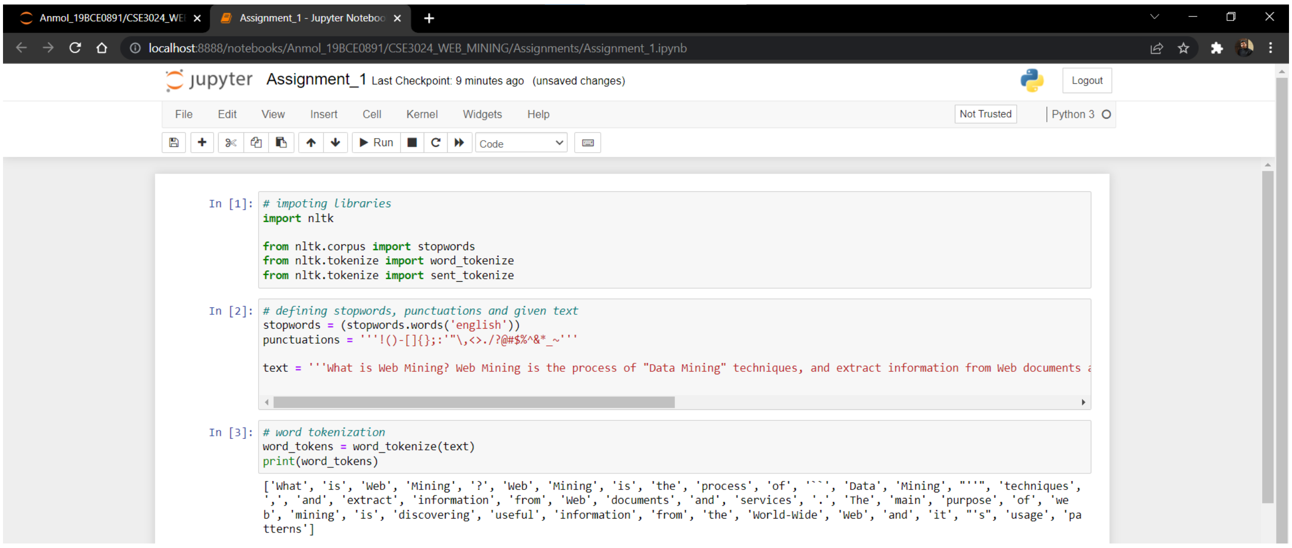
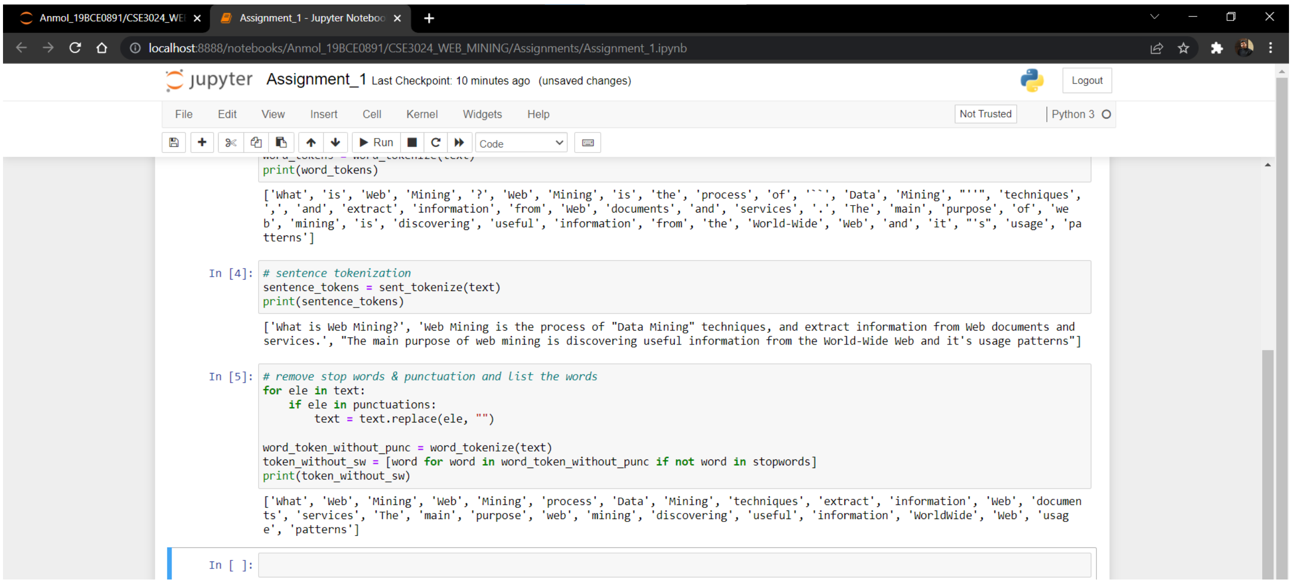
# CSE3024 - WEB MINING (L41 + L42)

**NAME – ANMOL**
**REG. NO. - 19BCE0891**

## DIGITAL ASSIGNMENT – 1


Anmol_19BCE0891/CSE3024_WE
Assignment_1 - Jupyter Noteboo
localhost:8888/notebooks/Anmol_19BCE0891/CSE3024_WEB_MINING/Assignments/Assignment_1.ipynb
jupyter Assignment_1 Last Checkpoint: 9 minutes ago (unsaved changes)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3
Run
Code

```
In [1]: # impoting libraries
        import nltk

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

In [2]: # defining stopwords, punctuations and given text
        stopwords = (stopwords.words('english'))
        punctuations = '''!()-[]{};:'"\,<>./?@#$%^&*_~'''

        text = '''What is Web Mining? Web Mining is the process of "Data Mining" techniques, and extract information from Web documents a

In [3]: # word tokenization
        word_tokens = word_tokenize(text)
        print(word_tokens)

        ['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', 'we
        b', 'mining', 'is', 'discovering', 'useful', 'information', 'from', 'the', 'World-Wide', 'Web', 'and', 'it', "'s", 'usage', 'pa
        tterns']
```


Anmol_19BCE0891/CSE3024_WE
Assignment_1 - Jupyter Noteboo
localhost:8888/notebooks/Anmol_19BCE0891/CSE3024_WEB_MINING/Assignments/Assignment_1.ipynb
jupyter Assignment_1 Last Checkpoint: 10 minutes ago (unsaved changes)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3
Run
Code

```
        print(word_tokens)

        ['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', 'we
        b', 'mining', 'is', 'discovering', 'useful', 'information', 'from', 'the', 'World-Wide', 'Web', 'and', 'it', "'s", 'usage', 'pa
        tterns']

In [4]: # sentence tokenization
        sentence_tokens = sent_tokenize(text)
        print(sentence_tokens)

        ['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"]

In [5]: # remove stop words & punctuation and list the words
        for ele in text:
            if ele in punctuations:
                text = text.replace(ele, "")

        word_token_without_punc = word_tokenize(text)
        token_without_sw = [word for word in word_token_without_punc if not word in stopwords]
        print(token_without_sw)

        ['What', 'Web', 'Mining', 'Web', 'Mining', 'process', 'Data', 'Mining', 'techniques', 'extract', 'information', 'Web', 'documen
        ts', 'services', 'The', 'main', 'purpose', 'web', 'mining', 'discovering', 'useful', 'information', 'WorldWide', 'Web', 'usag
        e', 'patterns']

In [ ]:
```

**CODE -**

```
# impoting libraries
import nltk

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


# defining stopwords, punctuations and given text
stopwords = (stopwords.words('english'))
punctuations = '''!()-[]{};:'"\,<>./?@#$%^&*_~'''

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 tokenization
word_tokens = word_tokenize(text)
print(word_tokens)


# sentence tokenization
sentence_tokens = sent_tokenize(text)
print(sentence_tokens)


# remove stop words & punctuation and list the words
for ele in text:
    if ele in punctuations:
        text = text.replace(ele, "")

word_token_without_punc = word_tokenize(text)
token_without_sw = [word for word in word_token_without_punc if not word in stopwords]
print(token_without_sw)
```