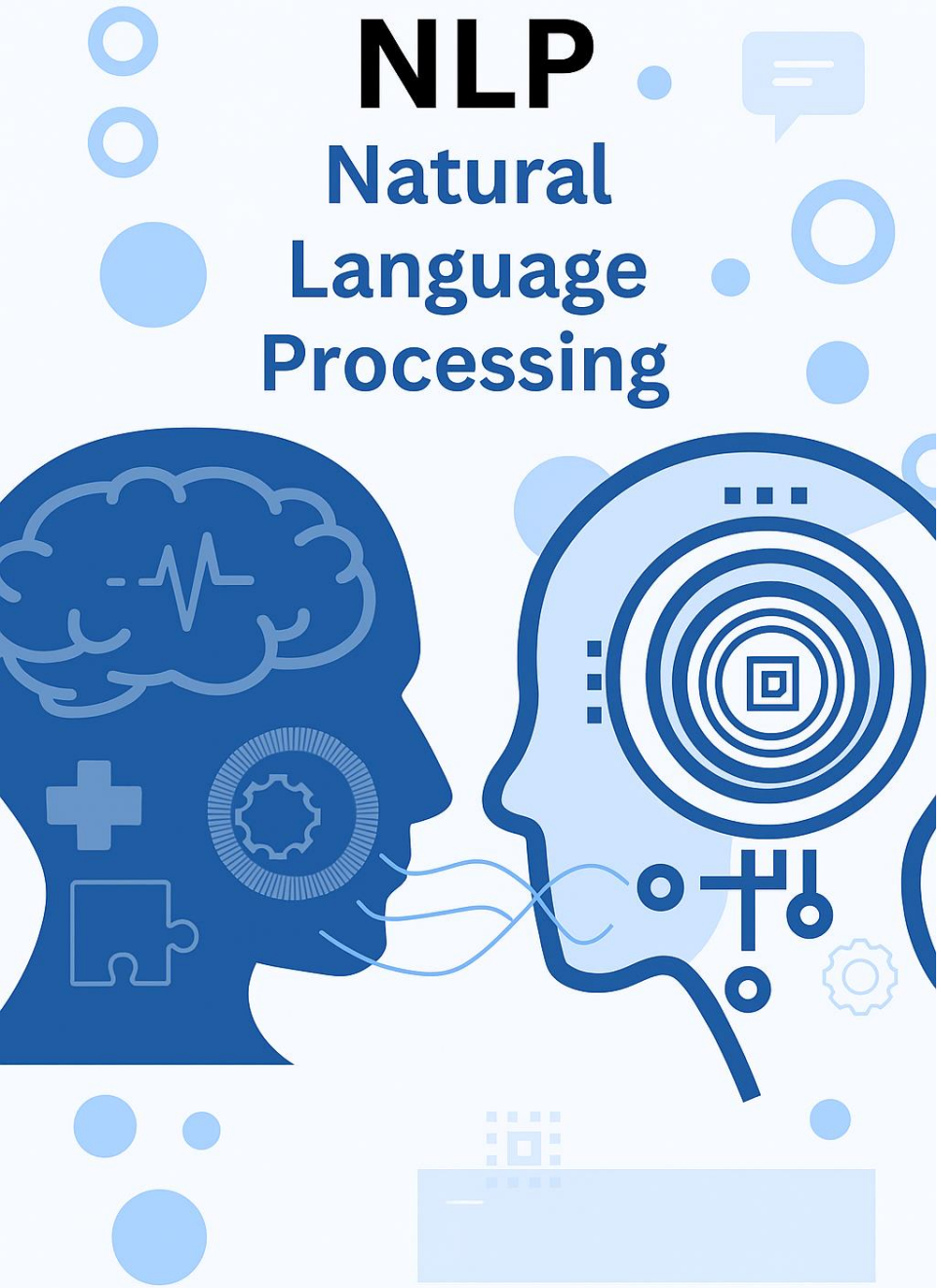


NATURAL LANGUAGE PROCESSING (NLP)

