Mohamed Raaiz

Project: Question-Answer Pairing (Text Labeling)

Core Idea / Introduction

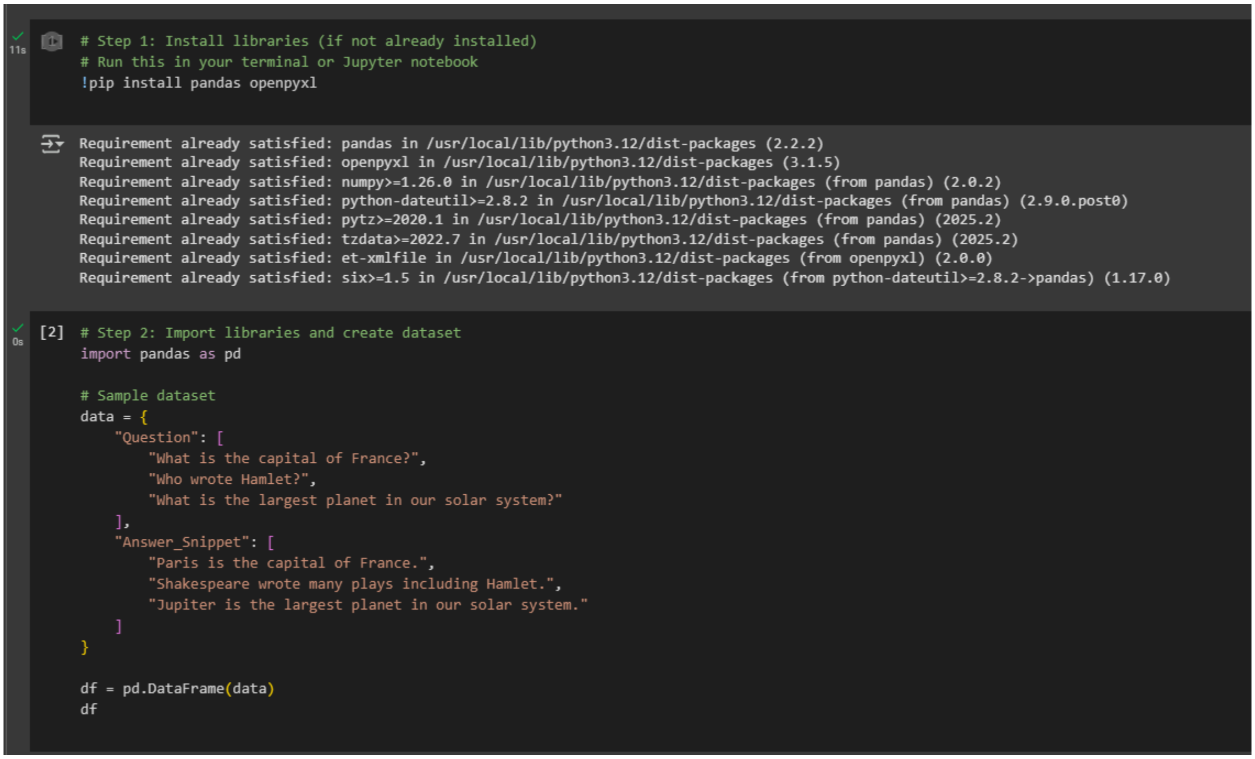
The goal of a Question-Answer Pairing project is to **identify whether a given text snippet correctly answers a specific question**. This type of labeling is crucial for training AI systems in **chatbots, search engines, and Q&A platforms**. Proper labeling ensures that the AI learns to provide **accurate and relevant answers**.

Key Points / Guidelines

1. **Understand the Question Clearly:**
   - Read the question carefully to grasp its intent.
   - Look for keywords or context that define what a correct answer would include.
2. **Check the Text Snippet:**
   - Determine if the snippet fully or partially answers the question.
   - Ignore irrelevant or misleading information.
3. **Labeling Categories:**
   - **Correct / Relevant:** The snippet directly answers the question.
   - **Incorrect / Irrelevant:** The snippet does not answer the question.
   - **Partially Correct / Ambiguous:** The snippet is related but incomplete or unclear.
4. **Consistency is Key:**
   - Follow the same criteria for all text snippets to maintain high-quality labeling.
5. **Optional Notes:**
   - You may add a brief comment for why a snippet was labeled ambiguous or partially correct.
   - Highlight phrases that match the question for easier review.

