**Day 11 — 3 July 2025 (Thursday)**

**Topic: Feature Extraction — Bag-of-Words & TF-IDF**

We learned how text becomes numbers for ML models. The trainer introduced **Bag-of-Words model** and **TF-IDF (Term Frequency-Inverse Document Frequency)**. These convert words into numerical vectors representing importance of terms in an email. TF-IDF is widely used in spam detection. Understanding these techniques is critical because ML models cannot read raw text — they need numeric features. Today’s session felt very practical as we saw how word frequency patterns reveal spam messages.

**Code Practiced:**

from sklearn.feature\_extraction.text import TfidfVectorizer

tfidf = TfidfVectorizer(max\_features=3000)

X = tfidf.fit\_transform(df['text']).toarray()

print(X.shape)