EX.NO:4 Text Preprocessing and analytics Pipeline

AIM:

To determine the test Preprocessing and data analytics Pipeline.

CODE:

```
import pandas as pd
import re
nlp = spacy.load("en_core_web")
df = pd.read_csv('amazon_reviews.csv')
Print(df[reviewText].load())
# function to clean the test casting Speech
def clean_text_space(text):
    if pd.isnull(text)
return[]
text = txt.lower()
text = re.sub(r'^[\w\s]_" kxt)
# Tokensize using spay
doc = nlp(text)
token = (token.text for token in doc if not
        token.is_stop and not token_is_pd
return tokens.
```

Top is frequent word in Amazon Reviews

[(" ,3203) ('hide' (447), ('int', 962) ('bude' 608)

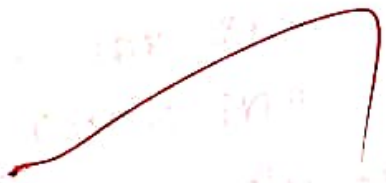
(kindes' 1561) (soown ,4.73) ('like', 452)

(reel', 434) (great' 422) ('use', 420)

('tv', 380) ('rduct', 347) ('good'), 529)

('device', 329) ('bo', 522)]

```
all_tokens = [token for token in df['cleaned_tokens'] for token in tokens]
Print ("\n TOP is frequent words in Amazon Review:")
Print (word_freq.most_convers(15))
```



RESULT:

The given Test Preprocessing and analytic pipeline has been excellent successfully.