

CSE-3024 Web Mining

Lab Assignment 1

Alokam Nikhitha

19BCE2555

Question 1

Problem statement:

1. Write a python program to remove the stopwords for any given paragraph.

Create a set of stop words given below and print the output

```
stop_words = ['.',',','a','they','the','his','so','and','were','from','that','of','in','only','with','to']
```

Procedure:

- At First, we import the text file in our work space. To do this we can use open method of python which reads the file into our workspace.
- Next, we read each word into a variable as string using a nested for loop wherein we split each word whenever we encounter a space.
- Next, using regex in python we remove the punctuations from our string input. This will make sure that tokens are free from sentence structure.
- We tokenize each token in our text using split() function of NumPy lists and save it in a list.
- Then we use NumPy's unique method to only include unique tokens from our identified set of tokens.
- A tentative list of stop words and using a nested for loop we check if the given token belongs to that list or not, to remove stopwords. If it doesn't then we save it else we discard it.
- Finally, we print our list that contains the resultant tokens after removal of stop words.

Code:

#Reading input from a text file and saving it as a string

```
text = ""
```

```
with open('test_file.txt') as file:
```

```
    for line in file:
```

```
        for word in line.split():
```

```
            text= text + " " + word
```

#Removing punctuations from our input file

```
import re
```

```
text = re.sub(r'^\w\s]', "", text)
```

```
text
```

#Printing each token

```
print(text.split())
```

#Printing unique tokens

```
import numpy as np
```

```
print(np.unique(text.split()))
```

#Removing StopWords

```
stopwords = ["i", "a", "am", "and", "at", "for", "in", "is", "my", "of", "this"]
```

```
res = []
```

```
for x in tokens:
```

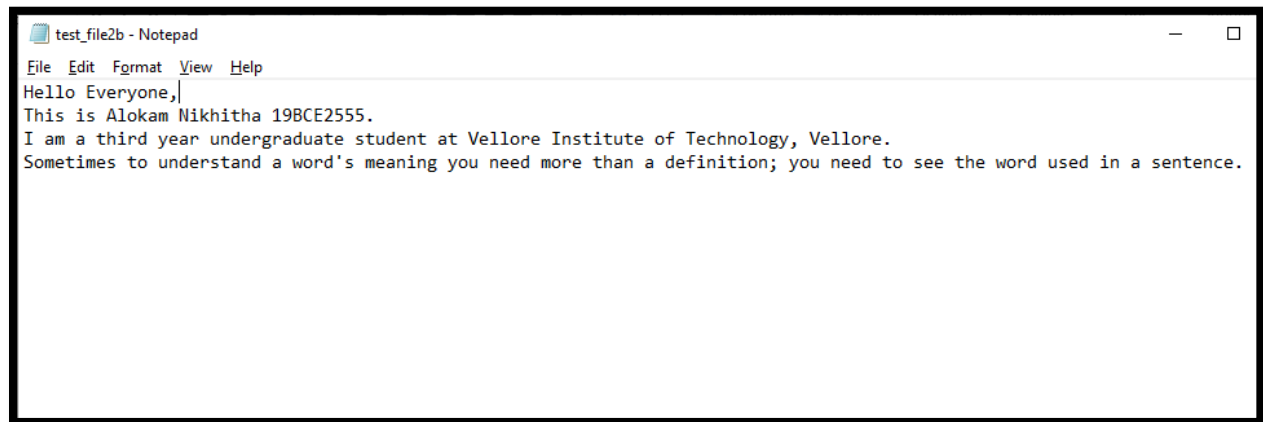
```
    if x not in stopwords:
```

```
        res.append(x)
```

#Printing cleaned tokens in our input text file

```
print(res)
```

Text File Taken as Input:



Code Snippets and Outputs:

```
In [3]: #Reading input from a text file and saving it as a string
text = ""
with open('test_file2b.txt') as file:
    for line in file:
        for word in line.split():
            text= text + " " + word.lower()
```

