Anuj Mutha - 41443

```
import pandas as pd
import numpy as np
import sklearn as sk
import math
import re
```

```
import nltk
!pip install textract
import docx2txt
```

```
my_text = "Millions of people in India took part in an annual tree planting drive Sunda
```

```
my_text
```

```
'Millions of people in India took part in an annual tree planting drive Sunday. More th
an 250 million saplings were p[illegible] across the country's most-populous
```

+ Code + Text

```
my_text= re.sub('[^A-Za-z0-9]+', ' ', my_text)
```

```
import nltk
nltk.download('punkt')
```

```
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
True
```

```
from nltk.tokenize import sent_tokenize
```

```
tokenized_text = sent_tokenize(my_text)
print(tokenized_text)
```

```
['Millions of people in India took part in an annual tree planting drive Sunday More tha
```

```
from nltk.tokenize import word_tokenize
```

```
tokenized_word = word_tokenize(my_text)
print(tokenized_word)
```

```
['Millions', 'of', 'people', 'in', 'India', 'took', 'part', 'in', 'an', 'annual', 'tree
```

```
import re
import nltk
nltk.download('stopwords')
from nltk.corpus import stopwords
```

```
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
```

```
stop_words = set(stopwords.words("english"))
print(stop_words)
```

```
{"needn't", 'of', 'who', 'if', 'an', "mustn't", 'yourselves', 'don', "she's", 'some', 'a
```

```
filtered_tokens = []
for w in tokenized_word:
    if w not in stop_words:
        filtered_tokens.append(w)

print("Tokenized Words:\n",tokenized_word)
print("\n\nFilterd Tokens:\n",filtered_tokens)
```

```
Tokenized Words:
 ['Millions', 'of', 'people', 'in', 'India', 'took', 'part', 'in', 'an', 'annual', 'tree


Filterd Tokens:
 ['Millions', 'people', 'India', 'took', 'part', 'annual', 'tree', 'planting', 'drive',
```

```
from nltk.stem import PorterStemmer


ps = PorterStemmer()
stemmed_words=[]

for w in filtered_tokens:
    stemmed_words.append(ps.stem(w))

print("Filtered Tokens After Removing Punctuations:\n",filtered_tokens)
print("\n\nStemmed Tokens:\n",stemmed_words)
```

```
Filtered Tokens After Removing Punctuations:
 ['Millions', 'people', 'India', 'took', 'part', 'annual', 'tree', 'planting', 'drive',
```

```
Stemmed Tokens:
 ['million', 'peopl', 'india', 'took', 'part', 'annual', 'tree', 'plant', 'drive', 'sunc
```


✓ 0s completed at 1:50 PM