EX.NO: 4
15/8/25

TEST PREPROCESSING AND ANALYTICS PIPELINE.

AIM:

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

PROGRAM CODE:

```
import pandas as pd
import re
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)
    text = text.encode('ascii', 'ignore').decode
                ('ascii')
    doc = nlp(text)
    tokens = [token.text for token in doc
              if not token.is_stop and not
                 token.is_punct]
    return tockens
df['cleaned_tokens'] = df['reviewText'].
                apply(clean_text_spacy)
Print(df[['reviewText','cleaned_tokens']].
                head(5))
```

OUTPUT:

me got this tips for my husband who is an (OTR):...

1) I'm professional OTR truck driver, and I bou...

2) Well, what can I say. I've had this unit in m...

3) Not going to write a long review, even thought...

4) I've had mine for a year and here's what me go..

Name: reviewText, dtype: object.

```
all_tokens = [token for tokens in df['cleaned_token']
              for token in tokens]

from collection import counter
word_freq = counter(all_tokens)

Print("\n Top 15 frequent words in Amazon
                                Reviews:")
Print( word_freq.most_common(15))
```

RESULT:

Thus the text cleaning performance has been executed successfully.