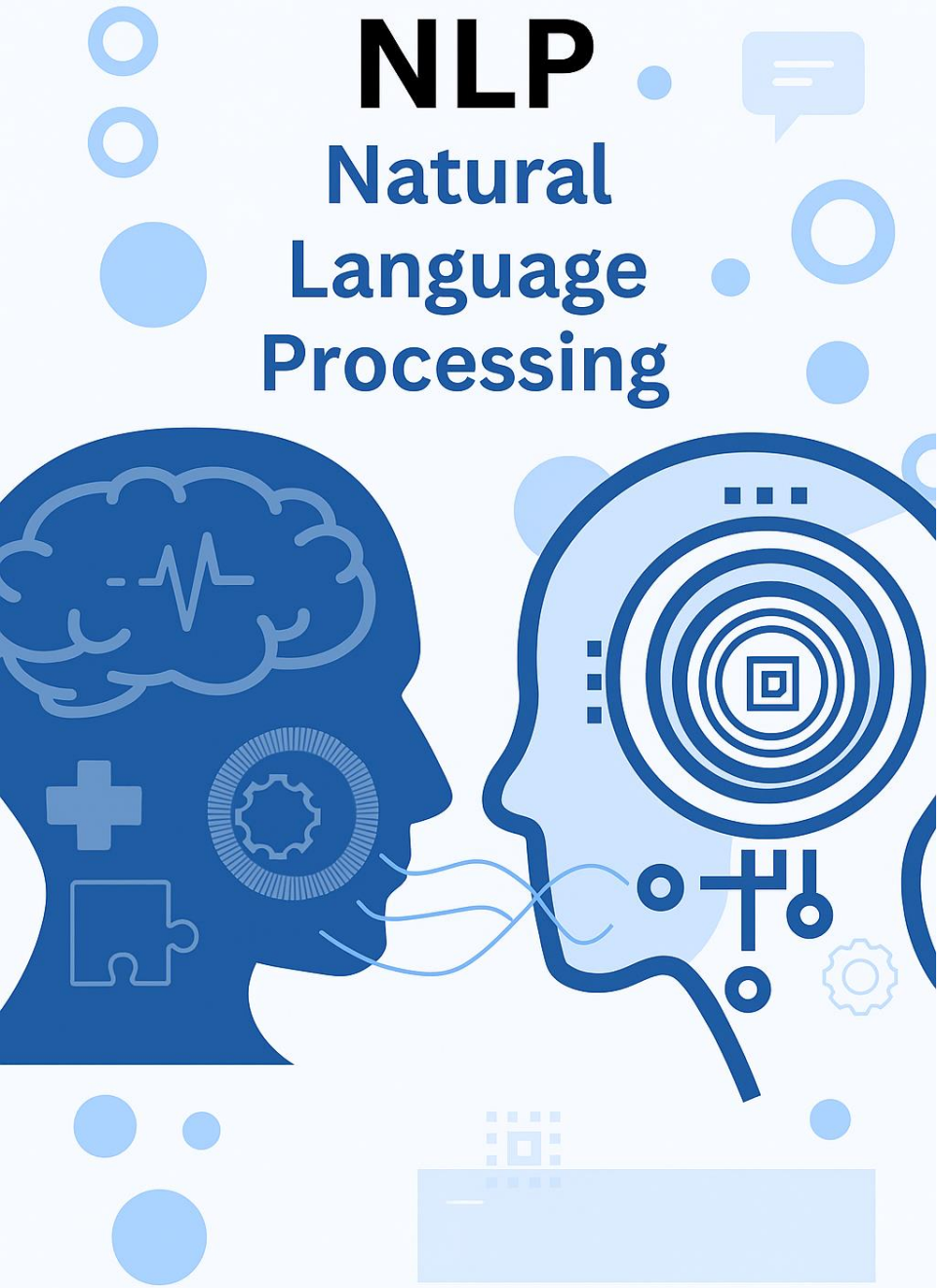


NATURAL LANGUAGE PROCESSING (NLP)

