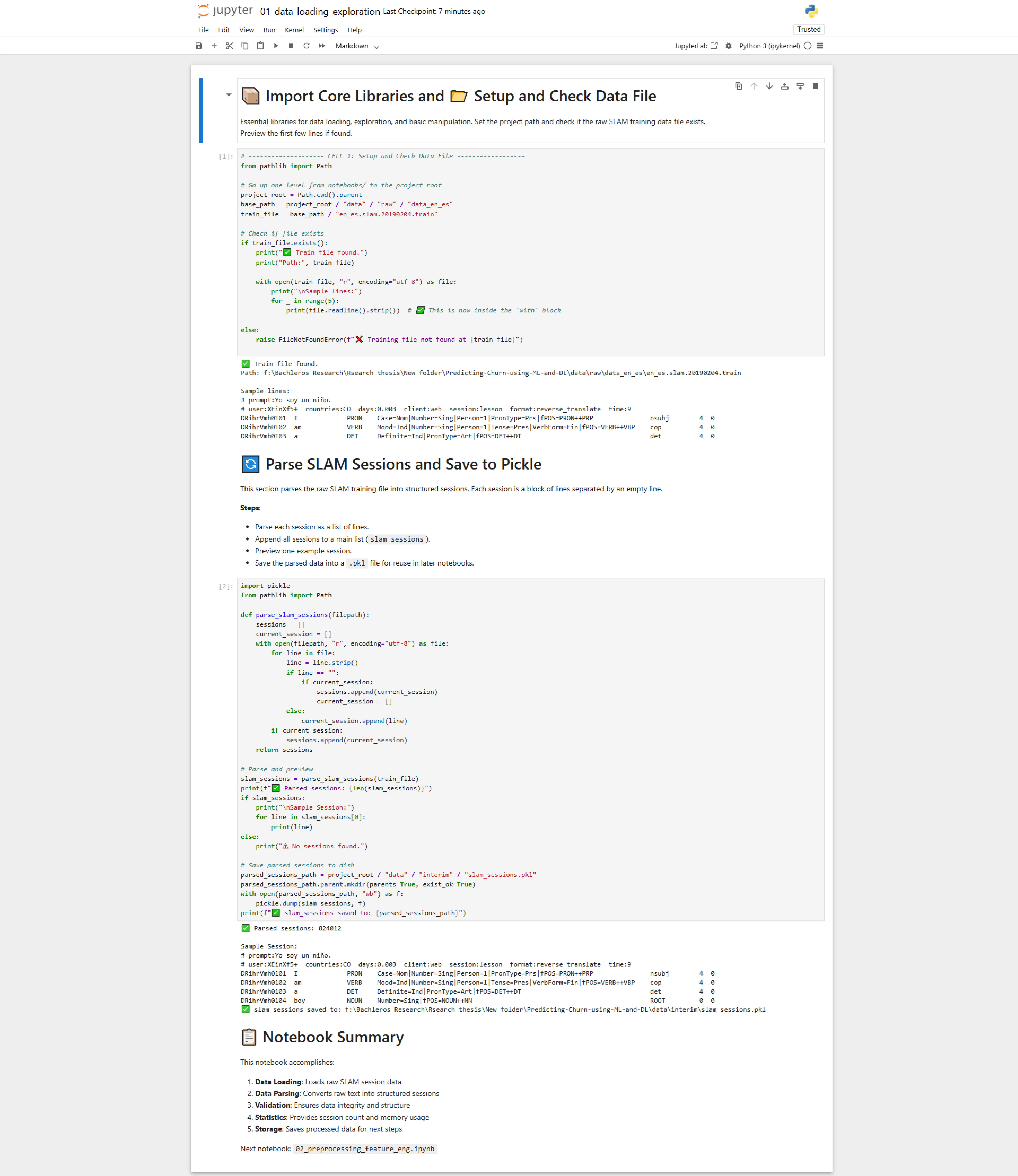
jupyter 01_data_loading_exploration Last Checkpoint: 7 minutes ago
File Edit View Run Kernel Settings Help
Trusted
Markdown
JupyterLab
Python 3 (ipykernel)

📦 Import Core Libraries and 📂 Setup and Check Data File

Essential libraries for data loading, exploration, and basic manipulation. Set the project path and check if the raw SLAM training data file exists. Preview the first few lines if found.

[1]:

```python
# ------------------- CELL 1: Setup and Check Data File -----------------
from pathlib import Path

# Go up one level from notebooks/ to the project root
project_root = Path.cwd().parent
base_path = project_root / "data" / "raw" / "data_en_es"
train_file = base_path / "en_es.slam.20190204.train"

# Check if file exists
if train_file.exists():
    print("✅ Train file found.")
    print("Path:", train_file)

    with open(train_file, "r", encoding="utf-8") as file:
        print("\nSample lines:")
        for _ in range(5):
            print(file.readline().strip())  # ✅ This is now inside the `with` block

else:
    raise FileNotFoundError(f"❌ Training file not found at {train_file}")
```

```
✅ Train file found.
Path: f:\Bachleros Research\Rsearch thesis\New folder\Predicting-Churn-using-ML-and-DL\data\raw\data_en_es\en_es.slam.20190204.train

Sample lines:
# prompt:Yo soy un niño.
# user:XEinXf5+  countries:CO  days:0.003  client:web  session:lesson  format:reverse_translate  time:9
DRihrVmh0101  I            PRON    Case=Nom|Number=Sing|Person=1|PronType=Prs|fPOS=PRON++PRP            nsubj       4  0
DRihrVmh0102  am           VERB    Mood=Ind|Number=Sing|Person=1|Tense=Pres|VerbForm=Fin|fPOS=VERB++VBP  cop         4  0
DRihrVmh0103  a            DET     Definite=Ind|PronType=Art|fPOS=DET++DT                                det         4  0
```

🔄 Parse SLAM Sessions and Save to Pickle

This section parses the raw SLAM training file into structured sessions. Each session is a block of lines separated by an empty line.

Steps:

- Parse each session as a list of lines.
- Append all sessions to a main list (`slam_sessions`).
- Preview one example session.
- Save the parsed data into a `.pkl` file for reuse in later notebooks.

[2]:

```python
import pickle
from pathlib import Path

def parse_slam_sessions(filepath):
    sessions = []
    current_session = []
    with open(filepath, "r", encoding="utf-8") as file:
        for line in file:
            line = line.strip()
            if line == "":
                if current_session:
                    sessions.append(current_session)
                    current_session = []
            else:
                current_session.append(line)
        if current_session:
            sessions.append(current_session)
    return sessions

# Parse and preview
slam_sessions = parse_slam_sessions(train_file)
print(f"✅ Parsed sessions: {len(slam_sessions)}")
if slam_sessions:
    print("\nSample Session:")
    for line in slam_sessions[0]:
        print(line)
else:
    print("⚠ No sessions found.")

# Save parsed sessions to disk
parsed_sessions_path = project_root / "data" / "interim" / "slam_sessions.pkl"
parsed_sessions_path.parent.mkdir(parents=True, exist_ok=True)
with open(parsed_sessions_path, "wb") as f:
    pickle.dump(slam_sessions, f)
print(f"✅ slam_sessions saved to: {parsed_sessions_path}")
```

```
✅ Parsed sessions: 824012

Sample Session:
# prompt:Yo soy un niño.
# user:XEinXf5+  countries:CO  days:0.003  client:web  session:lesson  format:reverse_translate  time:9
DRihrVmh0101  I            PRON    Case=Nom|Number=Sing|Person=1|PronType=Prs|fPOS=PRON++PRP            nsubj       4  0
DRihrVmh0102  am           VERB    Mood=Ind|Number=Sing|Person=1|Tense=Pres|VerbForm=Fin|fPOS=VERB++VBP  cop         4  0
DRihrVmh0103  a            DET     Definite=Ind|PronType=Art|fPOS=DET++DT                                det         4  0
DRihrVmh0104  boy          NOUN    Number=Sing|fPOS=NOUN++NN                                             ROOT        0  0
✅ slam_sessions saved to: f:\Bachleros Research\Rsearch thesis\New folder\Predicting-Churn-using-ML-and-DL\data\interim\slam_sessions.pkl
```

📋 Notebook Summary

This notebook accomplishes:

1. **Data Loading**: Loads raw SLAM session data
2. **Data Parsing**: Converts raw text into structured sessions
3. **Validation**: Ensures data integrity and structure
4. **Statistics**: Provides session count and memory usage
5. **Storage**: Saves processed data for next steps

Next notebook: `02_preprocessing_feature_eng.ipynb`