

CSE 3024

Web Mining

LAB ASSESSMENT - 1

NAME: Vibhu Kumar Singh

REG. NO: 19BCE0215

TEACHER: Mr. Hiteshwar Kumar Azad

1. Create a Python programme to tokenize the following using the NLTK toolkit:

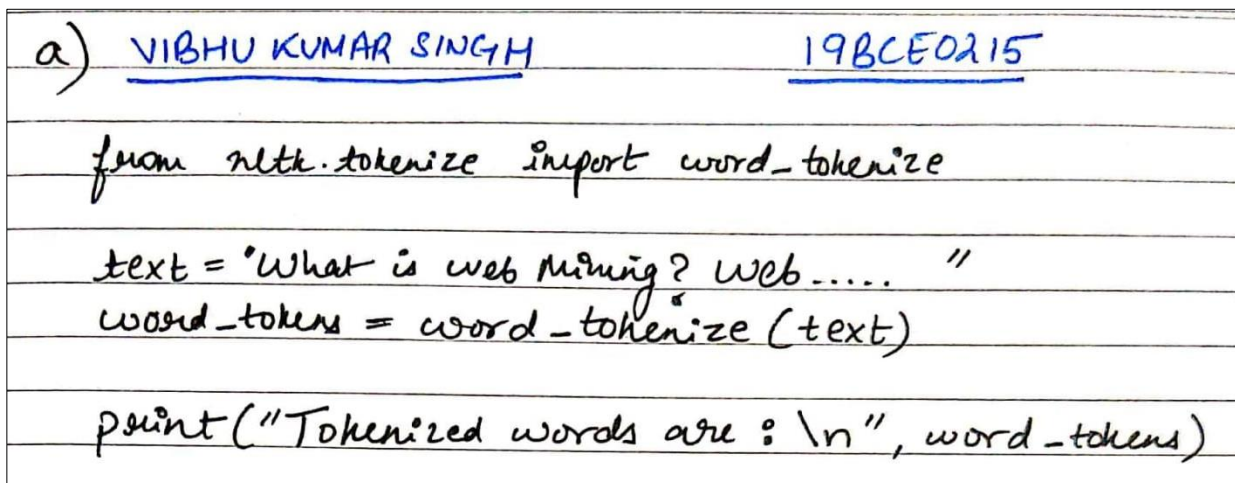
- a) word
- b) sentence
- c) remove stop words & punctuation and list the words.

Note: Take the input as “What is Web Mining? Web Mining is the process of “Data Mining” techniques, and extract information from Web documents and services. The main purpose of web mining is discovering useful information from the World-Wide Web and it’s usage patterns.”

Ans 1.

a) Word

HANDWRITTEN CODE:



```
a) VIBHU KUMAR SINGH 19BCE0215

from nltk.tokenize import word_tokenize

text = "What is web mining? web..... "
word_tokens = word_tokenize(text)

print("Tokenized words are : \n", word_tokens)
```

CODE:

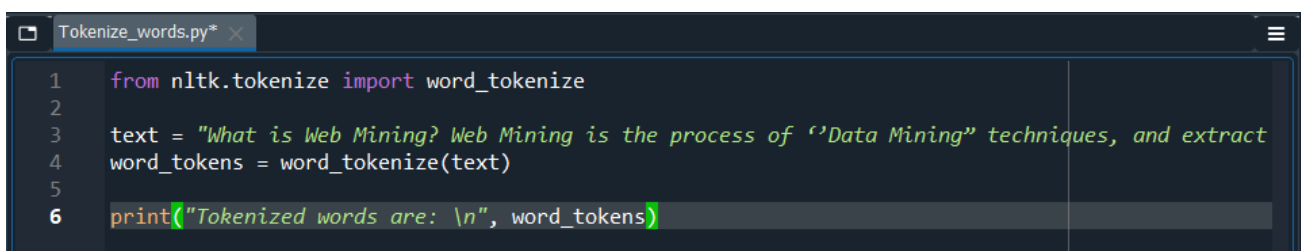
```
from nltk.tokenize import word_tokenize

text = "What is Web Mining? Web Mining is the process of “Data Mining” techniques,
and extract information from Web documents and services. The main purpose of web
mining is discovering useful information from the World-Wide Web and it’s usage
patterns"

word_tokens = word_tokenize(text)

print("Tokenized words are:\n", word_tokens)
```

CODE SCREENSHOT:



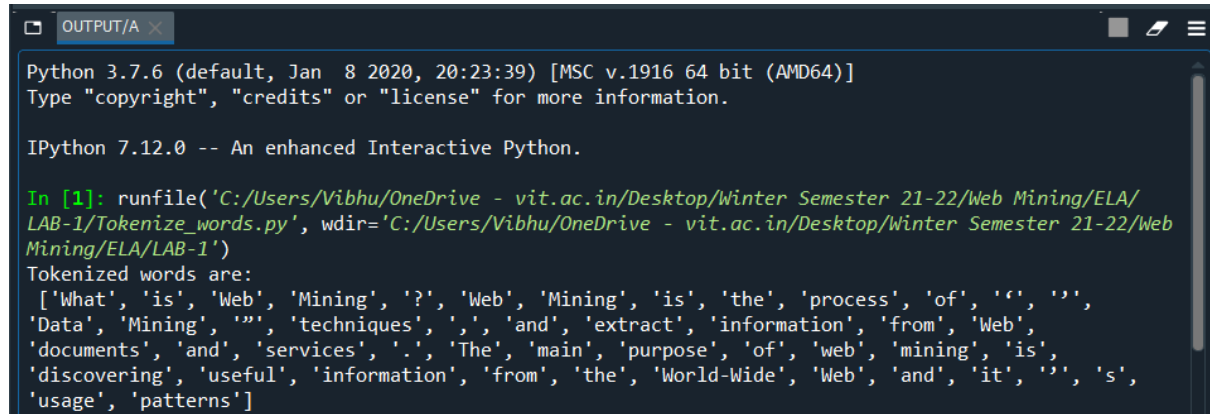
```
Tokenize_words.py x
1 from nltk.tokenize import word_tokenize
2
3 text = "What is Web Mining? Web Mining is the process of “Data Mining” techniques, and extract
4 word_tokens = word_tokenize(text)
5
6 print("Tokenized words are: \n", word_tokens)
```

OUTPUT:

Tokenized words are:

```
['What', 'is', 'Web', 'Mining', '?', 'Web', 'Mining', 'is', 'the', 'process', 'of', '"', '"', 'Data', 'Mining', '"', 'techniques', ',', 'and', 'extract', 'information', 'from', 'Web', 'documents', 'and', 'services', '.', 'The', 'main', 'purpose', 'of', 'web', 'mining', 'is', 'discovering', 'useful', 'information', 'from', 'the', 'World-Wide', 'Web', 'and', 'it', '"', 's', 'usage', 'patterns']
```

OUTPUT SCREENSHOT:



```
Python 3.7.6 (default, Jan 8 2020, 20:23:39) [MSC v.1916 64 bit (AMD64)]
Type "copyright", "credits" or "license" for more information.

IPython 7.12.0 -- An enhanced Interactive Python.

In [1]: runfile('C:/Users/Vibhu/OneDrive - vit.ac.in/Desktop/Winter Semester 21-22/Web Mining/ELA/LAB-1/Tokenize_words.py', wdir='C:/Users/Vibhu/OneDrive - vit.ac.in/Desktop/Winter Semester 21-22/Web Mining/ELA/LAB-1')
Tokenized words are:
['What', 'is', 'Web', 'Mining', '?', 'Web', 'Mining', 'is', 'the', 'process', 'of', '"', '"', 'Data', 'Mining', '"', 'techniques', ',', 'and', 'extract', 'information', 'from', 'Web', 'documents', 'and', 'services', '.', 'The', 'main', 'purpose', 'of', 'web', 'mining', 'is', 'discovering', 'useful', 'information', 'from', 'the', 'World-Wide', 'Web', 'and', 'it', '"', 's', 'usage', 'patterns']
```

b) Sentence

HANDWRITTEN CODE:

```
b) VIBHU KUMAR SINGH 19BCE0215

from nltk.tokenize import sent_tokenize

text = "What is web mining? Web.... "
sent_tokens = sent_tokenize(text)

print("Tokenized sentences are : \n", sent_tokens)
```

CODE:

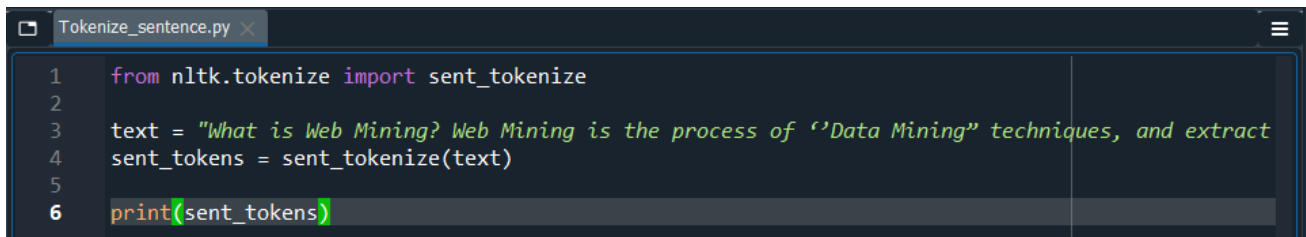
```
from nltk.tokenize import sent_tokenize
```

```
text = "What is Web Mining? Web Mining is the process of 'Data Mining' techniques, and extract  
information from Web documents and services. The main purpose of web mining is discovering  
useful information from the World-Wide Web and it's usage patterns"
```

```
sent_tokens = sent_tokenize(text)
```

```
print(sent_tokens)
```

