s1a1e5

Ex: 4

s1s125

Text Preprocessing And Analytic pipeline

AIM:

To perform text cleaning (remove stop words, special characters) and tokenisation 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'[^a-z\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['reviewsText'].apply(clean_text_spacy)
print(df['reviewText', 'cleaned_tokens']].head(5))
```

2025/10/25

OUTPUT:

We get this title for my husband who is an (OFW) :....

1) I'm professional OFW truck driver, and ...

2) Well what can I say. I have had this unit....

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

4) I've had mine for a year and what we go....

N

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

from collection import Counter
word_freq = Counter(all_tokens)
print("\nTop 15 frequent words in Amazon
                                reviews:")
print(word_freq.most_common(15))
```

RESULT:

Thus the text cleaning preformance has been executed succesfully.

2025/10/25