

Output:

```
#split text by White Space
```

```
C:\Users\Arjun\Python123\Cleaning_Text_Data>python text_cleaning.py
['i>The', 'Project', 'Gutenberg', 'EBook', 'of', 'Metamorphosis', 'by', 'Franz', 'Kafka', 'Translated', 'by', 'David',
 'Wyllie', 'One', 'morning', 'when', 'Gregor', 'Samsa', 'woke', 'from', 'troubled', 'dreams', 'he', 'found', 'himself',
 'transformed', 'in', 'his', 'bed', 'into', 'a', 'horrible', 'vermin.', 'He', 'lay', 'on', 'his', 'armour-like', 'back',
 'and', 'if', 'he', 'lifted', 'his', 'head', 'a', 'little', 'he', 'could', 'see', 'his', 'brown', 'belly', 'slightly',
 'domed', 'and', 'divided', 'by', 'arches', 'into', 'stiff', 'sections.', 'The', 'bedding', 'was', 'hardly', 'able',
 'to', 'cover', 'it', 'and', 'seemed', 'ready', 'to', 'slide', 'off', 'any', 'moment.', 'His', 'many', 'legs', 'pitifully',
 'thin', 'compared', 'with', 'the', 'size', 'of', 'the', 'rest', 'of', 'him', 'waved', 'about', 'helplessly', 'as',
 'he', 'looked.', '"what', 's', 'happened']
=====
```

```
us = 'lookseq' # MISC/2 # jabbbeuq }
```

```
#split by words
```

```
['i', 'The', 'Project', 'Gutenberg', 'EBook', 'of', 'Metamorphosis', 'by', 'Franz', 'Kafka', 'Translated', 'by', 'David',
 'Wyllie', 'One', 'morning', 'when', 'Gregor', 'Samsa', 'woke', 'from', 'troubled', 'dreams', 'he', 'found', 'himself',
 'transformed', 'in', 'his', 'bed', 'into', 'a', 'horrible', 'vermin.', 'He', 'lay', 'on', 'his', 'armour', 'like', 'back',
 'and', 'if', 'he', 'lifted', 'his', 'head', 'a', 'little', 'he', 'could', 'see', 'his', 'brown', 'belly', 'slightly',
 'domed', 'and', 'divided', 'by', 'arches', 'into', 'stiff', 'sections.', 'The', 'bedding', 'was', 'hardly', 'able', 'to',
 'cover', 'it', 'and', 'seemed', 'ready', 'to', 'slide', 'off', 'any', 'moment.', 'His', 'many', 'legs', 'pitifully', 't',
 'hin', 'compared', 'with', 'the', 'size', 'of', 'the', 'rest', 'of', 'him', 'waved', 'about', 'helplessly', 'as', 'he', 'l',
 'ooked']
```

```
lookseq.]
hin', compared ' with', the', size', of', the', rest', of', him', waved', about', helplessly', as', he', 'l',
ooked']
```

```
#split text by whitespace and remove punctuation from each word
```

```
['i>The', 'Project', 'Gutenberg', 'EBook', 'of', 'Metamorphosis', 'by', 'Franz', 'Kafka', 'Translated', 'by', 'David',
 'Wyllie', 'One', 'morning', 'when', 'Gregor', 'Samsa', 'woke', 'from', 'troubled', 'dreams', 'he', 'found', 'himself',
 'transformed', 'in', 'his', 'bed', 'into', 'a', 'horrible', 'vermin.', 'He', 'lay', 'on', 'his', 'armourlike', 'back', 'an',
 'd', 'if', 'he', 'lifted', 'his', 'head', 'a', 'little', 'he', 'could', 'see', 'his', 'brown', 'belly', 'slightly', 'dome',
 'd', 'and', 'divided', 'by', 'arches', 'into', 'stiff', 'sections.', 'The', 'bedding', 'was', 'hardly', 'able', 'to', 'cov',
 'er', 'it', 'and', 'seemed', 'ready', 'to', 'slide', 'off', 'any', 'moment.', 'His', 'many', 'legs', 'pitifully', 'thin',
 'compared', 'with', 'the', 'size', 'of', 'the', 'rest', 'of', 'him', 'waved', 'about', 'helplessly', 'as', 'he', 'looked',
 'whats', 'happened']
=====
```

```
us = 'lookseq' # MISC/2 # jabbbeuq }
```

```
#convert to lower case
```

```
['i>the', 'project', 'gutenberg', 'ebook', 'of', 'metamorphosis', 'by', 'franz', 'kafka', 'translated', 'by', 'david',
 'Wyllie', 'One', 'morning', 'when', 'Gregor', 'Samsa', 'woke', 'from', 'troubled', 'dreams', 'he', 'found', 'himself',
 'transformed', 'in', 'his', 'bed', 'into', 'a', 'horrible', 'vermin.', 'He', 'lay', 'on', 'his', 'armour-like', 'back',
 'and', 'if', 'he', 'lifted', 'his', 'head', 'a', 'little', 'he', 'could', 'see', 'his', 'brown', 'belly', 'slightly',
 'domed', 'and', 'divided', 'by', 'arches', 'into', 'stiff', 'sections.', 'The', 'bedding', 'was', 'hardly', 'able',
 'to', 'cover', 'it', 'and', 'seemed', 'ready', 'to', 'slide', 'off', 'any', 'moment.', 'his', 'many', 'legs', 'pitifully',
 'thin', 'compared', 'with', 'the', 'size', 'of', 'the', 'rest', 'of', 'him', 'waved', 'about', 'helplessly', 'as',
 'he', 'looked.', '"what', 's', 'happened']
=====
```

```
us = 'lookseq' # MISC/2 # jabbbeuq }
```

```
#cleaning text using nltk (Natural Language Toolkit)

#tokenixation
```

```
#tokenixation
```

