**Day 18 — 14 July 2025 (Monday)**

**Topic: Advanced NLP — Lemmatization & Text Normalization**

Today we learned deeper Natural Language Processing techniques to further refine text data. The instructor explained the difference between **stemming and lemmatization**, and why lemmatization preserves meaning better, which helps ML models understand email context more accurately. We applied text normalization including lowercasing, removing numeric patterns, email addresses, URLs, and special symbols. These techniques significantly enhance spam detection quality by eliminating noise.

**Hands-on work included:**

* Tokenizing sentences into words
* Removing repetitive characters
* Replacing URLs/emails with placeholders
* Lemmatizing words to base form

import re

from nltk.stem import WordNetLemmatizer

lemmatizer = WordNetLemmatizer()

text = "Get FREE entry!!! Visit our website now at http://offer.com"

text = re.sub(r'http\S+|www.\S+',' ', text)

clean = " ".join([lemmatizer.lemmatize(w) for w in text.split()])

print(clean)

This session strengthened the fundamentals of preparing robust text pipelines.