EX.NO.:4
5/8/25

TEXT PREPROCESSING AND ANALYSIS PIPELINE

Aim:

To perform text cleaning (remove stop words, special characters) and tokenization on a text dataset

Code

```
import pandas as pd
import spacy
nlp = spacy.load("en_core_web_sm")
df = pd.read_csv("amazon_reviews.csv")
print(df['reviewText'].head())

def clean_text_spacy(text):
    if pd.isnull(text):
    return []

    text = text.lower()
    text = re.sub(r'[^\w\s]', '', text)

    doc = nlp(text)
    tokens = [token.text for token in doc if not
              token.is_stop and not token.is_punct]

    return tokens
```

OUTPUT:

We got this GPS for my husband who is an (OTR)...

1 I'm a professional OTR truck driver, and I have...
2 well, what can I say. I've had this unit in...
3 Not going to write a long review, even though...
4 I've had mine for a year and here's what we g...

Name : reviewText, dtype : object

```
df['cleaned_news'] = df['sentence'].apply(clean_text_proc
print(df[['sentence', 'cleaned_tokens']].head(5))

all_tokens = [token for tokens in df['cleaned_tokens']
              for token in tokens]

from collections import Counter
word_freq = Counter(all_tokens)

print("In Top 15 frequent words in Amazon Reviews")
print(word_freq.most_common(15))
```

Results:

Once the program to perform text cleaning has been executed successfully