EX NO 4

5/8/25

TEST PREPROCESSING AND ANALYTICS PIPELINE

AIM:

To perform text 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_review.csv')
print(df['review'].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 tokens
df['cleaned_tokens'] = df['review_text'].apply(clean_text_spacy)
print(df[['review_text','cleaned_tokens']].head())
```

OUTPUT:

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

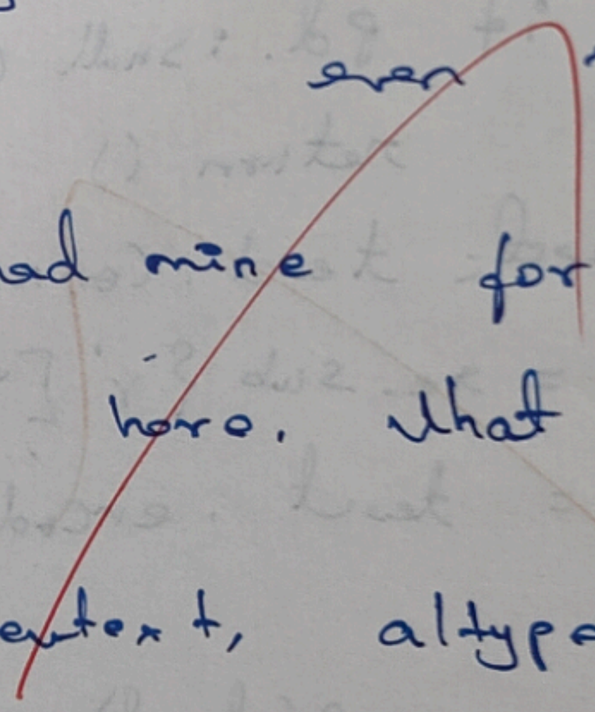
1) I'm professional OTR truck driver, and I bow

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. what we go...

name : varientext, altype, object.



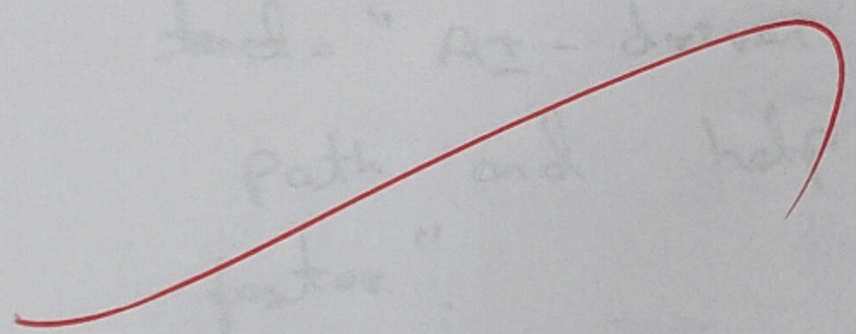
```
all - tokens = [token for tokens in df
                    for token in tokens]

from collection import Counter

word - freq = Counter (all - tokens)

print("\n Top 15 frequented words in
                    Amazon Reviews!:")

Print(word - freq. most_common(15))
```



Result:

Thus the text cleaning performance has been executed successfully.