Assignment-1

Ans1:

In [14]:
```
from nltk.book import *
text5.concordance("collocations")
```

```
*** Introductory Examples for the NLTK Book ***
Loading text1, ..., text9 and sent1, ..., sent9
Type the name of the text or sentence to view it.
Type: 'texts()' or 'sents()' to list the materials.
text1: Moby Dick by Herman Melville 1851
text2: Sense and Sensibility by Jane Austen 1811
text3: The Book of Genesis
text4: Inaugural Address Corpus
text5: Chat Corpus
text6: Monty Python and the Holy Grail
text7: Wall Street Journal
text8: Personals Corpus
text9: The Man Who Was Thursday by G . K . Chesterton 1908
No matches
```

Ans2:

In [7]:
```
my_set = ['This','is','a','set','of','words']
' '.join(my_set)
```

Out[7]:
```
'This is a set of words'
```

In [8]:
```
' '.join(my_set).split()
```

Out[8]:
```
['This', 'is', 'a', 'set', 'of', 'words']
```

Ans3:

In [121]:
```
text9.index('sunset')
```

Out[121]:
```
629
```

Ans4:

In [124]:
```
print(set(sent1))
print(set(sent2))
print(set(sent3))
print(set(sent4))
print(set(sent5))
print(set(sent6))
print(set(sent7))
print(set(sent8))
```

```
{'me', 'Ishmael', 'Call', '.'}
{'Dashwood', 'of', 'in', '.', 'long', 'Sussex', 'family', 'The', 'been', 'had',
'settled'}
{'God', 'earth', 'and', 'beginning', '.', 'In', 'heaven', 'created', 'the'}
{'of', 'and', 'Fellow', 'Representatives', 'House', '-', ':', 'Citizens', 'Senat
e', 'the'}
{'to', 'JOIN', 'problem', 'have', 'I', 'people', 'a', 'with', 'lol', 'me', 'PMin
g'}
{'ARTHUR', 'SCENE', ']', 'Whoa', ':', '!', 'clop', 'wind', 'KING', '[', 'there',
'1'}
{'nonexecutive', '.', 'old', 'board', ',', 'Vinken', '61', 'join', 'a', 'will',
'29', 'the', 'director', 'Pierre', 'as', 'Nov.', 'years'}
{'lady', 'older', 'attrac', '25', 'MALE', '.', 'encounters', 'for', ',', 'discre
et', 'single', 'seeks', 'SEXY'}
```

Ans5:

Difference - lets understand this from an example case

In [26]:
```
data = ['Abc', 'ABC', 'abc', 'aBc']
print(sorted(set([w.lower() for w in data])))
print((sorted([w.lower() for w in set(data)])))
```

```
['abc']
['abc', 'abc', 'abc', 'abc']
```

In [21]:
```
print(len(sorted(set([w.lower() for w in text1]))))
```

```
17231
```

In [30]:
```
print(len(sorted([w.lower() for w in set(text1)])))
```

```
19317
```

Ans6:

In [29]:
```
text2[-2:]
```

Out[29]:
```
['THE', 'END']
```

Ans7:

In [65]:
```
from operator import itemgetter

all_words = FreqDist(text5)
four_lettered = {key:value for key, value in sorted(all_words.items(), key=itemgett
er(1), reverse=True) if len(key)==4}
four_lettered
```

```
Out[65]: {'JOIN': 1021,
          'PART': 1016,
          'that': 274,
          'what': 183,
          'here': 181,
          '....': 170,
          'have': 164,
          'like': 156,
          'with': 152,
          'chat': 142,
          'your': 137,
          'good': 130,
          'just': 125,
          'lmao': 107,
          'know': 103,
          'room': 98,
          'from': 92,
          'this': 86,
          'well': 81,
          'back': 78,
          'hiya': 78,
          'they': 77,
          'dont': 75,
          'yeah': 75,
          'want': 71,
          'love': 60,
          'guys': 58,
          'some': 58,
          'been': 57,
          'talk': 56,
          'nice': 52,
          'time': 50,
          'when': 48,
          'haha': 44,
          'make': 44,
          'girl': 43,
          'need': 43,
          'U122': 42,
          'MODE': 41,
          'will': 40,
          'much': 40,
          'then': 40,
          'over': 39,
          'work': 38,
          'were': 38,
          'take': 37,
          'U121': 36,
          'U115': 36,
          'song': 36,
          'even': 35,
          'does': 35,
          'seen': 35,
          'U156': 35,
          'U105': 35,
          'more': 34,
          'damn': 34,
          'only': 33,
          'come': 33,
          'hell': 29,
          'long': 28,
          'them': 28,
          'name': 27,
          'tell': 27,
          'away': 26,
```

Ans8:

In [83]:
```
all_words = set(text6)
for word in all_words:
    if word == word.upper():
        print(word)
```

```
GUEST
.'
PARTY
CROWD
15
W
BRIDE
N
VILLAGER
SOLDIER
17
12
ARMY
ROBIN
MIDGET
HISTORIAN
SENTRY
10
24
OF
',
CARTOON
ENCHANTER
BEDEVERE
FRENCH
11
MONKS
,'
MASTER
[...
]
NI
CRONE
VOICE
HEADS
MAN
!)
,--
5
(
!]
OLD
DINGO
CRASH
PRISONER
?!
VILLAGERS
GIRLS
PRINCESS
GALAHAD
RIGHT
CHARACTERS
CAMERAMAN
I
KNIGHT
...
C
20
FATHER
KNIGHTS
LAUNCELOT
A
WINSTON
22
```

Ans9:

In [84]:
```
all_words = set(text6)
```

9(a)

In [93]:
```
new_set = {w for w in all_words if w.endswith('ize')}
new_set
```

Out[93]:
```
set()
```

9(b)

In [96]:
```
new_set = {w for w in all_words if 'z' in w}
new_set
```

Out[96]:
```
{'Fetchez', 'amazes', 'frozen', 'zhiv', 'zone', 'zoo', 'zoop', 'zoosh'}
```

In [98]:
```
new_set = {w for w in all_words if 'pt' in w}
new_set
```

Out[98]:
```
{'Chapter',
 'Thpppppt',
 'Thppppt',
 'Thpppt',
 'Thppt',
 'aptly',
 'empty',
 'excepting',
 'ptoo',
 'temptation',
 'temptress'}
```

Ans10:

In [99]:
```
sent = ['she', 'sells', 'sea', 'shells', 'by', 'the', 'sea', 'shore']
```

10(a)

In [103]:
```
word_set = set(sent)
for w in word_set:
    if w.startswith('sh'):
        print(w)
```

```
shells
shore
she
```

In [104]:
```
for w in word_set:
    if len(w)>4:
        print(w)
```

```
shells
shore
sells
```

Ans11:

In [107]:
```
# This code gives the total letter in the dataset
sum([len(w) for w in text1])
```

Out[107]: 999044

In [119]:
```
# Average word length of a text
sum([len(w) for w in text1])/len(text1)
```

Out[119]: 3.830411128023649

Ans12:

In [120]:
```
def vocab_size(text):
    return len(set(text))

vocab_size(text3)
```

Out[120]: 2789

Ans13:

In [118]:
```
def percent(word, text):
    return FreqDist(text).freq(word)*100

percent('monstrous', text1)
```

Out[118]: 0.003834076505162584