PMDS606L

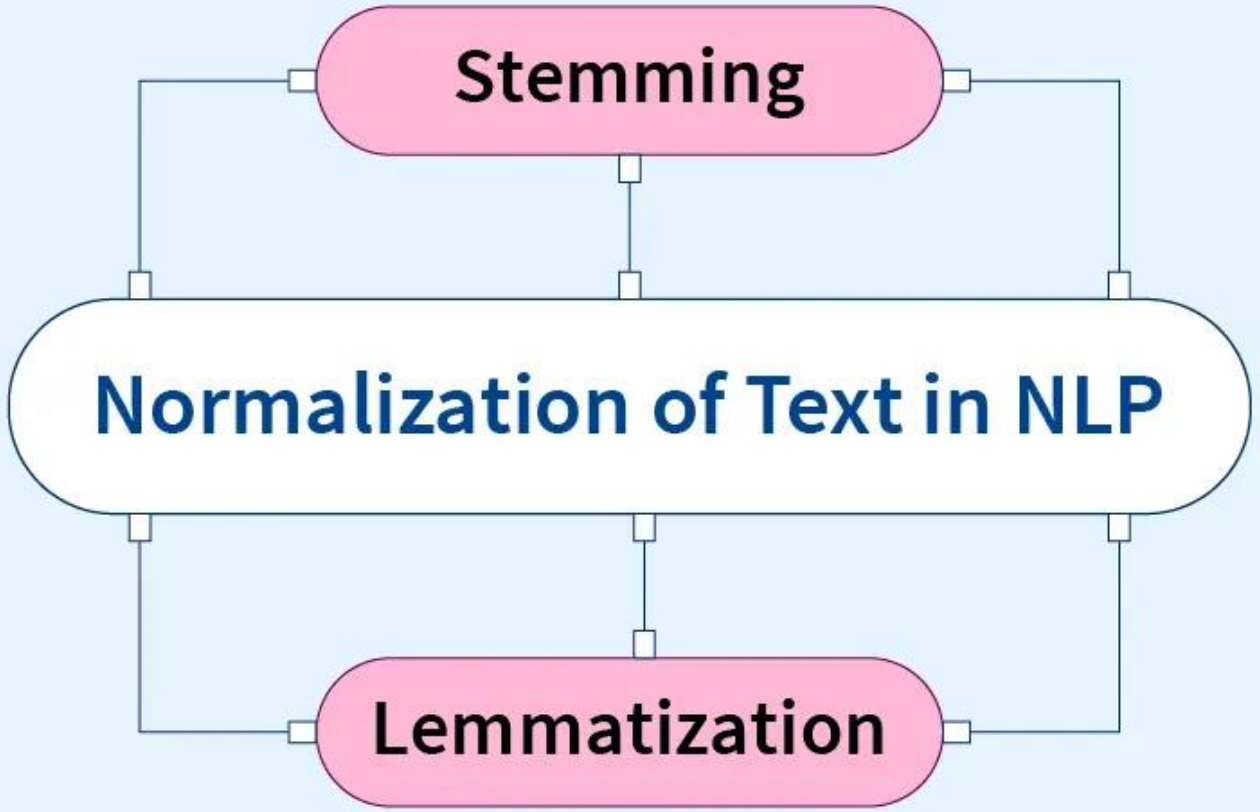
MODULE 2

LECTURE 6

Dr. Kamanasish Bhattacharjee

Assistant Professor

Dept. of Analytics, SCOPE, VIT


Stemming
Normalization of Text in NLP
Lemmatization

LEMMATIZATION

Lemmatization is a fundamental text normalization technique in Natural Language Processing (NLP) that reduces words to their **base or dictionary form**, called a **lemma**. Unlike stemming, lemmatization is **context-aware**—it considers the word's **Part-of-Speech (POS)** and ensures the output is a **valid English word**.

Natural language is rich with **inflections**—variations in a word's form depending on tense, number, gender, etc. Lemmatization helps:

- Standardize word forms (e.g., *"am", "is", "are"* → *"be"*)
- Improve **semantic meaning** consistency
- Enhance **text classification**, **machine translation**, **chatbots**, and **search systems**

STEMMING vs LEMMATIZATION

Word	Stemming	Lemmatization
running	run	run
better	better	good
studies	studi	study
was	wa	be
went	went	go

LEMMATIZATION TOOLS

Tool	Language Support	Strengths
WordNet Lemmatizer (NLTK)	English only	Simple and lightweight
spaCy	Multilingual	Fast and accurate
TextBlob	English	Easy-to-use wrapper
Stanford CoreNLP	Multilingual	Powerful linguistic tools
Stanza (by Stanford)	Multilingual	Neural network-based, modern

LEMMATIZATION STEPS

Step 1: Tokenization

Split the input text into individual tokens (words or phrases).

Input: "The striped bats are hanging on their feet for best.”

Output: ["The", "striped", "bats", "are", "hanging", "on", "their", "feet", "for", "best"]

LEMMATIZATION STEPS

Step 2: POS Tagging

Assign each token a part-of-speech (noun, verb, adjective, etc.) using POS taggers.

