose is to consolidate similar word forms so that they can be analyzed as a single item. This is particularly useful in tasks that involve text understanding and semantic analysis, where the meaning of the text is paramount.



What is the purpose of lemmatization

In other words, the total count of words before and after lemmatization typically remains the same, but the forms of some words change to make them more uniform.

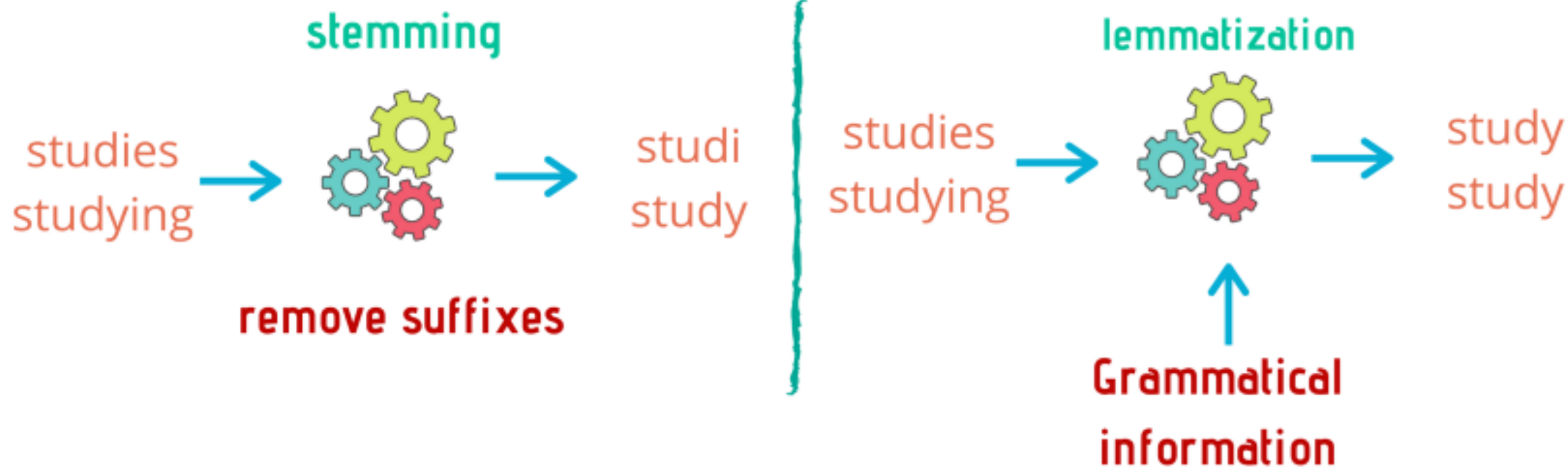


What is the purpose of lemmatization

This helps maintain semantic consistency across different forms of the same word, which is particularly important in NLP applications like search and retrieval, machine translation, and text analysis where consistent vocabulary is key.

▶ Stemming

STEMMING VS. LEMMATIZATION






Stemming vs Lemmatization

Unlike stemming, which often simply chops off word endings to reach a base form, lemmatization considers the morphological analysis of words. This makes lemmatization more sophisticated and accurate as it ensures that the resulting word lemma is a valid word according to the language.



How does Lemmatization really work

Lemmatization involves several steps, including identifying the part of speech of a word, understanding the context in which the word is used, and applying language-specific rules to deduce the base form of the word.



Examples

Improving → improve

Improvement → improve

Improver → improve

studying → study

student → study

studies → study



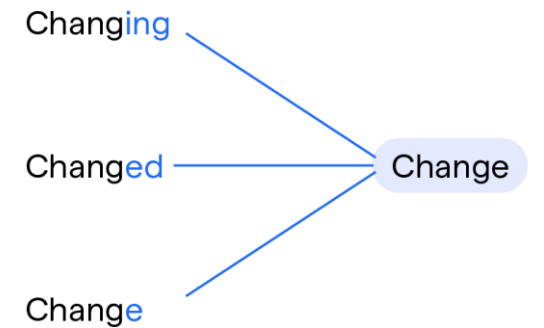
Language Dependency

Lemmatization is highly language-dependent. Each language has its own set of morphological rules, and thus, the lemmatizer must be specifically designed for each language to effectively apply its grammatical rules.

Use in NLP

In NLP, lemmatization is used in various applications including search engines, indexing, text analysis, and natural language understanding systems, where maintaining the semantic integrity of the language is crucial.

Lemmatization





What is the main benefit of Lemmatization

This practice's main benefit is its ability to accurately reduce words to their dictionary forms. This not only helps in maintaining the semantic meaning of the text but also improves the performance of various NLP tasks by reducing the complexity of text data.



Benefit: Maintaining Semantic Meaning

Lemmatization preserves the semantic meaning of text by normalizing different inflected forms of a word to a single base form, such as reducing "run," "runs," and "running" to "run." This uniformity is crucial for accurate semantic analysis in NLP applications, aiding tasks like machine translation by maintaining context across languages and improving information retrieval by matching query words to their variants in documents.



Benefit: Reducing Complexity of Text Data

Lemmatization simplifies text data by reducing words to their lemmas, which decreases the vocabulary size and enhances computational efficiency in NLP tasks. This simplification makes algorithms faster and more scalable, aiding text classification by reducing unique tokens and enhancing sentiment analysis by consolidating forms of sentiment-bearing words.



Benefit: Improving the Model's Performance

The use of lemmatization leads to more uniform datasets, which can significantly improve the learning process and performance of NLP models, especially in deep learning. This benefits models like BERT or GPT by enabling better generalization over text data and supports feature extraction by effectively identifying relevant topics or keywords.



Are there any drawbacks?

One major challenge of lemmatization is the need for extensive dictionaries and complex morphological analyzers, which can make lemmatization computationally expensive compared to simpler methods like stemming.

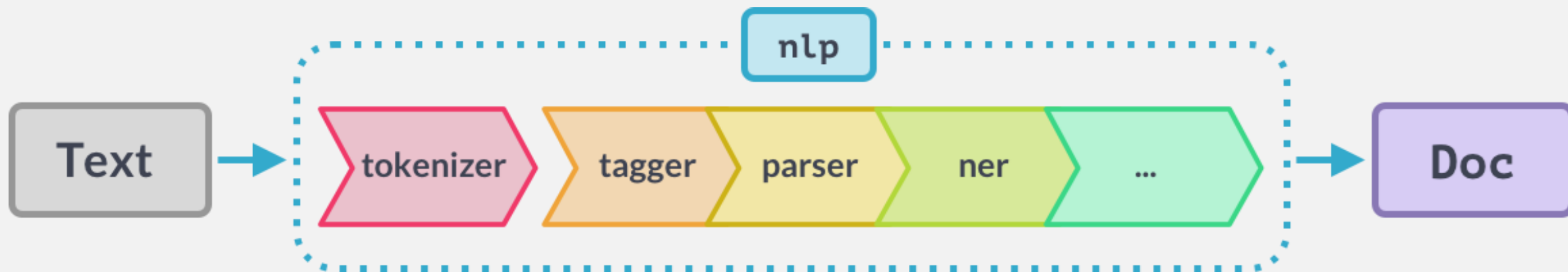


Lemmatization

is a powerful tool in the arsenal of NLP techniques.

Despite its challenges, the accuracy and depth of understanding it provides make it invaluable for any application where the true meaning of the text needs to be preserved.

spaCy NLP Pipeline



Let's code

NLP Pipeline



Appendix

THANK YOU



SDAIA
الهيئة السعودية للبيانات
والذكاء الاصطناعي
Saudi Data & AI Authority