In [1]:
```
import nltk
from nltk import FreqDist
import re
```

In [2]:
```
file1=nltk.corpus.gutenberg.fileids()[1]
austenpersuasiontext=nltk.corpus.gutenberg.raw(file1)
austenpersuasiontokens=nltk.wordpunct_tokenize(austenpersuasiontext)
austenpersuasionwords=[w.lower() for w in austenpersuasiontokens]
```

In [3]:
```
shortwords=austenpersuasionwords[11:111]
shortwords
```

Out[3]:
```
['elliot',
 ',',
 'of',
 'kellynch',
 'hall',
 ',',
 'in',
 'somersetshire',
 ',',
 'was',
 'a',
 'man',
 'who',
 ',',
 'for',
 'his',
 'own',
 'amusement',
 ',',
 'never',
 'took',
 'up',
 'any',
 'book',
 'but',
 'the',
 'baronetage',
 ';',
 'there',
 'he',
 'found',
 'occupation',
 'for',
 'an',
 'idle',
 'hour',
 ',',
 'and',
 'consolation',
 'in',
 'a',
 'distressed',
 'one',
 ';',
 'there',
 'his',
 'faculties',
 'were',
 'roused',
 'into',
 'admiration',
 'and',
 'respect',
 ',',
 'by',
 'contemplating',
 'the',
 'limited',
 'remnant',
 'of',
```

```
 'the',
 'earliest',
 'patents',
 ';',
 'there',
 'any',
 'unwelcome',
 'sensations',
 ',',
 'arising',
 'from',
 'domestic',
 'affairs',
 'changed',
 'naturally',
 'into',
 'pity',
 'and',
 'contempt',
 'as',
 'he',
 'turned',
 'over',
 'the',
 'almost',
 'endless',
 'creations',
 'of',
 'the',
 'last',
 'century',
 ';',
 'and',
 'there',
 ',',
 'if',
 'every',
 'other',
 'leaf',
 'were']
```

In [4]:
```
shortdist=FreqDist(shortwords)
shortdist.keys()
```

Out[4]:
```
dict_keys(['elliot', ',', 'of', 'kellynch', 'hall', 'in', 'somersetshire', 'was',
'a', 'man', 'who', 'for', 'his', 'own', 'amusement', 'never', 'took', 'up', 'any', 'b
ook', 'but', 'the', 'baronetage', ';', 'there', 'he', 'found', 'occupation', 'an', 'i
dle', 'hour', 'and', 'consolation', 'distressed', 'one', 'faculties', 'were', 'rouse
d', 'into', 'admiration', 'respect', 'by', 'contemplating', 'limited', 'remnant', 'ea
rliest', 'patents', 'unwelcome', 'sensations', 'arising', 'from', 'domestic', 'affair
s', 'changed', 'naturally', 'pity', 'contempt', 'as', 'turned', 'over', 'almost', 'en
dless', 'creations', 'last', 'century', 'if', 'every', 'other', 'leaf'])
```

In [5]:
```
for word in shortdist.keys():
    print(word,shortdist[word])
```

```
elliot 1
, 9
of 3
kellynch 1
hall 1
in 2
somersetshire 1
was 1
a 2
man 1
who 1
for 2
his 2
own 1
amusement 1
never 1
took 1
up 1
any 2
book 1
but 1
the 5
baronetage 1
; 4
there 4
he 2
found 1
occupation 1
an 1
idle 1
hour 1
and 4
consolation 1
distressed 1
one 1
faculties 1
were 2
roused 1
into 2
admiration 1
respect 1
by 1
contemplating 1
limited 1
remnant 1
earliest 1
patents 1
unwelcome 1
sensations 1
arising 1
from 1
domestic 1
affairs 1
changed 1
naturally 1
pity 1
contempt 1
as 1
turned 1
over 1
```

```
almost 1
endless 1
creations 1
last 1
century 1
if 1
every 1
other 1
leaf 1
```

In [6]:
```
stopwords = ['to', 'be', 'of', 'the', 'in', 'it', 'was',
'i', 'am', 'she', 'had', 'been', 'is', 'have','could', 'not',
'her', 'he', 'do', 'and', 'would', 'such', 'a', 'his', 'must']
```

In [7]:
```
def alphaStopFreqDist(words, stoplist):
    asdist=FreqDist()
    pattern=re.compile('.*[^a-z].*')
    for word in words:
        if not pattern.match(word):
            if not word in stoplist:
                asdist[word]+=1
    return asdist
```

In [8]:
```
asdist=alphaStopFreqDist(austenpersuasionwords,stopwords)
keys=list(asdist.keys())
```

In [9]:
```
print(keys[:50])
```

```
['persuasion', 'by', 'jane', 'austen', 'chapter', 'sir', 'walter', 'elliot', 'kellync
h', 'hall', 'somersetshire', 'man', 'who', 'for', 'own', 'amusement', 'never', 'too
k', 'up', 'any', 'book', 'but', 'baronetage', 'there', 'found', 'occupation', 'an',
'idle', 'hour', 'consolation', 'distressed', 'one', 'faculties', 'were', 'roused', 'i
nto', 'admiration', 'respect', 'contemplating', 'limited', 'remnant', 'earliest', 'pa
tents', 'unwelcome', 'sensations', 'arising', 'from', 'domestic', 'affairs', 'change
d']
```

In [10]:
```
for key in keys[:30]:
    print(key,asdist[key])
```

```
persuasion 7
by 418
jane 1
austen 1
chapter 24
sir 149
walter 141
elliot 289
kellynch 73
hall 28
somersetshire 4
man 134
who 190
for 707
own 163
amusement 10
never 155
took 19
up 91
any 199
book 8
but 664
baronetage 2
there 286
found 83
occupation 5
an 245
idle 3
hour 33
consolation 4
```

In [11]:
```
def bigramDist(words,stoplist):
    biDist=FreqDist()
    uniDist=alphaStopFreqDist(words,stoplist)
    for i in range(1,len(words)):
        if words[i-1] in uniDist and words[i] in uniDist:
            biword=words[i-1]+" "+words[i]
            biDist[biword]+=1
    return biDist
```

In [14]:
```
austenpersuasiondist=bigramDist(austenpersuasionwords,stopwords)
austenpersuasiondistkeys=list(austenpersuasiondist.keys())
for key in austenpersuasiondistkeys[:30]:
    print(key,austenpersuasiondist[key])
```

```
persuasion by 1
by jane 1
jane austen 1
sir walter 131
walter elliot 16
kellynch hall 25
man who 6
own amusement 2
never took 1
took up 4
up any 2
any book 1
book but 1
found occupation 1
occupation for 1
for an 9
an idle 1
idle hour 1
distressed one 1
faculties were 1
were roused 1
roused into 1
into admiration 1
by contemplating 1
limited remnant 1
earliest patents 1
there any 1
any unwelcome 1
unwelcome sensations 1
arising from 1
```

In [ ]: