LAB1Text

Annapoornima S
225229101

```
nltk.download("wordnet")
```

```
True
```

```
text="This is Andrew's text,isn't it?"
```

```
tokenizer= nltk.tokenize.WhitespaceTokenizer()
tokens=tokenizer.tokenize(text)
print(len(tokens))
print(tokens)
```

```
tokenizer= nltk.tokenize.TreebankWordTokenizer()
tokens=tokenizer.tokenize(text)
print(len(tokens))
print(tokens)
```

```
10
```

```
tokenizer= nltk.tokenize.WordPunctTokenizer()
tokens=tokenizer.tokenize(text)
print(len(tokens))
print(tokens)
```

```
12
```

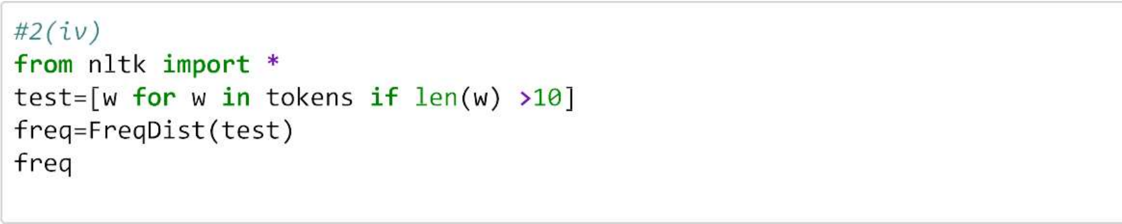
```
#1
filename = ("gift-of-magi.txt")
f=open (filename,'r')
text=f.read()
f.close()
```

```
ofore
```

```
 said,
Yshave
```

```
#2(i)
tokenizer= nltk.tokenize.WhitespaceTokenizer()
tokens=tokenizer.tokenize(text)
print(len(tokens))
```

```
2074
```



```
#2(iv)
from nltk import *
test=[w for w in tokens if len(w) >10]
freq=FreqDist(test)
freq
```

```
#2(v)
for i,j in freq.items():
    if len(i) > 10 and j>=2:
        print(i,j)
```

3

step

```
confidence,
```

step-2

```
True
```

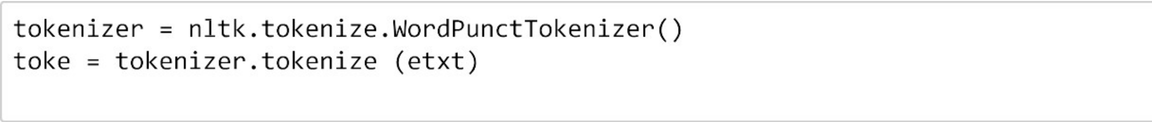
```
etypes =sorted(set(etoks))
etypes[-10:]
```

```
8000
```

```
efreq = nltk.FreqDist(etoks)
```

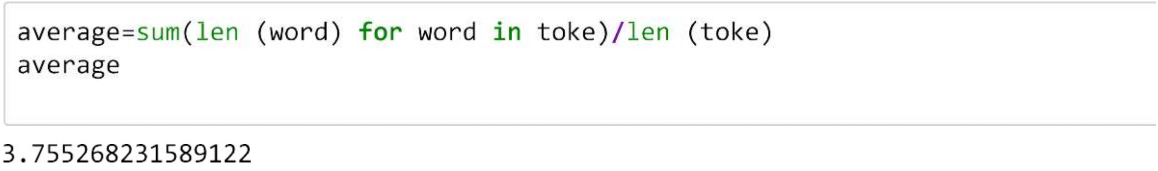
```
5198
```

with prefix and suffix



```
tokenizer = nltk.tokenize.WordPunctTokenizer()
toke = tokenizer.tokenize (etxt)
```

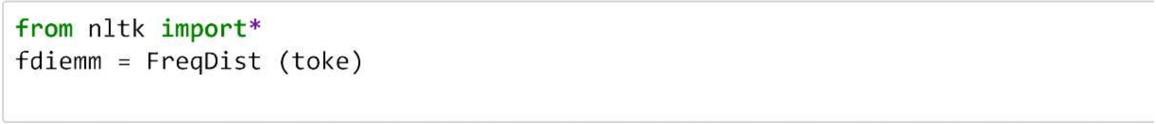
word



```
average=sum(len (word) for word in toke)/len (toke)
average
```

```
3.755268231589122
```

Word frequency

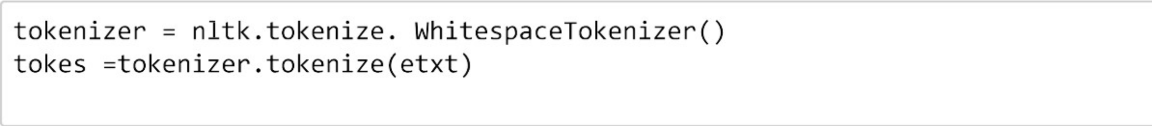


```
from nltk import*
fdiemm = FreqDist (toke)
```

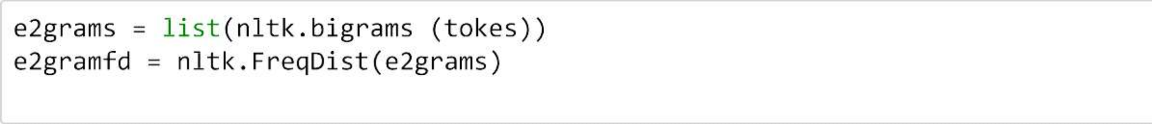
```
last_ten = FreqDist(dict(e2gramfd.most_common()[-10:]))
last_ten
```

```
FreqDist({(
          (
          (
          (
          (
          (
          (
          (
          (
          (
```

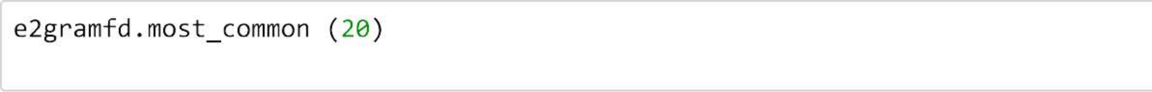
Bigram top frequency



```
tokenizer = nltk.tokenize. WhitespaceTokenizer()
tokes =tokenizer.tokenize(etxt)
```



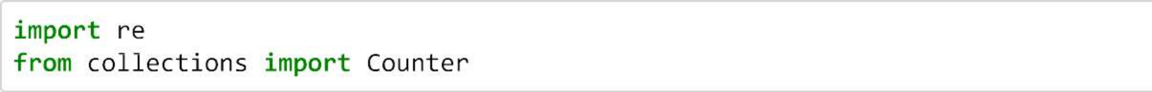
```
e2grams = list(nltk.bigrams (tokes))
e2gramfd = nltk.FreqDist(e2grams)
```



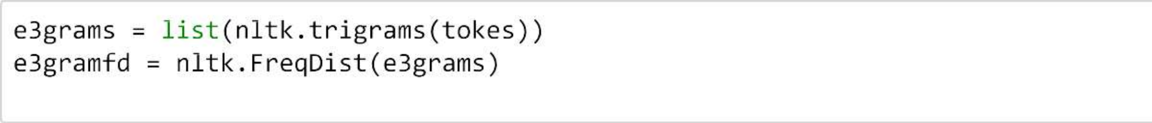
```
e2gramfd.most_common (20)
```

Bigram frequency count

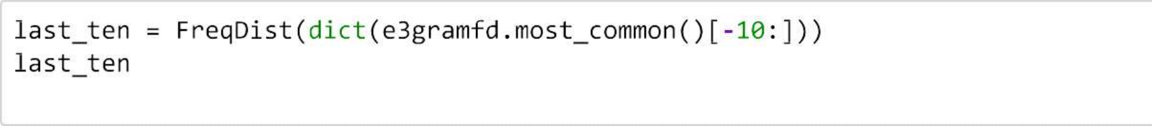
Word'so'



```
import re
from collections import Counter
```

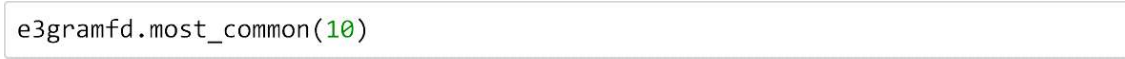


```
e3grams = list(nltk.trigrams(tokes))
e3gramfd = nltk.FreqDist(e3grams)
```



```
last_ten = FreqDist(dict(e3gramfd.most_common()[-10:]))
last_ten
```

Trigram top frequency



```
e3gramfd.most_common(10)
```

trigram frequency count

```
words1 = re.findall(r'so happy to \w+', open('austen-emma.txt').read())
print(words1)
```