**Day 9 — 1 July 2025 (Tuesday)**

**Topic: Intro to Natural Language Processing (NLP)**

We started NLP fundamentals today — understanding how machines interpret text. The trainer explained key NLP tasks like tokenization, stopword removal, stemming, and lemmatization. These steps help reduce noise in email text (remove “the”, “and”, punctuation, etc.). We learned how spam filtering relies heavily on text pre-processing. I practiced breaking email sentences into tokens and removing irrelevant text. Understanding NLP is foundational to designing our spam classifier, as spam patterns rely heavily on keyword frequency and structure.

**Code Practiced:**

from nltk.tokenize import word\_tokenize

from nltk.corpus import stopwords

text = "Congratulations! You have won a $1000 prize."

tokens = word\_tokenize(text)

filtered = [w for w in tokens if w.lower() not in stopwords.words("english")]

print(filtered)