PRINCE KUMAR SINGH

S25MCAG0001

BATCH- 1

LAB NO- 7

```
import nltk
from nltk.corpus import brown,stopwords
from nltk.tokenize import sent_tokenize, word_tokenize
from nltk.stem import SnowballStemmer, WordNetLemmatizer
from nltk.probability import FreqDist
from wordcloud import WordCloud
import matplotlib.pyplot as plt

nltk.download('brown')
nltk.download('punkt')
nltk.download('stopwords')
nltk.download('wordnet')
nltk.download('omw-1.4')
nltk.download('punkt_tab')
```

```
[nltk_data] Downloading package brown to /root/nltk_data...
[nltk_data]   Package brown is already up-to-date!
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
[nltk_data] Downloading package wordnet to /root/nltk_data...
[nltk_data]   Package wordnet is already up-to-date!
[nltk_data] Downloading package omw-1.4 to /root/nltk_data...
[nltk_data]   Package omw-1.4 is already up-to-date!
[nltk_data] Downloading package punkt_tab to /root/nltk_data...
[nltk_data]   Package punkt_tab is already up-to-date!
True
```

```
text = brown.raw(categories='news')
sentences = sent_tokenize(text)
words = word_tokenize(text)
fdist_before = FreqDist(words)
stop_words = set(stopwords.words('english'))
filtered_words = [w.lower() for w in words if w.isalpha() and w.lower() not in stop_words]
fdist_after = FreqDist(filtered_words)

print("top 10 words before stopwords removal:", fdist_before.most_common(10))
print("\ntop 10 words after stopwords removal:", fdist_after.most_common(10))
```

```
top 10 words before stopwords removal: [(',', 10376), ('/', 7850), ('the/at', 5558), ('.', 4454), ('./', 4012), ('of/in', 2716), ('

top 10 words after stopwords removal: [('u', 1)]
```

```
stemmer = SnowballStemmer('english')
stemmer_words = [stemmer.stem(w) for w in filtered_words]
fdist_stem = FreqDist(stemmer_words)
print("top 10 stemmed words: ", fdist_stem.most_common(10))
```

```
top 10 stemmed words:  [('u', 1)]
```

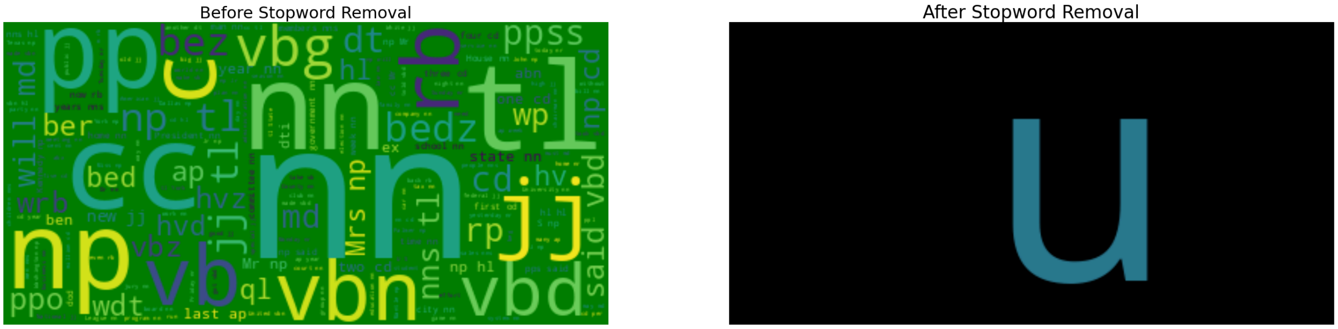
```
lemmatizer = WordNetLemmatizer()
lemmatizer_words = [lemmatizer.lemmatize(w) for w in filtered_words]
fdist_lamma = FreqDist(lemmatizer_words)
print("top 10 lemmatized words: ", fdist_lamma.most_common(10))
```

```
top 10 lemmatized words:  [('u', 1)]
```

```
wordcloud_before = WordCloud(width=400, height=200, background_color='green').generate(" ".join(words))
wordcloud_after = WordCloud(width=400, height=200, background_color='black').generate(" ".join(filtered_words))
plt.figure(figsize=(32,8))

plt.subplot(1,2,1)
plt.imshow(wordcloud_before, interpolation='bilinear')
plt.title("Before Stopword Removal", fontsize=22)
plt.axis("off")
plt.subplot(1,2,2)
plt.imshow(wordcloud_after, interpolation='bilinear')
plt.title("After Stopword Removal", fontsize=24)
plt.axis("off")

plt.show()
```



Before Stopword Removal

After Stopword Removal