In [48]:

```
import nltk
```

Chapter 1

In [1]:

```
from nltk.book import *
```

```
*** 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
```

In [2]:

```
text1
```

Out[2]:

```
<Text: Moby Dick by Herman Melville 1851>
```

In [3]:

```
text1.concordance("monstrous")
```

```
Building index...
Displaying 11 of 11 matches:
ong the former , one was of a most monstrous size . ... This came towards us ,
ON OF THE PSALMS . " Touching that monstrous bulk of the whale or ork we have r
ll over with a heathenish array of monstrous clubs and spears . Some were thick
d as you gazed , and wondered what monstrous cannibal and savage could ever hav
that has survived the flood ; most monstrous and most mountainous ! That Himmal
they might scout at Moby Dick as a monstrous fable , or still worse and more de
th of Radney .'" CHAPTER 55 Of the Monstrous Pictures of Whales . I shall ere l
ing Scenes . In connexion with the monstrous pictures of whales , I am strongly
ere to enter upon those still more monstrous stories of them which are to be fo
ght have been rummaged out of this monstrous cabinet there is no telling . But
of Whale - Bones ; for Whales of a monstrous size are oftentimes cast up dead u
```

In [4]:

```
text1.similar("monstrous")
```

```
Building word-context index...
abundant candid careful christian contemptible curious delightfully
determined doleful domineering exasperate fearless few gamesome
horrible impalpable imperial lamentable lazy loving
```

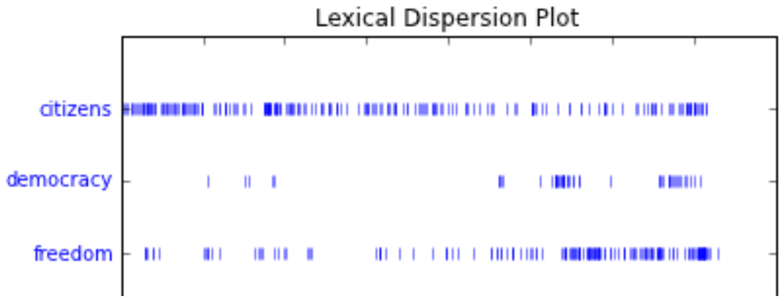
In [5]:

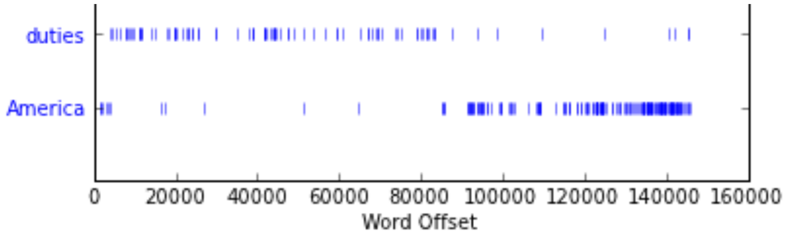
```
text2.common_contexts(["monstrous", "very"])
```

```
Building word-context index...
a_lucky a_pretty am_glad be_glad is_pretty
```

In [6]:

```
text4.dispersion_plot(["citizens", "democracy", "freedom", "duties", "America"])
```





```
In [7]: text3.generate()
```

```
Building ngram index...
In the same . And thither were all the rams of thy kindred , saying ,
Jacob ' s name Eve ; because there they mourned with a loud voi And it
came to pass , when he heard the voice of thy lord , but God meant it
unto his father ' s sake . And Seth lived after he begat Enos eight
hundred yea and he died . Now therefore when I came this day ; and go
thy way . And it came to pass , when his brethren , and herds , and do
ye ;
```

```
In [8]: len(text3)
```

```
Out[8]: 44764
```

```
In [26]: ?text3
```

```
In [30]: print sorted(set(text3))[:50]
```

```
['!', "'", '(', ')', ',', ',)', '.', '.)', ':', ';', ';)', '?', '?)', 'A', 'Abel',
'Abelmizraim', 'Abidah', 'Abide', 'Abimael', 'Abimelech', 'Abr', 'Abrah', 'Abraham',
'Abram', 'Accad', 'Achbor', 'Adah', 'Adam', 'Adbeel', 'Admah', 'Adullamite', 'After',
'Aholibamah', 'Ahuzzath', 'Ajah', 'Akan', 'All', 'Allonbachuth', 'Almighty', 'Almodad',
'Also', 'Alvah', 'Alvan', 'Am', 'Amal', 'Amalek', 'Amalekites', 'Ammon', 'Amorite',
'Amorites']
```

```
In [20]: len(set(text3))
```

```
Out[20]: 2789
```

```
In [22]: from __future__ import division
         len(text3) / len(set(text3))
```

```
Out[22]: 16.050197203298673
```

```
In [23]: text3.count("smote")
```

```
Out[23]: 5
```

```
In [24]: 100 * text4.count('a') / len(text4)
```

```
Out[24]: 1.4643016433938312
```

```
In [25]: 100 * text5.count('lol') / len(text5)
```

```
Out[25]: 1.5640968673628082
```

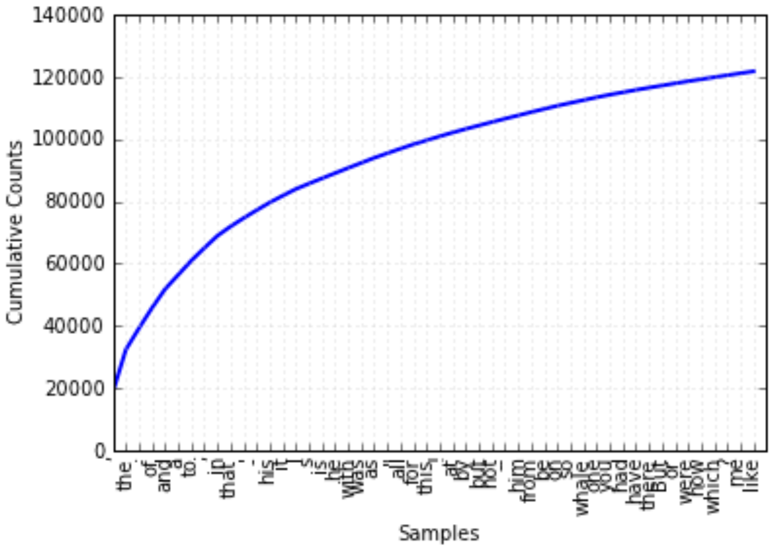
```
In [29]: fdist1 = FreqDist(text1)
         vocabulary1 = fdist1.keys()
         print vocabulary1[:50]
```

```
[',', 'the', '.', 'of', 'and', 'a', 'to', ';', 'in', 'that', "'", '-', 'his', 'it', 'I',
's', 'is', 'he', 'with', 'was', 'as', '"', 'all', 'for', 'this', '!', 'at', 'by', 'but',
'not', '--', 'him', 'from', 'be', 'on', 'so', 'whale', 'one', 'you', 'had', 'have',
'there', 'But', 'or', 'were', 'now', 'which', '?', 'me', 'like']
```

```
In [31]: fdist1['whale']
```

```
Out[31]: 906
```

```
In [32]: fdist1.plot(50, cumulative=True)
```



```
In [35]: print fdist1.hapaxes()[:50]
```

```
['!\'"', '!)"', '!*', '!--"', '"...', "',--", "';", '):', ');--', ',)', '--\'"', '---"', '-
--,', '."*', '."--', '.*--', '.--"', '100', '101', '102', '103', '104', '105', '106',
'107', '108', '109', '11', '110', '111', '112', '113', '114', '115', '116', '117', '118',
'119', '12', '120', '121', '122', '123', '124', '125', '126', '127', '128', '129', '130']
```

```
In [37]: V = set(text1)
         long_words = [w for w in V if len(w) > 15]
         print sorted(long_words)
```

```
['CIRCUMNAVIGATION', 'Physiognomically', 'apprehensiveness', 'cannibalistically',
'characteristically', 'circumnavigating', 'circumnavigation', 'circumnavigations',
'comprehensiveness', 'hermaphroditical', 'indiscriminately', 'indispensableness',
'irresistibleness', 'physiognomically', 'preternaturalness', 'responsibilities',
'simultaneousness', 'subterraneousness', 'supernaturalness', 'superstitiousness',
'uncomfortableness', 'uncompromisedness', 'undiscriminating', 'uninterpenetratingly']
```

```
In [39]: fdist5 = FreqDist(text5)
         print sorted([w for w in set(text5) if len(w) > 7 and fdist5[w] > 7])
```

```
['#14-19teens', '#talkcity_adults', '((((((((((', '........', 'Question', 'actually',
'anything', 'computer', 'cute.-ass', 'everyone', 'football', 'innocent', 'listening',
'remember', 'seriously', 'something', 'together', 'tomorrow', 'watching']
```

```
In [40]: bigrams(['more', 'is', 'said', 'than', 'done'])
```

```
Out[40]: [('more', 'is'), ('is', 'said'), ('said', 'than'), ('than', 'done')]
```

```
In [41]: text4.collocations()
```

```
Building collocations list
United States; fellow citizens; four years; years ago; Federal
Government; General Government; American people; Vice President; Old
World; Almighty God; Fellow citizens; Chief Magistrate; Chief Justice;
God bless; every citizen; Indian tribes; public debt; one another;
```

```
foreign nations; political parties
```

```
In [42]: fdist = FreqDist([len(w) for w in text1])
         fdist.keys()

Out[42]: [3, 1, 4, 2, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 20]
```

```
In [44]: print fdist.items()

         [(3, 50223), (1, 47933), (4, 42345), (2, 38513), (5, 26597), (6, 17111), (7, 14399), (8,
         9966), (9, 6428), (10, 3528), (11, 1873), (12, 1053), (13, 567), (14, 177), (15, 70), (16,
         22), (17, 12), (18, 1), (20, 1)]
```

```
In [45]: fdist.freq(3)

Out[45]: 0.19255882431878046
```

```
In [46]: ?fdist
```

Chapter 2

```
In [50]: print nltk.corpus.gutenberg.fileids()

         ['austen-emma.txt', 'austen-persuasion.txt', 'austen-sense.txt', 'bible-kjv.txt', 'blake-
         poems.txt', 'bryant-stories.txt', 'burgess-busterbrown.txt', 'carroll-alice.txt',
         'chesterton-ball.txt', 'chesterton-brown.txt', 'chesterton-thursday.txt', 'edgeworth-
         parents.txt', 'melville-moby_dick.txt', 'milton-paradise.txt', 'shakespeare-caesar.txt',
         'shakespeare-hamlet.txt', 'shakespeare-macbeth.txt', 'whitman-leaves.txt']
```

```
In [51]: emma = nltk.corpus.gutenberg.words('austen-emma.txt')
```

```
In [52]: len(emma)

Out[52]: 192427
```

```
In [54]: ?nltk.corpus.gutenberg
```

```
 In [ ]:
```