Experiment - 4

Text preprocessing and analytics pipeline

Aim:
To determine the text preprocessing and analytics pipeline.

Code:

```
import pandas as pd
import
import spacy
# load .spacy tokenizer.
nlp = spacy.load("en_core_web_sm")
# Read the User Review csv file
df = pd.read_csv("user_review.csv")
print(df['review Text'].head())
def clean_text_spacy(text):
    if pd.isnull(text):
        return []
text = text.lower()
text = re.sub(r'[\d\s]', '', text)
```

Tokenize using spacy

doc = nlp(text)

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

return tokens

df['tokens'] = [token for token in df['cleaned
token] for token in tokens]

print("In Top 15 frequent word of
linear Regression")

print(word_freq.most_common(15))

Result: The given text preprocessing and Analytic pipelines has been executed successfully.

Top 15 frequent words in Review

[('.', 3203) ('the', 1447), ('to', 962)
('books', 603) ('kindle', 561) ('the', 452)
('futted 347) ('good', 327) ('device', 327)

OUTPUT

1. Im a Professional OTR trucke driver and I
2. Well ; what can I say I've had th
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