y David willie Project Gutenberg eBook of Metamorphosis by Franz Kafka Translated b
y found himself transformed in his bed into a horrible vermin from troubled dreams
his armour-like back and if he lifted his head a little he could see hi
s brown belly slightly domed and divided by arches into stiff sections The
bedding was hardly able to cover it and seemed ready to slide off any mome
nt His many legs pitifully thin compared with the size of]

ic : his : red : billion : rufu : cowbleg : wifu : cue : site : ol :

```
#remove all token which ar not alphabetic
```

i, 'Project', 'Gutenberg', 'EBook', 'of', 'Metamorphosis', 'by', 'Franz', 'Kafka', 'Translated', 'by', 'David', 'Wyll
 le', 'One', 'morning', 'when', 'Gregor', 'Samsa', 'woke', 'from', 'troubled', 'dreams', 'he', 'found', 'himself', 'trans
 formed', 'in', 'his', 'bed', 'into', 'a', 'horrible', 'vermin'. 'He', 'lay', 'on', 'his', 'back', 'and', 'if', 'he', 'li
 fted', 'his', 'head', 'a', 'little', 'he', 'could', 'see', 'his', 'brown', 'belly', 'slightly', 'domed', 'and', 'divided
 by', 'arches', 'into', 'stiff', 'sections'. 'The', 'bedding', 'was', 'hardly', 'able', 'to', 'cover', 'it', 'and'
 seemed', 'ready', 'to', 'slide', 'off', 'any', 'moment'. 'His', 'many', 'legs', 'pitifully', 'thin', 'compared', 'with',
 the', 'size', 'of', 'the', 'rest', 'of', 'him', 'waved', 'about', 'helplessly', 'as', 'he', 'looked', 'what', 'happene
 d', 'to']
 =====
 3. co.]

```
#printing stopword
```

