Portfolio: Data Annotation & Data Cleaning Skills (Day 1 - Day 3)

Day 1: Annotation Fundamentals

- Learned core annotation concepts: labels, classes, guidelines, edge cases.
- Understood importance of consistency and high-quality ground truth.
- Practiced text classification and sentiment labeling.
- Gained awareness of how annotation quality affects ML performance.

Day 2: Label Studio Basics

- Created and configured annotation projects in Label Studio.
- Imported datasets and designed labeling interfaces.
- Annotated text samples using structured workflows.
- Exported annotations in JSON/CSV formats for ML pipelines.

Day 3: Python and Pandas for Data Cleaning

- Installed and configured Python and pandas.
- Loaded CSV datasets using pandas.
- Cleaned text data: lowercasing, trimming whitespace, regex cleaning.
- Removed duplicates and missing values.
- Exported cleaned datasets for annotation or ML training.
- Verified dataset quality using value counts and inspection methods.

Code Snippets (with explanations)

Importing pandas

```
import pandas as pd
```

Loads the pandas library for data manipulation.

Loading a CSV file

```
df = pd.read_csv("annotations.csv")
```

Reads the dataset into a DataFrame.

Viewing the first rows

```
df.head()
```

Displays the first 5 rows.

Cleaning text

```
df['text'] = df['text'].str.lower().str.strip()
```

```
df['text'] = df['text'].str.replace(r"[^a-zA-Z0-9\s]", "", regex=True)
```

Normalizes text and removes special characters.

Removing duplicates and missing values

```
df = df.drop_duplicates()
```

```
df = df.dropna()
```

Ensures dataset consistency.

Removing empty rows

```
df = df[df["text"] != ""]
```

Removes rows with empty text after cleaning.

Exporting cleaned data

```
df.to_csv("cleaned_annotations.csv", index=False)
```

Saves the cleaned dataset.