exp4. Test preprocessing and Analytics pipeline

Aim:

To determine the text preprocessing and analytics pipline.

Code:

```
import pandas as pd
import re
nlp = spacy.load('en_core_web_sm')
df = pd.read_csv('amazon_revsm.csv')
print(df['reviewText'].head())
# function to clean the text casing speech.
def clean_text_space(text):
    if pd.isnull(text)
    return []
text = text.lower() # convert to lowercase
text = re.sub(r'[^\w\s]', '', text) # Reverse function.
# Tokenize using spacy
doc = nlp(text)
token = [token.text for token in doc if not
        token.is_stop and not token.is_punct)
return token_s.
```

Top 15 frequent word in Amazon Review

[('', 3203) ('nude' 1447), ('int', 962) ('bads' 608)
('kindle', 561) (sonon, 473), ('like', 452).
('real', 434), (great', 422) ('use', 420)
('tv', 380), ('tablet', 347) ('good', 329)
('device', 329), ('bn', 321)]

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

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

Result:

The given Text preprocessing and Analytic pipeline has been Excellent successfully.