

# Text Processing and Normalization

Dr. Sambit Praharaj  
Assistant Professor (II)

# Text Processing

- Text processing is the process of converting raw text into a clean and structured form so that machines can understand and analyze it effectively.
- Used in NLP, IR, sentiment analysis, text classification, etc.

# Need for Normalization & Cleaning

- Removes noise and inconsistencies
- Ensures uniform representation
- Reduces vocabulary size
- Improves model accuracy
- Essential for real-world text data

# Unicode Normalization

- Same characters can have multiple Unicode forms
- Normalization converts text into canonical form
- Example:
- Café → Café
- Why needed?
- Avoids duplicate tokens

# Encoding Normalization

- Text can use different encodings (ASCII, UTF-8, UTF-16)
- UTF-8 is the most common

Why needed?

- Prevents garbled characters
- Ensures correct multilingual text display

# Whitespace Normalization

- Removes extra spaces, tabs, newlines
- Ensures consistent tokenization
- Example:
- I love NLP → I love NLP

# Punctuation Handling

- Includes . , ! ? ; : etc.
- Removed or retained based on task
- Example:
- Hello, world! → Hello world
- Important for noise reduction

# Case Folding

- Converts text to a single case (lowercase)
- Reduces vocabulary size
- Example:
- AI, Ai, ai → ai
- Note: Avoided in NER tasks

# Summary

- Text processing prepares text for ML/NLP
- Normalization ensures consistency
- Cleaning improves efficiency and accuracy