Here we are reading the text file using open method in python. Then reading each line we split each word and append it to a string variable with a space in between.

```
In [4]: #Removing punctuations from our input file
import re
text = re.sub(r'^\w\s]', '', text)
text
```

```
Out[4]: ' hello everyone this is alokam nikhitha 19bce2555 i am a third y
ear undergraduate student at vellore institute of technology vell
ore sometimes to understand a words meaning you need more than a
definition you need to see the word used in a sentence'
```

Here we are removing punctuations from our input file. This is done using regex, where we keep only alphanumeric inputs in our text string. We can see all the periods and commas from original input files are removed here.

```
In [5]: #Printing each token
tokens = text.split()
print(tokens)

['hello', 'everyone', 'this', 'is', 'alokam', 'nikhitha', '19bce2555', 'i', 'am', 'a', 'third', 'year', 'undergraduate', 'student', 'at', 'vellore', 'institute', 'of', 'technology', 'vellore', 'sometimes', 'to', 'understand', 'a', 'words', 'meaning', 'you', 'need', 'more', 'than', 'a', 'definition', 'you', 'need', 'to', 'see', 'the', 'word', 'used', 'in', 'a', 'sentence']
```

Next, we are splitting each word in our string using space character. Clearly, they form a token and hence we print each token.

```
In [6]: #Printing unique tokens
import numpy as np
tokens = np.unique(tokens)
print(tokens)

['19bce2555' 'a' 'alokam' 'am' 'at' 'definition' 'everyone' 'hello' 'i' 'in' 'institute' 'is' 'meaning' 'more' 'need' 'nikhitha' 'of' 'see' 'sentence' 'sometimes' 'student' 'technology' 'than' 'the' 'third' 'this' 'to' 'undergraduate' 'understand' 'used' 'vellore' 'word' 'words' 'year' 'you']
```

Here we are only printing unique tokens from all the generated tokens using split method. This is done using NumPy's unique function, which identifies all the unique elements from a list.

```
In [7]: #Removing StopWords
stopwords = ["i", "a", "am", "and", "at", "for", "in", "is", "my", "of", "this"]
res = []
for x in tokens:
    if x not in stopwords:
        res.append(x)

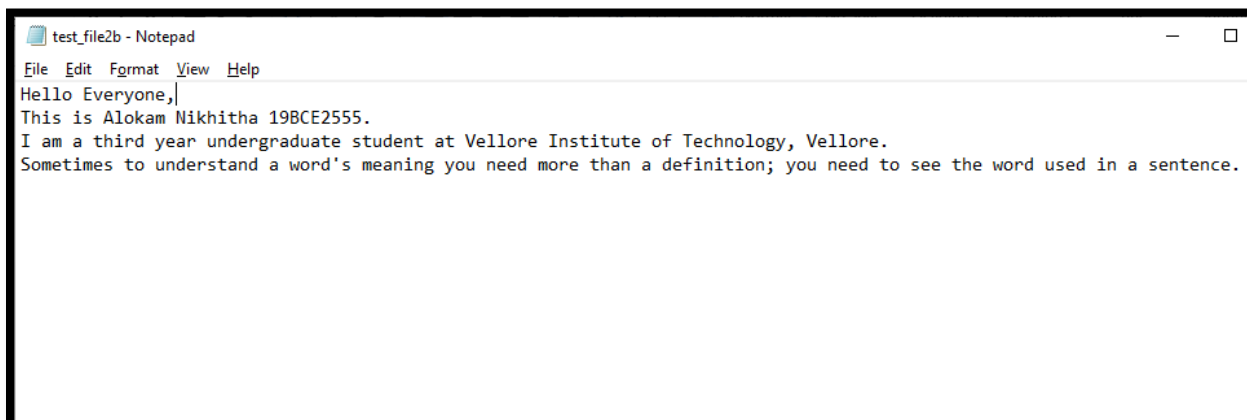
In [8]: #Printing cleaned tokens in our input text file
print(res)

['19bce2555', 'alokam', 'definition', 'everyone', 'hello', 'institute', 'meanin', 'more', 'need', 'nikhitha', 'see', 'sentence', 'sometimes', 'student', 'technology', 'than', 'the', 'third', 'to', 'undergraduate', 'understand', 'used', 'vellore', 'word', 'words', 'year', 'you']
```

Here we remove all the stop words from a self-defined list of stop words. We use nested loop to check if given token belongs to both tokens list and stopwords list. If it does, we don't add it to our result else we add it to our results.

