Ex5b: Information Extraction using spaCy

```
import pandas as pd
import spacy
import string
```

```
from google.colab import drive
drive.mount('/content/drive')
```

```
Mounted at /content/drive
```

```
import pandas as pd
file_path = "/content/drive/MyDrive/Reviews.csv"
```

```
df = pd.read_csv(file_path)
reviews = df['Text'].dropna().head(10000)
print(reviews.head())
```

```
0    I have bought several of the Vitality canned d...
1    Product arrived labeled as Jumbo Salted Peanut...
2    This is a confection that has been around a fe...
3    If you are looking for the secret ingredient i...
4    Great taffy at a great price.  There was a wid...
Name: Text, dtype: object
```

```
def preprocess(text):
    text = text.lower()
    text = text.translate(str.maketrans("", "", string.punctuation))
    return text
reviews = reviews.apply(preprocess)
```

```
import spacy
nlp = spacy.load("en_core_web_sm")
doc = nlp(reviews.iloc[0])
tokens = [token.text for token in doc if token.is_alpha and not token.is_stop]
print("Tokens:\n", tokens)
```

```
Tokens:
 ['bought', 'vitality', 'canned', 'dog', 'food', 'products', 'found', 'good', 'quality', 'product', 'looks', 'like', 'stew', 'p
```

```
pos_tags = [(token.text, token.pos_) for token in doc if token.is_alpha]

print("POS Tags:\n", pos_tags)
```

```
POS Tags:
 [('i', 'PRON'), ('have', 'AUX'), ('bought', 'VERB'), ('several', 'ADJ'), ('of', 'ADP'), ('the', 'DET'), ('vitality', 'NOUN'),
```

```
entities = [(ent.text, ent.label_) for ent in doc.ents]

print("Named Entities:\n", entities)
```

```
Named Entities:
 []
```

```
Start coding or generate with AI.
```