```
import nltk
nltk.download('punkt')
nltk.download('stopwords')
from nltk.corpus import wordnet as wn
import sklearn
```

```
[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!
```

```
from nltk.corpus.reader import wordnet
# Uploading the file
my_file = open("Anomaly.txt", "rt") # open lorem.txt for reading text
contents = my_file.read()         # read the entire file to string
```

```
#Cleaning the Text

# Below the code for removal of meta-deta
import re
from nltk.corpus import stopwords
from nltk.stem.porter import PorterStemmer  # For Steming the word
from nltk.stem import WordNetLemmatizer # It's an Word Limitizer

# Now Create Objects for cleaning Purposes

ps = PorterStemmer()
wordNet  =WordNetLemmatizer()
sentences = nltk.sent_tokenize(contents)
corpus = []
for i in range(len(sentences)):
  review = re.sub('[^a-zA-Z]', ' ', sentences[i])
  review = review.lower()
  review = review.split()
  review = [word for word in review if word not in stopwords.words('english')]
  #review = [wordnet.lemmatize(word) for word in review if not word in set (stopwords.words('
  review = ' '.join(review)
  corpus.append(review)
  review
```

```
type(contents)
```

```
str
```

```
len(contents)
```

```
817
```

#splitting the paragraph in sentences.

```
from nltk.tokenize import sent_tokenize

#splitting the paragraph in sentences.

len(sent_tokenize(contents)[:])

print(sent_tokenize(contents)[:])
```

```
["Anomaly detection (aka outlier analysis) is a step in data mining that identifies data
```

```
from nltk.tokenize import word_tokenize
text = contents
print(word_tokenize(text))
```

```
['Anomaly', 'detection', '(', 'aka', 'outlier', 'analysis', ')', 'is', 'a', 'step', 'in
```