Assignment 3 Set A2

March 23, 2024

1 Consider any text paragraph. Remove the stopwords. Tokenize the paragraph to extract words and sentences. Calculate the word frequency distribution and plot the frequencies. Plot the wordcloud of the text.

[1]:

```
import nltk
from nltk.tokenize import word_tokenize
from nltk.tokenize import sent_tokenize

text="""Hello all, Welcome to Python Programming Academy. Python Programming␣
 ↪Academy is a nice platform to learn
new programming skills. It is difficult to get enrolled in this Academy."""

#Tokenizing the paragraph to extract words and sentences
tokenized_text_data=sent_tokenize(text)
print("Tokenized Sentences : \n", tokenized_text_data, "\n")
tokenized_words=word_tokenize(text)
```

```
Tokenized Sentences :
 ['Hello all, Welcome to Python Programming Academy.', 'Python Programming
Academy is a nice platform to learn \nnew programming skills.', 'It is difficult
to get enrolled in this Academy.']
```

[2]:

```
#Removing Stopwords
from nltk.corpus import stopwords

stop_words_data=set(stopwords.words("english"))

filtered_words_list=[]
for words in tokenized_words:
    if words not in stop_words_data:
        filtered_words_list.append(words)
print("Tokenized Words : \n",tokenized_words,"\n")
print("Filtered Words : \n",filtered_words_list,"\n")
```

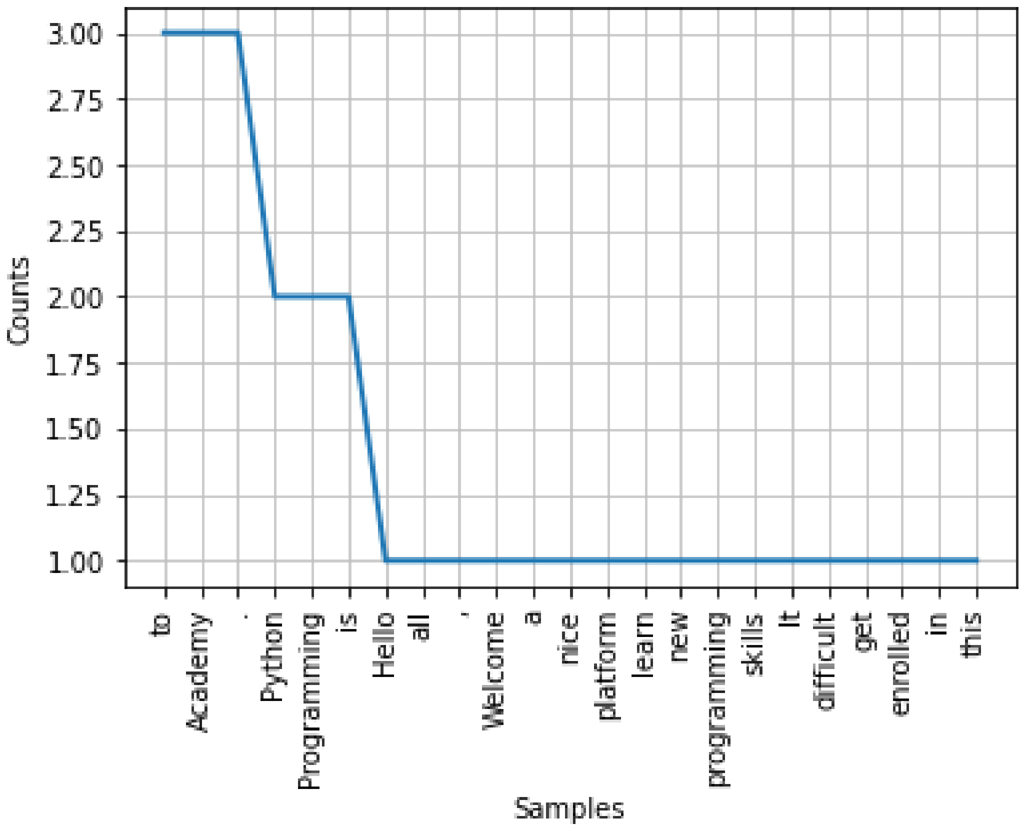
```
Tokenized Words :
 ['Hello', 'all', ',', 'Welcome', 'to', 'Python', 'Programming', 'Academy', '.',
'Python', 'Programming', 'Academy', 'is', 'a', 'nice', 'platform', 'to',
'learn', 'new', 'programming', 'skills', '.', 'It', 'is', 'difficult', 'to',
'get', 'enrolled', 'in', 'this', 'Academy', '.']

Filtered Words :
 ['Hello', ',', 'Welcome', 'Python', 'Programming', 'Academy', '.', 'Python',
'Programming', 'Academy', 'nice', 'platform', 'learn', 'new', 'programming',
'skills', '.', 'It', 'difficult', 'get', 'enrolled', 'Academy', '.']
```

[3]:
```python
#Word Frequency Distribution

from nltk.probability import FreqDist

frequency_distribution=FreqDist(tokenized_words)
print(frequency_distribution)
frequency_distribution
```

```
<FreqDist with 23 samples and 32 outcomes>
```

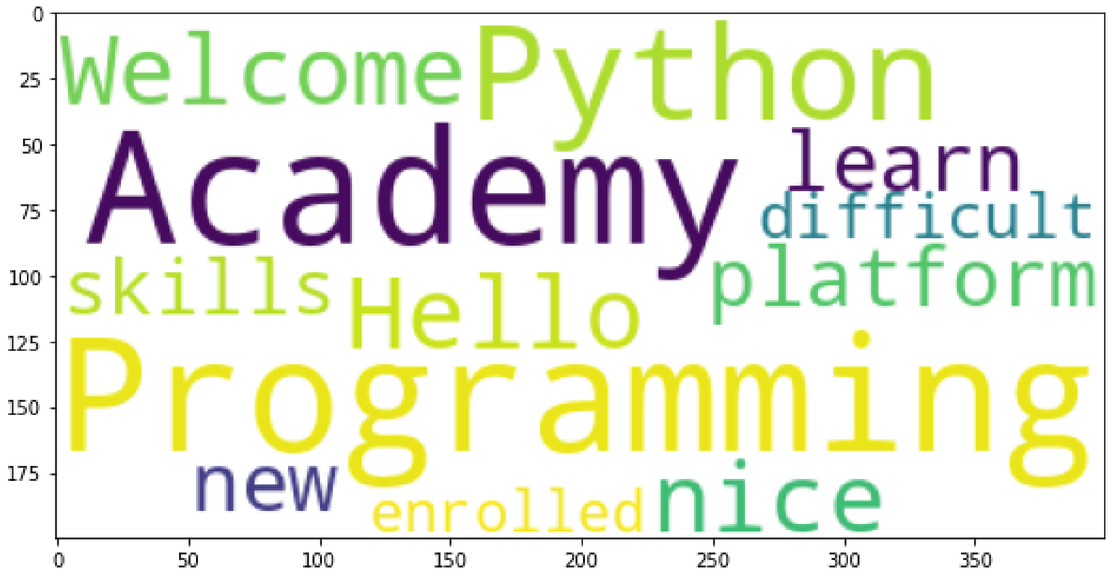
[3]:
```
FreqDist({'to': 3, 'Academy': 3, '.': 3, 'Python': 2, 'Programming': 2, 'is': 2,
'Hello': 1, 'all': 1, ',': 1, 'Welcome': 1, …})
```

[4]:
```python
#Plotting Frequencies
import matplotlib.pyplot as plt
frequency_distribution.plot(32,cumulative=False)
plt.show()
```



[5]:

```
#Wordcloud of the text
from wordcloud import WordCloud
import matplotlib.pyplot as plt
from PIL import Image, ImageFont
from wordcloud import WordCloud, STOPWORDS, ImageColorGenerator
#font = ImageFont.truetype("arial.ttf", 15)
wc=WordCloud(collocations = False, background_color = 'white').generate(text)
plt.figure(figsize=(10,10))
plt.imshow(wc)
plt.show()
```


Welcome
Python
Academy
learn
difficult
skills
Hello
platform
Programming
new
enrolled
nice
0
25
50
75
100
125
150
175
0
50
100
150
200
250
300
350

[ ]: