IDENTIFICATION PROJECT DATA LABELING

MOHAMED RAAIZ

Mohamed Raaiz

Paraphrase Identification Project

Core Introduction

Paraphrase Identification is a text labeling task where the goal is to determine **whether two sentences convey the same meaning**. This is widely used in **chatbots, search engines, plagiarism detection, and question-answering systems**. Accurate labeling ensures models can **understand semantic similarity** and provide relevant responses.

Key Points / Guidelines

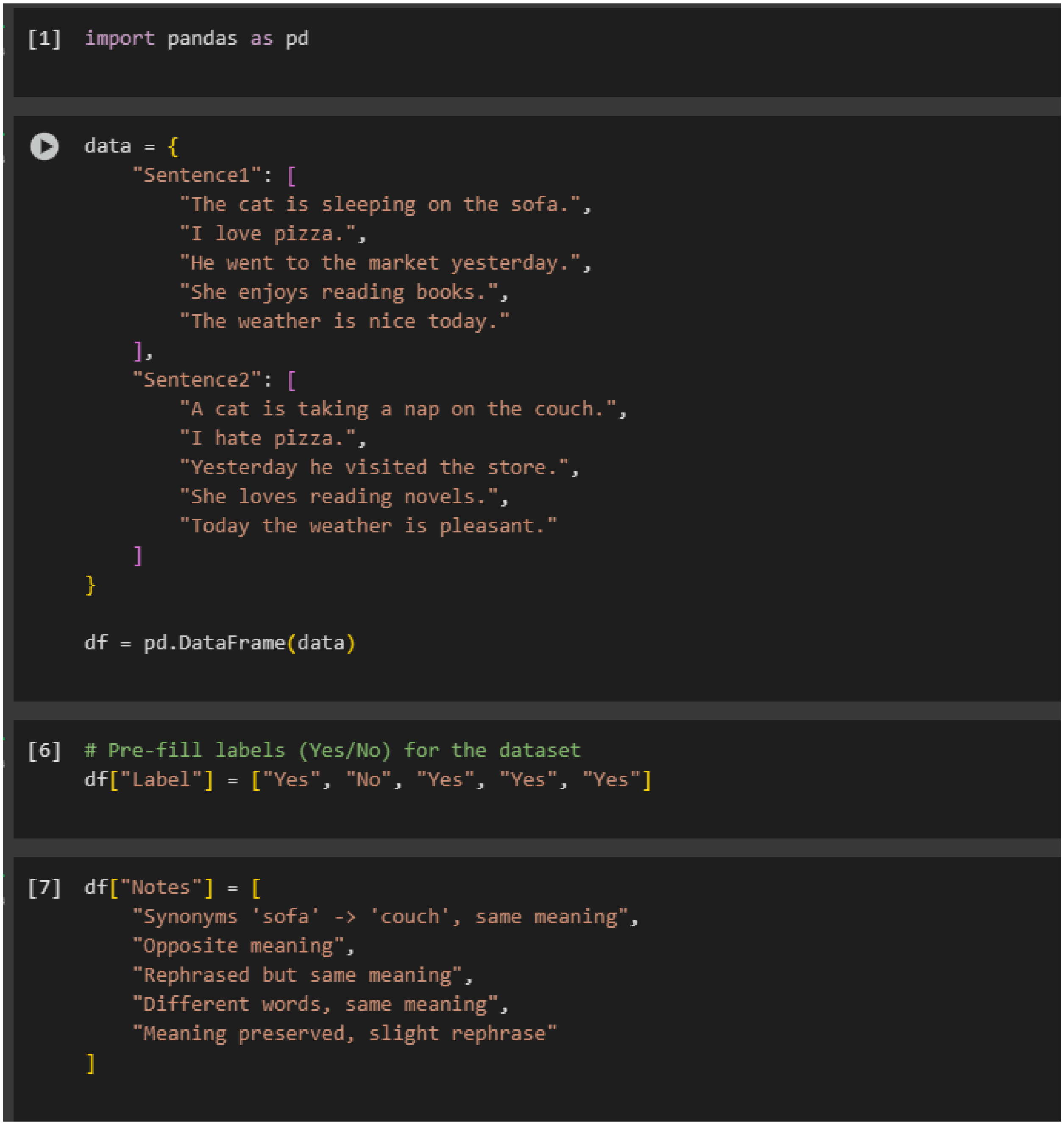
1. **Read Both Sentences Carefully:**
   - Understand the meaning, not just the words.
   - Watch for synonyms, reordering, or subtle changes.
2. **Label Categories:**
   - **Yes / Paraphrase:** Both sentences mean the same thing.
   - **No / Not Paraphrase:** Sentences convey different meanings.
3. **Consider Context:**
   - Some sentences may have ambiguous meaning. Mark carefully.
4. **Consistency:**
   - Use the same criteria across all examples to avoid noisy data.

Example

Sentence 1	**Sentence 2**	**Label**	**Notes**
"The cat is sleeping on the sofa."	"A cat is taking a nap on the couch."	Yes	Synonyms “sofa” → “couch”, same meaning.
"I love pizza."	"I hate pizza."	No	Opposite meaning.
"He went to the market yesterday."	"Yesterday he visited the store."	Yes	Meaning preserved, slight rephrasing.

