UNIT 5

The complexity of minimax algorithm is…….if ”b” is the branching factor and “d” is the depth.

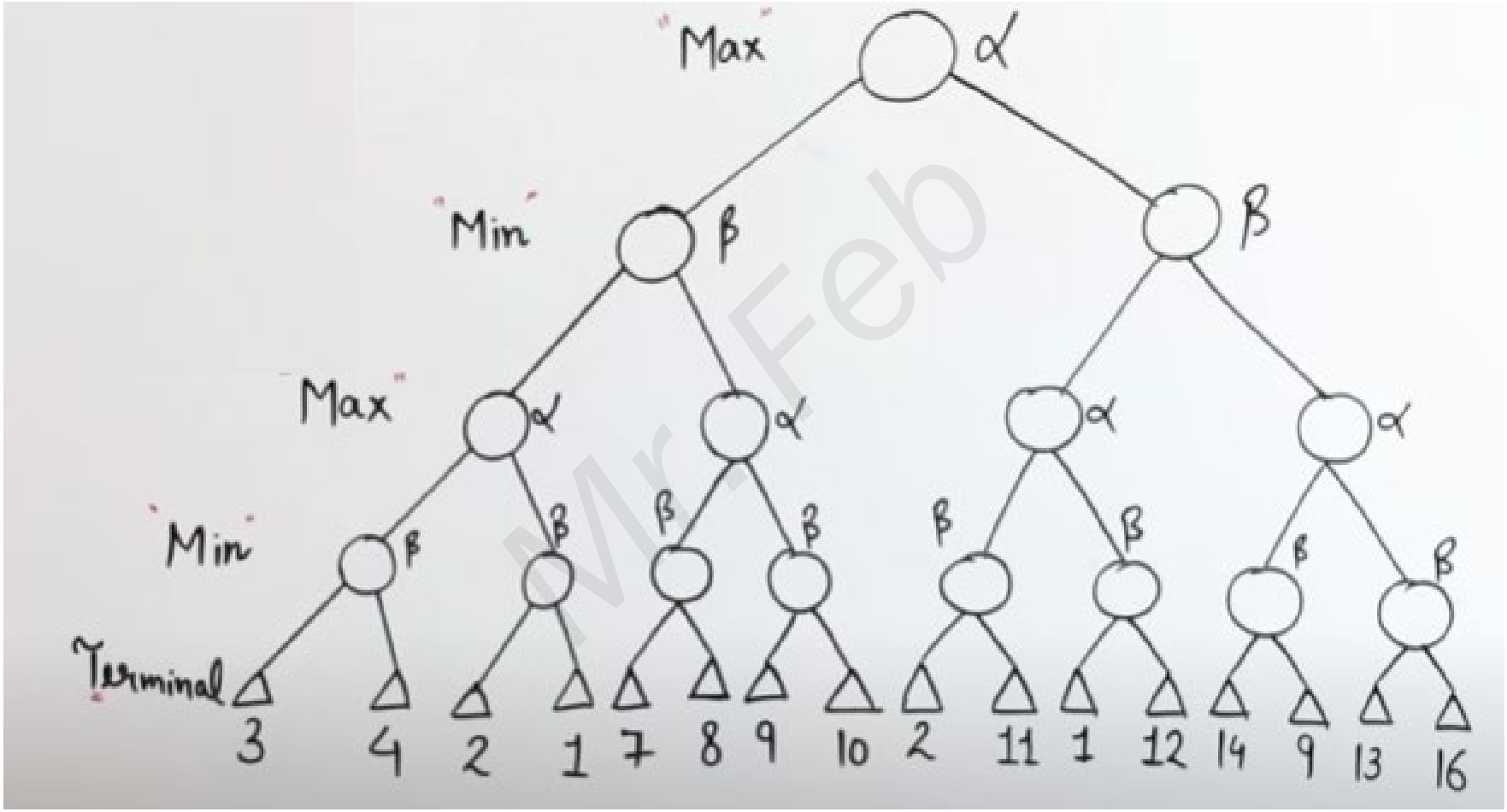
A. Same as of DFS
B. Space – O(bd) and time – O(bd)
C. Time – O(bd) and space – O(b^d)
D. Same as BFS

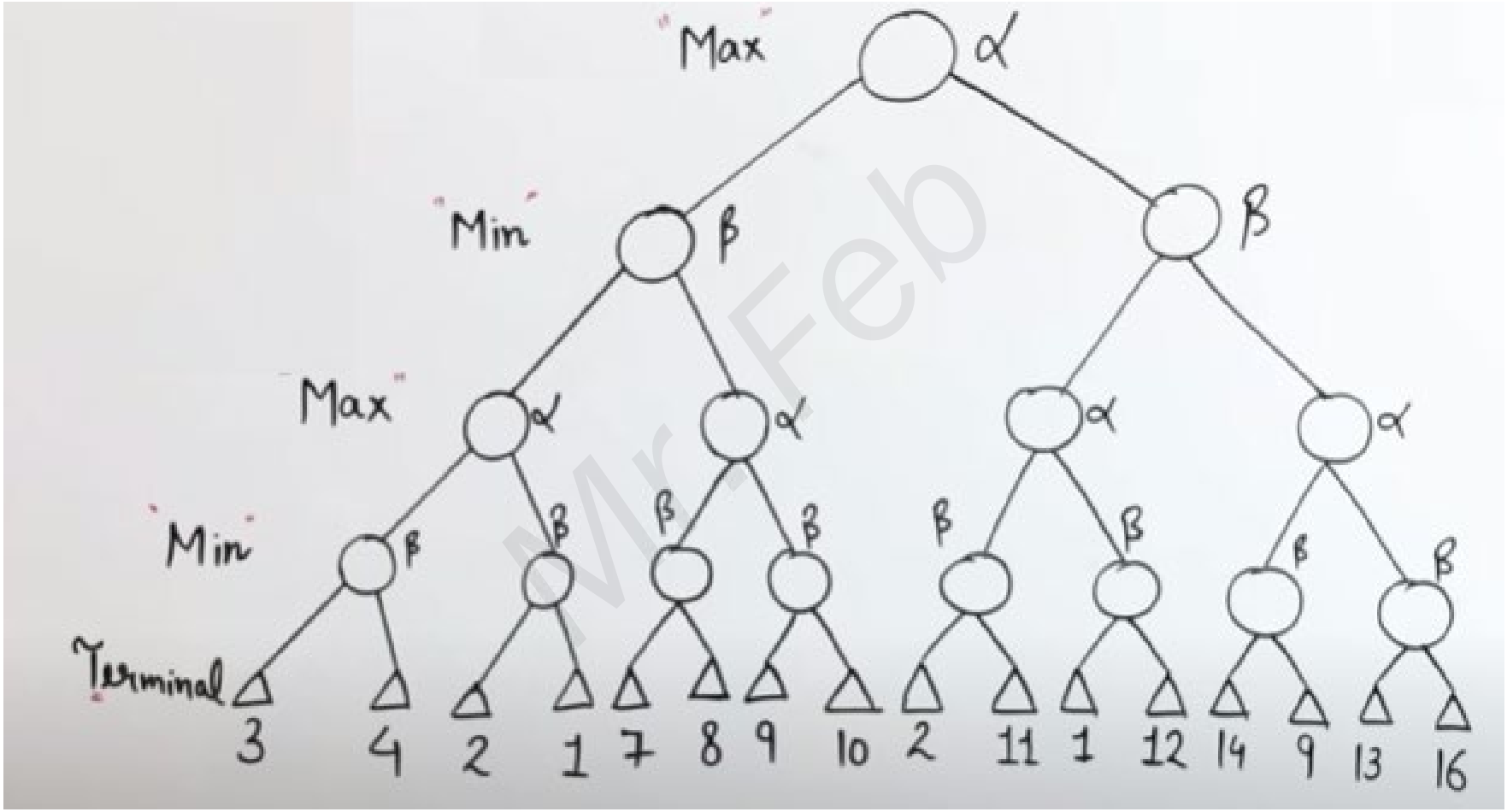
The complexity of minimax algorithm is…….if ”b” is the branching factor and “d” is the depth.

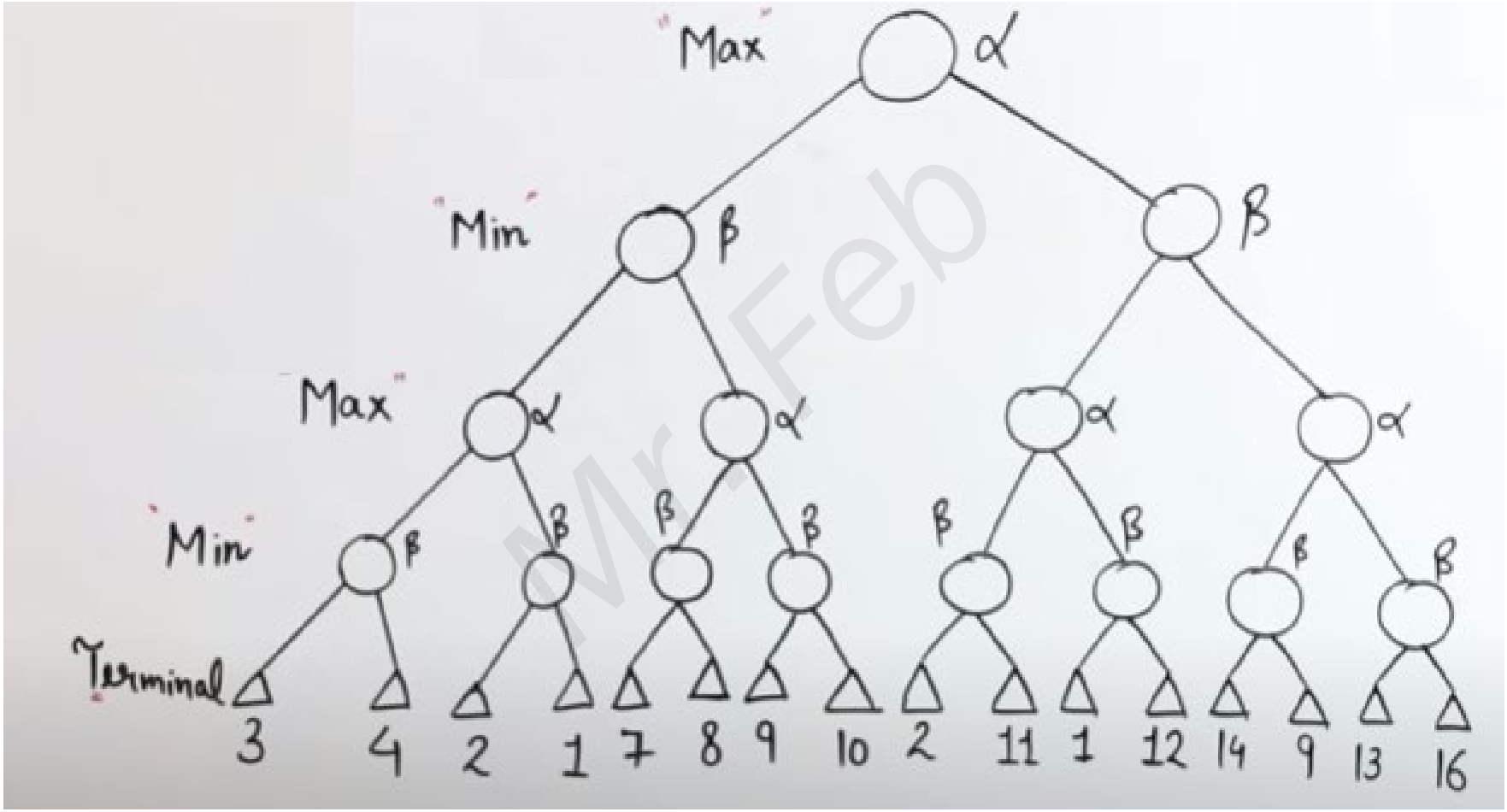
A. Time – O(bd) and space – O(b^d)
B. Space – O(bd) and time – O(bd)
C. Time – O(b^d) and space – O(bd)
D. Same as BFS

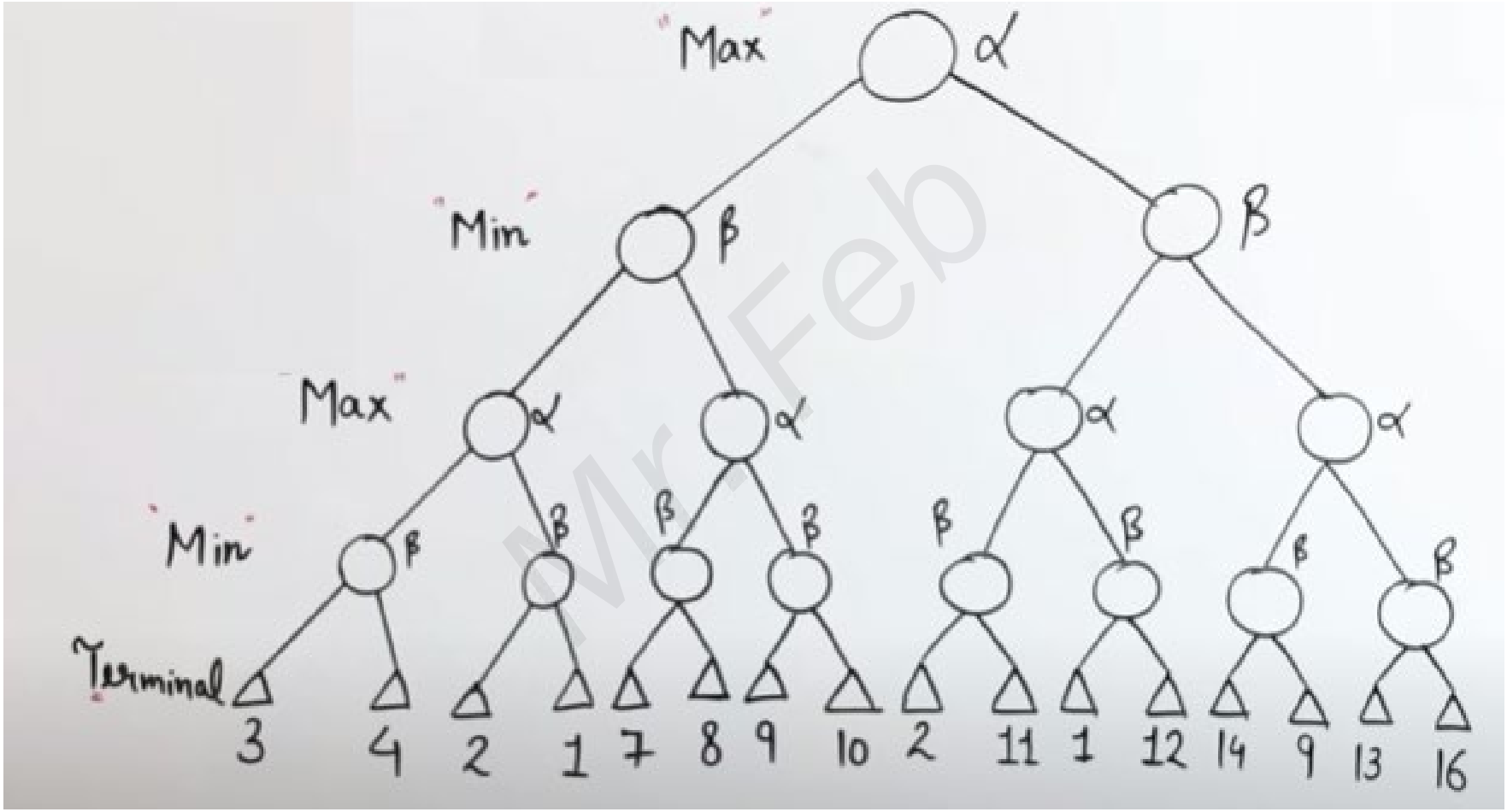
Time complexity of alpha-beta pruning algorithm is ……in ideal case and ……in worst case.

1. $O(b^d)$, $O(bd)$
2. $O(b^d)$, $O(b^{d/2})$
3. $O(b^{d/2})$, $O(b^d)$
4. $O(bd/2)$, $O(bd)$


Max
α
Min
β
β
Max
α
α
α
α
Min
β
β
β
β
β
β
β
β
Terminal
3
4
2
1
7
8
9
10
2
11
1
12
14
9
13
16


Max
α
Min
β
β
Max
α
α
α
α
Min
β
β
β
β
β
β
β
β
Terminal
3
4
2
1
7
8
9
10
2
11
1
12
14
9
13
16

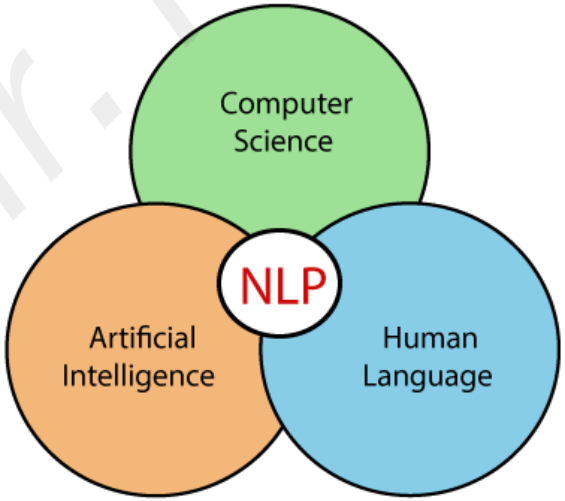

Max
α
Min
β
β
Max
α
α
α
α
Min
β
β
β
β
β
β
β
β
Terminal
3
4
2
1
7
8
9
10
2
11
1
12
14
9
13
16


Max
α
Min
β
β
Max
α
α
α
α
Min
β
β
β
β
β
β
β
β
Terminal
3
4
2
1
7
8
9
10
2
11
1
12
14
9
13
16

Natural Language Processing

What is NLP?

NLP stands for **Natural Language Processing**, which is a part of **Computer Science, Human language,** and **Artificial Intelligence**. It is the technology that is used by machines to understand, analyse, manipulate, and interpret human's languages. It helps developers to organize knowledge for performing tasks such as **translation, automatic summarization, Named Entity Recognition (NER), speech recognition, relationship extraction,** and **topic segmentation**.



Applications of NLP

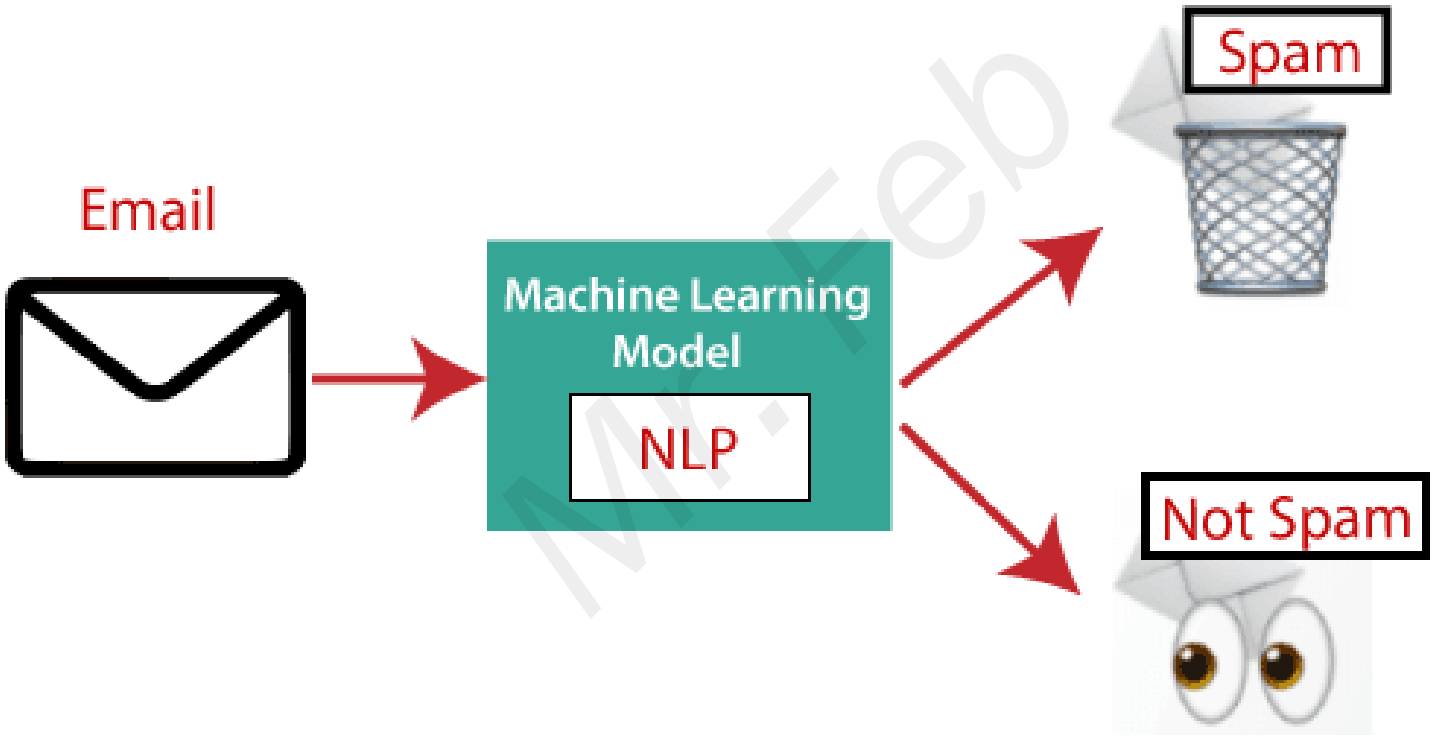
1. Question Answering

Question Answering focuses on building systems that automatically answer the questions asked by humans in a natural language.

Applications of NLP

2. Spam Detection

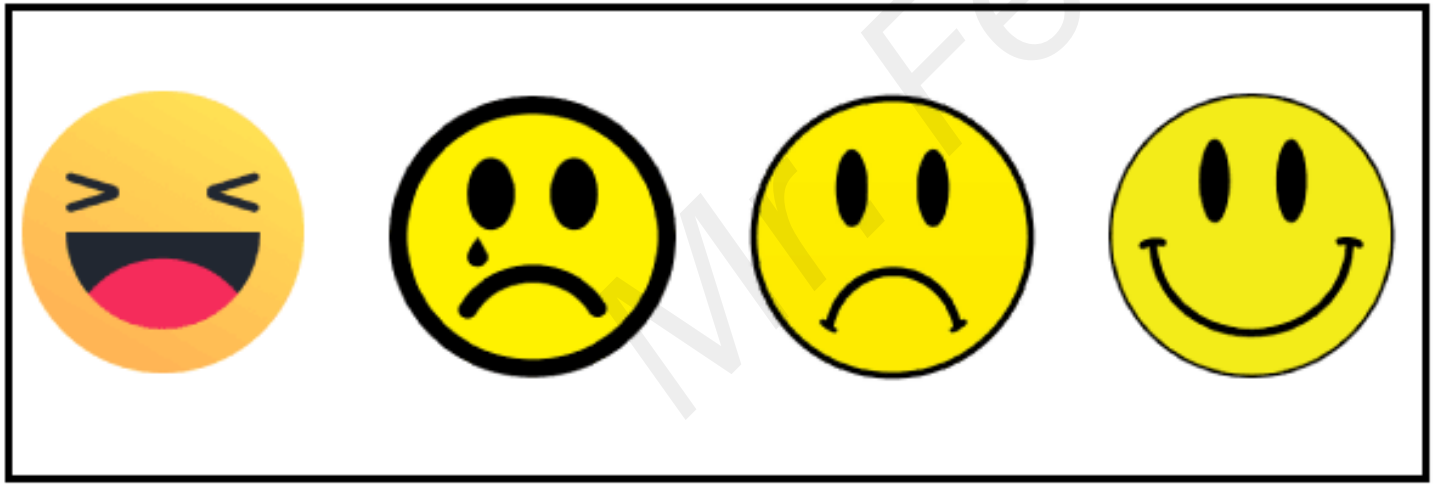
Spam detection is used to detect unwanted e-mails getting to a user's inbox.



Applications of NLP

3. Sentiment Analysis

Sentiment Analysis is also known as **opinion mining**. It is used on the web to analyse the attitude, behaviour, and emotional state of the sender. This application is implemented through a combination of NLP (Natural Language Processing) and statistics by assigning the values to the text (positive, negative, or natural), identify the mood of the context (happy, sad, angry, etc.)



Applications of NLP

4. Machine Translation

Machine translation is used to translate text or speech from one natural language to another natural language.

Like: Google Translator

5. Spelling correction

Microsoft Corporation provides word processor software like MS-word, PowerPoint for the spelling correction.

Applications of NLP

6. Speech Recognition

Speech recognition is used for converting spoken words into text. It is used in applications, such as mobile, home automation, video recovery, dictating to Microsoft Word, voice biometrics, voice user interface, and so on.

7. Chatbot

Implementing the Chatbot is one of the important applications of NLP. It is used by many companies to provide the customer's chat services.

Applications of NLP

8. Information extraction

Information extraction is one of the most important applications of NLP. It is used for extracting structured information from unstructured or semi-structured machine-readable documents.

9. Natural Language Understanding (NLU)

It converts a large set of text into more formal representations such as first-order logic structures that are easier for the computer programs to manipulate notations of the natural language processing.

Advantages of NLP

- NLP helps users to ask questions about any subject and get a direct response within seconds.
- NLP offers exact answers to the question means it does not offer unnecessary and unwanted information.
- NLP helps computers to communicate with humans in their languages.
- It is very time efficient.
- Most of the companies use NLP to improve the efficiency of documentation processes, accuracy of documentation, and identify the information from large databases.

Disadvantages of NLP

A list of disadvantages of NLP is given below:

- NLP may not show context.
- NLP is unpredictable
- NLP may require more keystrokes.
- NLP is unable to adapt to the new domain, and it has a limited function that's why NLP is built for a single and specific task only.

1. Natural Language Understanding (NLU)

Natural Language Understanding (NLU) helps the machine to understand and analyse human language by extracting the metadata from content such as concepts, entities, keywords, emotion, relations, and semantic roles.

NLU mainly used in Business applications to understand the customer's problem in both spoken and written language.

NLU involves the following tasks -

- It is used to map the given input into useful representation.
- It is used to analyze different aspects of the language.

2. Natural Language Generation (NLG)

It is the process of producing meaningful phrases and sentences in the form of natural language from some internal representation.

It involves −

•Text planning − It includes retrieving the relevant content from knowledge base.

•Sentence planning − It includes choosing required words, forming meaningful phrases, setting tone of the sentence.

•Text Realization − It is mapping sentence plan into sentence structure.

Difference between NLU and NLG

NLU	NLG
NLU is the process of reading and interpreting language.	NLG is the process of writing or generating language.
It produces non-linguistic outputs from natural language inputs.	It produces constructing natural language outputs from non-linguistic inputs.

Why NLP is difficult?

NLP is difficult because Ambiguity and Uncertainty exist in the language. There are the following three ambiguity -

1. Lexical Ambiguity

Lexical Ambiguity exists in the presence of two or more possible meanings of the sentence within a single word.

Example:

Manya is looking for a **match**.

In the above example, the word match refers to that either Manya is looking for a partner or Manya is looking for a match. (Cricket or other match)

2. Syntactic Ambiguity

Syntactic Ambiguity exists in the presence of two or more possible meanings within the sentence.

Example:

“He lifted the beetle with red cap.” − Did he use cap to lift the beetle or he lifted a beetle that had red cap?

3. Referential Ambiguity

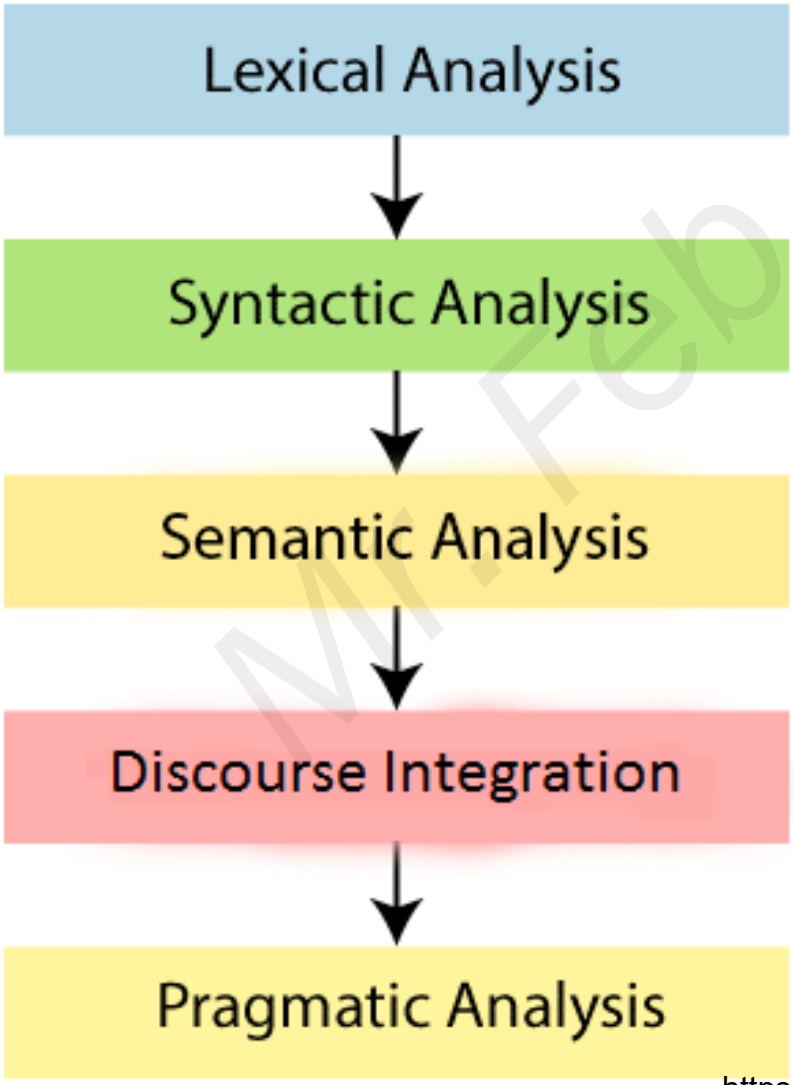
Referential Ambiguity exists when you are referring to something using the pronoun.

Example: Kiran went to Sunita. She said, "I am hungry."

In the above sentence, you do not know that who is hungry, either Kiran or Sunita.

Phases of NLP

There are the following five phases of NLP:



1. Lexical Analysis − It involves identifying and analyzing the structure of words. Lexicon of a language means the collection of words and phrases in a language. Lexical analysis is dividing the whole chunk of txt into paragraphs, sentences, and words.

2. Syntactic Analysis (Parsing) − It involves analysis of words in the sentence for grammar and arranging words in a manner that shows the relationship among the words. The sentence such as “The school goes to boy” is rejected by English syntactic analyzer.

3. Semantic Analysis − It draws the exact meaning or the dictionary meaning from the text. The text is checked for meaningfulness. It is done by mapping syntactic structures and objects in the task domain. The semantic analyzer disregards sentence such as “hot ice-cream”.

4. Discourse Integration − The meaning of any sentence depends upon the meaning of the sentence just before it. In addition, it also brings about the meaning of immediately succeeding sentence.

5. Pragmatic Analysis − During this, what was said is re-interpreted on what it actually meant. It involves deriving those aspects of language which require real world knowledge.

For Example: "Open the door" is interpreted as a request instead of an order.

What is the field of Natural Language Processing (NLP)?

a) Computer Science
b) Artificial Intelligence
c) Linguistics
d) All of the mentioned

What is the field of Natural Language Processing (NLP)?

a) Computer Science
b) Artificial Intelligence
c) Linguistics
d) All of the mentioned

What is the main challenge/s of NLP?

a) Handling Ambiguity of Sentences

b) Handling Tokenization

c) Handling POS-Tagging

d) All of the mentioned

What is the main challenge/s of NLP?

a) Handling Ambiguity of Sentences

b) Handling Tokenization

c) Handling POS-Tagging

d) All of the mentioned

Choose form the following areas where NLP can be useful.

a) Automatic Text Summarization
b) Automatic Question-Answering Systems
c) Information Retrieval
d) All of the mentioned

Choose form the following areas where NLP can be useful.

a) Automatic Text Summarization
b) Automatic Question-Answering Systems
c) Information Retrieval
d) All of the mentioned

What is Machine Translation?

a) Converts one human language to another

b) Converts human language to machine language

c) Converts any human language to English

d) Converts Machine language to human language

What is Machine Translation?

a) Converts one human language to another

b) Converts human language to machine language

c) Converts any human language to English

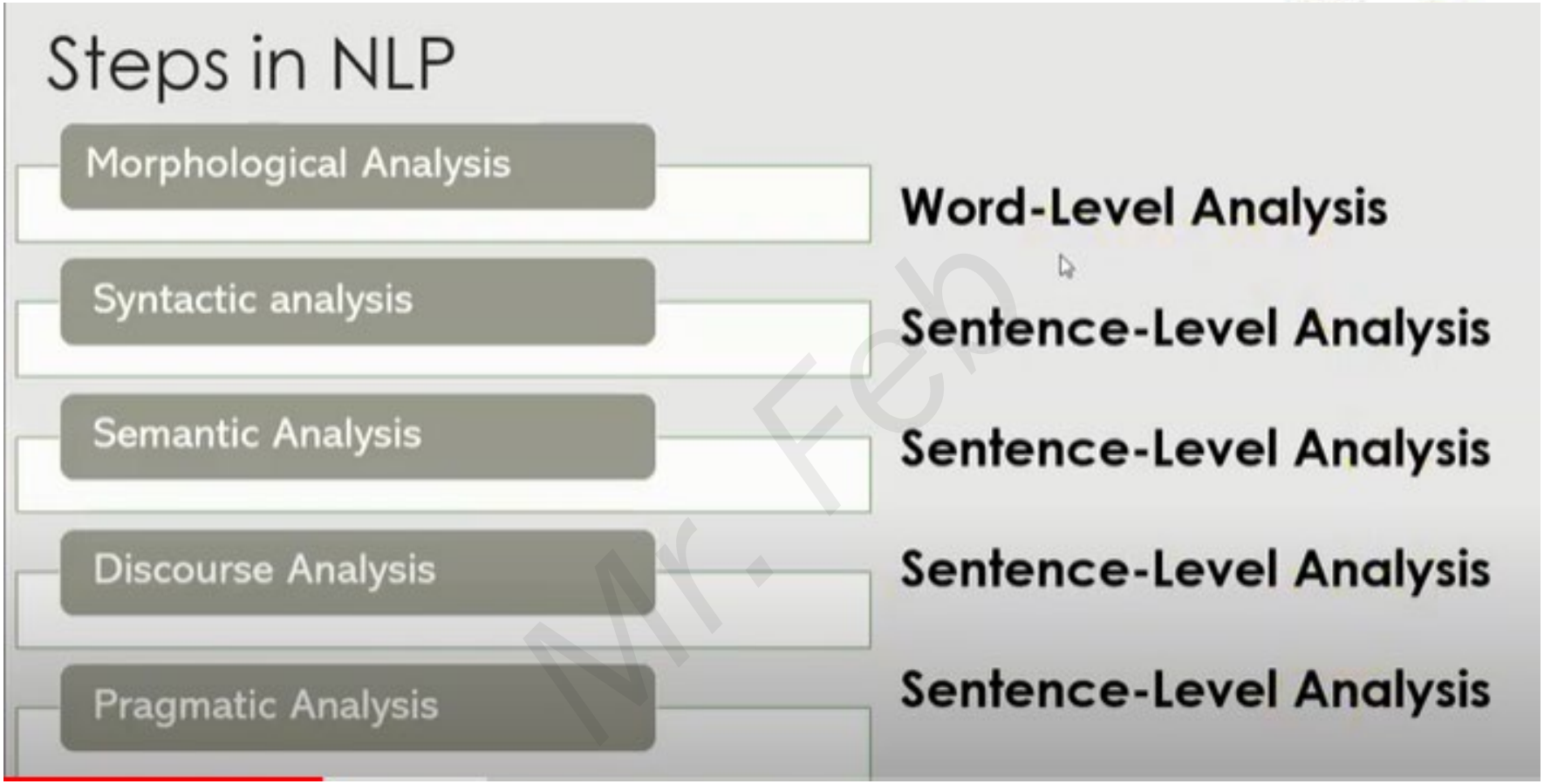
d) Converts Machine language to human language

Natural language processing is divided into the two subfields of -

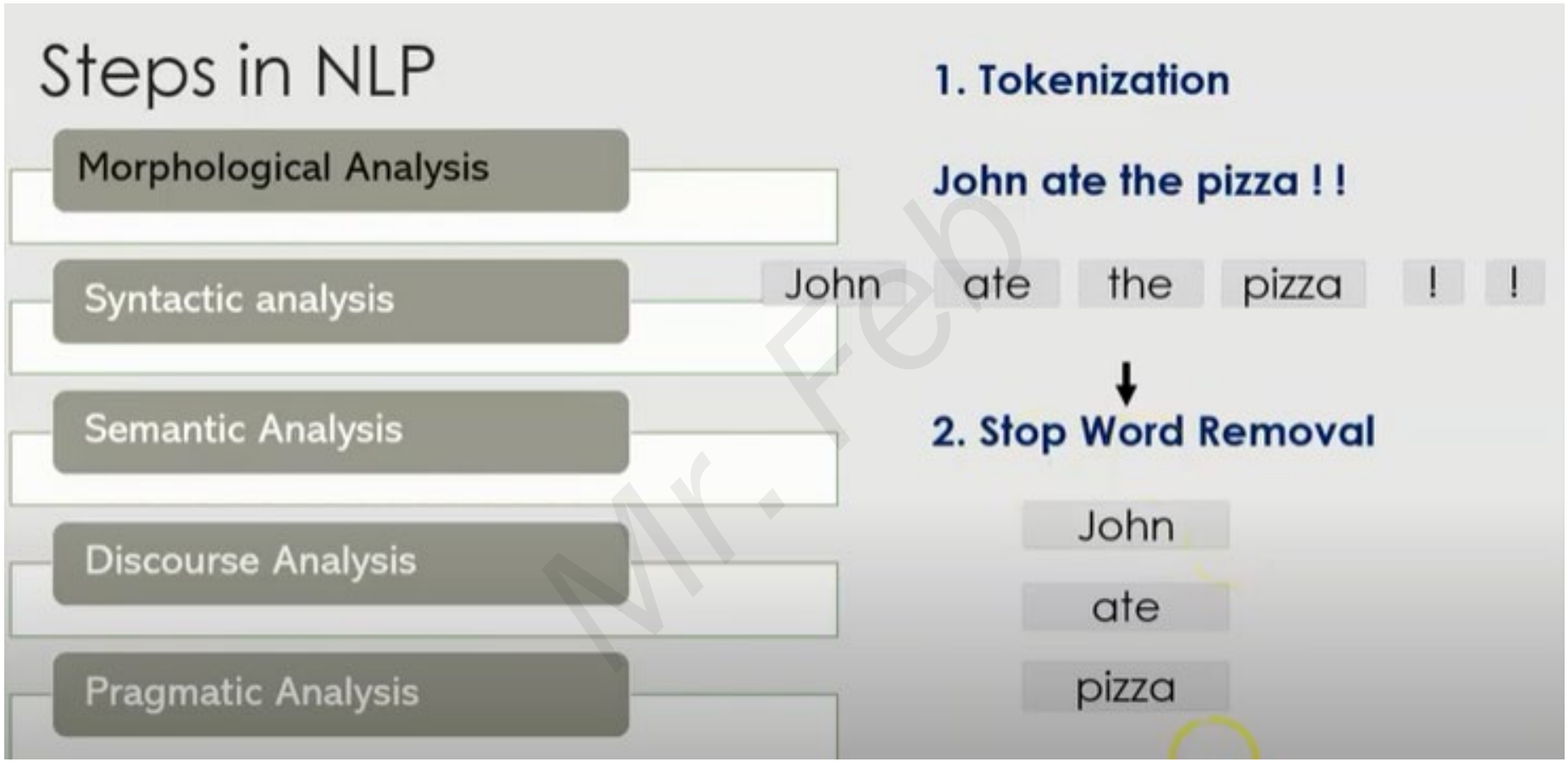
A. symbolic and numeric
B. algorithmic and heuristic
C. time and motion
D. understanding and generation

Natural language processing is divided into the two subfields of -

A. symbolic and numeric
B. algorithmic and heuristic
C. time and motion
D. understanding and generation


Steps in NLP
Morphological Analysis
Word-Level Analysis
Syntactic analysis
Sentence-Level Analysis
Semantic Analysis
Sentence-Level Analysis
Discourse Analysis
Sentence-Level Analysis
Pragmatic Analysis
Sentence-Level Analysis

1. Lexical Analysis/Morphological Analysis − It involves identifying and analyzing the structure of words. Lexicon of a language means the collection of words and phrases in a language. Lexical analysis is dividing the whole chunk of txt into paragraphs, sentences, and words.



Steps in NLP

- Morphological Analysis
- Syntactic analysis
- Semantic Analysis
- Discourse Analysis
- Pragmatic Analysis

3. Stemming

Stemming is a process of reducing words into its base form (Root form/stem form)

John->John
Ate -> eat
Pizza->Pizza

- car, cars -> car
- run, ran, running -> run
- stemmer, stemming, stemmed -> stem

Steps in NLP

- Morphological Analysis
- Syntactic analysis
- Semantic Analysis
- Discourse Analysis
- Pragmatic Analysis

4. N-Gram Language Model

1-gram-	John	Ate	the	Pizza
Bigram -	John	Ate	the	Pizza
Trigram -	John	Ate	the	Pizza
4-Gram -	John	Ate	the	Pizza

John Ate the _?

2. Syntactic Analysis (Parsing) − It involves analysis of words in the sentence for grammar and arranging words in a manner that shows the relationship among the words. The sentence such as “The school goes to boy” is rejected by English syntactic analyzer.

Steps in NLP

- Morphological Analysis
- Syntactic analysis
- Semantic Analysis
- Discourse Analysis
- Pragmatic Analysis

John Ate the Apple ✓

Ate the Apple John ✗

Steps in NLP

- Morphological Analysis
- Syntactic analysis
- Semantic Analysis
- Discourse Analysis
- Pragmatic Analysis

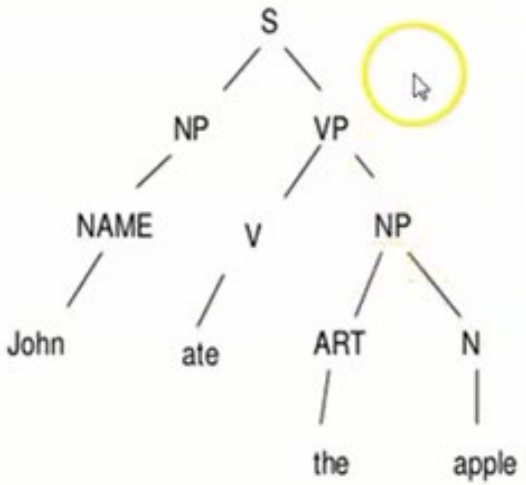
John Ate the Apple ✓

Ate the Apple John ✗

- A parse tree :

John ate the apple.

1. S -> NP VP
2. VP -> V NP
3. NP -> NAME
4. NP -> ART N
5. NAME -> John
6. V -> ate
7. ART-> the
8. N -> apple



What is Morphological Segmentation?

a) Does Discourse Analysis

b) Separate words into individual morphemes and identify the class of the morphemes

c) Is an extension of propositional logic

d) None of the mentioned

What is Morphological Segmentation?

a) Does Discourse Analysis

b) Separate words into individual morphemes and identify the class of the morphemes

c) Is an extension of propositional logic

d) None of the mentioned

OCR (Optical Character Recognition) uses NLP.

a) True

b) False

OCR (Optical Character Recognition) uses NLP.

a) True

b) False

In linguistic morphology _____ is the process for reducing inflected words to their root form.

a) Rooting
b) Stemming
c) Text-Proofing
d) Both Rooting & Stemming

In linguistic morphology _____ is the process for reducing inflected words to their root form.

a) Rooting

b) Stemming

c) Text-Proofing

d) Both Rooting & Stemming

3. Semantic Analysis − It draws the exact meaning or the dictionary meaning from the text. The text is checked for meaningfulness. It is done by mapping syntactic structures and objects in the task domain. The semantic analyzer disregards sentence such as “hot ice-cream”.

Steps in NLP

- Morphological Analysis
- Syntactic analysis
- Semantic Analysis
- Discourse Analysis
- Pragmatic Analysis

She drank Some Milk ✓

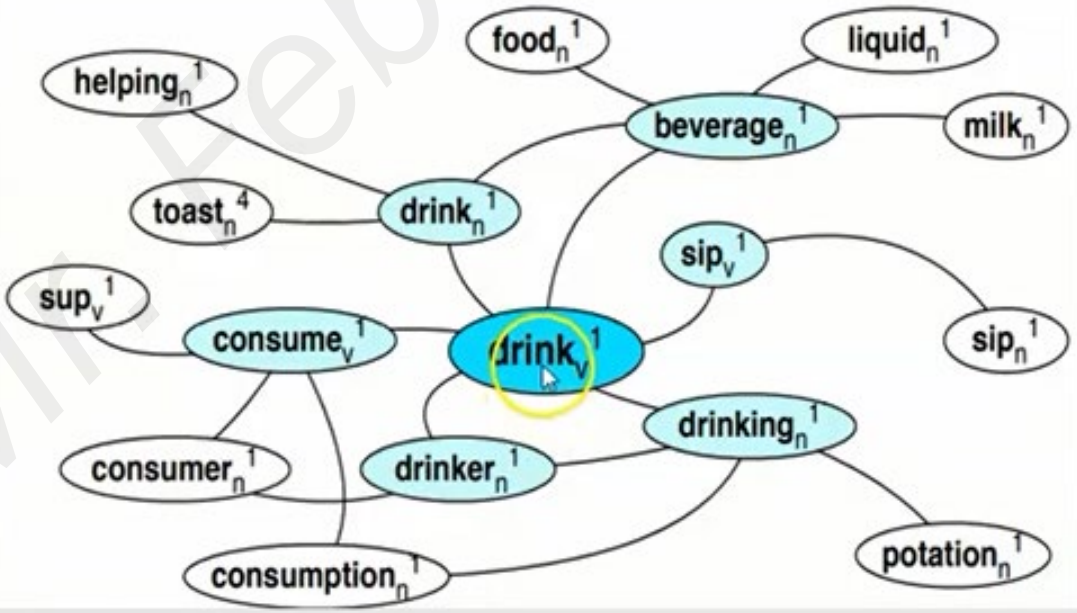
She drank Some books ✗

Steps in NLP

- Morphological Analysis
- Syntactic analysis
- Semantic Analysis
- Discourse Analysis
- Pragmatic Analysis

She drank Some Milk ✓

She drank Some books ✗



4. Discourse Integration − The meaning of any sentence depends upon the meaning of the sentence just before it. In addition, it also brings about the meaning of immediately succeeding sentence.

Steps in NLP

- Morphological Analysis
- Syntactic analysis
- Semantic Analysis
- Discourse Analysis
- Pragmatic Analysis

- **Monkeys Eat Banana, when they Wake up.**

  Who is they here?
  -Monkey

- **Monkeys eat Banana, when they are ripe.**

  Who is they here?
  -Banana

5. Pragmatic Analysis − During this, what was said is re-interpreted on what it actually meant. It involves deriving those aspects of language which require real world knowledge.

For Example: "Open the door" is interpreted as a request instead of an order.

Steps in NLP

- Morphological Analysis
- Syntactic analysis
- Semantic Analysis
- Discourse Analysis
- Pragmatic Analysis

Close the Door

-Order

Please Close the Door

- Request ,affirmation

Quiz

Semantic analysis is done after

A. Morphological phase
B. Syntactic phase
C. Discourse
D. None of these

Quiz

Semantic analysis is done after

A. Morphological phase

B. Syntactic phase

C. Discourse

D. None of these

Quiz

"I taught am teaching." which phase generates error

A. Pragmatic

B. Semantic

C. Syntactic

D. Lexical

Quiz

"I taught am teaching." which phase generates error

A. Pragmatic

B. Semantic

C. Syntactic

D. Lexical

N-grams are defined as the combination of N keywords together. How many bi-grams can be generated from a given sentence:

“Analytics Vidhya is a great source to learn data science”

A) 7

B) 8

C) 9

D) 10

N-grams are defined as the combination of N keywords together. How many bi-grams can be generated from a given sentence:

“Analytics Vidhya is a great source to learn data science”

A) 7

B) 8

C) 9

D) 10

Parse Tree

A parser is a program, that accepts as input a sequence of words in a natural language and breaks them up into parts (nouns, verbs, and their attributes), to be managed by other programming.

- Parsing can be defined as the act of analyzing the grammaticality an utterance according to some specific grammar.
- Parsing is the process to check, that a particular sequence of words in a sentence correspond to a language defined by its grammar.
- Parsing means show how we can get from the start symbol of the grammar to the sequence of words using the production rules.
- The output of a parser is *a Parse tree*.

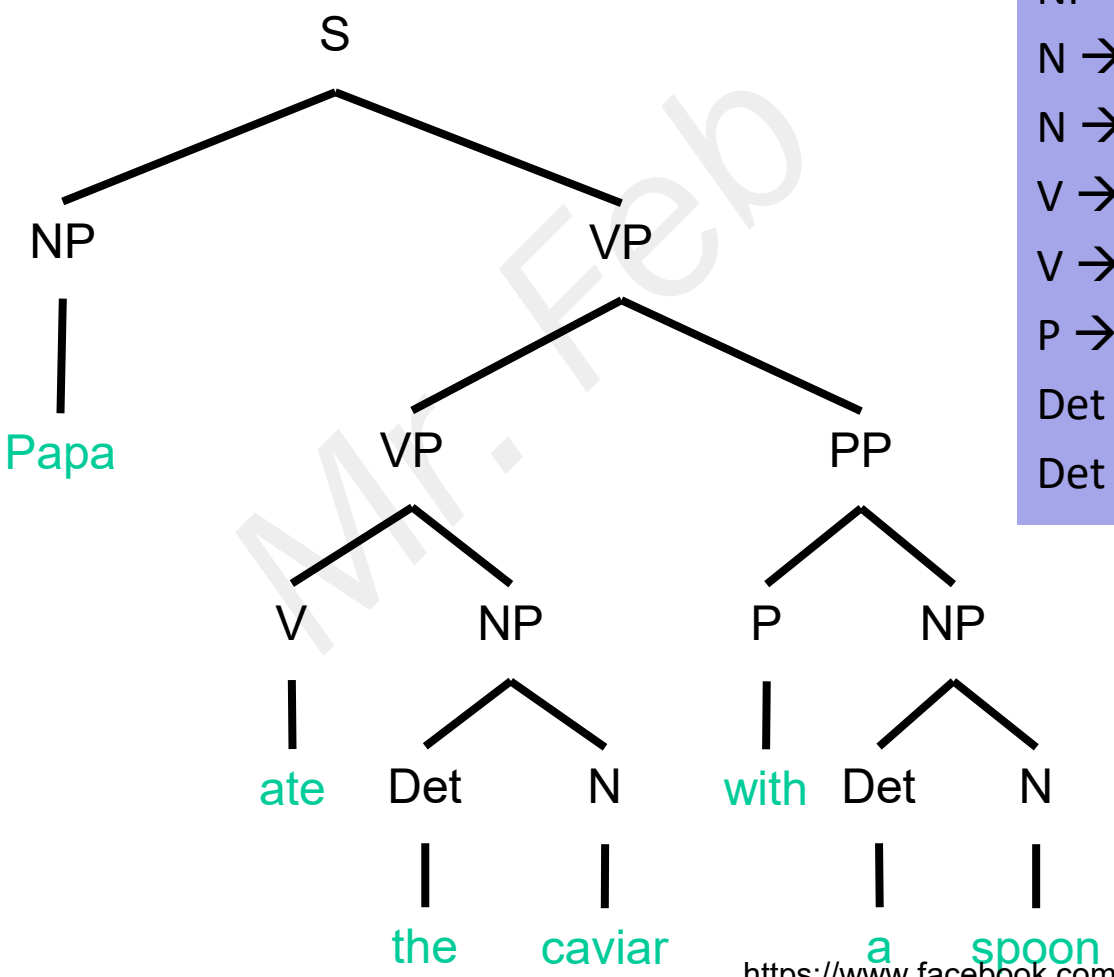
- Syntactic processing have **two** main components:
  - A ***declarative representation, called a grammar***
  - A ***procedure, called parser,*** that compares the grammar against input sentences to produce parsed structures.

Parse Tree is a way of representing the output of a parser.

- Each phrasal constituent found during parsing becomes a branch node of the parse tree;
- the words of the sentence become the leaves of the parse tree;
- there can be more than one parse tree for a single sentence;

What is Parsing?

S → NP VP
NP → Det N
NP → NP PP
VP → V NP
VP → VP PP
PP → P NP



NP → Papa
N → caviar
N → spoon
V → spoon
V → ate
P → with
Det → the
Det → a

Parsing

To parse a sentence, it is necessary to find a way in which the sentence could have been generated from the start symbol. There two ways to do : One, Top-Down Parsing and the other, Bottom-UP Parsing.

- **Top-Down Parsing**

  Begin with the start symbol and apply the grammar rules forward until the symbols at the terminals of the tree corresponds to the components of the sentence being parsed.

- **Bottom-UP Parsing**

  Begin with the sentence to be parsed and apply the grammar rules backward until a single tree whose terminals are the wards of the sentence and whose top node is the start symbol has been produced.

Tokenization

Tokenization is essentially splitting a phrase, sentence, paragraph, or an entire text document into smaller units, such as individual words or terms. Each of these smaller units are called tokens.

Why is Tokenization required in NLP?

- Before processing a natural language, we need to identify the *words* that constitute a string of characters. That's why tokenization is the most basic step to proceed with NLP (text data). **This is important because the meaning of the text could easily be interpreted by analyzing the words present in the text.**
- Let's take an example. Consider the below string:

"This is a cat."

What do you think will happen after we perform tokenization on this string?

['This', 'is', 'a', cat'].

- There are numerous uses of doing this. We can use this tokenized form to:
- Count the number of words in the text
- Count the frequency of the word, that is, the number of times a particular word is present

Give the “isa” representation of the sentence, “All Punjabis are Indians.”

A. isa(Punjabi, Indian)
B. isa(Indian, Punjabi)
C. ∀x: isa(x, Punjabi) → isa(x, Indian)
D. isa(x, Punjabi) → isa(x, Indian)

Which of the following is the correct meaning of “girl(Catherine)” in predicate logic?

A. Catherine is a girl.
B. Catherine was a girl.
C. Both “Catherine is a girl” and “Catherine was a girl” are correct.
D. Above representation is not in line with predicate logic.

S1 ⇔ S2 in propositional logic denotes

A. Implication of S1 to S2.
B. Implication of S2 to S1.
C. Implication of S1 to S2 or Implication of S2 to S1.
D. Biconditional between S1 and S2.

Bag of Words Model

Bag of Words (BoW) **model** is a simple algorithm used in **Natural Language Processing**. In BoW **model** a sentence or a document is considered as a '**Bag**' containing **words**. It will take into account the **words** and their frequency of occurrence in the sentence or the document disregarding semantic relationship in the sentences.

Text Mining (TM)

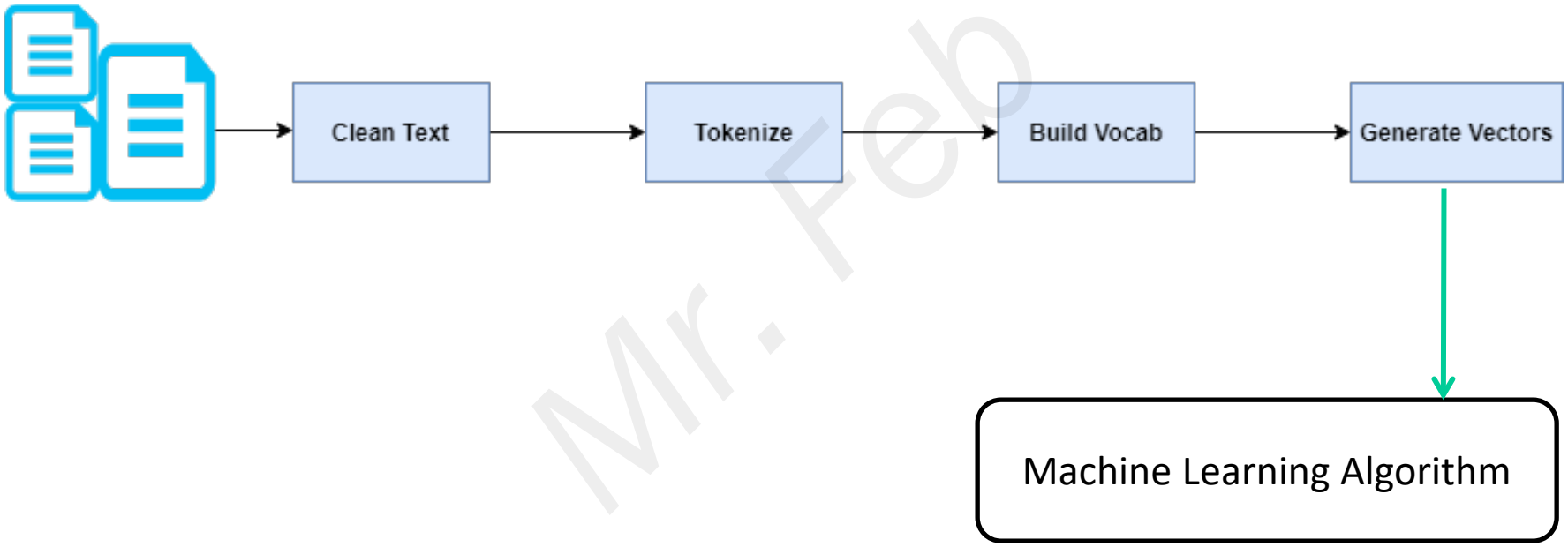
Bag Of Words Representation Of Text

- Considers text a simple set/bag of words
- Based on the following (unrealistic) assumptions:
  - words are mutually independent,
  - word order in text is irrelevant
- But, **highly effective**, and is often used in TM

- BOW extracts features from text documents.
- These features can be used for training machine learning algorithms.
- It creates a vocabulary of unique words occurring in all the documents in the training set.
- it's a collection of words to represent a sentence with word count and mostly disregarding the order in which they appear.

BOW is an approach widely used with:

- Natural language processing
- Information retrieval from documents
- Document classifications


Clean Text
Tokenize
Build Vocab
Generate Vectors
Machine Learning Algorithm

Example of the Bag-of-Words Model

Step 1: Collect Data

- Below is a snippet of the first few lines of text from the book "A Tale of Two Cities" by Charles Dickens, taken from Project Gutenberg.
- *It was the best of times,*
  it was the worst of times,
  it was the age of wisdom,
  it was the age of foolishness,
- For this small example, let's treat each line as a separate "document" and the 4 lines as our entire corpus of documents.

Step 2: Design the Vocabulary

Now we can make a list of all of the words in our model vocabulary.

- The unique words here (ignoring case and punctuation) are:

"it"

"was"

"the"

"best"

"of"

"times"

"worst"

"age"

"wisdom"

"foolishness"

- That is a vocabulary of 10 words from a corpus containing 24 words.

Step 3: Create Document Vectors

- The next step is to score the words in each document.
- The objective is to turn each document of free text into a vector that we can use as input or output for a machine learning model.
- Because we know the vocabulary has 10 words, we can use a fixed-length document representation of 10, with one position in the vector to score each word.
- The simplest scoring method is to mark the presence of words as a boolean value, 0 for absent, 1 for present.
- Using the arbitrary ordering of words listed above in our vocabulary, we can step through the first document ("*It was the best of times*") and convert it into a binary vector.

- The scoring of the document would look as follows:

"it" = 1

"was" = 1

"the" = 1

"best" = 1

"of" = 1

"times" = 1

"worst" = 0

"age" = 0

"wisdom" = 0

"foolishness" = 0

As a binary vector, this would look as follows:

```
1 [1, 1, 1, 1, 1, 1, 0, 0, 0, 0]
```

The other three documents would look as follows:

```
1 "it was the worst of times" = [1, 1, 1, 0, 1, 1, 1, 0, 0, 0]
2 "it was the age of wisdom" = [1, 1, 1, 0, 1, 0, 0, 1, 1, 0]
3 "it was the age of foolishness" = [1, 1, 1, 0, 1, 0, 0, 1, 0, 1]
```

Consider the below two sentences.

1. "John likes to watch movies. Mary likes movies too.“

2. "John also likes to watch football games."

These two sentences can be also represented with a collection of words.

1. ['John', 'likes', 'to', 'watch', 'movies.', 'Mary', 'likes', 'movies', 'too.']

2. ['John', 'also', 'likes', 'to', 'watch', 'football', 'games']

Remove multiple occurrences and use the word count.

1. {"John":1,"likes":2,"to":1,"watch":1,"movies":2,"Mary":1,"too":1}

2. {"John":1,"also":1,"likes":1,"to":1,"watch":1,"football":1, "games":1}

- Tabular form for all documents

Words	Frequencies
John	2
Likes	3
To	2
Watch	2
Movies	2
Mary	1
Too	1
also	1
Football	1
games	1

Create a vector whose length is equals to total length of vocabulary

“John likes to watch movies. Mary likes movies too”
[1, 2, 1, 1, 2, 1, 1, 0, 0, 0]

“John also likes to watch football games”
[1, 1, 1, 1, 0, 0, 0, 1, 1, 1]

t1	t2	t3	t4	t5	t6	….	…..	tn
w1	w2	w3	w4	w5	w6	…..	……	wn

Words	Frequencies
John	2
Likes	3
To	2
Watch	2
Movies	2
Mary	1
Too	1
also	1
Football	1
games	1

The Bag-of-Words approach: NLP

A. keeps word order, disregards word multiplicity
B. keeps word order, keeps word multiplicity
C. disregards word order, keeps word multiplicity
D. disregards word order, disregards word multiplicity

The Bag-of-Words approach: NLP

A. keeps word order, disregards word multiplicity
B. keeps word order, keeps word multiplicity
C. disregards word order, keeps word multiplicity
D. disregards word order, disregards word multiplicity

Which are multiple word sequences?

A. Tokenization
B. N-grams
C. Stopwords
D. Corpus

Which are multiple word sequences?

A. Tokenization
B. N-grams
C. Stopwords
D. Corpus

What are the possible features of a text corpus in NLP?

a. Count of the word in a document
b. Vector notation of the word
c. Part of Speech Tag
d. All of the above

What are the possible features of a text corpus in NLP?

a. Count of the word in a document
b. Vector notation of the word
c. Part of Speech Tag
d. All of the above

Which one of the following is not a pre-processing technique in NLP

a. Stemming and Lemmatization
b. removing punctuations
c. removal of stop words
d. Sentiment analysis

Which one of the following is not a pre-processing technique in NLP

a. Stemming and Lemmatization

b. removing punctuations

c. removal of stop words

d. Sentiment analysis

What is morphology?

A. The study of the rules governing the sounds that form words
B. The study of the rules governing sentence formation
C. The study of the rules governing word formation
D. The study of the rules governing the sounds that form sentence

What is morphology?

A. The study of the rules governing the sounds that form words
B. The study of the rules governing sentence formation
C. The study of the rules governing word formation
D. The study of the rules governing the sounds that form sentence

_____ *is not a module in question answering system*

A. Question Analysis
B. Answer Selection
C. Sentiment Analysis
D. Information Retrieval

_____ *is not a module in question answering system*

A. Question Analysis
B. Answer Selection
C. Sentiment Analysis
D. Information Retrieval

- All ordering of the words is nominally discarded and we have a consistent way of extracting features from any document in our corpus, ready for use in modeling.
- New documents that overlap with the vocabulary of known words, but may contain words outside of the vocabulary, can still be encoded, where only the occurrence of known words are scored and unknown words are ignored.
- You can see how this might naturally scale to large vocabularies and larger documents.

Managing Vocabulary

- As the vocabulary size increases, so does the vector representation of documents.
- In the previous example, the length of the document vector is equal to the number of known words.

There are simple text cleaning techniques that can be used as a first step, such as:

- Ignoring case
- Ignoring punctuation
- Ignoring frequent words that don't contain much information, called stop words, like "a," "of," etc.

All of the following are challenges associated with natural language processing except

A. dividing up a text into individual words in English.
B. understanding the context in which something is said.
C. recognizing typographical or grammatical errors in texts
D. distinguishing between words that have more than one meaning.

All of the following are challenges associated with natural language processing except

A. dividing up a text into individual words in English.
B. understanding the context in which something is said.
C. recognizing typographical or grammatical errors in texts
D. distinguishing between words that have more than one meaning.

Fixing misspelled words.

- Reducing words to their stem (e.g. “play” from “playing”) using stemming algorithms.
- A more sophisticated approach is to create a vocabulary of grouped words. This both changes the scope of the vocabulary and allows the bag-of-words to capture a little bit more meaning from the document.
- In this approach, each word or token is called a “gram”. Creating a vocabulary of two-word pairs is, in turn, called a bigram model. Again, only the bigrams that appear in the corpus are modeled, not all possible bigrams.

Which approach is used for spelling error detection and correction

A. Script Validation
B. Tokenization
C. N-gram
D. Filteration

Which approach is used for spelling error detection and correction

A. Script Validation
B. Tokenization
C. N-gram
D. Filteration

How given sentence represented using Bigram model? "I want to eat Indian food"

A. {(I, want), (want, to), (to, eat), (eat, Indian),(Indian, food)}
B. {(I), (want, to), (to, eat), (eat, Indian),(Indian, food),(food, I)}
C. {(I, want, to), (want, to, eat), (to, eat, Indian), (eat, Indian, food)}
D. {(I), (want), (to), (eat), (Indian), (food)}

How given sentence represented using Bigram model? "I want to eat Indian food"

A. {(I, want), (want, to), (to, eat), (eat, Indian),(Indian, food)}
B. {(I), (want, to), (to, eat), (eat, Indian),(Indian, food),(food, I)}
C. {(I, want, to), (want, to, eat), (to, eat, Indian), (eat, Indian, food)}
D. {(I), (want), (to), (eat), (Indian), (food)}

What is the single morpheme of word "Boxes"?

A. Box
B. Boxes
C. Boxses
D. Boxing

What is the single morpheme of word "Boxes"?

A. Box

B. Boxes

C. Boxses

D. Boxing

SPELL Checking

SPELL Checking

- A Spell checker is one of the basic tools required for language processing.
- Spell checking involves
  - identifying **words** and **non words**
  - also suggesting the possible alternatives for its correction.
- used in
  - Word processing
  - Character or text recognition
  - Speech recognition and generation.

SPELL Checking

- Most available spell checkers focus on processing **isolated** words and do not take into account the context.
  - "Henry **sar** on the box"
  - "Henry **at** on the box"

Spelling Errors

Three cause of error are:

- **Insertion:** Insertion of extra letter while typing. E.g. maximum typed as *max**ii**mum.*
- **Deletion:** A case of a letter missing or not typed in a word. E.g. *netwrk* instead of netw**o**rk.
- **Substitution:** Typing of a letter in place of the correct one. E.g. *inte**ll**ugence.*

Spelling errors may be classified into following types:

- **Typographic errors:**
  - Cause due to mistakes committed while **typing.**
  - E.g. netwrk in place of network.
- **Orthographic errors:**
  - Result due to a lack of **comprehension** of the concerned language on part of user.
  - E.g. arithmatic, wellcome, accomodation.
- **Phonetic errors:**
  - result due to poor cognition on part of **listener.**
  - E.g. the word **rough** could be spelt as **ruff.**
  - Listen as lisen, piece as peace or peas, reed as read.

Spell checking techniques

Spell checking techniques can be broadly classified into **three** categories-

(a) **Non-Word** error detection:

(b) **Isolated-word** error correction:

(c) **Context dependent** error detection and correction:

Spell checking techniques

(a) **Non-Word error detection:** This process involves the detection of misspelled words or non-words.

- E.g. the word sopar is a non-word ; its correct form is super or sober.
- The most commonly used techniques to detect such errors are the
  - **N-gram analysis**
  - **Dictionary look-up.**

Spell checking techniques

N-Gram Analysis:

- Make use of probabilities of occurrence of N-grams in a large corpus of text to decide on the error in the word.
- N-gram to be sequence of letters rather than words.
- Try to predict next letter rather than next words.
- Used in text (handwritten or printed) recognition.

Spell checking techniques

Dictionary look-up

involves the use of an efficient dictionary lookup coupled with pattern-matching algorithm (such as hashing technique, Finite state automata), dictionary portioning schemes and morphological processing methods.

Spell checking techniques

(b) **Isolated-word error correction:**

- Focus on the correction of an isolated non-words by finding its nearest and meaningful word and make an attempt to rectify the error.
- It thus transform the word “soper” into super.
- Isolated word correction may be looked upon as a combination of three sub-problems

  Error detection,

  candidate (correct word) generation,

  ranking of correct candidates.

Minimum Edit distance technique:

- Wanger[1974] define the minimum edit distance between the misspelled word and the possible correct candidate
- minimum number of edit operations needed to transform the misspelled word to the correct candidate.
- Edit operation-insertion, deletion, and substitution of a single character.
- The minimum number of such operations required to affect the transformation is commonly known as ***Levenshtein distance.***

(c) Context dependent error detection and correction:

- In addition to detect errors, try to find whether the corrected word fits in to context of the sentence.
- More complex to implement.
- "**Peace** comes from within", "**Piece** comes from within" ; **first** word in both sentence is a correct word.
- This involves correction of real-word errors or those that result in another valid error.

Soundex

- Soundex is a phonetic algorithm for indexing names by sound, as pronounced in English.
- The goal is for homophones to be encoded to the same representation so that they can be matched despite minor differences in spelling.

Soundex: Problems with Names

- Names can be misspelt: Rossner
- Same name can be spelt in different ways Kirkop; Chircop
- Same name appears differently in different cultures: Tchaikovsky; Chaicowski
- To solve this problem, we need *phonetically oriented* algorithms which can find similar sounding terms and names.
- Just such a family of algorithms exist and are called SoundExes, after the first patented version.

Soundex – typical algorithm

- Turn every token to be indexed into a 4-character reduced form
- Do the same with query terms
- Build and search an index on the reduced forms
  - (when the query calls for a soundex match)

Soundex – typical algorithm

1. Retain the first letter of the word.
2. Change all occurrences of the following letters to '0' (zero):
   'A', E', 'I', 'O', 'U', 'H', 'W', 'Y'.
3. Change letters to digits as follows:
   - B, F, P, V $\rightarrow$ 1
   - C, G, J, K, Q, S, X, Z $\rightarrow$ 2
   - D,T $\rightarrow$ 3
   - L $\rightarrow$ 4
   - M, N $\rightarrow$ 5
   - R $\rightarrow$ 6

Soundex continued

4. Remove all pairs of consecutive digits.
5. Remove all zeros from the resulting string.
6. Pad the resulting string with trailing zeros and return the first four positions, which will be of the form <uppercase letter> <digit> <digit> <digit>.

E.g., ***Herman*** becomes H655.

Will ***hermann*** generate the same code?

Word	Soundex code
Grate, Great	
Network, network	
Henry, Henary	
Torn	
Worn	
Horn	

- Soundex code for some words:

Word	Soundex code
Grate, Great	G630
Network, network	N362
Henry, Henary	H560
Torn	T650
Worn	W650
Horn	H650

- Used to measure similarity of two words.