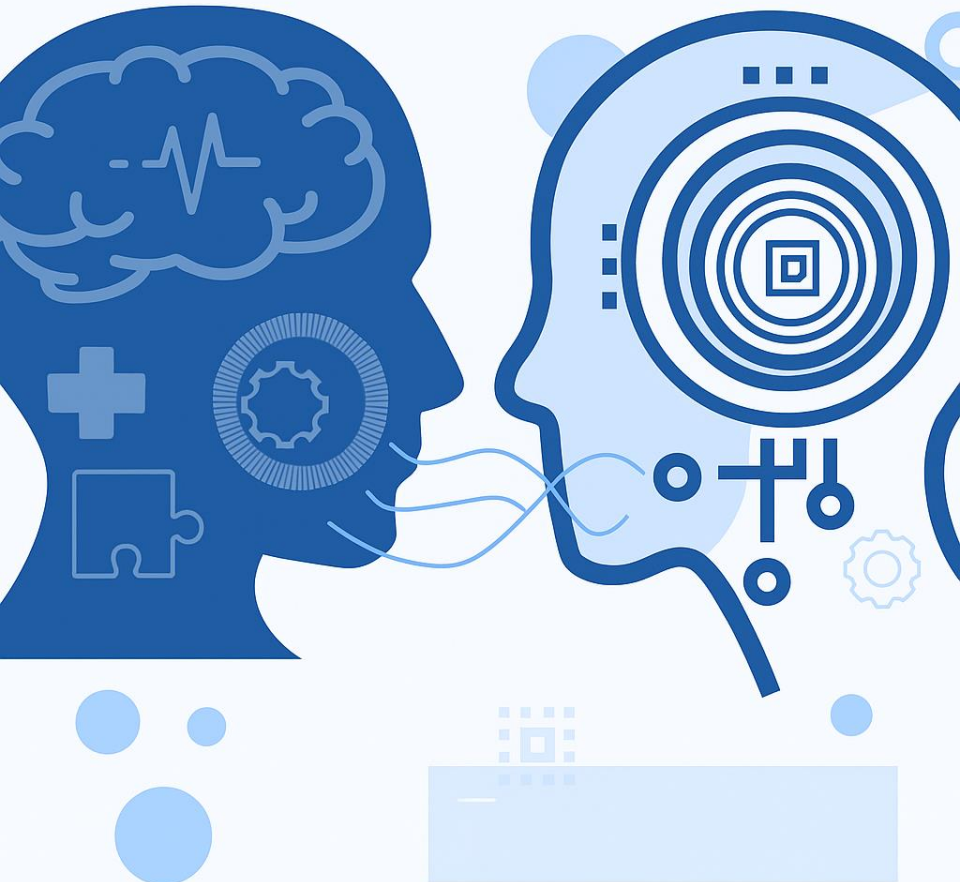


# NLP

Natural  
Language  
Processing



# NATURAL LANGUAGE PROCESSING (NLP)

## PMDS606L

MODULE 2

LECTURE 4

***Dr. Kamanasish Bhattacharjee***

*Assistant Professor*

*Dept. of Analytics, SCOPE, VIT*



# REGEX SYMBOLS

Symbol	Meaning	Example
.	Any character except newline	a.b matches "acb", "a7b"
^	Start of string	^Hello matches "Hello there"
\$	End of string	world\$ matches "Hello world"
*	0 or more repetitions	lo* matches "l", "lo", "loo"
+	1 or more repetitions	lo+ matches "lo", "loo"
?	0 or 1 repetition	colou?r matches "color" and "colour"
{n}	Exactly n repetitions	\d{3} matches "123"
[]	Any one character in brackets	[aeiou] matches a vowel
\d	Any digit	\d+ matches numbers
\w	Any alphanumeric character	\w+ matches "word"
\s	Any whitespace	\s+ matches spaces, tabs
,	,	OR
()	Grouping	(\d{2})/(\d{2})/(\d{4}) matches dates like 12/05/2023

