Text Mining

```
import nltk
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize
from nltk.probability import FreqDist
```

```
#read text file
with open('/content/sample_data/README.md', 'r') as file:
  text = file.read()
```

```
#Tokenize the text into indivisual words
token = word_tokenize(text)
```

```
#remove stop word from the tokenised list
stop_word = set(stopwords.words('english'))
filtered_token = [word for word in token if not word.lower() in stop_word]
```

```
#calculate the frequency distribution of the remaining word
fdist = FreqDist(filtered_token)
```

```
#print the 10 most common word and their frequencies
for word, frequency in fdist.most_common(10):
  print(f'{word}: {frequency}')
```

```
:: 8
.: 7
`: 6
*: 5
(: 5
): 5
https: 4
[: 3
]: 3
sample: 2
```