1

November 3, 2023

1. Write a program for pre-processing of a text document such as stop word removal, stemming.

[2]:
```
pip install nltk
```

```
Requirement already satisfied: nltk in c:\users\sarthak\new_anaconda\lib\site-
packages (3.8.1)
Requirement already satisfied: click in c:\users\sarthak\new_anaconda\lib\site-
packages (from nltk) (8.0.4)
Requirement already satisfied: joblib in c:\users\sarthak\new_anaconda\lib\site-
packages (from nltk) (1.2.0)
Requirement already satisfied: regex>=2021.8.3 in
c:\users\sarthak\new_anaconda\lib\site-packages (from nltk) (2022.7.9)
Requirement already satisfied: tqdm in c:\users\sarthak\new_anaconda\lib\site-
packages (from nltk) (4.65.0)
Requirement already satisfied: colorama in
c:\users\sarthak\new_anaconda\lib\site-packages (from click->nltk) (0.4.6)
Note: you may need to restart the kernel to use updated packages.
```

[3]:
```
import nltk
```

[4]:
```
nltk.download('punkt')
```

```
[nltk_data] Downloading package punkt to
[nltk_data]     C:\Users\Sarthak\AppData\Roaming\nltk_data…
[nltk_data]   Unzipping tokenizers\punkt.zip.
```

[4]:
```
True
```

[5]:
```
import nltk
from nltk.corpus import stopwords
from nltk.stem import PorterStemmer
from nltk.tokenize import word_tokenize, sent_tokenize

# Sample random text (100 words)
random_text = """
Data processing encompasses a series of operations that convert raw data into␣
 ↪structured
and organized information. This process begins with data collection, where data␣
 ↪is gathered
```

```
from various sources such as sensors, databases, forms, or external systems. Once collected,
the data can be in various formats, including text, numbers, images, or multimedia.
The next step in data processing is data cleaning and validation. This involves identifying
and correcting errors, inconsistencies, and missing values in the data. Clean and accurate data
is essential for reliable analysis and decision-making. Data cleaning often involves techniques
like outlier detection and data imputation.
After data cleaning, data transformation is performed. This includes tasks like data normalization,
aggregation, and summarization. Normalization ensures that data is on a consistent scale, while
aggregation and summarization reduce data complexity by generating statistics or aggregating data into meaningful groups.
Data processing also includes data integration, where data from multiple sources is combined
into a unified dataset. Integration can be challenging due to differences in data structures and
formats. Techniques like data mapping and data warehousing are used to facilitate integration.
"""

# Tokenize the text into words
words = word_tokenize(random_text)

# Initialize the NLTK Porter Stemmer
stemmer = PorterStemmer()

# Get the English stop words
nltk.download('stopwords')
stop_words = set(stopwords.words("english"))

# Initialize a list to store the preprocessed words
preprocessed_words = []

# Perform text preprocessing
for word in words:
    # Remove punctuation and convert to lowercase
    word = word.lower()
    word = word.strip('.,?!-()[]{}"\'')

    # Check if the word is not a stop word
    if word not in stop_words:
```

```
        # Stem the word
        word = stemmer.stem(word)

        # Add the preprocessed word to the list
        preprocessed_words.append(word)

# Join the preprocessed words back into a text
preprocessed_text = " ".join(preprocessed_words)

# Print the original text and preprocessed text
print("Original Text:")
print(random_text)
print("\nPreprocessed Text:")
print(preprocessed_text)
```

```
Original Text:

Data processing encompasses a series of operations that convert raw data into
structured
and organized information. This process begins with data collection, where data
is gathered
from various sources such as sensors, databases, forms, or external systems.
Once collected,
the data can be in various formats, including text, numbers, images, or
multimedia.
The next step in data processing is data cleaning and validation. This involves
identifying
and correcting errors, inconsistencies, and missing values in the data. Clean
and accurate data
is essential for reliable analysis and decision-making. Data cleaning often
involves techniques
like outlier detection and data imputation.
After data cleaning, data transformation is performed. This includes tasks like
data normalization,
aggregation, and summarization. Normalization ensures that data is on a
consistent scale, while
aggregation and summarization reduce data complexity by generating statistics or
aggregating data into meaningful groups.
Data processing also includes data integration, where data from multiple sources
is combined
into a unified dataset. Integration can be challenging due to differences in
data structures and
formats. Techniques like data mapping and data warehousing are used to
facilitate integration.


Preprocessed Text:
```

```
data process encompass seri oper convert raw data structur organ inform  process
begin data collect  data gather variou sourc sensor  databas  form  extern
system  collect  data variou format  includ text  number  imag  multimedia  next
step data process data clean valid  involv identifi correct error  inconsist
miss valu data  clean accur data essenti reliabl analysi decision-mak  data
clean often involv techniqu like outlier detect data imput  data clean  data
transform perform  includ task like data normal  aggreg  summar  normal ensur
data consist scale  aggreg summar reduc data complex gener statist aggreg data
meaning group  data process also includ data integr  data multipl sourc combin
unifi dataset  integr challeng due differ data structur format  techniqu like
data map data wareh use facilit integr
```

```
[nltk_data] Downloading package stopwords to
[nltk_data]     C:\Users\Sarthak\AppData\Roaming\nltk_data…
[nltk_data]   Package stopwords is already up-to-date!
```

[ ]: