Module 3: Text Preprocessing

What is Text Processing?

In NLP, text preprocessing is the initial step of cleaning and transforming raw text data into a structured and analyzable form. Since most NLP algorithms and models work on numerical or standardized representations, preprocessing ensures the **text is consistent**, **meaningful**, and **ready for analysis or modeling**.

Key Aspects of Text Preprocessing

1. **Cleaning the Text:** Removing punctuation, numbers, stopwords.
2. **Tokenization:** Splitting text into words, subwords, or sentences.
3. **Normalization:** Converting to lowercase, stemming, and lemmatization.

Lowercasing

- **Lowercasing** converts all text to lowercase to reduce duplication.
- Example: “Customer Service” $\rightarrow$ “customer service”.
- Example: “ELON MUSK” $\rightarrow$ “elon musk”.

Tokenization

- We have seen it before in Module 2.
- Essentuially, tokens split text into smaller units: Words, subwords, and sentences.
- Example: “I love NLP” $\rightarrow$ [“I”, “love”, “NLP”].

Removing Punctuation and Special Characters

- Cleans the text by removing characters like !, @, #, $, etc.
- Example: "Great product!!!" $\rightarrow$ "Great product".

Removing Stopwords

- **Stopwords** are common words with little semantic meaning.
- Examples: “the”, “is”, “and”, etc.

Stemming

- **Stemming** is the process of reducing words to their root words.
- Example: “running”, “runs”, “ran” $\rightarrow$ “run”.
- Example: “connection”, “connects”, “connected”, “connecting”, “connections” $\rightarrow$ “connect”.
- **Note:** The stem may not always be a valid word in the English language.
- Example: “argue”, “argued”, “argument”, “arguing”, “arguer” $\rightarrow$ “argu”.

Lemmatization

- **Lemmatization** is the process of reducing words to their dictionary form, which makes them more accurate and meaningful.
- Example: “better” $\rightarrow$ “good”.
- Example: “ate”, “eaten”, “eating” $\rightarrow$ “eat”.

Handling Numbers / Dates / URLs / Emails

- Replace or remove domain-specific entities if not needed for analysis.
- Example: “Order #12345” $\rightarrow$ “order”.

Text Normalization

- Converts special characters, accents, or multiple spaces into standard forms.
- Example: “café” $\rightarrow$ “cafe”.

The Importance of Text Preprocessing

- **Reduces Noise:** Removes irrelevant or misleading characters.
- **Improves Consistency:** Ensures similar words are treated the same way.
- **Enhances Model Performance:** Models learn better on clean, standardized input.
- **Enables Feature Extraction:** Such as frequency counts, embeddings, or sentiment features.

Lab 3

In this lab, we shall apply the text preprocessing skills learned in this module to real-world business scenarios.