

## Experiment - 4.

### Text preprocessing and analytics pipeline

Aim: To determine the text preprocessing and analytics pipeline.

Code:

```
import pandas as pd
import re
import spacy
# load spacy tokenizer.
np = spacy.load("en-core-web-sm")
# load the data from csv-file
df = pd.read_csv("car-review.csv")
# print (df[['review', 'text']].head())
def clean_text_spacy(text):
    if pd.isnull(text):
        return []
    text = text.lower()
    text = re.sub(r"[^\w\s]", "", text)
```

```
text = text.lower()
```

```
text = re.sub(r"[^\w\s]", "", text)
```

Tokenizer using spacy

doc = ls (text -

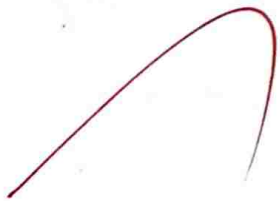
token = [token.text for token in doc  
if not token.is\_stop and not  
token.is\_punct]

return tokens

df\_tokens = [token for i, token in df[df['token']  
for token in tokens]

print ("In Top 15 frequent word or  
Linear Regression")

print (word\_freq.most\_common(15))



Result: The given text preprocessing  
and analytic pipelines has been  
created successfully.

Top 15 frequent word in Review  
 [ ('', 3203) ('u', 1447) , (w, 962)  
 (books, 6032) (kindly, 561) (wh, 452)  
 (fulled 342) (good, 327) (device, 327)

## Output

1. I'm a Professional OTR truck driver and
2. Well, what can I say I've had this
3. Not going to write --

## review Txt

1. I'm a Professional
2. well, what can I say
3. Not going to write along reviews