Lab -1

Name: Adithi Shinde

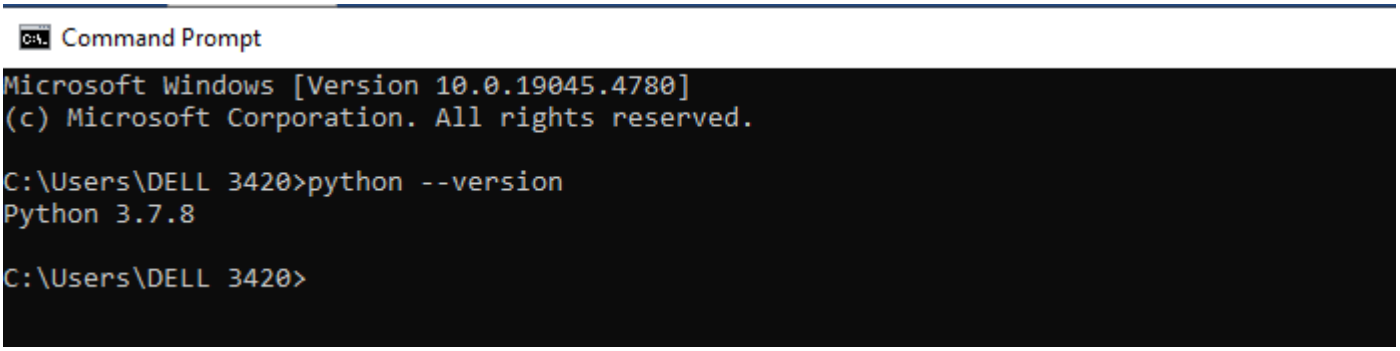
HT no: 2203A54032

Stream: CSE-Data Science

Batch:40

---

1. Download and install Python from the official website:python.org



2. LoadingTextDatasetsfromDifferentResources

(i) LoadingaTextFile

.TextDataset

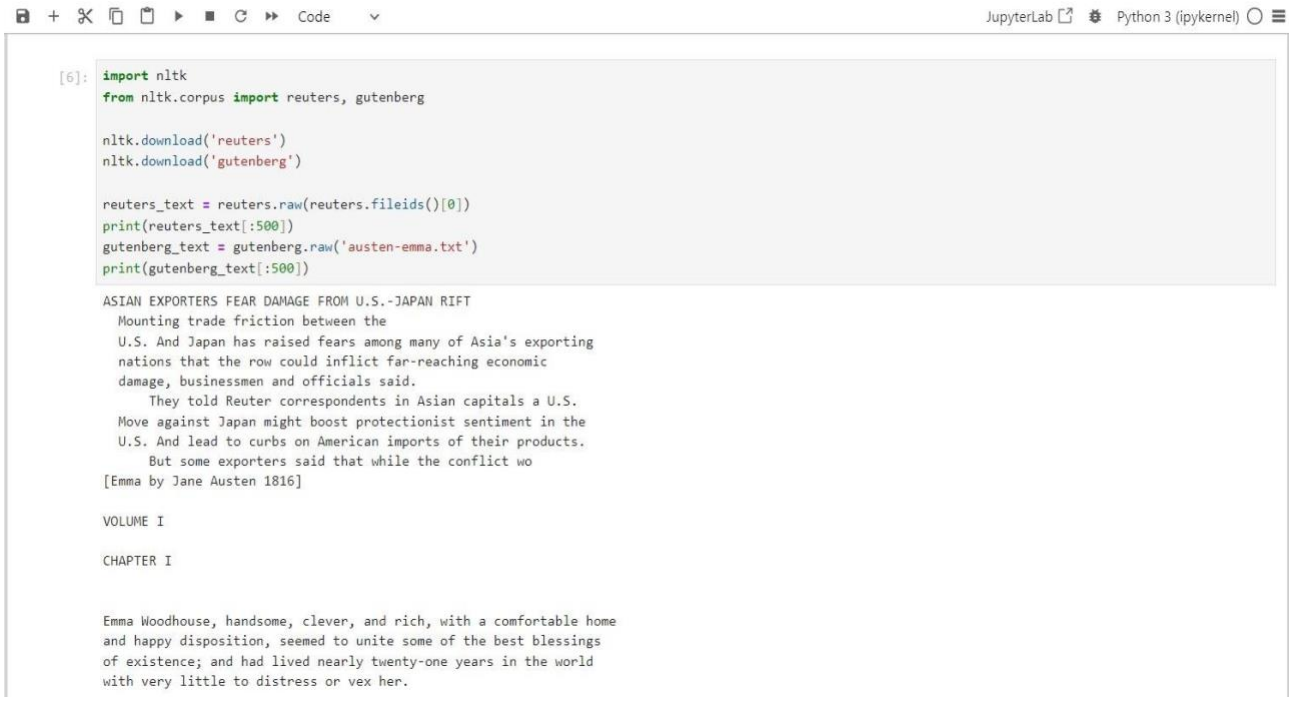
(ii) LoadingTextDatafromaCSVFile

.CSVfileexampleusingpandaslibrary

```
import pandas as pd
d=pd.read_csv('Bigfoot.csv')
print(d.head())
```

```
  Report Type    Class  Year Season      Month       State
0      Report  Class A  2021   Fall    October  Washington
1      Report  Class B  2021   Fall    October        Utah
2      Report  Class B  2021   Fall  September       Texas
3      Report  Class B  2021   Fall  September       Texas
4      Report  Class B  2021   Fall    October    Oklahoma
```

