Assignment 2

Part I

In [ ]:
```
import pandas as pd

path = 'E:/Noman_Ali/Learning/GenerativeAI/Class3/matches.csv'

df = pd.read_csv(path)
```

In [ ]:
```
df.head(10)
```

Out[ ]:

	id	season	city	date	team1	team2	toss_winner	toss_decision	result	dl_applied	winner	win_by_ru
0	1	2017	Hyderabad	2017-04-05	Sunrisers Hyderabad	Royal Challengers Bangalore	Royal Challengers Bangalore	field	normal	0	Sunrisers Hyderabad	
1	2	2017	Pune	2017-04-06	Mumbai Indians	Rising Pune Supergiant	Rising Pune Supergiant	field	normal	0	Rising Pune Supergiant	
2	3	2017	Rajkot	2017-04-07	Gujarat Lions	Kolkata Knight Riders	Kolkata Knight Riders	field	normal	0	Kolkata Knight Riders	
3	4	2017	Indore	2017-04-08	Rising Pune Supergiant	Kings XI Punjab	Kings XI Punjab	field	normal	0	Kings XI Punjab	
4	5	2017	Bangalore	2017-04-08	Royal Challengers Bangalore	Delhi Daredevils	Royal Challengers Bangalore	bat	normal	0	Royal Challengers Bangalore	
5	6	2017	Hyderabad	2017-04-09	Gujarat Lions	Sunrisers Hyderabad	Sunrisers Hyderabad	field	normal	0	Sunrisers Hyderabad	
6	7	2017	Mumbai	2017-04-09	Kolkata Knight Riders	Mumbai Indians	Mumbai Indians	field	normal	0	Mumbai Indians	
7	8	2017	Indore	2017-04-10	Royal Challengers Bangalore	Kings XI Punjab	Royal Challengers Bangalore	bat	normal	0	Kings XI Punjab	
8	9	2017	Pune	2017-04-11	Delhi Daredevils	Rising Pune Supergiant	Rising Pune Supergiant	field	normal	0	Delhi Daredevils	

	id	season	city	date	team1	team2	toss_winner	toss_decision	result	dl_applied	winner	win_by_ru
9	10	2017	Mumbai	2017-04-12	Sunrisers Hyderabad	Mumbai Indians	Mumbai Indians	field	normal	0	Mumbai Indians	

Part II

In [ ]:
```
import nltk
from nltk.corpus import gutenberg
from nltk.tokenize import word_tokenize, sent_tokenize
from nltk.stem import PorterStemmer
from nltk.stem import WordNetLemmatizer
from nltk.corpus import stopwords
```

Load the Corpus

In [ ]:
```
# Download necessary NLTK data
nltk.download("gutenberg")
nltk.download("punkt")
nltk.download("wordnet")
nltk.download("stopwords")

# Load "Alice in Wonderland" text
alice_text = gutenberg.raw("carroll-alice.txt")
```

```
[nltk_data] Downloading package gutenberg to
[nltk_data]     C:\Users\ProBook\AppData\Roaming\nltk_data...
[nltk_data]   Unzipping corpora\gutenberg.zip.
[nltk_data] Downloading package punkt to
[nltk_data]     C:\Users\ProBook\AppData\Roaming\nltk_data...
[nltk_data]   Unzipping tokenizers\punkt.zip.
[nltk_data] Downloading package wordnet to
[nltk_data]     C:\Users\ProBook\AppData\Roaming\nltk_data...
[nltk_data] Downloading package stopwords to
[nltk_data]     C:\Users\ProBook\AppData\Roaming\nltk_data...
[nltk_data]   Unzipping corpora\stopwords.zip.
```

Tokenizing

In [ ]:
```python
# Sentence Tokenization
sentences = sent_tokenize(alice_text)
print("Number of sentences:", len(sentences))

# Word Tokenization
words = word_tokenize(alice_text)
print("Number of words:", len(words))
```

```
Number of sentences: 1625
Number of words: 33494
```

Steaming

In [ ]:
```python
# Initialize the Porter Stemmer
stemmer = PorterStemmer()

# Perform Stemming
stemmed_words = [stemmer.stem(word) for word in words]
print("First 10 stemmed words:", stemmed_words[:10])
```

```
First 10 stemmed words: ['[', 'alic', "'s", 'adventur', 'in', 'wonderland', 'by', 'lewi', 'carrol', '1865']
```

Lematiztion

In [ ]:
```python
# Initialize the WordNet Lemmatizer
lemmatizer = WordNetLemmatizer()

# Perform Lemmatization
lemmatized_words = [lemmatizer.lemmatize(word) for word in words]
print("First 10 lemmatized words:", lemmatized_words[:10])
```

```
First 10 lemmatized words: ['[', 'Alice', "'s", 'Adventures', 'in', 'Wonderland', 'by', 'Lewis', 'Carroll', '1865']
```

Stop Word Removal

In [ ]:
```python
# Load the list of stop words
stop_words = set(stopwords.words("english"))

# Perform Stop Word Removal
filtered_words = [word for word in lemmatized_words if word.lower() not in stop_words]
print("First 10 words after stop word removal:", filtered_words[:10])
```

```
First 10 words after stop word removal: ['[', 'Alice', "'s", 'Adventures', 'Wonderland', 'Lewis', 'Carroll', '1865',
']', 'CHAPTER']
```