# REGEX SYMBOLS

[abc] – a or b or c

[^abc] – neither a or b or c

[a-z] – a to z

[A-Z] – A to Z

[a-zA-Z] – a to z, A to Z

[0-9] – 0 to 9

# REGEX SYMBOLS

$[ ]?$  – occurs 0 or 1 time

$[ ]^+$  – occurs 1 or more times

$[ ]^*$  – occurs 0 or more times

$[ ]\{n\}$  – occurs n times

$[ ]\{n, \}$  – occurs n or more times

$[ ]\{m,n\}$  – occurs at least m times, at most n times

# REGEX SYMBOLS

`\d` – [0-9]

`\D` – [^0-9]

`\w` – [a-zA-Z\_0-9]

`\W` – [^\w]

# REGEX EXAMPLES

Regex	Matches	Example Matches
\d+	One or more digits	123, 42, 9
\w+	One or more word characters (letters, digits, underscore)	hello, user_123, AI2025
[A-Za-z]+	One or more English letters	chat, GPT, Data
[^a-zA-Z0-9\s]	Any character <b>not</b> a letter, digit, or space (used for removing punctuation)	!, @, #, ?
\s+	One or more whitespace characters	spaces, tabs, newlines
\b\w+\b	Full words (word boundaries)	NLP, model, theory
[0-9]{4}	Exactly 4 digits	1990, 2024
\d{2}/\d{2}/\d{4}	Dates in DD/MM/YYYY format	01/01/2025, 15/08/2023
\b[A-Z][a-z]+	Capitalized words (starting with uppercase, followed by lowercase)	India, London
[A-Za-z0-9._%+-]+@[A-Za-z0-9.-]+\.[A-Za-z]{2,}	Email addresses	name@example.com, user123@domain.in
https?://[^\s]+	URLs starting with http or https	http://example.com, https://openai.com
`\b(?:Mr	Mrs	Dr).?\s[A-Z][a-z]+'`
(?i)covid	Case-insensitive match for “covid”	Covid, COVID, covid
(\w+)\s+\1	Repeated words	go go, hello hello
(?<=@)\w+	Word following @ symbol (e.g., Twitter handle)	user123 from @user123

# REGEX EXAMPLES

1. Mobile number that starts with 8 or 9 and total digit is 10.
2. First character uppercase, contains lower case alphabet and only one digit allowed in between
3. Email ID

# REGEX

## Example:

Match **"cat"** or **"dog"**:

cat|dog

## Example:

Match **"http"** or **"https"**:

https?



# REGEX


## Example:

Match **"apple"** only if followed by " pie":

apple(=? pie)

## Matches:

apple in "apple pie" 

Not matched in "apple juice" 

# REGEX

## Example:

Match **"world"** only if preceded by "hello ":

(?<=hello )world

## Matches:

"hello world" 

"goodbye world" 

# REGEX

**(?(condition)yes|no)**

If group 1 was matched earlier, match "X", else "Y":

**(A)?(?(1)X|Y)**

"AX"  (group 1 matched "A", so expect "X")

"Y"  (group 1 didn't match, so expect "Y")

"AY"  (group 1 matched, but "Y" not allowed)

**(abc)?(?(1)def|xyz)**

If group 1 ((abc)) exists, then match def

Else match xyz

# REGEX EXAMPLES

**Instructions:** Write a regular expression to match the following patterns.

A 4-digit year (e.g., 1998, 2025)

**Answer:** \_\_\_\_\_

An email address (e.g., student.name@univ.edu)

**Answer:** \_\_\_\_\_

A word that starts with a capital letter (e.g., India, Python)

**Answer:** \_\_\_\_\_

A date in the format DD/MM/YYYY (e.g., 25/12/2023)

**Answer:** \_\_\_\_\_

# REGEX EXAMPLES

A URL starting with http or https

**Answer:** \_\_\_\_\_

Any word followed by the **same word again** (e.g., go go, bye bye)

**Answer:** \_\_\_\_\_

Hashtags (e.g., #AI, #NLP2025)

**Answer:** \_\_\_\_\_

Twitter handles (e.g., @OpenAI, @student\_123)

**Answer:** \_\_\_\_\_