

expt.

Text preprocessing and Analytics pipeline

Aim:

To determine the text preprocessing and analytics pipeline.

Code:

```
import pandas as pd, time, re, nltk
import nltk

nlp = nltk.load('en-core-web-tr')

df = pd.read_csv('amazon-reviews.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) # Remove function.
    # Tokenize using spacy

    doc = nlp(text)

    tokens = [token.text for token in doc if not
               token.is_stop and not token.is_punct]
    return tokens
```

Top 15 frequent word in Amazon Review

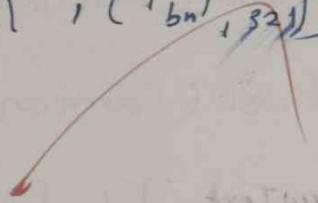
[(' ', 3203) ('hude', 1447), ('int', 962) ('buds', 602)

('kinder', 561) ('soreen', 473), ('like', 452).

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

('tv', 380), ('tallit', 347) ('good', 329)

('device', 329), ('bn', 321)]



Tokenize using word

class: nlp (class)

token = [token, text for token in text if token != ' ']
token = stop word not token (is - end)

return token

```
all_tokens = [token for token in df ['clean-  
tokens']] for token in tokens]
```

```
print (" In Top 15 frequent words in Amazon  
Review:")
```

```
print (word_freq.most_common(15))
```

Result:

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