Day 16 Training Report

14 July 2025

Text Preprocessing — Tokenization, Stopwords Removal, Stemming, Lemmatization

On **Day 16**, students began exploring **Natural Language Processing (NLP)**, focusing on **text preprocessing**, which is a critical first step before applying any NLP model. Raw text data is often **noisy and unstructured**, so cleaning it is essential for accurate model performance.

---

1. Objectives of the Session

- Understand **why text preprocessing is important** in NLP.
- Learn **techniques to clean and standardize text data**.
- Prepare textual data for **numerical feature extraction**.

---

2. Key Techniques

a) Tokenization

- Splits text into smaller units (words or sentences).
- Converts raw text into **tokens** that can be processed.

Example:

```
from nltk.tokenize import word_tokenize

text = "Natural Language Processing is amazing!"
tokens = word_tokenize(text)
print(tokens)
```

Output: ['Natural', 'Language', 'Processing', 'is', 'amazing', '!']

---

b) Stopwords Removal

- Removes common words that **do not carry significant meaning** (e.g., “the”, “is”, “and”).
- Helps reduce **dimensionality and noise** in models.

Example:

```
from nltk.corpus import stopwords
```

```
stop_words = set(stopwords.words('english'))
filtered_tokens = [word for word in tokens if word.lower() not in stop_words]
print(filtered_tokens)
```

Output: ['Natural', 'Language', 'Processing', 'amazing', '!']

---

c) Stemming

- Reduces words to their **root form** by removing suffixes.
- Example: “running” → “run”, “jumps” → “jump”

d) Lemmatization

- Similar to stemming, but uses **dictionary-based approach** for more accurate root forms.
- Example: “better” → “good”, “was” → “be”

Example:

```
from nltk.stem import PorterStemmer, WordNetLemmatizer

stemmer = PorterStemmer()
lemmatizer = WordNetLemmatizer()

print(stemmer.stem("running"))        # Output: run
print(lemmatizer.lemmatize("running")) # Output: running
```

---

3. Hands-on Practice

- Cleaned sample text data from **tweets and reviews**.
- Applied **tokenization, stopwords removal, stemming, and lemmatization** sequentially.
- Compared **differences between stemming and lemmatization**.