Results and Output

- Input text:



test_file2b - Notepad

File Edit Format View Help

Hello Everyone,
This is Alokam Nikhitha 19BCE2555.
I am a third year undergraduate student at Vellore Institute of Technology, Vellore.
Sometimes to understand a word's meaning you need more than a definition; you need to see the word used in a sentence.

- Tokens of input text:

```
['19bce2555' 'a' 'alokam' 'am' 'at' 'definition' 'everyone' 'hello' 'i'  
'in' 'institute' 'is' 'meaning' 'more' 'need' 'nikhitha' 'of' 'see'  
'sentence' 'sometimes' 'student' 'technology' 'than' 'the' 'third' 'this'  
'to' 'undergraduate' 'understand' 'used' 'vellore' 'word' 'words' 'year'  
'you']
```

-Tokens after removal of stop words:

```
['19bce2555', 'alokam', 'definition', 'everyone', 'hello', 'institute', 'meanin  
g', 'more', 'need', 'nikhitha', 'see', 'sentence', 'sometimes', 'student', 'tec  
hnology', 'than', 'the', 'third', 'to', 'undergraduate', 'understand', 'used',  
'vellore', 'word', 'words', 'year', 'you']
```

Question 2

Problem statement:

2. Write a python program to tokenize
a) A sentence b) Multiple sentences (Without Nltk)

Procedure:

- Firstly, we import the text file in our work space. To do this we can use open method of python which reads the file into our workspace.
- Next, we read each word into a variable as string. This can be done using a nested for loop wherein we split each word whenever we encounter a space.
- Next, using regex in python we remove the punctuations from our string input. This will make sure that tokens are free from sentence structure.
- Finally, we will print each token and then unique tokens.

a)

Code:

```
#Reading input from a text file and saving it as a string
text = ""
with open('test_file2a.txt') as file:
    for line in file:
        for word in line.split():
            text= text + " " + word
```

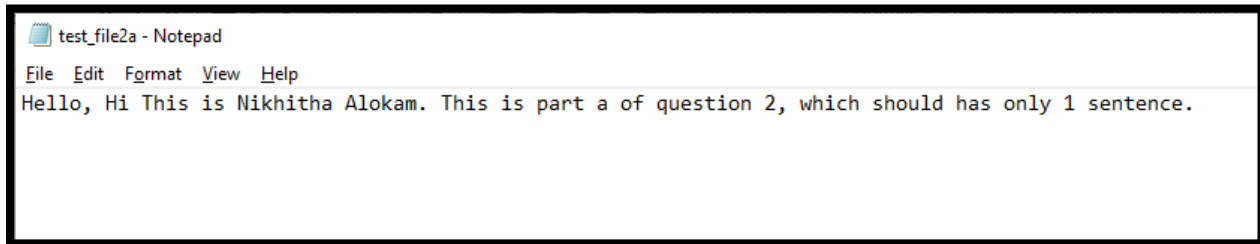


```
#Removing punctuations from our input file
import re
text = re.sub(r'^\w\s]', "", text)
text
```

```
#Printing each token
print(text.split())
```

```
#Printing unique tokens
import numpy as np
print(np.unique(text.split()))
```

Text File Taken as Input:



Code Snippets and Outputs:

```
In [1]: #Reading input from a text file and saving it as a string
text = ""
with open('test_file2a.txt') as file:
    for line in file:
        for word in line.split():
            text= text + " " + word
```

Here we are reading the text file using open method in python. Later we are reading each line we split each word and append it to a string variable with a space in between.

```
In [2]: #Removing punctuations from our input file
import re
text = re.sub(r'^\w\s', '', text)
text
```

```
Out[2]: ' Hello Hi This is Nikhitha Alokam This is part a of question 2 which should has only 1 sentence'
```

Here we are removing punctuations from our input file using regex, where we keep only alphanumeric inputs in our text string. We can see all the periods and commas from original input files are removed here.

```
In [3]: #Printing each token
print(text.split())
```

```
['Hello', 'Hi', 'This', 'is', 'Nikhitha', 'Alokam', 'This', 'is', 'part', 'a', 'of', 'question', '2', 'which', 'should', 'has', 'only', '1', 'sentence']
```

Next, we are splitting each word in our string using space character. Clearly, they form a token and hence we print each token.

```
In [4]: import numpy as np
print(np.unique(text.split()))
```

```
['1' '2' 'Alokam' 'Hello' 'Hi' 'Nikhitha' 'This' 'a' 'has' 'is' 'of' 'only' 'part' 'question' 'sentence' 'should' 'which']
```

Here we are only printing unique tokens from all the generated tokens using split method. This is done using NumPy's unique function, which identifies all the unique elements from a list.

Results and Output

- Input text:

Hello, Hi This is Nikhitha Alokam. This is part a of question 2, which should has only 1 sentence.

- Tokens of input text:

```
['1' '2' 'Alokam' 'Hello' 'Hi' 'Nikhitha' 'This' 'a' 'has' 'is' 'of' 'only' 'part' 'question' 'sentence' 'should' 'which']
```

b)

Code:

```
#Reading input from a text file and saving it as a string
```

```
text = ""
```

```
with open('test_file2b.txt') as file:
```

```
    for line in file:
```

```
        for word in line.split():
```

```
            text= text + " " + word
```

```
#Removing punctuations from our input file
```

```
import re
```

```
text = re.sub(r'^\w\s', "", text)
```

```
text
```

```
#Printing each token
```

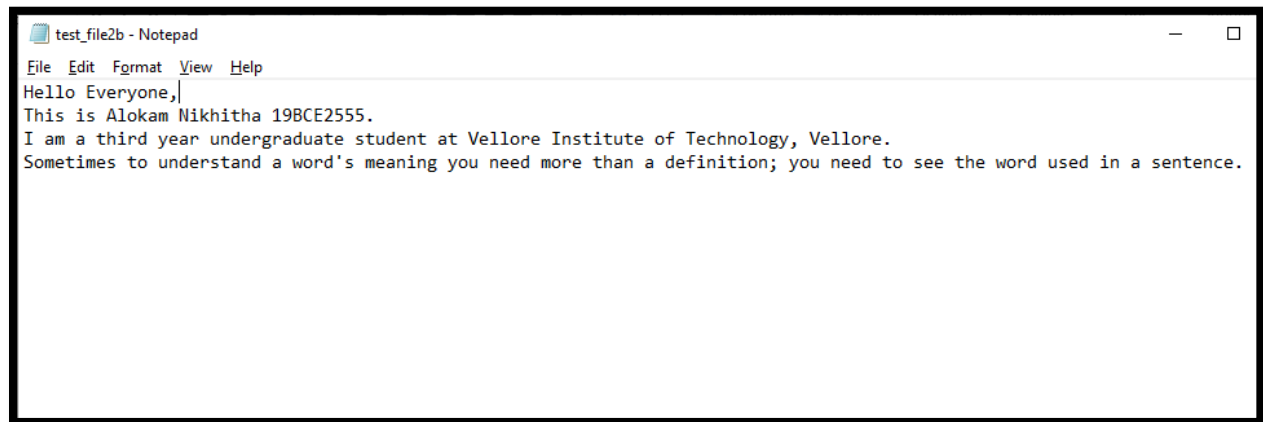
```
print(text.split())
```

```
#Printing unique tokens
```

```
import numpy as np
```

```
print(np.unique(text.split()))
```

Text File Taken as Input:



Code Snippets and Outputs:

```
In [5]: #Reading input from a text file and saving it as a string
text = ""
with open('test_file2b.txt') as file:
    for line in file:
        for word in line.split():
            text = text + " " + word
```

Here we are reading the text file using open method in python. Then reading each line we split each word and append it to a string variable with a space in between.

```
In [6]: #Removing punctuations from our input file
import re
text = re.sub(r'^\w\s', '', text)
text
```

```
Out[6]: ' Hello Everyone This is Alokam Nikhitha 19BCE2555 I am a third y
ear undergraduate student at Vellore Institute of Technology Vell
ore Sometimes to understand a words meaning you need more than a
definition you need to see the word used in a sentence'
```

Here we are removing punctuations from our input file. This is done using regex, where we keep only alphanumeric inputs in our text string. We can see all the periods and commas from original input files are removed here.

```
In [7]: #Printing each token
print(text.split())

['Hello', 'Everyone', 'This', 'is', 'Alokam', 'Nikhitha', '19BCE2555', 'I', 'am', 'a', 'third', 'year', 'undergraduate', 'student', 'at', 'Vellore', 'Institute', 'of', 'Technology', 'Vellore', 'Sometimes', 'to', 'understand', 'a', 'words', 'meaning', 'you', 'need', 'more', 'than', 'a', 'definition', 'you', 'need', 'to', 'see', 'the', 'word', 'used', 'in', 'a', 'sentence']
```

Next, we are splitting each word in our string using space character. Clearly, they form a token and hence we print each token.

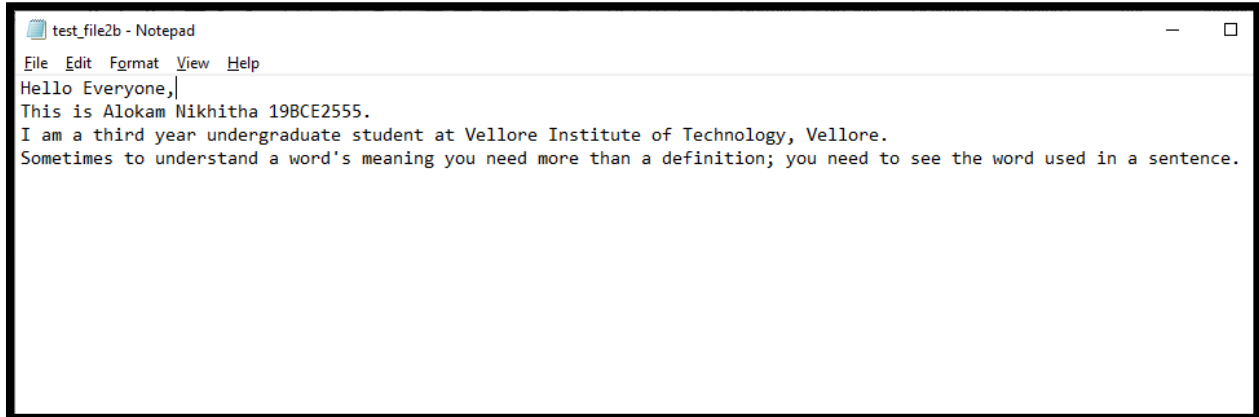
```
In [8]: import numpy as np
print(np.unique(text.split()))

['19BCE2555' 'Alokam' 'Everyone' 'Hello' 'I' 'Institute' 'Nikhitha' 'Sometimes' 'Technology' 'This' 'Vellore' 'a' 'am' 'at' 'definition' 'in' 'is' 'meaning' 'more' 'need' 'of' 'see' 'sentence' 'student' 'than' 'the' 'third' 'to' 'undergraduate' 'understand' 'used' 'word' 'words' 'year' 'you']
```

Here we are only printing unique tokens from all the generated tokens using split method. This is done using NumPy's unique function, which identifies all the unique elements from a list.

Results and Output:

- Input text:



```
test_file2b - Notepad
File Edit Format View Help
Hello Everyone,
This is Alokam Nikhitha 19BCE2555.
I am a third year undergraduate student at Vellore Institute of Technology, Vellore.
Sometimes to understand a word's meaning you need more than a definition; you need to see the word used in a sentence.
```

- Tokens of input text:

```
['19BCE2555' 'Alokam' 'Everyone' 'Hello' 'I' 'Institute' 'Nikhitha'
'Sometimes' 'Technology' 'This' 'Vellore' 'a' 'am' 'at' 'definition' 'in'
'is' 'meaning' 'more' 'need' 'of' 'see' 'sentence' 'student' 'than' 'the'
'third' 'to' 'undergraduate' 'understand' 'used' 'word' 'words' 'year'
'you']
```