simplilearn | PURDUE UNIVERSITY

Natural Language Processing

Natural Language Understanding Techniques

Learning Objectives

By the end of this lesson, you will be able to:

- Define parts-of-speech tagging
- Explain the different parsing methods
- Apply fuzzy search to identify similar words
- Get the polarity of the given survey for a product
- Extract city and person name from a text

DATA AND
ARTIFICIAL INTELLIGENCE

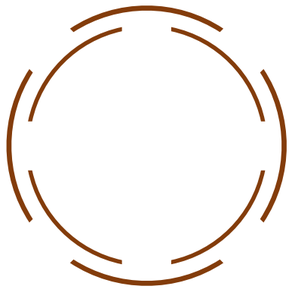
Parts-of-Speech Tagging

Parts-of-Speech Tagging

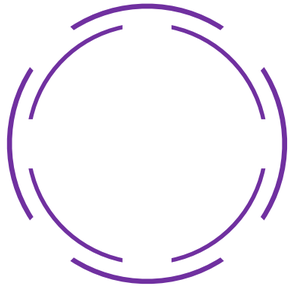
- It is used to understand the POS in a phrase or sentence.
- It assigns POS tag to each word in a phrase or sentence.
- POS tags include:
  - NN (Noun, singular)
  - VB (Verb, base form)
  - CD (Cardinal number)
  - RB (Adverb)
- A set of all POS tags used in a corpus is called a tagset.

Parts-of-Speech Tagging

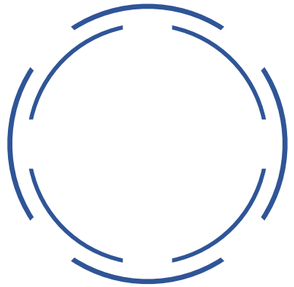
The parts-of-speech in the English language are:



Noun

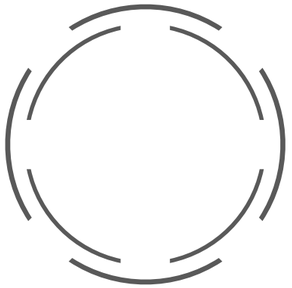


Pronoun

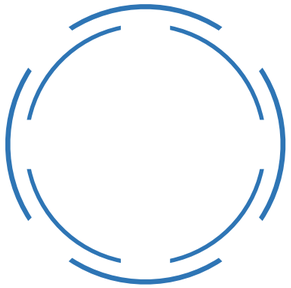


Verb

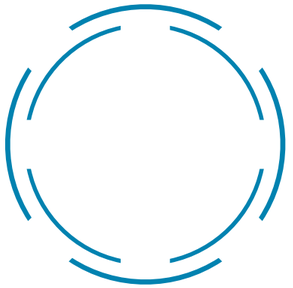
Adjective



Adverb



Preposition



Conjunction

Parts-of-Speech Tagging: Advantages

1. Has better understanding of the context of the sentence
2. Identifies the relationship between words
3. Recognizes cross-references

Parts-of-Speech Tagging: Tags

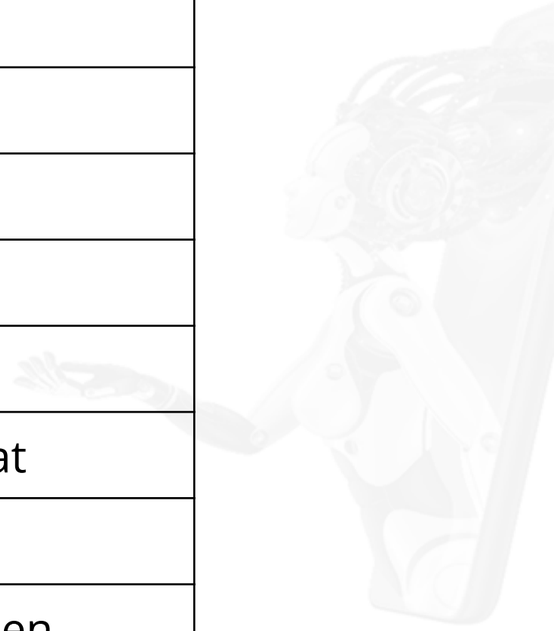
Tag	Description	Example
NN	Noun, Singular	Chair
NNS	Noun, Plural	Chairs
NNP	Proper Noun, Singular	Mathew
PRP	Pronoun Personal	He
PRPS	Pronoun Possessive	His
VB	Verb, Base form	Run
JJ	Adjective	Good
JJR	Adjective Comparative	Better
JJS	Adjective Superlative	Best
RB	Adverb	Naturally
CC	Coordinating conjunction	And
DT	Determiner	The
CD	Cardinal number	I
POS	Possessive ending	Mathew's

Parts-of-Speech Tagging: Tags

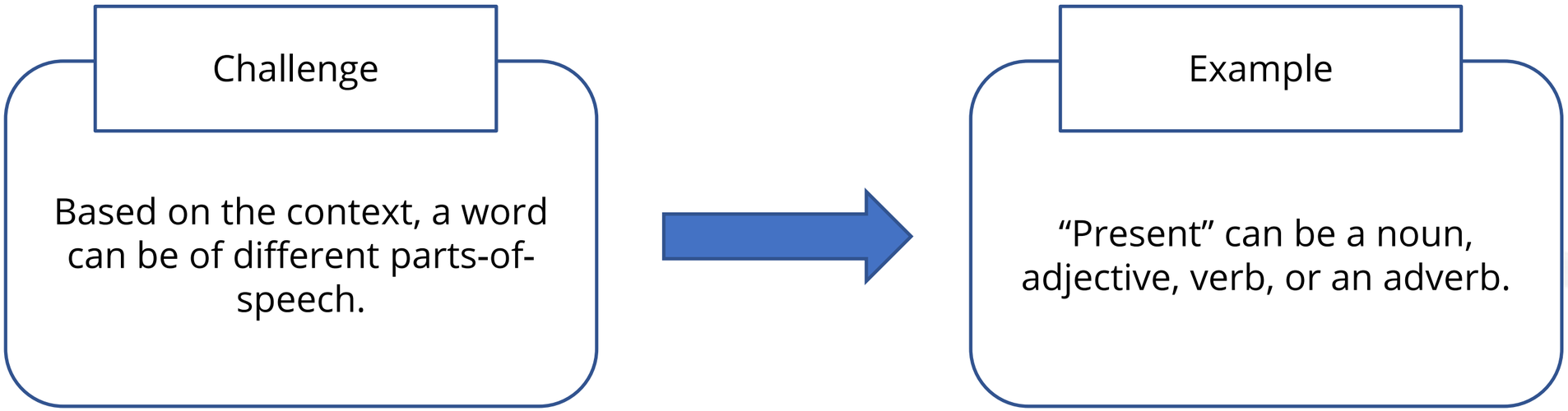
Tag	Description	Example
EX	Existential there	there is
FW	Foreign word	d'hoevre
IN	Preposition or subordinating conjunction	in, of, like
LS	List item marker	1)
MD	Modal	could, will
NNPS	Proper noun, plural	doors
PDT	Predeterminer	both the boys
RBR	Adverb, comparative	better
RBS	Adverb, superlative	best
RP	Particle	give up
SYM	Symbol	@, *
TO	*to*	to go, to him
UH	Interjection	uhhuhhuhh
VBD	Verb, past tense	took

Parts-of-Speech Tagging: Tags

Tag	Description	Example
VBG	Verb, gerund or present participle	taking
VBN	Verb, past participle	taken
VBP	Verb, non-3rd person singular present	take
VBZ	Verb, 3rd person singular present	takes
WDT	Wh-determiner	which
WP	Wh-pronoun	who, what
WP$	Possessive wh-pronoun	whose
WRB	Wh-adverb	where, when

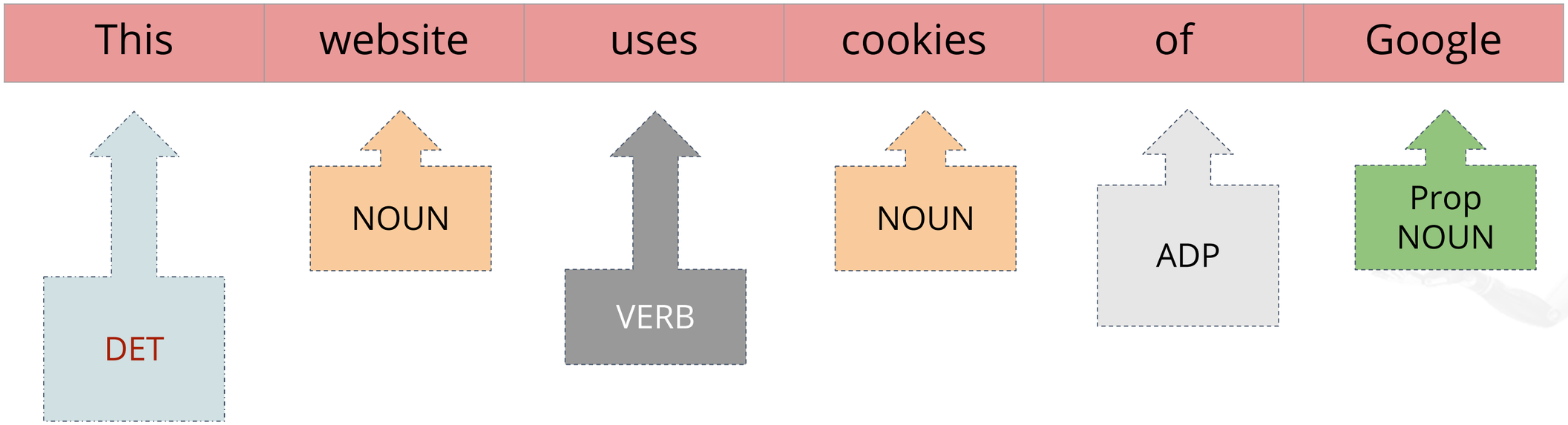


Parts-of-Speech Tagging: Challenges

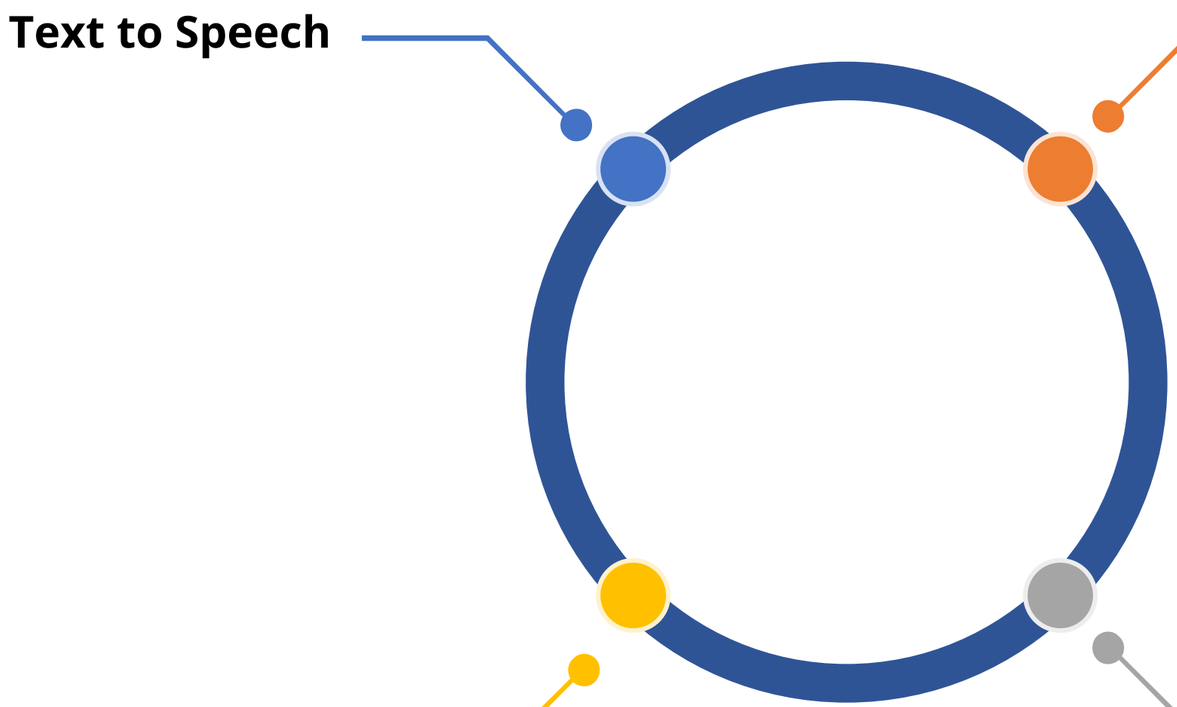


Parts-of-Speech Tagging: Example

Sentence: This website uses cookies of Google.



Parts-of-Speech Tagging: Usage



Apply Fuzzy Search to Identify Similar Words

Problem Statement: Parts-of-speech plays a very crucial role in creating a grammatically correct sentence. To understand the sentence structure we can use POS tag identification techniques. Find a noun word from the given sentence and use this keyword to find similar documents.

Access: Click on the **Practice Labs** tab on the left side panel of the LMS. Copy or note the username and password that is generated. Click on the **Launch Lab** button. On the page that appears, enter the username and password in the respective fields, and click **Login**.

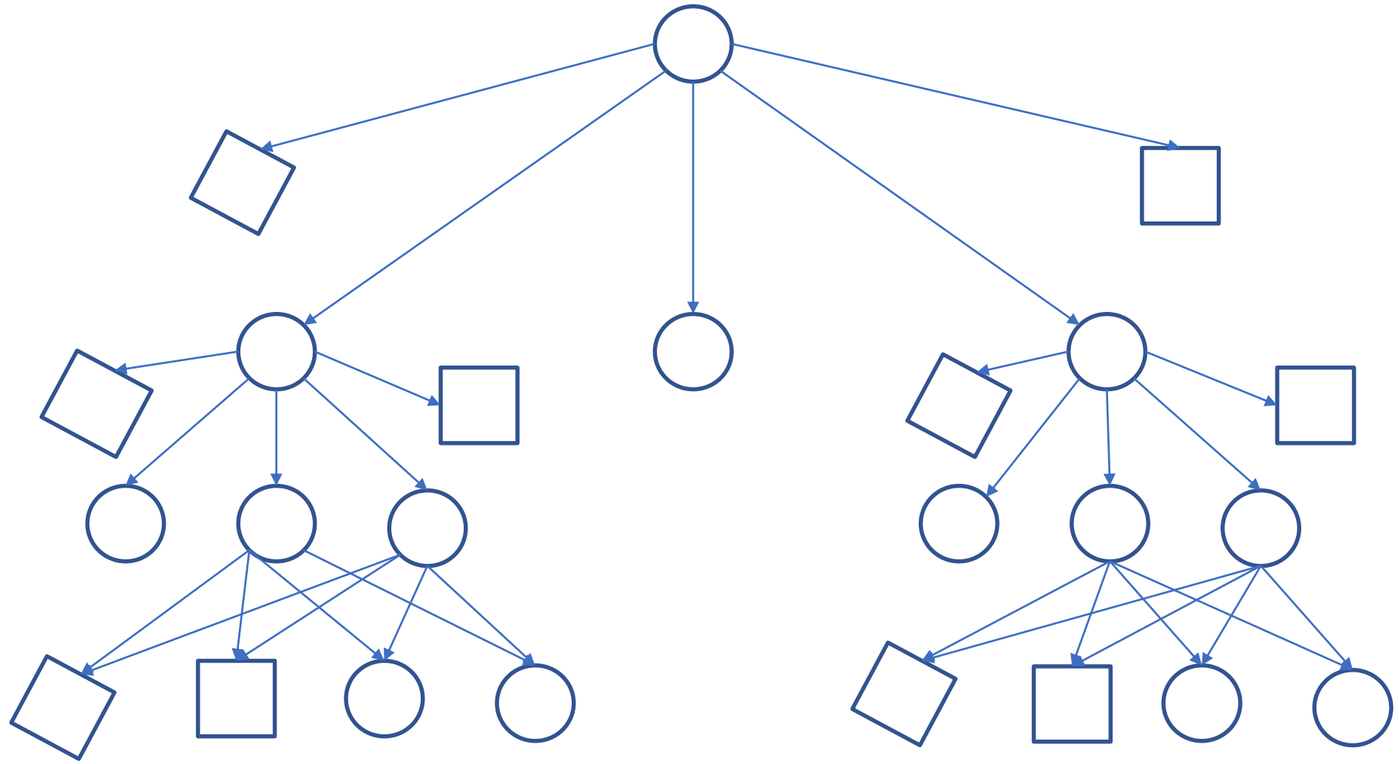
ASSISTED PRACTICE

DATA AND
ARTIFICIAL INTELLIGENCE

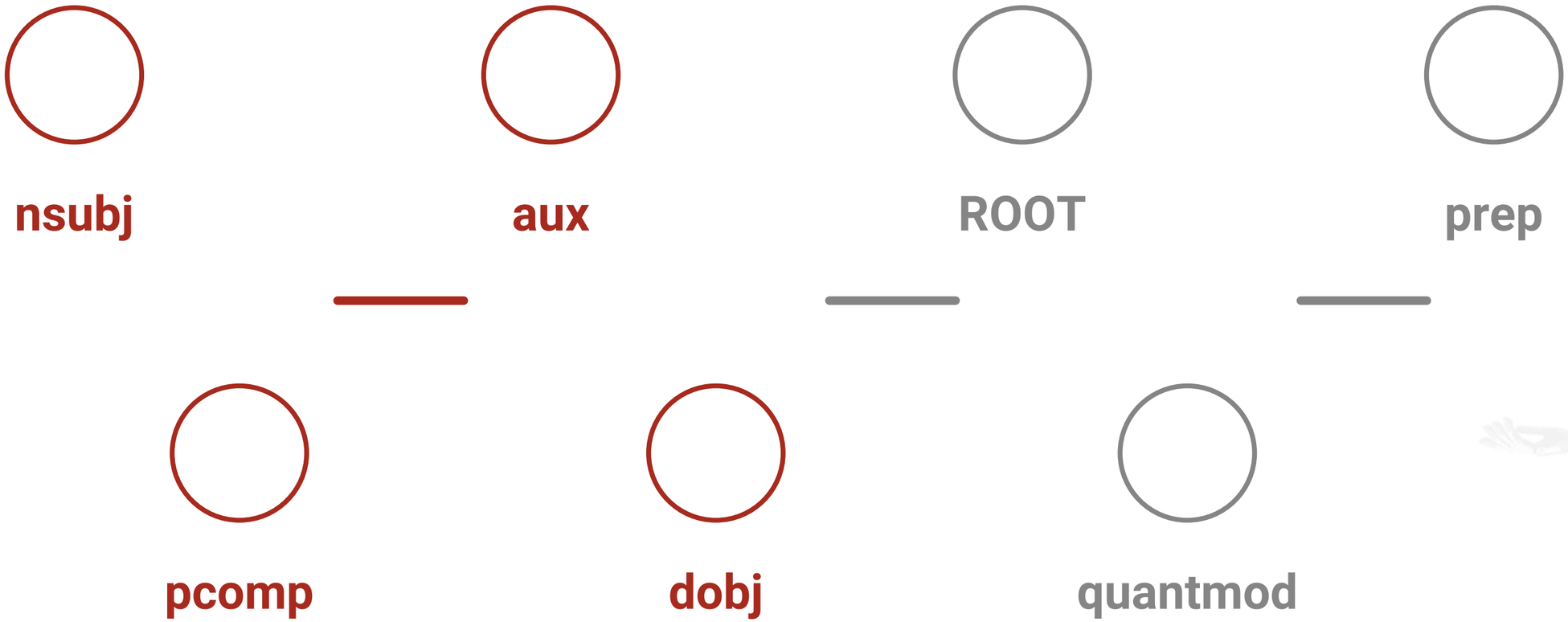
Dependency Parsing

Dependency Parsing

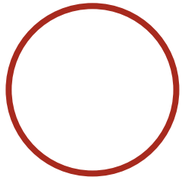
- A dependency parser analyzes the grammatical structure of a sentence.
- It finds relationships between **head** words and words which modify those heads.
- Relations among the words are illustrated in a tree called the dependency tree.



Dependency Parsing: Modifiers



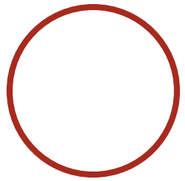
Dependency Parsing: Modifiers



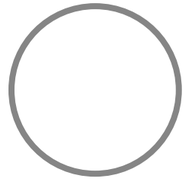
nsubj

Auxiliary

Expresses categories such as tense, aspect, and mood



aux



ROOT

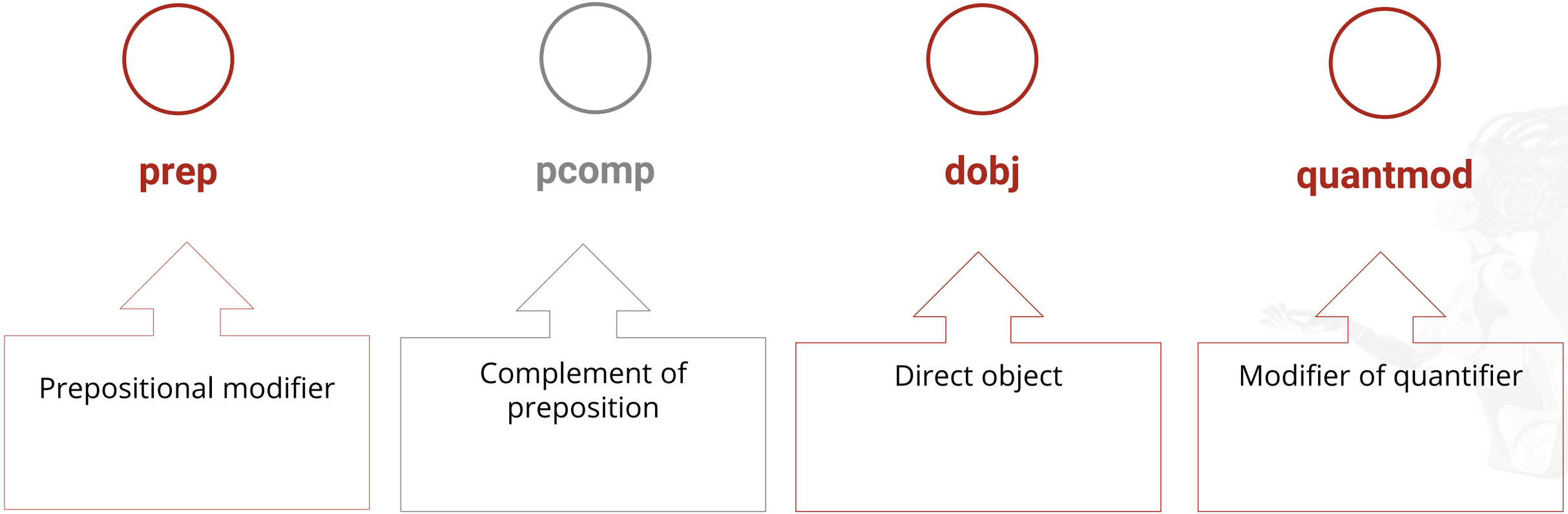
Nominal Subject

- Is the syntactic subject and the proto-agent of a clause
- Is only applied to semantic arguments of a predicate

Root

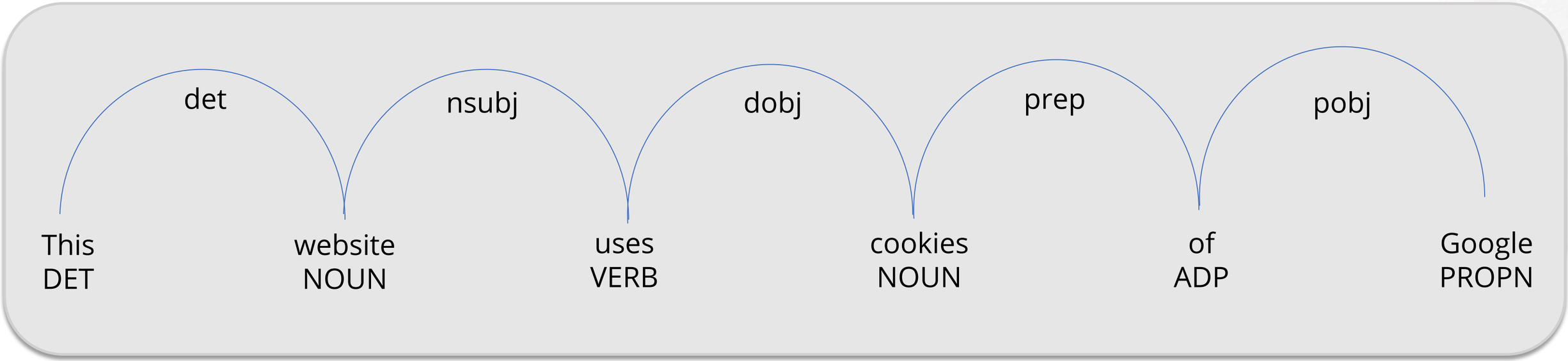
- Is indexed with 0
- Is the only modifier with root dependency

Dependency Parsing: Modifiers



Dependency Parsing: Example

Sentence: **This website uses cookies of Google.**



Dependency Parsing: Types

Grammar-Driven (Rule-based)

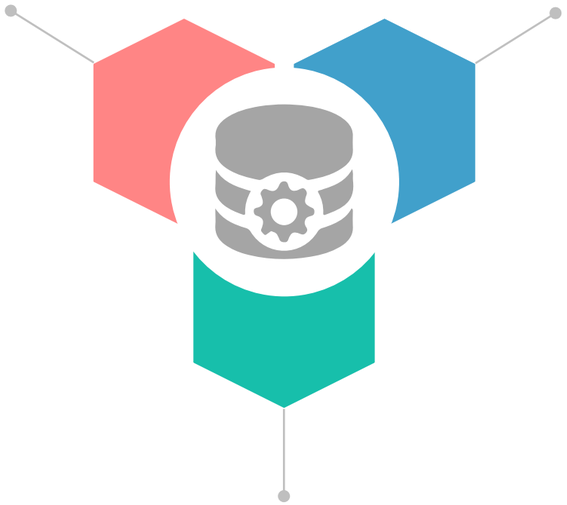
Data-Driven (Machine learning based)

Dependency Parsing: Pros and Cons

Below are the pros and cons of dependency parsing:

Word Order

- Structure is independent of word order
- Suitable for free word order languages



Limited Expressivity

- One node per word
- Hard to find phrase modification and head modification in unlabeled dependency structure

Transparency

- Direct encoding of predicate-argument structure

Dependency Parsing: Usage

Constituency Parsing

Constituency Parsing

1. It extracts a constituency-based parse tree from a sentence.
2. It denotes the syntactic structure according to phrase structure grammar.
3. Nonterminals are a type of phrases.
4. Terminals are words.

Constituency Parsing

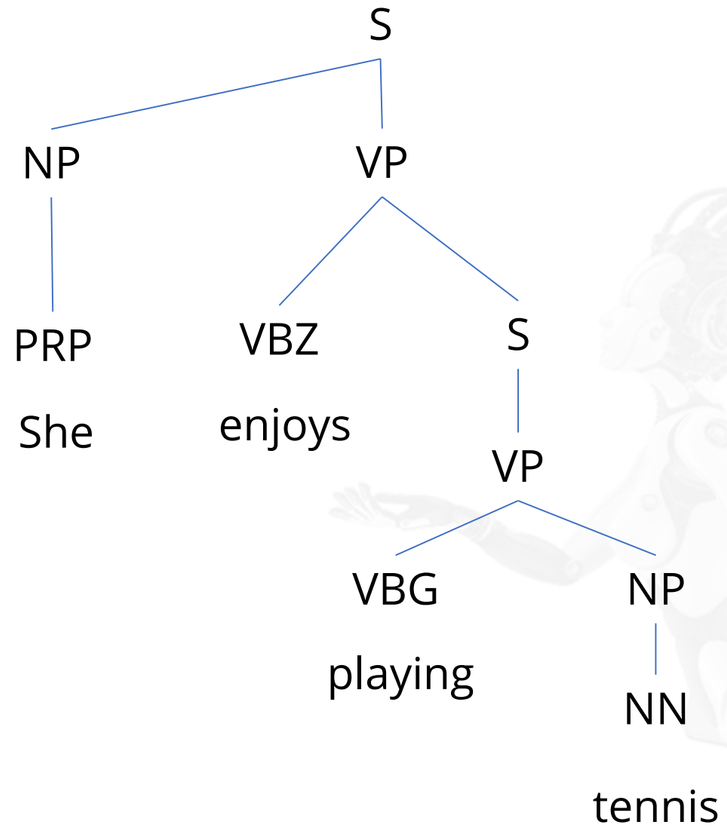
Syntactic categories in NLP:

- NP: Noun Phrase
- VP: Verb Phrase
- S: Sentence
- DET: Determiner
- N: Noun
- TV: Transitive Verb
- IV: Intransitive Verb
- PRP: Preposition
- PP: Prepositional Phrase
- ADJ: Adjective

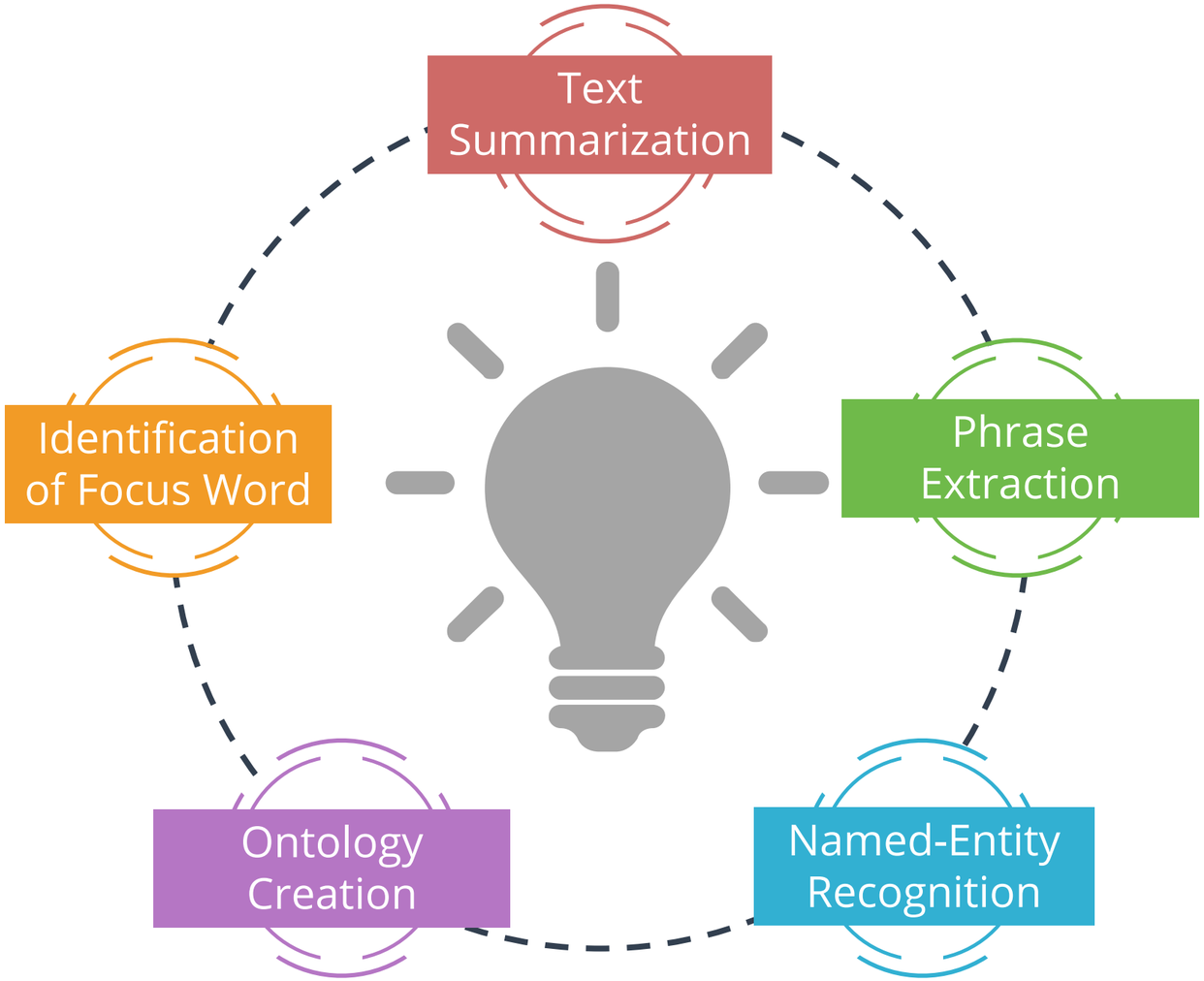
Constituency Parsing

Sentence : She enjoys playing tennis

- The word “she” is personal pronoun and noun phrase.
- VP is verb phrase and S is sentence. VP has three words: enjoys, playing, tennis.



Constituency Parsing: Usage



DATA AND
ARTIFICIAL INTELLIGENCE

Morphological Techniques

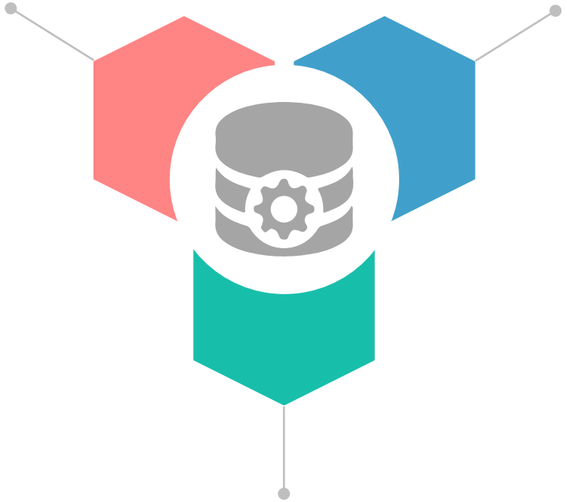
Morphological Techniques

- Morphology is the study of structure and formation of words.
- Every important unit is called a morpheme.
- Example: UNHAPPINESS
  It has three parts and therefore, three morphemes: **UN HAPPY NESS**

Morphological Techniques: Classes

Classification of morphological structural types:

Isolating Languages



Agglutinative Languages

Inflecting Languages

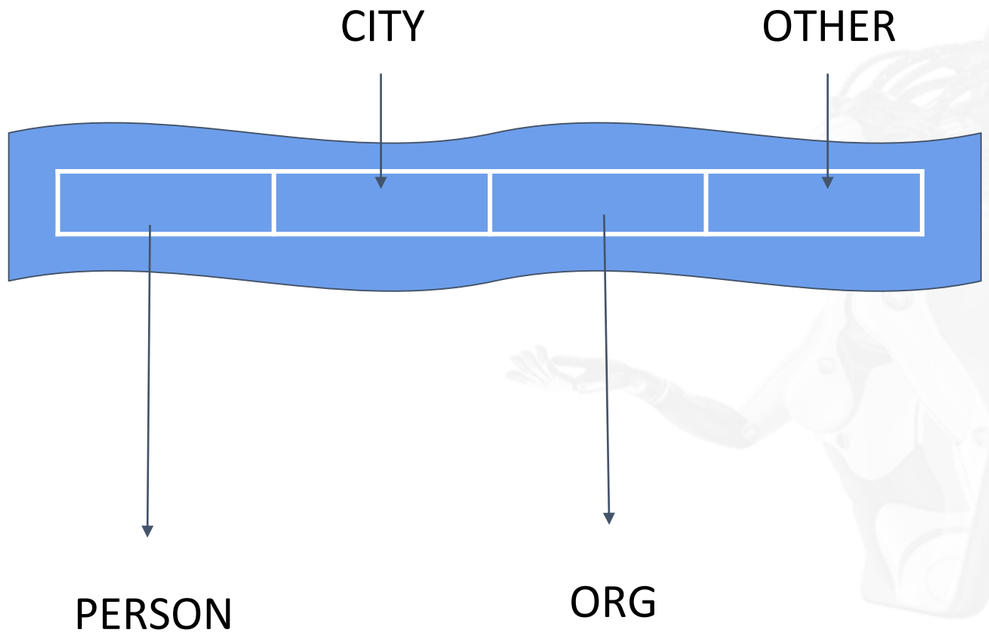
Morphological Techniques: Process

DATA AND
ARTIFICIAL INTELLIGENCE

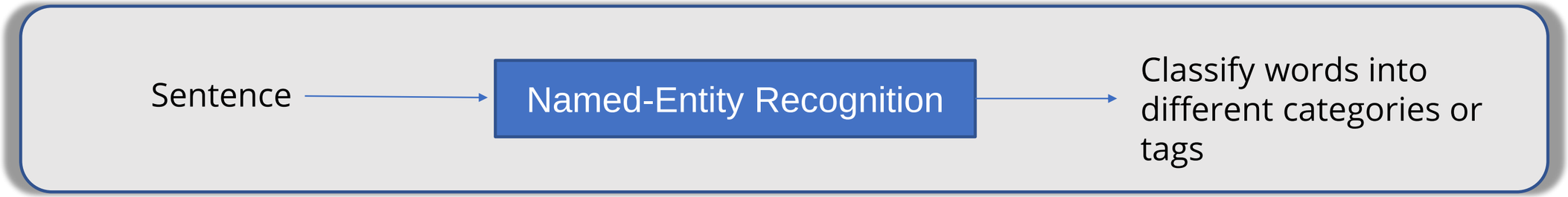
Named-Entity Recognition

Named-Entity Recognition

- It is a process for information extraction that targets to locate and classify named entities.
- Named entities include person, object, city, and the like.
- It includes entity identification, entity chunking, and entity extraction.
- IOB technique is used to train the model.



Named-Entity Recognition



Example:

Trump became the President of the U.S.

Names entities or tags are as below:

- Name : **Trump**
- Place : **U.S.**

Categories Example:

- Companies
- Places
- Organizations
- Cities
- Money

Named-Entity Recognition: IOB Technique

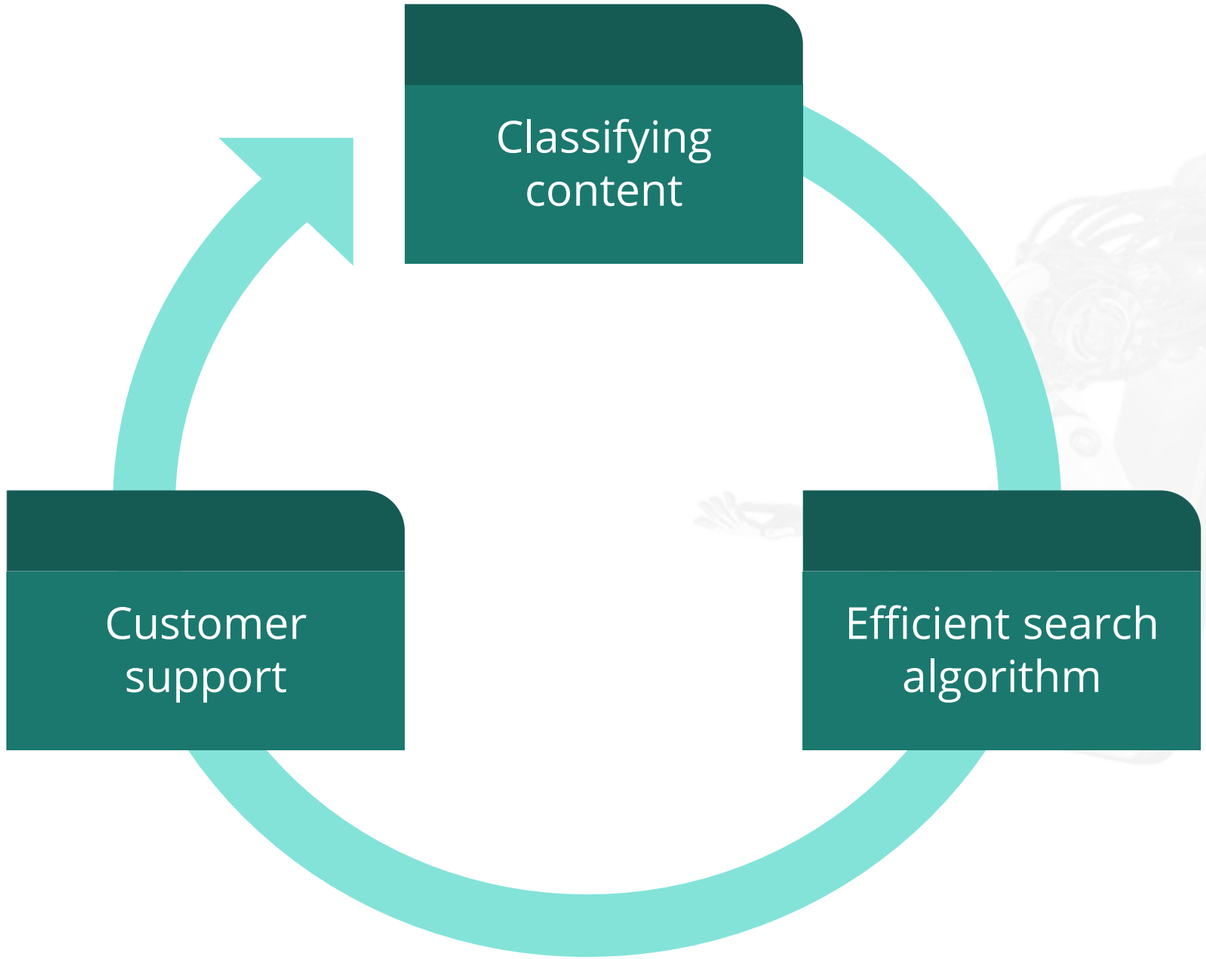
Inside–Outside–Beginning (Tagging)

- It is a common tagging format for tagging tokens.
- I-prefix indicates that the tag is inside a chunk.
- B-prefix indicates that the tag is the beginning of a chunk.
- An O tag indicates that a token belongs to no chunk (outside).

SBI	internet	banking	enables	fund	transfers
B-NP	**I-NP**	**I-NP**			

Named-Entity Recognition: Use Case

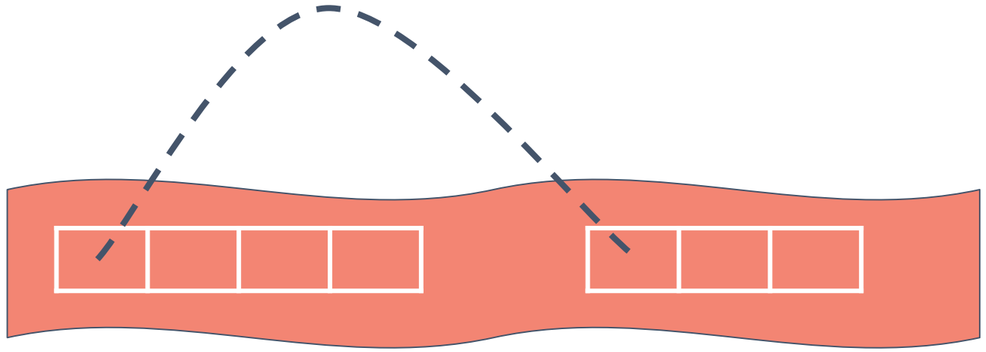
Google ORG made a late push into the hardware of Apple ORG's product available on iPhone and Amazon ORG's Alexa which runs on its Echo product and Dot product devices. These have clear leads in consumer adoption.



DATA AND
ARTIFICIAL INTELLIGENCE

Coreference Resolution

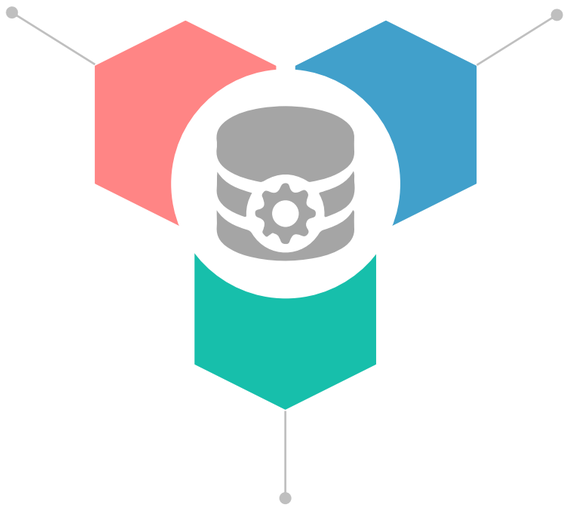
Coreference Resolution



- It is a task to find expressions that refer to the same entity in a text.
- It is useful in document summarization, question answering, and information extraction.
- It is used to find the context in a conversation and relating the terms with immediate previous context.

Coreference Resolution: Applications

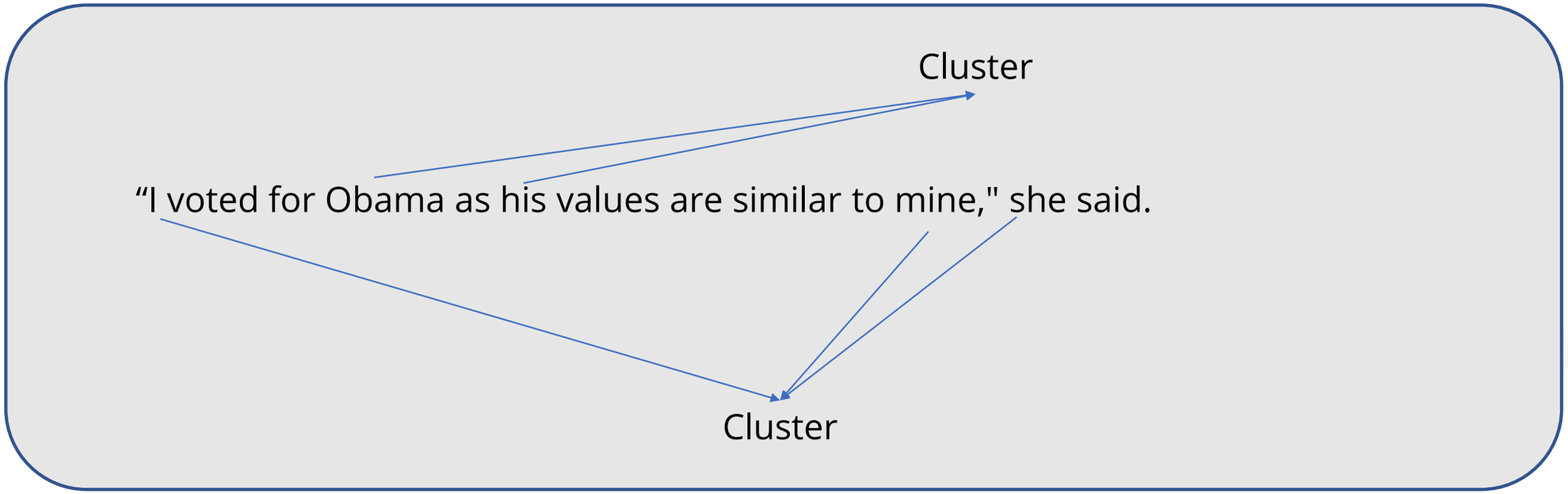
Document Summarization



Question Answering

Information Extraction

Coreference Resolution: Example



DATA AND
ARTIFICIAL INTELLIGENCE

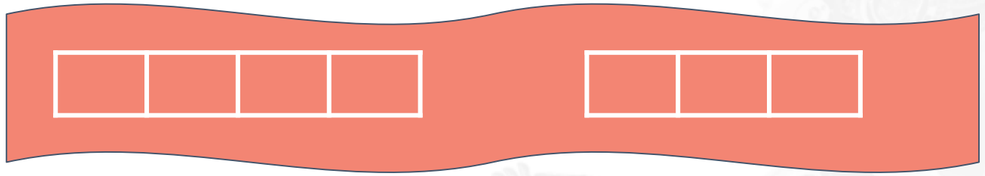
Word-Sense Disambiguation

Word-Sense Disambiguation

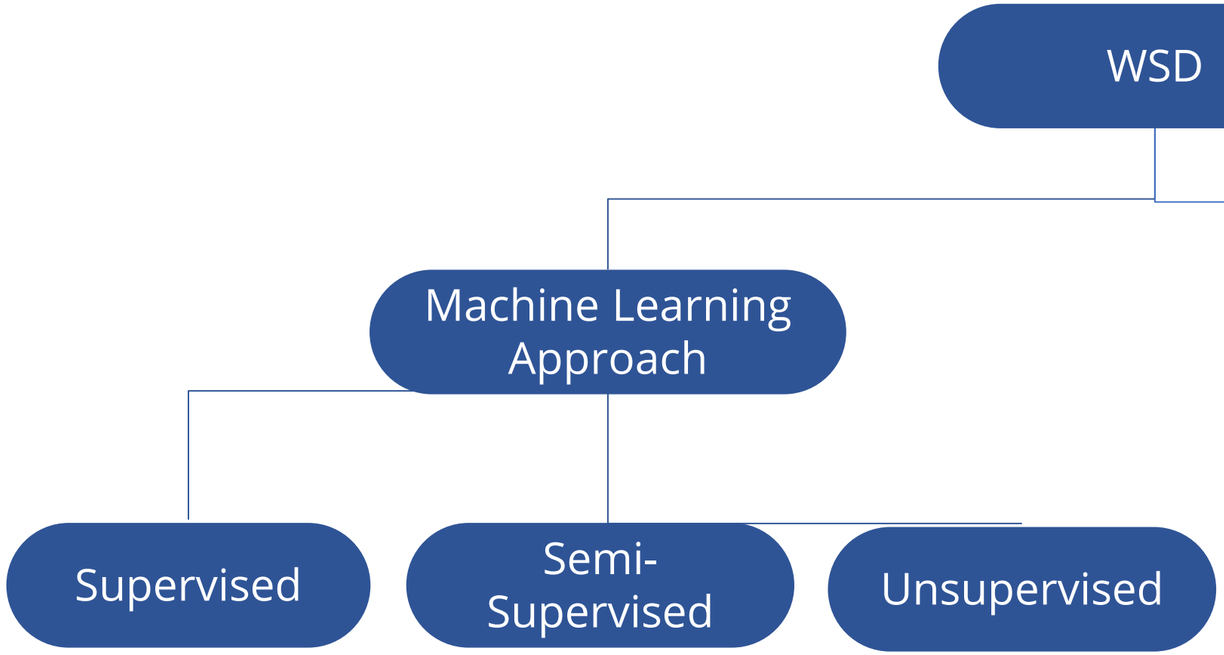
- It is used to understand the sense of meaning in the text.
- It arises due to different meanings of words in different contexts.

Example:

- The bank is a financial institution.
- I was sitting near the bank of the river.



Word-Sense Disambiguation: Techniques



Word-Sense Disambiguation: Applications

Machine Translation

Information Retrieval

Lexicography

Word-Sense Disambiguation: Challenges

1. Requires huge data
2. Does not have context
3. Does not have proper grammatical structure

DATA AND
ARTIFICIAL INTELLIGENCE

Fuzzy Search

Fuzzy Search

1. Fuzzy search is a process that identifies the relevant web pages or texts for searched argument.
2. It is done by fuzzy match program which returns a list of results even though the search argument spelling is not correct.
3. Fuzzy string is also known as approximate string matching.

Fuzzy Search: Techniques

Regular Expression

Edit Distance

Wildcard in SQL Matching

DATA AND
ARTIFICIAL INTELLIGENCE

Document and Sentence Similarity

Document and Sentence Similarity

It determines the similarity between the two pieces of text or documents and also, used to cluster the document.

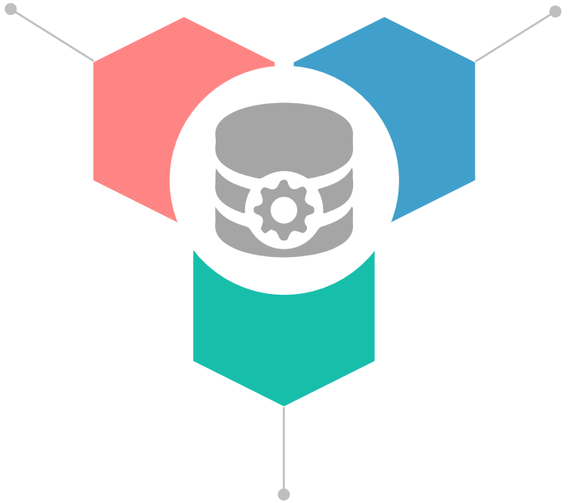
The following are the types of similarity determined:

Lexical

Semantic

Document and Sentence Similarity: Techniques

LSI + Cosine Similarity



Jaccard Similarity

Word embeddings + Word Mover Distance

Document and Sentence Similarity: Example

- "The mouse ate the cat's food"
- "The cat ate the mouse"

How similar are these?

- On Lexical similarity, they are very close.
- On Semantic similarity, they convey totally different meanings.

Document and Sentence Similarity: Applications

Quora

stackoverflow

Customer Service

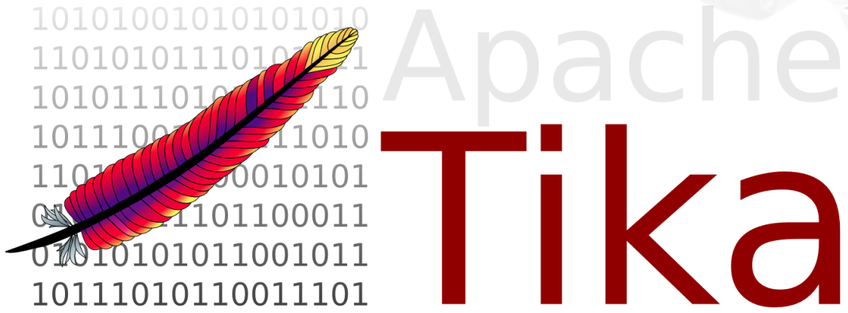
DATA AND
ARTIFICIAL INTELLIGENCE

Document Indexing

Document Indexing

1. It is the process of associating or tagging document or file with various search terms.
2. It is used for search and retrieval purpose in the future.
3. It creates searchable information.
4. Index information is stored in documents or records management system.
5. **Full Text Indexing** and **Date-Based Indexing** are the two types of indexing.

Document Indexing: Tools



Document Indexing: Use Cases

Searching books in digital library

Searching the relevant web page on search engine

Document tracking system

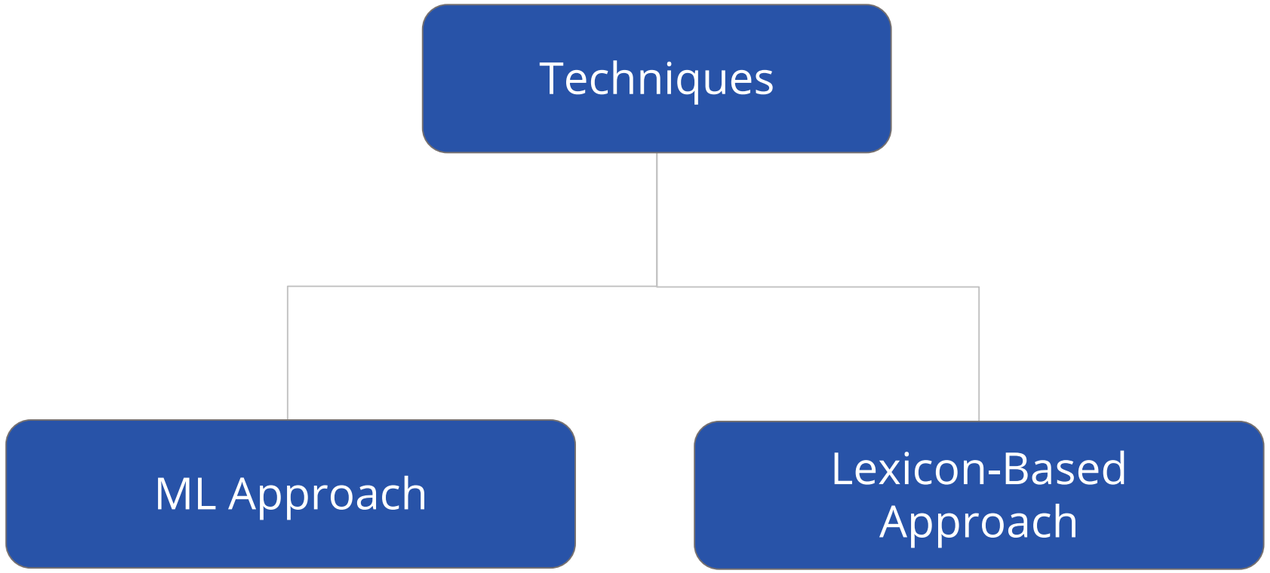
DATA AND
ARTIFICIAL INTELLIGENCE

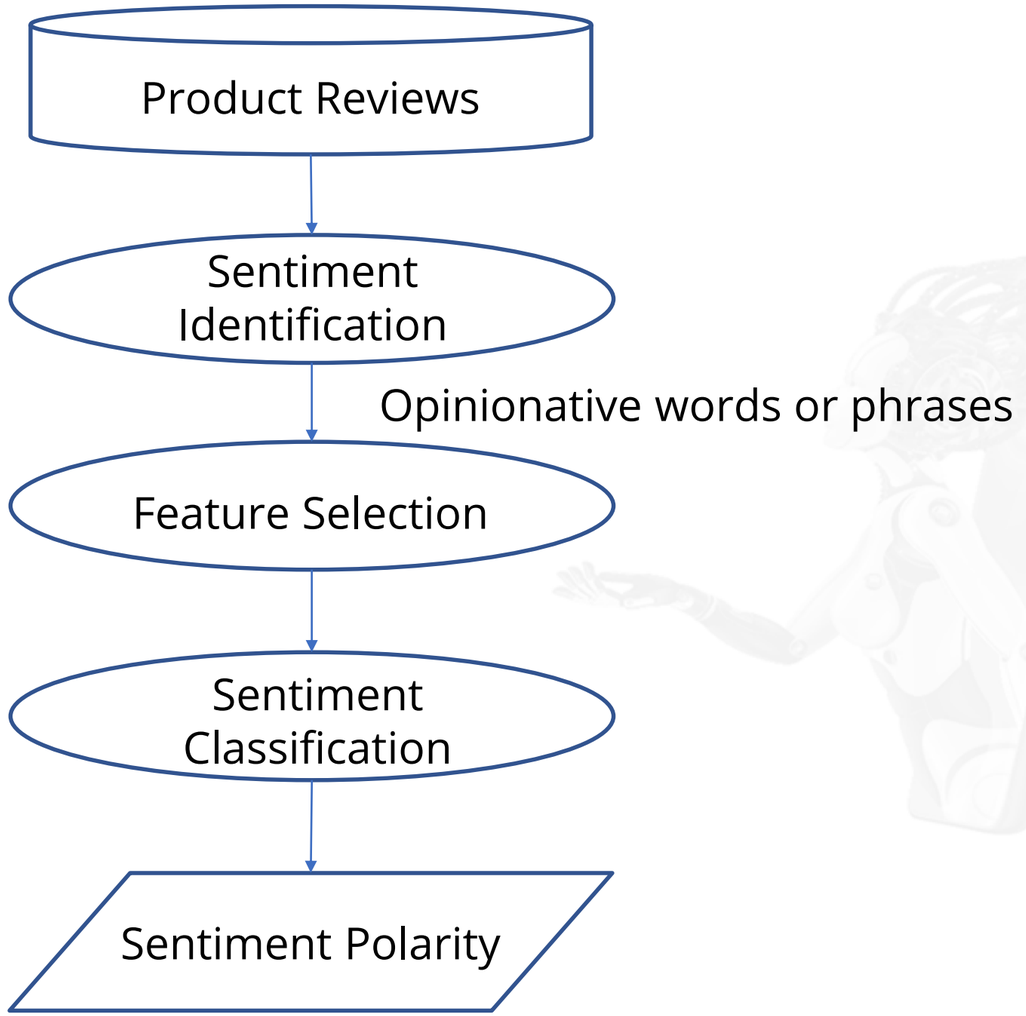
Sentiment Analysis

Sentiment Analysis

- Sentiment analysis is a technique used for text analysis using NLP or ML to assign weighted sentiment scores to the opinion mentioned in the text.
- Sentiment is a view or opinion that is held or expressed.
- It is widely applied to review and survey responses online and on social media, for providing good customer service and marketing.

Sentiment Analysis: Techniques





Sentiment Analysis: Techniques

Below are the steps involved in sentiment analysis:

1. User gives the feedback or comment on the portal
2. Categorization of possible sentiments is done
3. Feature selection is applied after noise removal and text-processing
4. ML or Lexicon-based approach will be applied for sentiment classification
5. Score polarity is defined

Sentiment Analysis: Applications

DATA AND
ARTIFICIAL INTELLIGENCE

Spacy

Spacy: Introduction

- Is a library for advanced Natural Language Processing in Python and Cython
- Supports 49+ languages
- Is easy to install and has simple and productive API
- Has efficient binary serialization
- Has easy model packaging and deployment

Spacy: Syntax and Library

System Requirements:

- Operating System: macOS or OS X, Linux, Windows (Cygwin, MinGW, Visual Studio)
- Python Version: Python 2.7, 3.5+ (only 64 bit)

```
import spacy

nlp = spacy.load("en_core_web_sm")

doc = nlp(u"This is a training.")
```

Spacy: POS Tagging

```
#spacy model load
import spacy

#loading english model
nlp = spacy.load('en')
doc = nlp("Simplilearn is one of the
world's leading certification providers.")

#printing POS tag for tokens
for token in doc:
    print(token.text, " --- ", token.pos_)
```

Spacy: POS Tagging

Output:

Simplilearn --- PROPN
is --- VERB
one --- NUM
of --- ADP
the --- DET
world --- NOUN
's --- PART
leading --- VERB
certification --- NOUN
providers --- NOUN
. --- PUNCT

In [28]:

```
#spacy model load
import spacy

#loading english model
nlp = spacy.load('en')
doc = nlp("Simplilearn is one of the world's leading certification providers.")

#printing POS tag for tokens
for token in doc:
    print(token.text, " --- ", token.pos_)
```

```
Simplilearn  ---  PROPN
is  ---  VERB
one  ---  NUM
of  ---  ADP
the  ---  DET
world  ---  NOUN
's  ---  PART
leading  ---  VERB
certification  ---  NOUN
providers  ---  NOUN
.  ---  PUNCT
```

Spacy: Dependency Parsing

```
#spacy model load
import spacy

#loading english model
nlp = spacy.load('en')
doc = nlp("Simplilearn is one of the world's leading
certification providers.")

#printing dependency for tokens
for token in doc:
    print(token.text," --> " ,token.lemma_, token.pos_,
token.tag_, token.dep_,
            token.shape_, token.is_alpha, token.is_stop)
```

Spacy: Dependency Parsing

Output:

Simplilearn --> Simplilearn PROPN NNP nsubj Xxxxx True False
is --> be VERB VBZ ROOT xx True True
one --> one NUM CD attr xxx True True
of --> of ADP IN prep xx True True
the --> the DET DT det xxx True True
world --> world NOUN NN poss xxxx True False
's --> 's PART POS case 'x False True
leading --> lead VERB VBG amod xxxx True False
certification --> certification NOUN NN compound xxxx True False
providers --> provider NOUN NNS pobj xxxx True False
. --> . PUNCT . punct . False False

In [23]:

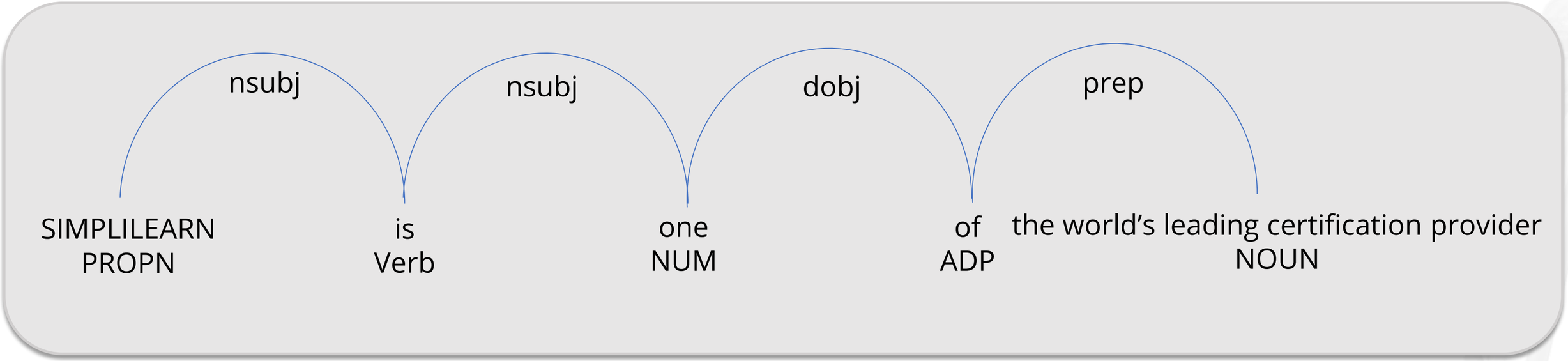
```
#spacy model load
import spacy

#loading english model
nlp = spacy.load('en')
doc = nlp("Simplilearn is one of the world's leading certification providers.")

#printing dependency for tokens
for token in doc:
    print(token.text," --> " ,token.lemma_, token.pos_, token.tag_, token.dep_,
            token.shape_, token.is_alpha, token.is_stop)
```

```
Simplilearn  -->  Simplilearn PROPN NNP nsubj Xxxxx True False
is  -->  be VERB VBZ ROOT xx True True
one  -->  one NUM CD attr xxx True True
of  -->  of ADP IN prep xx True True
the  -->  the DET DT det xxx True True
world  -->  world NOUN NN poss xxxx True False
's  -->  's PART POS case 'x False True
leading  -->  lead VERB VBG amod xxxx True False
certification  -->  certification NOUN NN compound xxxx True False
providers  -->  provider NOUN NNS pobj xxxx True False
.  -->  . PUNCT . punct . False False
```

Spacy: Dependency Parsing



Spacy: Named-Entity Recognition

```
#loading english model
nlp = spacy.load('en')
doc = nlp("Apple is looking at buying U.K. startup for
$1 billion")

#printing Named Entities
for ent in doc.ents:
    print(ent.text, ent.start_char, ent.end_char,
ent.label_)
```

Spacy: Named-Entity Recognition

Output:

Apple 0 5 ORG
U.K. 27 31 GPE
$1 billion 44 54 MONEY

In [27]:

```
#loading english model
nlp = spacy.load('en')
doc = nlp("Apple is looking at buying U.K. startup for $1 billion")

#printing Named Entities
for ent in doc.ents:
    print(ent.text, ent.start_char, ent.end_char, ent.label_)
```

```
Apple 0 5 ORG
U.K. 27 31 GPE
$1 billion 44 54 MONEY
```

Spacy: Limitations

1. Only few models available for processing
2. Tough to install in windows environment
3. Annotation tool is not good

Get the Polarity of the Given Survey for a Product

Problem Statement: Sentiment analysis is one of the most common applications in natural language processing. With sentiment analysis, we can decide with what emotion a text is written. Using sentiment analysis, get the polarity of the given survey for a product.

Access: Click on the **Practice Labs** tab on the left side panel of the LMS. Copy or note the username and password that is generated. Click on the **Launch Lab** button. On the page that appears, enter the username and password in the respective fields, and click **Login**.

ASSISTED PRACTICE

Extract City and Person Name from Text

Problem Statement: While conversing with other users, we try to understand what is being said and the entities mentioned in the sentence. Machine also has the capability to understand these entities and we can achieve this by using named-entity recognition. Apply existing NER library to extract city and person name from the given text.

Access: Click on the **Practice Labs** tab on the left side panel of the LMS. Copy or note the username and password that is generated. Click on the **Launch Lab** button. On the page that appears, enter the username and password in the respective fields, and click **Login**.

ASSISTED PRACTICE

Identifying Top Product Features from User Reviews

Objective: Analyze the reviews of a product on Amazon to identify the features of the product that customers are talking about the most. Use POS tagging and extracting the right parts of speech to analyze.

Problem Statement: Amazon Tap is a product made by Amazon. It is a portable Bluetooth and Wi-Fi-enabled speaker that gives you rich, full-range sound. Customers can use Alexa Voice Service that performs the role of a virtual assistant. Amazon is planning to launch a new version of Tap but, need to understand what features of the products are important for the customers. Amazon wants to analyze the reviews gathered on the product so far, to assess which of the product's features are most talked about by the customers.

LESSON-END PROJECT

Knowledge Check

Knowledge Check

1

Dependency parsing is used for _____.

a. Semantic understanding

b. Finding relation between words

c. Understanding modifiers

d. All of the above

Knowledge Check

1

Dependency parsing is used for _____.

a. Semantic understanding

b. Finding relation between words

c. Understanding modifiers

d. All of the above

The correct answer is **d.**

It helps in finding semantic structure with relation and modifiers.

Knowledge Check

2

I would like to extract the person's name from the paragraph. Which of the following techniques is suitable?

a. POS Tagging

b. Regular Expression

c. Word Sense Disambiguation

d. Named-Entity Recognition

Knowledge Check

2

I would like to extract the person's name from the paragraph. Which of the following techniques is suitable?

a. POS Tagging

b. Regular Expression

c. Word Sense Disambiguation

d. Named-Entity Recognition

The correct answer is **d.**

Named-entity recognition understands the names and gives you the named-entity words.

Knowledge Check

3

POS tags are the parts of speech tags which indicate the tag only for _____.

a. Single word

b. Consecutive word

c. N-gram phrases

d. Complete sentence

Knowledge Check

3

POS tags are the parts of speech tags which indicate the tag only for _____.

a. Single word

b. Consecutive word

c. N-gram phrases

d. Complete sentence

The correct answer is **a.**

POS tag indicates the tag only for single words. The combined tags are only for phrases.

Knowledge Check

4

“Sky is Blue". Which technique is best suitable to understand the meaning of "Blue"?

a. Word Sense Disambiguation

b. Dependency Parsing

c. Fuzzy Search

d. Coreference Resolution

Knowledge Check

4

“Sky is Blue". Which technique is best suitable to understand the meaning of "Blue"?

a. Word Sense Disambiguation

b. Dependency Parsing

c. Fuzzy Search

d. Coreference Resolution

The correct answer is **a.**

A word has many meanings in different contexts. WSD has the technique to overcome this problem.

Knowledge Check

5

What is the main application of sentiment analysis?

a. Understanding user feedback

b. Understanding syntactic structure

c. Fuzzy Search

d. Coreference Resolution

Knowledge Check

5

What is the main application of sentiment analysis?

a. Understanding user feedback

b. Understanding syntactic structure

c. Fuzzy Search

d. Coreference Resolution

The correct answer is **a.**

User's feedback is used to improve services and sentiment analysis plays a major part in this process.

Key Takeaways

You are now able to:

- Define parts-of-speech tagging
- Explain the different parsing methods
- Apply fuzzy search to identify similar words
- Get the polarity of the given survey for a product
- Extract city and person name from a text