EDA & Preprocessing Notes (Cheat Sheet)

==========

Exploratory Data Analysis

==========

- df.shape Dataset dimensions
- df.dtypes Data types
- df.head(), df.tail() Preview data
- df.describe() Summary stats (mean, std, etc.)
- df.isnull().sum() Missing values check
- df['target'].value_counts() Class balance
- sns.pairplot(), heatmap Data visualization

==========

Preprocessing Overview

==========

1. Missing Values:
   - df.dropna()
   - df['col'].fillna(mean/median)

2. Scaling:
   - StandardScaler()
   - MinMaxScaler()
   - RobustScaler()

3. Encoding:
   - LabelEncoder
   - OneHotEncoder
   - OrdinalEncoder

4. Outlier Removal:
   - Z-Score: |z| > 3 (remove)

5. Transformation:

- np.log1p() fix skewed data

6. Text Preprocessing:

- Tokenization: word_tokenize()
- Stopword removal
- Stemming: PorterStemmer

7. Imbalanced Data:

- SMOTE() creates synthetic samples

==========

Summary Table

==========

Task	Purpose
EDA	Understand data
Missing Values	Fix gaps
Scaling	Normalize features
Encoding	Convert text to numbers
Outliers	Remove anomalies
Transformation	Normalize skewed data
Tokenization	Prepare text for NLP
Stopwords/Stemming	Clean text
SMOTE	Fix class imbalance