PMDS606L

MODULE 2

LECTURE 5

Dr. Kamanasish Bhattacharjee

Assistant Professor

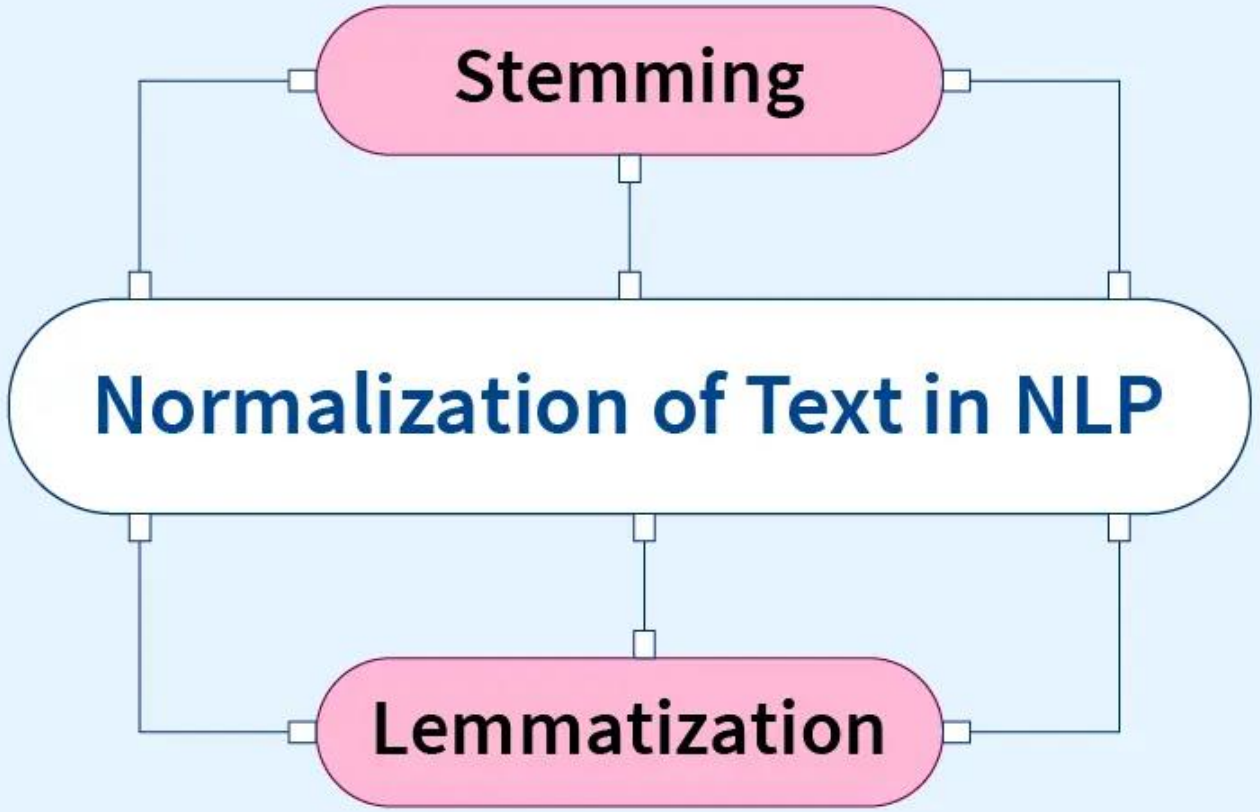
Dept. of Analytics, SCOPE, VIT

TEXT NORMALIZATION

Text Normalization in NLP is the process of transforming text into a canonical (standard or normalized) form so that it can be analyzed more effectively by computational models. It helps reduce variations in the text that are irrelevant to meaning and improves consistency across the dataset. This step is crucial for downstream NLP tasks such as text classification, sentiment analysis, machine translation, and information retrieval.

NEED OF TEXT NORMALIZATION

- **Different form of same word –** e.g., Go, Went, Gone
- **Variability in natural language** – e.g., “colour” vs “color”
- **Inconsistent casing** – e.g., “Apple” vs “apple”
- **Non-standard spellings or abbreviations** – e.g., “u” for “you”, “lol”, “gr8”
- **Typographical errors** – e.g., “teh” instead of “the”
- **Unnecessary symbols or punctuation** – e.g., “!!!”, “@#%”

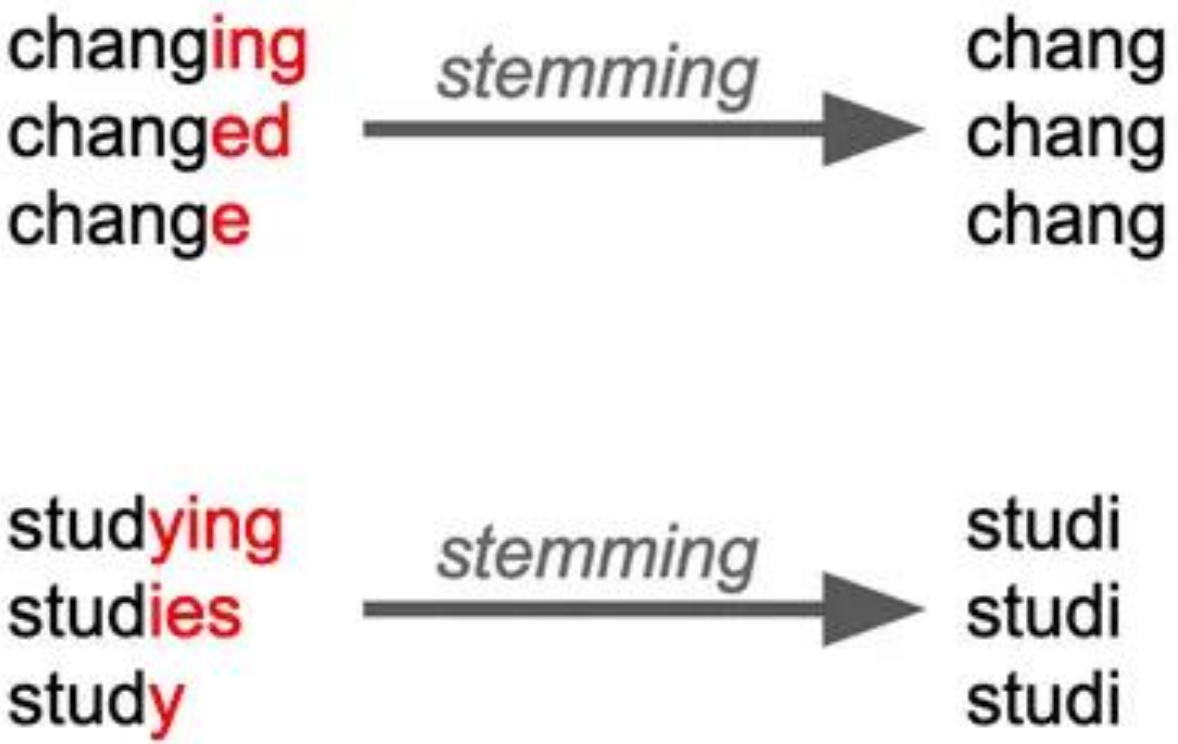

Stemming
Normalization of Text in NLP
Lemmatization

STEMMING EXAMPLE

Word	Stem Output
connects	connect
connected	connect
connecting	connect
connection	connect
connections	connect

Word	Stem Output
collection	collect
collective	collect
collectively	collect
collects	collect
collected	collect

STEMMING EXAMPLE



NEED OF TEXT NORMALIZATION

Stemming uses **heuristic rules** (rather than deep linguistic analysis) to remove common endings such as:

Remove ***-ed, -ing, -s, -es, -ly, -ment***, etc.

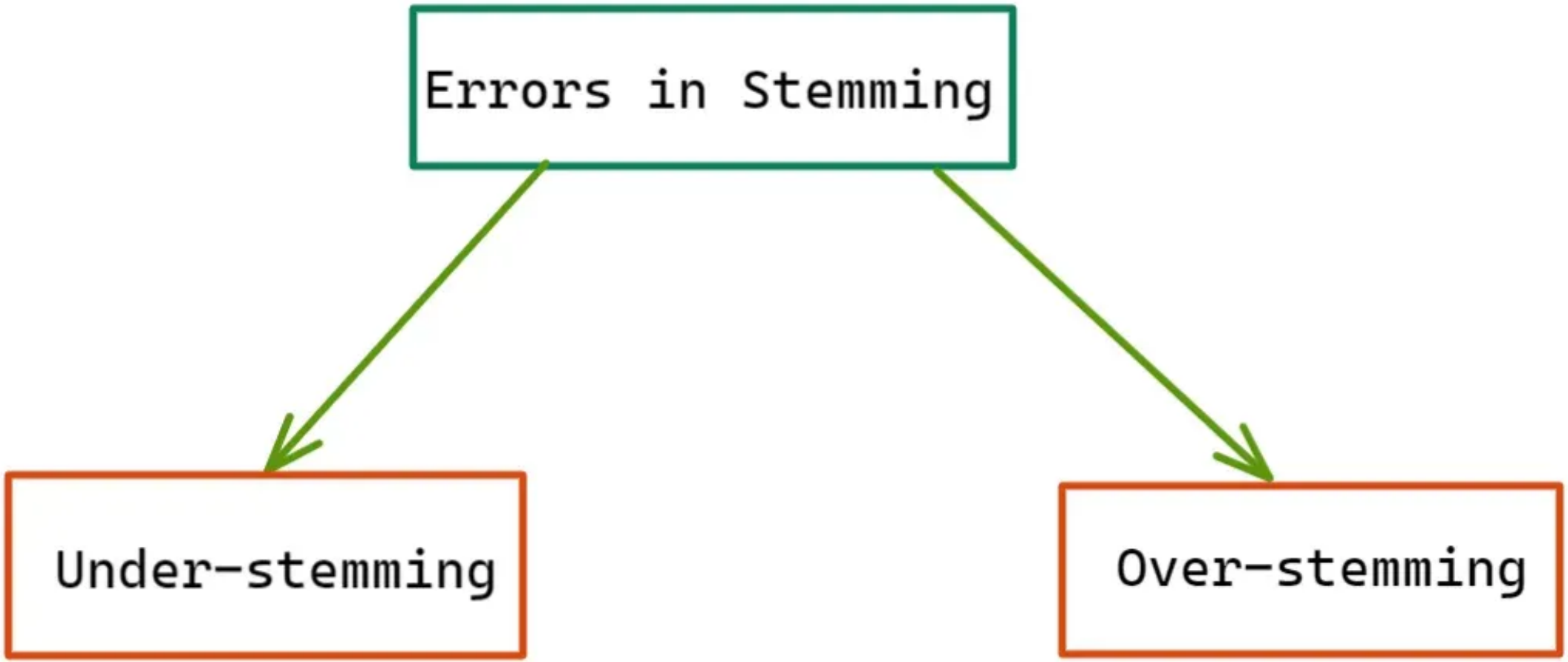
It does **not** necessarily return a valid English word.

It does not consider the semantic meaning of the word.

Drawback: The resulting stem may not be a meaningful word.

Example: "laziness" → "lazi" (not "lazy")

Can lead to loss of meaning or incorrect interpretations.

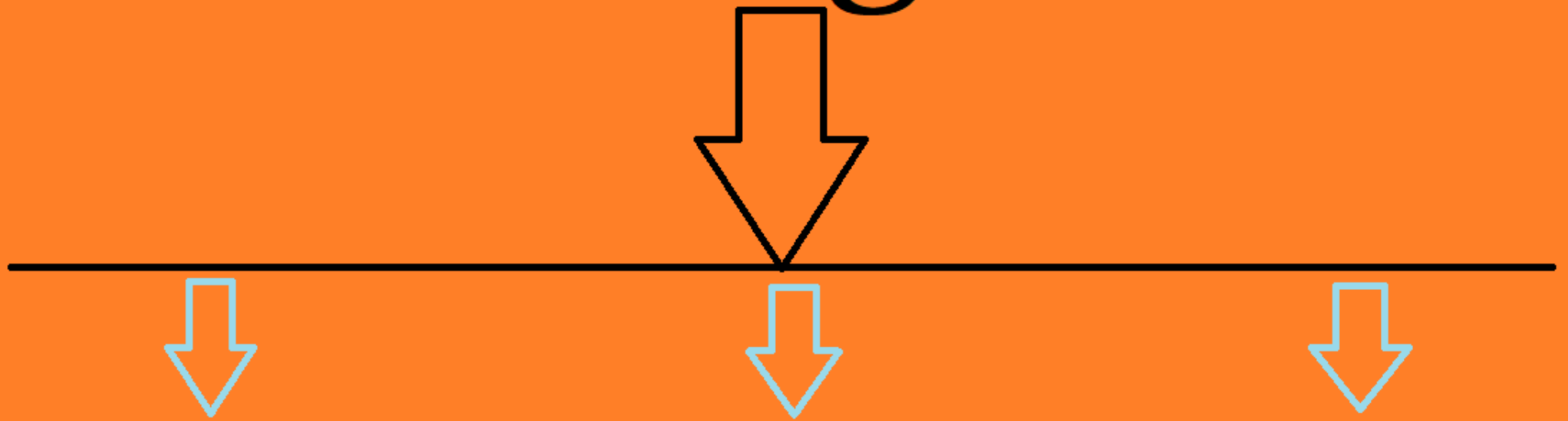

Errors in Stemming
Under-stemming
Over-stemming

ERRORS IN STEMMING

UNDER-STEMMING(not reducing enough)
Occurs when the stemmer **does not reduce words enough**.
Results in **related words being treated as different**.
Example: "data" → "dat" and "datum" → "datu“ (instead of the same stem “dat”)
The **stemmer fails** to recognize that **both words come** from the **same root**

OVER-STEMMING (too much chopping)
Over-stemming occurs when the stemmer **removes too many characters** from a word.
This leads to **different words being incorrectly treated as the same**.
Example: "university" and "universe" both reduced to "univers".
This **falsely implies** they are **semantically** the **same**, which they are not.

Stemming in NLP



PORTER STEMMER ALGORITHM

- https://vijinimallawaarachchi.com/2017/05/09/porter-stemming-algorithm/
- https://people.scs.carleton.ca/~armyunis/projects/KAPI/porter.pdf
- https://www.youtube.com/watch?v=GQ1sXx8hH4k
- https://www.youtube.com/watch?v=W34VpI7jXpY