Example POS Tags:
The/DT striped/JJ bats/NNS are/VBP hanging/VBG on/IN their/PRP$ feet/NNS for/IN best/JJS

LEMMATIZATION STEPS

Step 3: Map POS Tags to Lemmatizer Format

POS tags (like VBG, NN, etc.) need to be converted to a format compatible with the lemmatizer (typically WordNet format).

Treebank POS	WordNet POS
NN, NNS	wordnet.NOUN
VB, VBG, VBD	wordnet.VERB
JJ, JJR	wordnet.ADJ
RB, RBR	wordnet.ADV

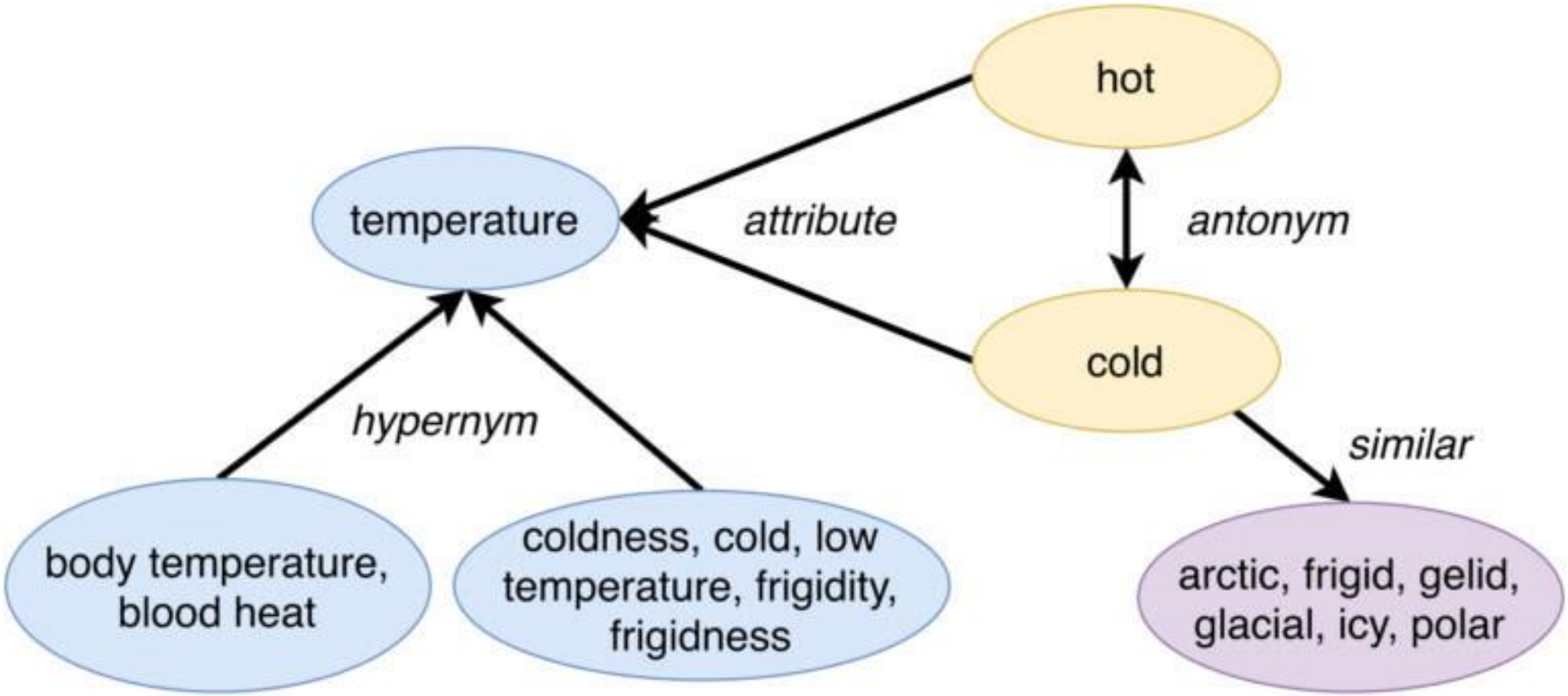
LEMMATIZATION STEPS

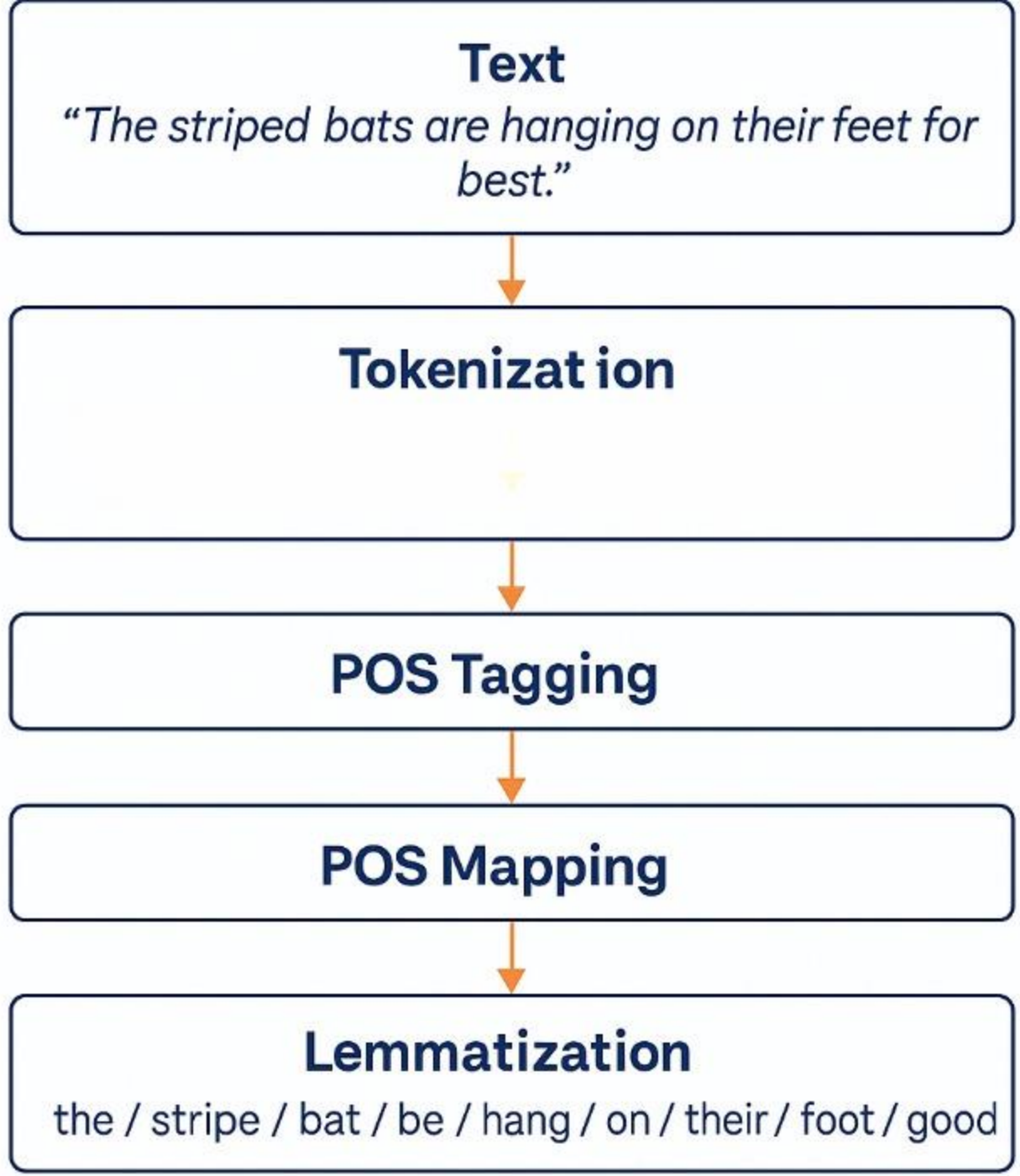
Step 4: Lemmatization Using Dictionary Lookup

The lemmatizer checks the **token + POS** against a dictionary (e.g., **WordNet**) and returns the lemma.

Examples:
"bats" (plural noun) → "bat"
"hanging" (verb) → "hang"
"feet" (plural noun) → "foot"
"best" (superlative adjective) → "good”

This step **preserves context** and produces meaningful, correct lemmas.


hot
temperature
attribute
antonym
cold
hypernym
similar
body temperature, blood heat
coldness, cold, low temperature, frigidity, frigidness
arctic, frigid, gelid, glacial, icy, polar


Text
“The striped bats are hanging on their feet for best.”
Tokenizat ion
POS Tagging
POS Mapping
Lemmatization
the / stripe / bat / be / hang / on / their / foot / good

Stemming vs Lemmatization

Stemming

achieve -> achiev
achieving -> achiev

- Can reduce words to a stem that is not an existing word
- Operates on a single word without knowledge of the context
- Simpler and faster

Lemmatization

achieve -> achieve
achieving -> achieve

- Reduces inflected words to their lemma, which is always an existing word
- Can leverage context to find the correct lemma of a word
- More accurate but slower