NATURAL LANGUAGE PROCESSING LAB ASSIGNMENT - 1

Course Code: CSE3015 Course Title: Natural Language Processing

Professor: Prof.V SRIKANTH REDDY Slot: L47+L48

Name: P. Sathwika Reg. No.: 21BCE8118

```
#1.Read the paragraph and obtain the frequency of words.from collections import Counter

paragraph = "It was the best of times,it was the worst of times, it was the age of wisdom, it was the age of foolishness"
words = paragraph.split()
word_frequency = Counter(words)
print(word_frequency)
```

```
Counter({'was': 4, 'the': 4, 'of': 4, 'it': 2, 'age': 2, 'It': 1, 'best': 1, 'times,it': 1, 'worst': 1, 'times,': 1, 'wisdom,': 1, 'foolishness': 1})
```

```
#2. Read the content from a web page and extract the tokens /expression/words/number.

!pip install requests
!pip install beautifulsoup4
import requests
from bs4 import BeautifulSoup
import re

url = "https://vitap.ac.in/"
response = requests.get(url)
html_content = response.text
soup = BeautifulSoup(html_content, 'html.parser')
tokens = re.findall(r'\b\w+\b', soup.get_text())
print(tokens)
```

```
Requirement already satisfied: requests in /usr/local/lib/python3.10/dist-packages (2.31.0)
Requirement already satisfied: charset-normalizer<4,>=2 in /usr/local/lib/python3.10/dist-packages (from requests) (3.3.2)
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.10/dist-packages (from requests) (3.6)
Requirement already satisfied: urllib3<3,>=1.21.1 in /usr/local/lib/python3.10/dist-packages (from requests) (2.0.7)
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.10/dist-packages (from requests) (2024.2.2)
Requirement already satisfied: beautifulsoup4 in /usr/local/lib/python3.10/dist-packages (4.12.3)
Requirement already satisfied: soupsieve>1.2 in /usr/local/lib/python3.10/dist-packages (from beautifulsoup4) (2.5)
['VIT', 'AP', 'Apply', 'Knowledge', 'Improve', 'Life', 'Menu', 'VIT', 'Campuses', 'VIT', 'Vellore', 'VIT', 'Chennai', 'VIT', 'Bhopal', 'VIT', 'Bangalore', 'Admissions', 'Overview', 'E
```

```
#3. Read only the word content from the webpage and display their frequency?

from collections import Counter
word_frequency_webpage = Counter(tokens)
print(word_frequency_webpage)
```

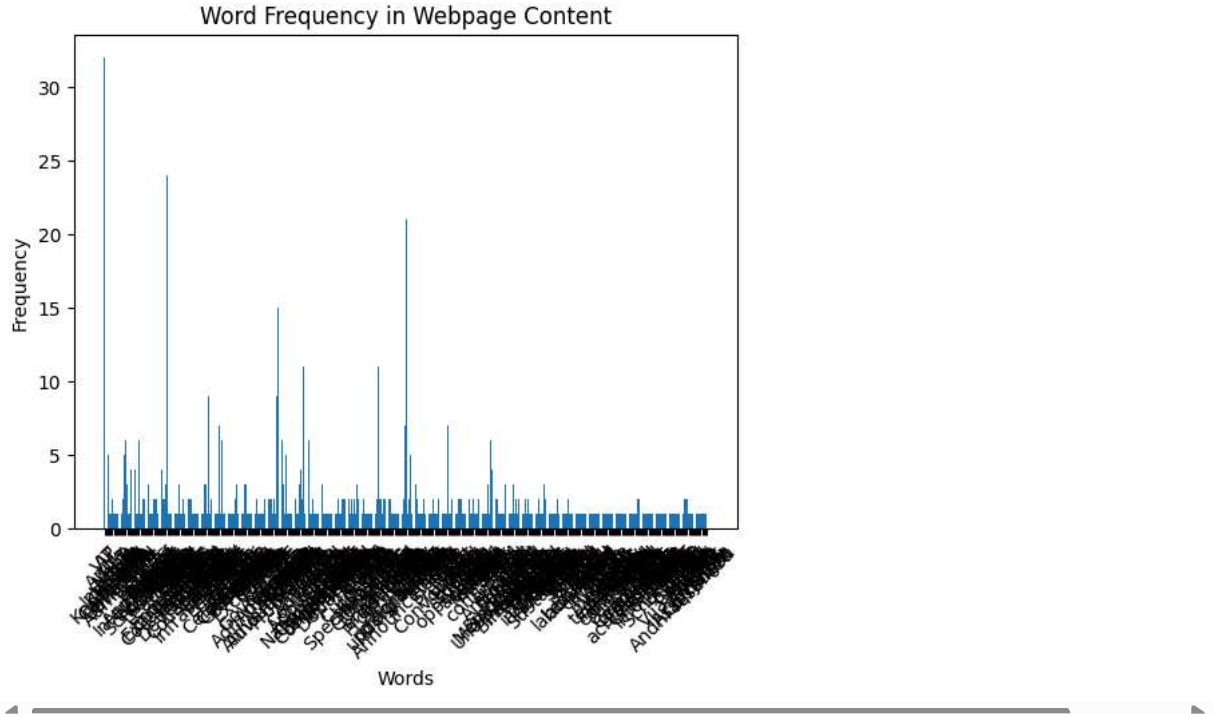
```
Counter({'VIT': 32, 'AP': 26, 'the': 26, 'of': 24, 'More': 21, 'for': 15, '2024': 11, 'University': 11, 'and': 9, 'Advertisement': 9, 'in': 8, 'to': 7, 'Learn': 7, 'View': 7, 'B': 6,
```

```
#4.Plot the frequency for count obtained in question 3.

!pip install matplotlib
```

```
import matplotlib.pyplot as plt
words, counts = zip(*word_frequency_webpage.items())
plt.bar(words, counts)
plt.xlabel('Words')
plt.ylabel('Frequency')
plt.title('Word Frequency in Webpage Content')
plt.xticks(rotation=45)
plt.show()
```

```
Requirement already satisfied: matplotlib in /usr/local/lib/python3.10/dist-packages (3.7.1)
Requirement already satisfied: contourpy>=1.0.1 in /usr/local/lib/python3.10/dist-packages (from matplotlib) (1.2.0)
Requirement already satisfied: cycler>=0.10 in /usr/local/lib/python3.10/dist-packages (from matplotlib) (0.12.1)
Requirement already satisfied: fonttools>=4.22.0 in /usr/local/lib/python3.10/dist-packages (from matplotlib) (4.49.0)
Requirement already satisfied: kiwisolver>=1.0.1 in /usr/local/lib/python3.10/dist-packages (from matplotlib) (1.4.5)
Requirement already satisfied: numpy>=1.20 in /usr/local/lib/python3.10/dist-packages (from matplotlib) (1.25.2)
Requirement already satisfied: packaging>=20.0 in /usr/local/lib/python3.10/dist-packages (from matplotlib) (23.2)
Requirement already satisfied: pillow>=6.2.0 in /usr/local/lib/python3.10/dist-packages (from matplotlib) (9.4.0)
Requirement already satisfied: pyparsing>=2.3.1 in /usr/local/lib/python3.10/dist-packages (from matplotlib) (3.1.1)
Requirement already satisfied: python-dateutil>=2.7 in /usr/local/lib/python3.10/dist-packages (from matplotlib) (2.8.2
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.10/dist-packages (from python-dateutil>=2.7->matplotl
```




```
import nltk
nltk.download('all')
```

```
[nltk_data]    |   Unzipping corpora/swadesh.zip.
[nltk_data]    | Downloading package switchboard to /root/nltk_data...
[nltk_data]    |   Unzipping corpora/switchboard.zip.
[nltk_data]    | Downloading package tagsets to /root/nltk_data...
[nltk_data]    |   Unzipping help/tagsets.zip.
[nltk_data]    | Downloading package timit to /root/nltk_data...
[nltk_data]    |   Unzipping corpora/timit.zip.
[nltk_data]    | Downloading package toolbox to /root/nltk_data...
[nltk_data]    |   Unzipping corpora/toolbox.zip.
[nltk_data]    | Downloading package treebank to /root/nltk_data...
[nltk_data]    |   Unzipping corpora/treebank.zip.
[nltk_data]    | Downloading package twitter_samples to
[nltk_data]    |     /root/nltk_data...
[nltk_data]    |   Unzipping corpora/twitter_samples.zip.
[nltk_data]    | Downloading package udhr to /root/nltk_data...
[nltk_data]    |   Unzipping corpora/udhr.zip.
[nltk_data]    | Downloading package udhr2 to /root/nltk_data...
[nltk_data]    |   Unzipping corpora/udhr2.zip.
[nltk_data]    | Downloading package unicode_samples to
[nltk_data]    |     /root/nltk_data...
[nltk_data]    |   Unzipping corpora/unicode_samples.zip.
[nltk_data]    | Downloading package universal_tagset to
[nltk_data]    |     /root/nltk_data...
[nltk_data]    |   Unzipping taggers/universal_tagset.zip.
[nltk_data]    | Downloading package universal_treebanks_v20 to
[nltk_data]    |     /root/nltk_data...
[nltk_data]    | Downloading package vader_lexicon to
[nltk_data]    |     /root/nltk_data...
[nltk_data]    | Downloading package verbnet to /root/nltk_data...
[nltk_data]    |   Unzipping corpora/verbnet.zip.
[nltk_data]    | Downloading package verbnet3 to /root/nltk_data...
[nltk_data]    |   Unzipping corpora/verbnet3.zip.
[nltk_data]    | Downloading package webtext to /root/nltk_data...
[nltk_data]    |   Unzipping corpora/webtext.zip.
[nltk_data]    | Downloading package wmt15_eval to /root/nltk_data...
[nltk_data]    |   Unzipping models/wmt15_eval.zip.
[nltk_data]    | Downloading package word2vec_sample to
[nltk_data]    |     /root/nltk_data...
[nltk_data]    |   Unzipping models/word2vec_sample.zip.
[nltk_data]    | Downloading package wordnet to /root/nltk_data...
[nltk_data]    | Downloading package wordnet2021 to /root/nltk_data...
[nltk_data]    | Downloading package wordnet2022 to /root/nltk_data...
[nltk_data]    |   Unzipping corpora/wordnet2022.zip.
[nltk_data]    | Downloading package wordnet31 to /root/nltk_data...
[nltk_data]    | Downloading package wordnet_ic to /root/nltk_data...
[nltk_data]    |   Unzipping corpora/wordnet_ic.zip.
[nltk_data]    | Downloading package words to /root/nltk_data...
[nltk_data]    |   Unzipping corpora/words.zip.
[nltk_data]    | Downloading package ycoe to /root/nltk_data...
[nltk_data]    |   Unzipping corpora/ycoe.zip.
[nltk_data]    |
[nltk_data]  Done downloading collection all
True
```

```
import nltk
nltk.download('punkt')
from nltk import word_tokenize, sent_tokenize
sent = "My name is sathwika. I am studying in VIT-AP"

print(sent_tokenize(sent))
print(word_tokenize(sent))
```

```
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Unzipping tokenizers/punkt.zip.
['My name is sathwika.', 'I am studying in VIT-AP']
['My', 'name', 'is', 'sathwika', '.', 'I', 'am', 'studying', 'in', 'VIT-AP']
```

```
from nltk import pos_tag
from nltk import word_tokenize

text = "NLP is a specialization subject in CSE Branch."
tokenized_text = word_tokenize(text)
tags = tokens_tag = pos_tag(tokenized_text)
tags
```

```
[('NLP', 'NNP'),
 ('is', 'VBZ'),
 ('a', 'DT'),
 ('specialization', 'NN'),
 ('subject', 'NN'),
 ('in', 'IN'),
 ('CSE', 'NNP'),
 ('Branch', 'NNP'),
 ('.', '.')]
```

```
from nltk.stem import WordNetLemmatizer

lemmatizer = WordNetLemmatizer()
print(lemmatizer.lemmatize("sits", 'v'))
print(lemmatizer.lemmatize("sat", 'v'))
print(lemmatizer.lemmatize("sit", 'v'))
print(lemmatizer.lemmatize("sitting", 'v'))
```

```
sit
sit
sit
sit
```

```
from nltk.stem import PorterStemmer

porter = PorterStemmer()
print(porter.stem("sit"))
print(porter.stem("sitting"))
print(porter.stem("sits"))
print(porter.stem("sat"))
```

```
sit
sit
sit
sat
```