CODE SCREENSHOT:

A screenshot of a code editor window titled 'Tokenize_sentence.py'. The code is as follows:

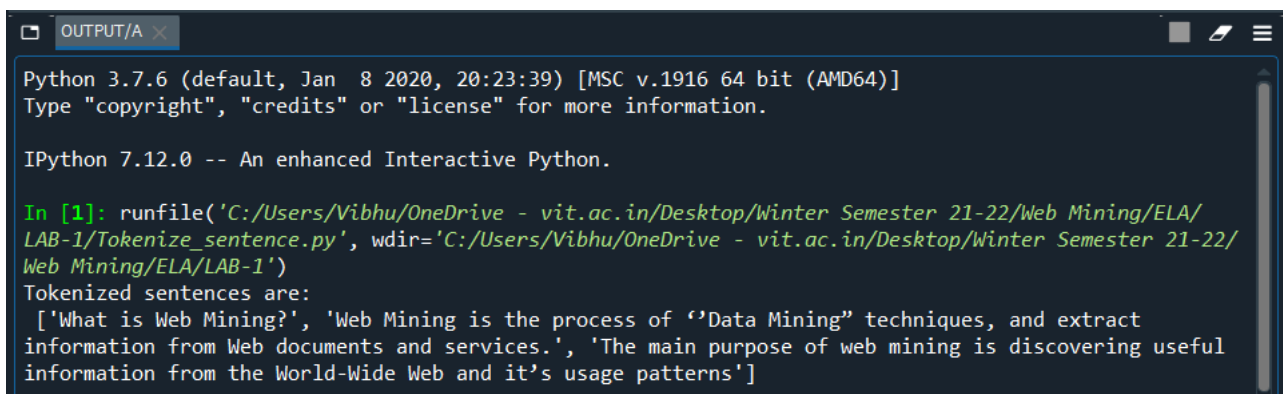
```
1 from nltk.tokenize import sent_tokenize
2
3 text = "What is Web Mining? Web Mining is the process of 'Data Mining' techniques, and extract
4 sent_tokens = sent_tokenize(text)
5
6 print(sent_tokens)
```

OUTPUT:

Tokenized sentences are:

```
['What is Web Mining?', 'Web Mining is the process of "Data Mining" techniques, and extract  
information from Web documents and services.', 'The main purpose of web mining is discovering useful  
information from the World-Wide Web and it's usage patterns']
```

OUTPUT SCREENSHOT:

A screenshot of a Python IPython shell window titled 'OUTPUT/A'. The output shows the execution of the script and the resulting tokenized sentences:

```
Python 3.7.6 (default, Jan 8 2020, 20:23:39) [MSC v.1916 64 bit (AMD64)]
Type "copyright", "credits" or "license()" for more information.

IPython 7.12.0 -- An enhanced Interactive Python.

In [1]: runfile('C:/Users/Vibhu/OneDrive - vit.ac.in/Desktop/Winter Semester 21-22/Web Mining/ELA/
LAB-1/Tokenize_sentence.py', wdir='C:/Users/Vibhu/OneDrive - vit.ac.in/Desktop/Winter Semester 21-22/
Web Mining/ELA/LAB-1')
Tokenized sentences are:
['What is Web Mining?', 'Web Mining is the process of "Data Mining" techniques, and extract
information from Web documents and services.', 'The main purpose of web mining is discovering useful
information from the World-Wide Web and it's usage patterns']
```

c) Remove stop words & punctuation and list the words

HANDWRITTEN CODE:

c) VIBHU KUMAR SINGH 19BCE0215

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

```
text = "What is web mining? web ....."
```

```
punctuations = '@#!%^&*()_+-= '~? \ ' : \ | " / . , < > ; [ ] '  
without_punctuation = ""
```

```
for char in text:
```

```
    if char not in punctuations:
```

```
        without_punctuation += char
```

```
stop_words = set(stopwords.words('english'))
```

```
word_tokens = word_tokenize(without_punctuation)
```

```
filtered_words = [w for w in word_tokens if not  
                    w.lower() in stop_words]
```

```
filtered_words = []
```

```
for w in word_tokens:
```

```
    if w not in stop_words:
```

```
        filtered_words.append(w)
```

```
print("words without stop-words and punctuation  
are: \n", filtered_words)
```

CODE:

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

text = "What is Web Mining? Web Mining is the process of 'Data Mining' techniques,
and extract information from Web documents and services. The main purpose of web
mining is discovering useful information from the World-Wide Web and it's usage
patterns"

punctuations = '!@#$$%^&*()_+=-`~?\'":\|"/.,<>[ ]'
without_punctuation = ""

for char in text:
    if char not in punctuations:
        without_punctuation+=char

stop_words = set(stopwords.words('english'))

word_tokens = word_tokenize(without_punctuation)

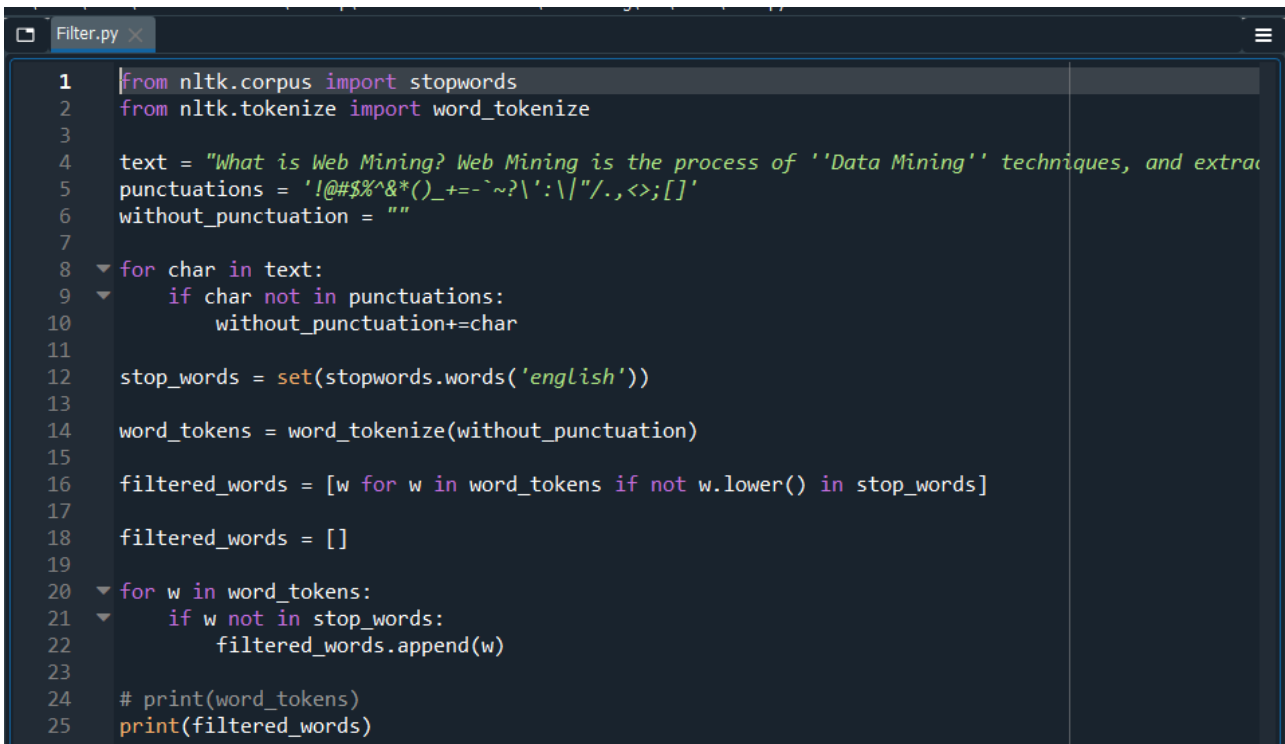
filtered_words = [w for w in word_tokens if not w.lower() in stop_words]

filtered_words = []

for w in word_tokens:
    if w not in stop_words:
        filtered_words.append(w)

print(filtered_words)
```

CODE SCREENSHOT:

A screenshot of a code editor window titled 'Filter.py'. The code is written in Python and uses the NLTK library to filter stopwords from a text document. The code is as follows:

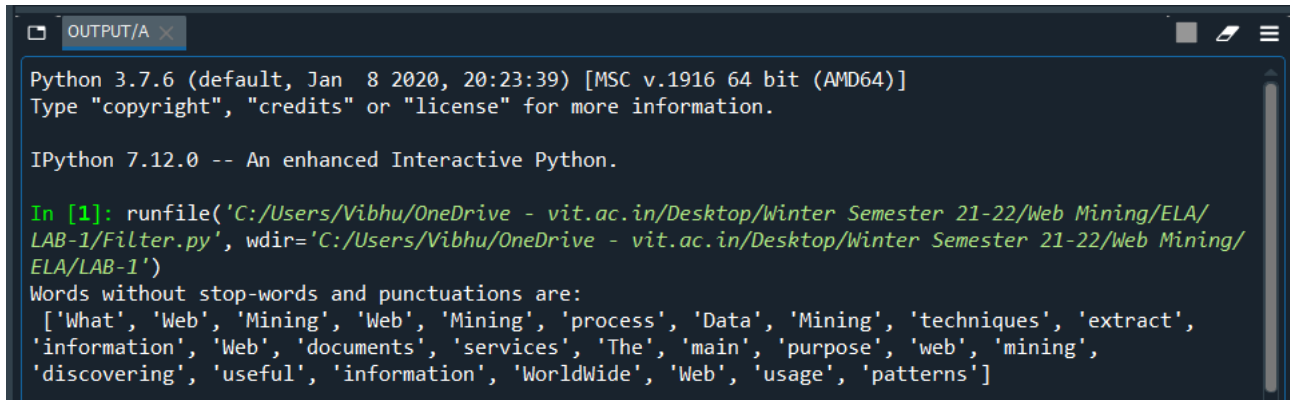
```
1 from nltk.corpus import stopwords
2 from nltk.tokenize import word_tokenize
3
4 text = "What is Web Mining? Web Mining is the process of 'Data Mining' techniques, and extra
5 punctuations = '!@#$$%^&*()_+=-`~?\'":\|"/.,<>[ ]'
6 without_punctuation = ""
7
8 for char in text:
9     if char not in punctuations:
10         without_punctuation+=char
11
12 stop_words = set(stopwords.words('english'))
13
14 word_tokens = word_tokenize(without_punctuation)
15
16 filtered_words = [w for w in word_tokens if not w.lower() in stop_words]
17
18 filtered_words = []
19
20 for w in word_tokens:
21     if w not in stop_words:
22         filtered_words.append(w)
23
24 # print(word_tokens)
25 print(filtered_words)
```


OUTPUT:

Words without stop-words and punctuations are:

['What', 'Web', 'Mining', 'Web', 'Mining', 'process', 'Data', 'Mining', 'techniques', 'extract', 'information', 'Web', 'documents', 'services', 'The', 'main', 'purpose', 'web', 'mining', 'discovering', 'useful', 'information', 'WorldWide', 'Web', 'usage', 'patterns']

OUTPUT SCREENSHOT:

A screenshot of a Python IPython terminal window. The window title is "OUTPUT/A". The terminal shows the Python version (3.7.6) and the IPython version (7.12.0). The user has run a script named "Filter.py" located at "C:/Users/Vibhu/OneDrive - vit.ac.in/Desktop/Winter Semester 21-22/Web Mining/ELA/LAB-1/Filter.py". The output of the script is displayed in the terminal, showing a list of words without stop-words and punctuations: ['What', 'Web', 'Mining', 'Web', 'Mining', 'process', 'Data', 'Mining', 'techniques', 'extract', 'information', 'Web', 'documents', 'services', 'The', 'main', 'purpose', 'web', 'mining', 'discovering', 'useful', 'information', 'WorldWide', 'Web', 'usage', 'patterns'].

```
Python 3.7.6 (default, Jan 8 2020, 20:23:39) [MSC v.1916 64 bit (AMD64)]
Type "copyright", "credits" or "license" for more information.

IPython 7.12.0 -- An enhanced Interactive Python.

In [1]: runfile('C:/Users/Vibhu/OneDrive - vit.ac.in/Desktop/Winter Semester 21-22/Web Mining/ELA/
LAB-1/Filter.py', wdir='C:/Users/Vibhu/OneDrive - vit.ac.in/Desktop/Winter Semester 21-22/Web Mining/
ELA/LAB-1')
Words without stop-words and punctuations are:
['What', 'Web', 'Mining', 'Web', 'Mining', 'process', 'Data', 'Mining', 'techniques', 'extract',
'information', 'Web', 'documents', 'services', 'The', 'main', 'purpose', 'web', 'mining',
'discovering', 'useful', 'information', 'WorldWide', 'Web', 'usage', 'patterns']
```