Example

Question	Text Snippet	Label	Notes
What is the capital of France?	Paris is the capital of France.	Correct	Exact match.
What is the capital of France?	France is in Europe.	Incorrect	Does not answer the question.
What is the capital of France?	France has many cities like Lyon and Marseille.	Partially Correct	Mentions cities, but not the capital.



```
# Step 1: Install libraries (if not already installed)
# Run this in your terminal or Jupyter notebook
!pip install pandas openpyxl
```

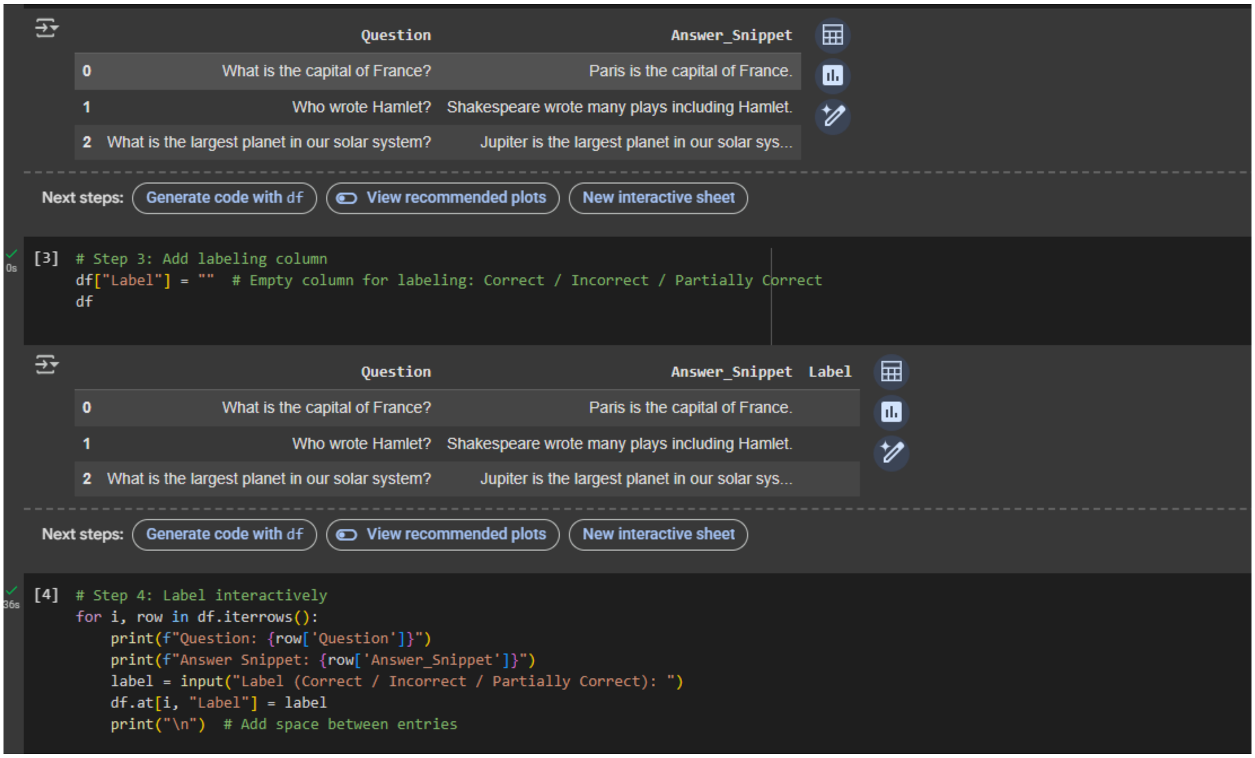
```
Requirement already satisfied: pandas in /usr/local/lib/python3.12/dist-packages (2.2.2)
Requirement already satisfied: openpyxl in /usr/local/lib/python3.12/dist-packages (3.1.5)
Requirement already satisfied: numpy>=1.26.0 in /usr/local/lib/python3.12/dist-packages (from pandas) (2.0.2)
Requirement already satisfied: python-dateutil>=2.8.2 in /usr/local/lib/python3.12/dist-packages (from pandas) (2.9.0.post0)
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.12/dist-packages (from pandas) (2025.2)
Requirement already satisfied: tzdata>=2022.7 in /usr/local/lib/python3.12/dist-packages (from pandas) (2025.2)
Requirement already satisfied: et-xmlfile in /usr/local/lib/python3.12/dist-packages (from openpyxl) (2.0.0)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.12/dist-packages (from python-dateutil>=2.8.2->pandas) (1.17.0)
```

[2]

```
# Step 2: Import libraries and create dataset
import pandas as pd

# Sample dataset
data = {
    "Question": [
        "What is the capital of France?",
        "Who wrote Hamlet?",
        "What is the largest planet in our solar system?"
    ],
    "Answer_Snippet": [
        "Paris is the capital of France.",
        "Shakespeare wrote many plays including Hamlet.",
        "Jupiter is the largest planet in our solar system."
    ]
}

df = pd.DataFrame(data)
df
```

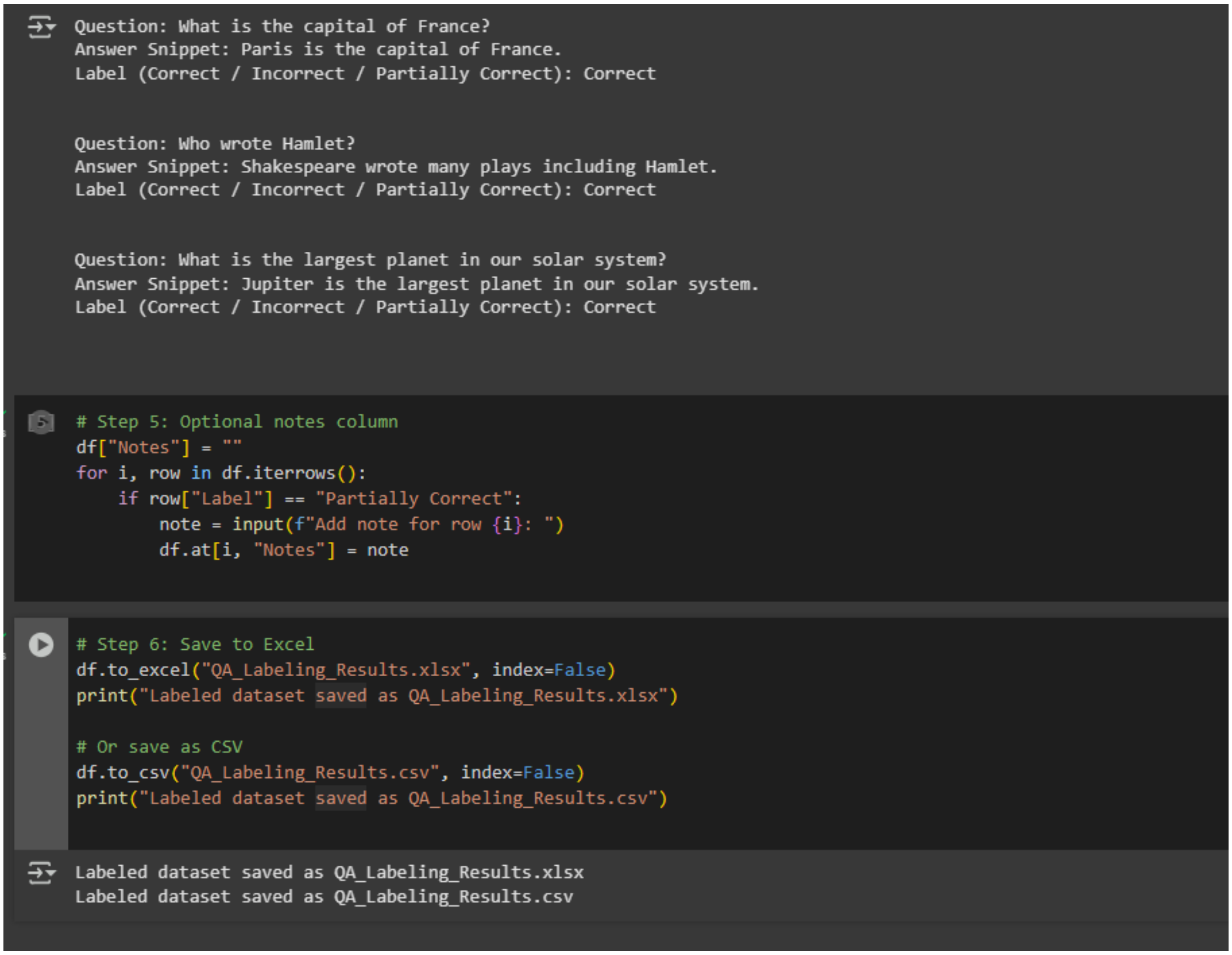


	Question	Answer_Snippet
0	What is the capital of France?	Paris is the capital of France.
1	Who wrote Hamlet?	Shakespeare wrote many plays including Hamlet.
2	What is the largest planet in our solar system?	Jupiter is the largest planet in our solar sys...

```
# Step 3: Add labeling column
df["Label"] = ""  # Empty column for labeling: Correct / Incorrect / Partially Correct
df
```

	Question	Answer_Snippet	Label
0	What is the capital of France?	Paris is the capital of France.	
1	Who wrote Hamlet?	Shakespeare wrote many plays including Hamlet.	
2	What is the largest planet in our solar system?	Jupiter is the largest planet in our solar sys...	

```
# Step 4: Label interactively
for i, row in df.iterrows():
    print(f"Question: {row['Question']}")
    print(f"Answer Snippet: {row['Answer_Snippet']}")
    label = input("Label (Correct / Incorrect / Partially Correct): ")
    df.at[i, "Label"] = label
    print("\n")  # Add space between entries
```



```
Question: What is the capital of France?
Answer Snippet: Paris is the capital of France.
Label (Correct / Incorrect / Partially Correct): Correct


Question: Who wrote Hamlet?
Answer Snippet: Shakespeare wrote many plays including Hamlet.
Label (Correct / Incorrect / Partially Correct): Correct


Question: What is the largest planet in our solar system?
Answer Snippet: Jupiter is the largest planet in our solar system.
Label (Correct / Incorrect / Partially Correct): Correct
```

```python
# Step 5: Optional notes column
df["Notes"] = ""
for i, row in df.iterrows():
    if row["Label"] == "Partially Correct":
        note = input(f"Add note for row {i}: ")
        df.at[i, "Notes"] = note
```

```python
# Step 6: Save to Excel
df.to_excel("QA_Labeling_Results.xlsx", index=False)
print("Labeled dataset saved as QA_Labeling_Results.xlsx")

# Or save as CSV
df.to_csv("QA_Labeling_Results.csv", index=False)
print("Labeled dataset saved as QA_Labeling_Results.csv")
```

```
Labeled dataset saved as QA_Labeling_Results.xlsx
Labeled dataset saved as QA_Labeling_Results.csv
```

Conclusion

Question-Answer Pairing ensures AI systems **understand and match questions with accurate answers**. Accurate labeling improves:

- **User satisfaction** in chatbots.
- **Search relevance** in knowledge retrieval.
- **Training quality** for NLP models.

Tip: Always check for **relevance, accuracy, and completeness**. Ambiguity should be labeled carefully to avoid model confusion.