Tanay Jhawar

BE4 Roll No:22

EXPERIMENT 3

AIM: Implement Filteration, Script Validation, Stop Word Removal in Python

SOURCE CODE:

```
from nltk.tokenize import word_tokenize
from nltk.corpus import stopwords
f = open("tanay.txt","r")

#wordtokenizer
for line in f:
   print("Tokenization without NLTK \n")
   print(line.split())
   print("\n")
   print("Length of Tokenization without NLTK:")
   print(len(line.split()))
   print("\n")

   print("Tokenization with NLTK \n")
   print(word_tokenize(line))
   print("\n")
   print("Length of Tokenization with NLTK:")
   print(len(word_tokenize(line)))
   print("\n")

#filteration
   print("Filteration \n")
   bad_chars = [';',',','!',':','*','#','<','>','?','@','.']
   words = word_tokenize(line)
```

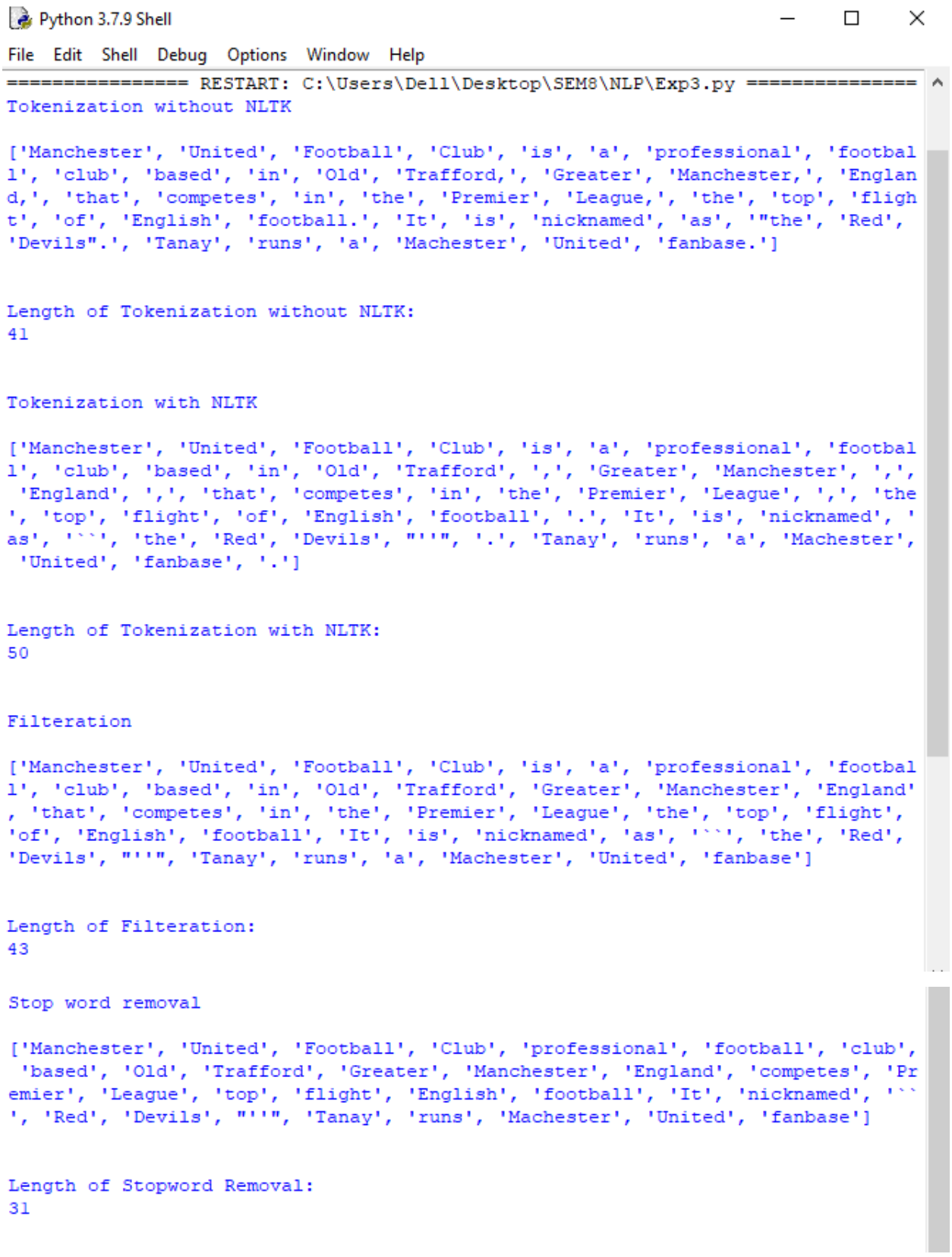
```
    print(list(filter(lambda i: i not in bad_chars,words)))
    print("\n")
    print("Length of Filteration:")
    print(len(list(filter(lambda i: i not in bad_chars,words))))
    print("\n")

#stop word removal
    print("Stop word removal \n")
    stop_words = set(stopwords.words("English"))
    without_stop_words = [word for word in words if not word in stop_words]
    with_stopwords = list(filter(lambda i: i not in bad_chars, without_stop_words))
    print(with_stopwords)
    print("\n")
    print("Length of Stopword Removal:")
    print(len(with_stopwords))
    print("\n")
```

INPUT:

Manchester United Football Club is a professional football club based in Old Trafford, Greater Manchester, England, that competes in the Premier League, the top flight of English football. It is nicknamed as "the Red Devils". Tanay runs a Machester United fanbase.

OUTPUT:



Python 3.7.9 Shell

File Edit Shell Debug Options Window Help

```
========== RESTART: C:\Users\Dell\Desktop\SEM8\NLP\Exp3.py ==========
Tokenization without NLTK

['Manchester', 'United', 'Football', 'Club', 'is', 'a', 'professional', 'footbal
l', 'club', 'based', 'in', 'Old', 'Trafford,', 'Greater', 'Manchester,', 'Englan
d,', 'that', 'competes', 'in', 'the', 'Premier', 'League,', 'the', 'top', 'fligh
t', 'of', 'English', 'football.', 'It', 'is', 'nicknamed', 'as', '"the', 'Red',
'Devils".', 'Tanay', 'runs', 'a', 'Machester', 'United', 'fanbase.']


Length of Tokenization without NLTK:
41


Tokenization with NLTK

['Manchester', 'United', 'Football', 'Club', 'is', 'a', 'professional', 'footbal
l', 'club', 'based', 'in', 'Old', 'Trafford', ',', 'Greater', 'Manchester', ',',
 'England', ',', 'that', 'competes', 'in', 'the', 'Premier', 'League', ',', 'the
', 'top', 'flight', 'of', 'English', 'football', '.', 'It', 'is', 'nicknamed', '
as', '``', 'the', 'Red', 'Devils', "''", '.', 'Tanay', 'runs', 'a', 'Machester',
 'United', 'fanbase', '.']


Length of Tokenization with NLTK:
50


Filteration

['Manchester', 'United', 'Football', 'Club', 'is', 'a', 'professional', 'footbal
l', 'club', 'based', 'in', 'Old', 'Trafford', 'Greater', 'Manchester', 'England'
, 'that', 'competes', 'in', 'the', 'Premier', 'League', 'the', 'top', 'flight',
'of', 'English', 'football', 'It', 'is', 'nicknamed', 'as', '``', 'the', 'Red',
'Devils', "''", 'Tanay', 'runs', 'a', 'Machester', 'United', 'fanbase']


Length of Filteration:
43


Stop word removal

['Manchester', 'United', 'Football', 'Club', 'professional', 'football', 'club',
 'based', 'Old', 'Trafford', 'Greater', 'Manchester', 'England', 'competes', 'Pr
emier', 'League', 'top', 'flight', 'English', 'football', 'It', 'nicknamed', '``
', 'Red', 'Devils', "''", 'Tanay', 'runs', 'Machester', 'United', 'fanbase']


Length of Stopword Removal:
31
```