(i 'me' 'my' 'myself' 'we' 'our' 'ours' 'ourselves' 'you' 'you're' 'you've' 'you'll' 'you'd' 'your' 'you
 rs' 'yourself' 'yourselves' 'he' 'him' 'his' 'himself' 'she' 'she's' 'her' 'hers' 'herself' 'it' 'it's' 'i
 t' 'itself' 'they' 'them' 'their' 'theirs' 'themselves' 'what' 'which' 'who' 'whom' 'this' 'that' 'that'll' 'i
 s' 'these' 'those' 'am' 'is' 'are' 'was' 'were' 'be' 'been' 'being' 'have' 'has' 'had' 'having' 'do' 'doe
 s' 'did' 'doing' 'a' 'an' 'the' 'and' 'but' 'if' 'or' 'because' 'as' 'until' 'while' 'of' 'at' 'by' 'fo
 r' 'with' 'about' 'against' 'between' 'into' 'through' 'during' 'before' 'after' 'above' 'below' 'to' 'from
 up' 'down' 'in' 'out' 'on' 'off' 'over' 'under' 'again' 'further' 'then' 'once' 'here' 'there' 'when' 'w
 here' 'why' 'how' 'all' 'any' 'both' 'each' 'few' 'more' 'most' 'other' 'some' 'such' 'no' 'nor' 'not'
 'only' 'own' 'same' 'so' 'than' 'too' 'very' 's' 't' 'can' 'will' 'just' 'don' 'don't' 'should' 'should
 ve' 'now' 'd' 'll' 'm' 'o' 're' 've' 'y' 'ain' 'aren' 'aren't' 'couldn' 'couldn't' 'didn' 'didn't' 'doe
 sn't' 'doesn't' 'hadn' 'hadn't' 'hasn' 'hasn't' 'haven' 'haven't' 'isn' 'isn't' 'ma' 'mightn' 'mightn't' 'mus
 tn' 'mustn't' 'needn' 'needn't' 'shan' 'shan't' 'shouldn' 'shouldn't' 'wasn' 'wasn't' 'weren' 'weren't' 'won
 ' 'won't' 'wouldn' 'wouldn't']
 =====
 mou ㄇㄡˊ 'monjqu' 'monjqu ㄇㄡˊ
 tu ㄊㄨˊ 'tu ㄊㄨˊ ㄟ 'shonjqu' 'shonjqu ㄟ 'mɿʈu' 'mɿʈu ㄟ 'mɿleu' 'mɿleu ㄟ 'mou

```
#filter out stopword
```

```
[ ] the project, gutenberg ebook of 'metamorphosis' by franz kafka, translated by david willie, one morning when gregor samsa woke from troubled dreams, he found himself transformed in his bed into a horrible vermin. he lay on his armour-like back, and if he lifted his head a little, he could see his brown belly, slightly domed and divided by arches into stiff sections; the bedding was hardly able to cover it and seemed ready to slide off any moment. his many legs, pitifully thin compared with the size of the rest of him, waved about helplessly as he looked. "what's happened?"
```

#stemming of words

```
[('i', '>', 'the', 'project', 'gutenberg', 'ebook', 'of', 'metamorphosi', 'by', 'franz', 'kafka', 'translat', 'by',  
'david', 'wylli', 'one', 'morn', 'when', 'gregor', 'samsa', 'woke', 'from', 'troubl', 'dream', 'he', 'fo  
und', 'himself', 'transform', 'in', 'hi', 'bed', 'into', 'a', 'horribl', 'vermin', 'he', 'lay', 'on', 'hi', 'armour  
-lik', 'back', 'and', 'if', 'he', 'lift', 'hi', 'head', 'a', 'littl', 'he', 'could', 'see', 'hi', 'brown', 'belli',  
abl', 'slightli', 'dome', 'and', 'divid', 'by', 'arch', 'into', 'stiff', 'section', 'the', 'bed', 'wa', 'hardli',  
'piti', 'thin', 'compar', 'with', 'the', 'size', 'of']  
biri', 'cuni', 'combi', 'mipu', 'cpe', 'size', 'of']
```