(iii) LoadingTextDatafromanOnlineSource



```
[6]: import nltk
from nltk.corpus import reuters, gutenberg

nltk.download('reuters')
nltk.download('gutenberg')

reuters_text = reuters.raw(reuters.fileids()[0])
print(reuters_text[:500])
gutenberg_text = gutenberg.raw('austen-emma.txt')
print(gutenberg_text[:500])
```

```
ASIAN EXPORTERS FEAR DAMAGE FROM U.S.-JAPAN RIFT
  Mounting trade friction between the
  U.S. And Japan has raised fears among many of Asia's exporting
  nations that the row could inflict far-reaching economic
  damage, businessmen and officials said.
      They told Reuter correspondents in Asian capitals a U.S.
  Move against Japan might boost protectionist sentiment in the
  U.S. And lead to curbs on American imports of their products.
      But some exporters said that while the conflict wo
[Emma by Jane Austen 1816]

VOLUME I

CHAPTER I


Emma Woodhouse, handsome, clever, and rich, with a comfortable home
and happy disposition, seemed to unite some of the best blessings
of existence; and had lived nearly twenty-one years in the world
with very little to distress or vex her.
```

(iv) LoadingBuilt-inTextDatasetswithNLTK

```
[7]: import requests
from bs4 import BeautifulSoup

url = 'https://www.flipkart.com/'
response = requests.get(url)
html_content = response.text
soup = BeautifulSoup(html_content, 'html.parser')
text = soup.get_text()
print("Extracted Text:\n", text[:500])
```

```
Extracted Text:


Online Shopping India Mobile, Cameras, Lifestyle & more Online @ Flipkart.com
```

(v) LoadingTextDataUsingHuggingFaceDatasets

```
[55]: from datasets import load_dataset

dataset = load_dataset('ag_news', split='train')
print(dataset[0])
```

```
{'text': "Wall St. Bears Claw Back Into the Black (Reuters) Reuters - Short-sellers, Wall Street's dwindling\\band of ultra-cynics, are seeing green agai
n.", 'label': 2}
```

3. Take your own text or take text as “The bank can guarantee deposits will eventuallycoverfuturetuitioncostsbecauseitinvestsinadjustable-ratemortgage securities." Implement Ambiguity Removal in the text.

```
[62]: import pandas as pd

data = {
    'Text': ["The bank can guarantee deposits will eventually cover future tuition costs because it involves minimal risk."]
}

df = pd.DataFrame(data)
print("Original dataset:")
print(df.head())

print("\nMissing values in each column:")
print(df.isnull().sum())

df_cleaned = df.copy()
print("\nNumber of duplicate rows:", df_cleaned.duplicated().sum())

df_cleaned = df_cleaned.drop_duplicates()
print("\nData after removing duplicates:")
print(df_cleaned.head())

df_cleaned['Text'] = df_cleaned['Text'].str.lower()
print("\nCleaned data after standardization:")
print(df_cleaned.head())

cleaned_file_path = 'cleaned_sample_dataset.csv'
df_cleaned.to_csv(cleaned_file_path, index=False)
```

```
Original dataset:
                                                Text
0  The bank can guarantee deposits will eventuall...

Missing values in each column:
Text    0
dtype: int64

Number of duplicate rows: 0

Data after removing duplicates:
                                                Text
0  The bank can guarantee deposits will eventuall...

Cleaned data after standardization:
                                                Text
0  the bank can guarantee deposits will eventuall...
```

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Text												
2	the bank can guarantee deposits will eventually cover future tuition costs because it invests in adjustable-rate mortgage securities.												
3													
4													
5													

4. Take your own text or take text as “Hellothere! How areyoudoingtoday?NLPis fascinating." Implement Sentence Segmentation in the text.

Code JupyterLab Python 3 (ipykernel)

[70]:
```
import nltk
from nltk.tokenize import word_tokenize, sent_tokenize
import pandas as pd

data = {
    'Text': ["Hello there! How are you doing today? NLP is fascinating."]
}

df = pd.DataFrame(data)
print(df.head())

df['Sentences'] = df['Text'].apply(sent_tokenize)
print(df[['Text', 'Sentences']].head())

df['Words'] = df['Text'].apply(word_tokenize)
print(df[['Text', 'Words']].head())

segmented_file_path = 'segmented_text_dataset.csv'
df.to_csv(segmented_file_path, index=False)
```

```
                                                Text
0  Hello there! How are you doing today? NLP is f...
                                                Text  \
0  Hello there! How are you doing today? NLP is f...

                                           Sentences
0  [Hello there!, How are you doing today?, NLP i...
                                                Text  \
0  Hello there! How are you doing today? NLP is f...

                                               Words
0  [Hello, there, !, How, are, you, doing, today,...
```

Text	Sentences	Words
Hello there! How are you doing today? NLP is fascinating.	"['Hello there!', 'How are you doing today?', 'NLP is fascinating.']"	"['Hello', 'there', '!', 'How', 'are', 'you', 'doing', 'today', '?', 'NLP', 'is', 'fascinating', '.']"