Mohamed Raaiz

Coding



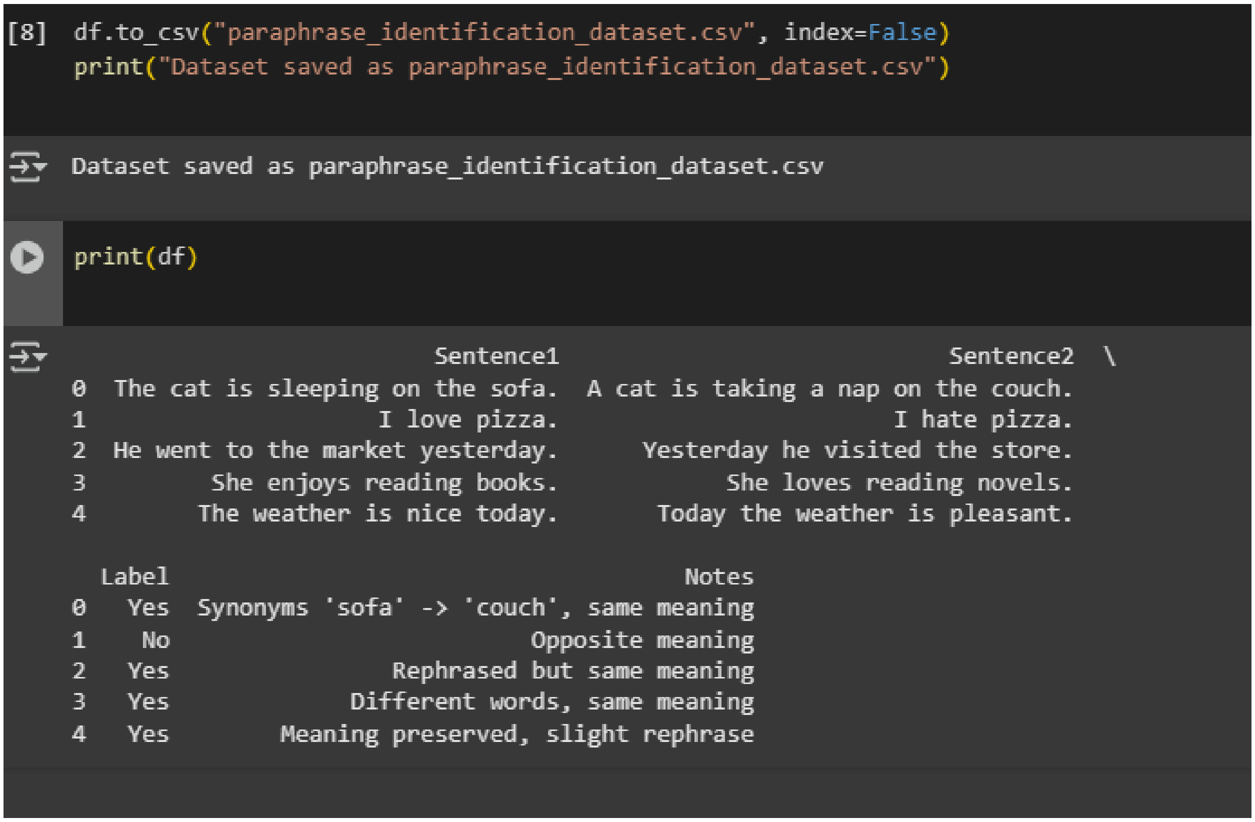
```
[1]  import pandas as pd
```

```
     data = {
         "Sentence1": [
             "The cat is sleeping on the sofa.",
             "I love pizza.",
             "He went to the market yesterday.",
             "She enjoys reading books.",
             "The weather is nice today."
         ],
         "Sentence2": [
             "A cat is taking a nap on the couch.",
             "I hate pizza.",
             "Yesterday he visited the store.",
             "She loves reading novels.",
             "Today the weather is pleasant."
         ]
     }

     df = pd.DataFrame(data)
```

```
[6]  # Pre-fill labels (Yes/No) for the dataset
     df["Label"] = ["Yes", "No", "Yes", "Yes", "Yes"]
```

```
[7]  df["Notes"] = [
         "Synonyms 'sofa' -> 'couch', same meaning",
         "Opposite meaning",
         "Rephrased but same meaning",
         "Different words, same meaning",
         "Meaning preserved, slight rephrase"
     ]
```



```
[8] df.to_csv("paraphrase_identification_dataset.csv", index=False)
    print("Dataset saved as paraphrase_identification_dataset.csv")
```

```
Dataset saved as paraphrase_identification_dataset.csv
```

```
print(df)
```

```
                         Sentence1                            Sentence2  \
0  The cat is sleeping on the sofa.  A cat is taking a nap on the couch.
1                    I love pizza.                        I hate pizza.
2  He went to the market yesterday.      Yesterday he visited the store.
3         She enjoys reading books.           She loves reading novels.
4        The weather is nice today.      Today the weather is pleasant.

  Label                                  Notes
0   Yes  Synonyms 'sofa' -> 'couch', same meaning
1    No                       Opposite meaning
2   Yes              Rephrased but same meaning
3   Yes             Different words, same meaning
4   Yes        Meaning preserved, slight rephrase
```

Conclusion

Paraphrase Identification is essential for **natural language understanding**. Proper labeling improves:

- **Search relevance** by matching similar queries.
- **Chatbot responses** by recognizing equivalent user inputs.
- **AI training** for semantic similarity and paraphrase detection models