AUTOR: Isaac Reyes

```
import nltk
nltk.download('webtext')
```

```
[nltk_data] Downloading package webtext to /root/nltk_data...
[nltk_data]   Package webtext is already up-to-date!
True
```

```
from nltk.corpus import webtext
print(webtext.fileids())
```

```
['firefox.txt', 'grail.txt', 'overheard.txt', 'pirates.txt', 'singles.txt', 'wine.txt']
```

```
for fileid in webtext.fileids():
  print(fileid, webtext.raw(fileid)[:65], '...')
```

```
firefox.txt Cookie Manager: "Don't allow sites that set removed cookies to se ...
grail.txt SCENE 1: [wind] [clop clop clop]
KING ARTHUR: Whoa there!  [clop ...
overheard.txt White guy: So, do you have any plans for this evening?
Asian girl ...
pirates.txt PIRATES OF THE CARRIBEAN: DEAD MAN'S CHEST, by Ted Elliott & Terr ...
singles.txt 25 SEXY MALE, seeks attrac older single lady, for discreet encoun ...
wine.txt Lovely delicate, fragrant Rhone wine. Polished leather and strawb ...
```

```
from nltk.corpus import nps_chat
nltk.download('nps_chat')
```

```
[nltk_data] Downloading package nps_chat to /root/nltk_data...
[nltk_data]   Unzipping corpora/nps_chat.zip.
True
```

```
chatroom = nps_chat.posts('10-19-20s_706posts.xml') #chatroom[123]
chatroom[:5]
```

```
[['now', 'im', 'left', 'with', 'this', 'gay', 'name'],
 [':P'],
 ['PART'],
 ['hey', 'everyone'],
 ['ah', 'well']]
```

```
nltk.download('brown')
```

```
[nltk_data] Downloading package brown to /root/nltk_data...
[nltk_data]   Unzipping corpora/brown.zip.
True
```

```
from nltk.corpus import brown
brown.categories()
```

```
['adventure',
 'belles_lettres',
 'editorial',
 'fiction',
 'government',
 'hobbies',
 'humor',
 'learned',
 'lore',
 'mystery',
 'news',
 'religion',
 'reviews',
 'romance',
 'science_fiction']
```

```
brown.words(categories='news')
```

```
['The', 'Fulton', 'County', 'Grand', 'Jury', 'said', ...]
```

```
brown.sents(categories=['news','editorial','reviews'])
```

```
[['The', 'Fulton', 'County', 'Grand', 'Jury', 'said', 'Friday', 'an', 'investigation', 'of', "Atlanta's", 'recent', 'primary',
'election', 'produced', '``', 'no', 'evidence', "''", 'that', 'any', 'irregularities', 'took', 'place', '.'], ['The', 'jury',
'further', 'said', 'in', 'term-end', 'presentments', 'that', 'the', 'City', 'Executive', 'Committee', ',', 'which', 'had', 'over-all',
'charge', 'of', 'the', 'election', ',', '``', 'deserves', 'the', 'praise', 'and', 'thanks', 'of', 'the', 'City', 'of', 'Atlanta', "''",
'for', 'the', 'manner', 'in', 'which', 'the', 'election', 'was', 'conducted', '.'], ...]
```

```
news_text = brown.words(categories='news')
fdist = nltk.FreqDist(w.lower() for w in news_text)
modals = ['can','could']
for m in modals: print(m +':', fdist[m], end = ' ')
```

```
can: 94 could: 87
```

```
cfd = nltk.ConditionalFreqDist((genre, word)
for genre in brown.categories()
for word in brown.words(categories = genre))
```

```
genres = ['news','religion','hobbies','editorial','fiction','adventure']
modals = ['can','could','may','the','might','must','will']
cfd.tabulate(conditions = genres, samples = modals)
```

	can	could	may	the	might	must	will
news	93	86	66	5580	38	50	389
religion	82	59	78	2295	12	54	71
hobbies	268	58	131	4300	22	83	264
editorial	121	56	74	3508	39	53	233
fiction	37	166	8	3423	44	55	52
adventure	46	151	5	3370	58	27	50

```
import nltk
nltk.download('inaugural')
from nltk.corpus import inaugural
```

```
[nltk_data] Downloading package inaugural to /root/nltk_data...
[nltk_data]   Unzipping corpora/inaugural.zip.
```

```
inaugural.fileids()
```

```
['1789-Washington.txt',
 '1793-Washington.txt',
 '1797-Adams.txt',
 '1801-Jefferson.txt',
 '1805-Jefferson.txt',
 '1809-Madison.txt',
 '1813-Madison.txt',
 '1817-Monroe.txt',
 '1821-Monroe.txt',
 '1825-Adams.txt',
 '1829-Jackson.txt',
 '1833-Jackson.txt',
 '1837-VanBuren.txt',
 '1841-Harrison.txt',
 '1845-Polk.txt',
 '1849-Taylor.txt',
 '1853-Pierce.txt',
 '1857-Buchanan.txt',
 '1861-Lincoln.txt',
 '1865-Lincoln.txt',
 '1869-Grant.txt',
 '1873-Grant.txt',
 '1877-Hayes.txt',
 '1881-Garfield.txt',
 '1885-Cleveland.txt',
 '1889-Harrison.txt',
 '1893-Cleveland.txt',
 '1897-McKinley.txt',
 '1901-McKinley.txt',
 '1905-Roosevelt.txt',
 '1909-Taft.txt',
 '1913-Wilson.txt',
 '1917-Wilson.txt',
 '1921-Harding.txt',
 '1925-Coolidge.txt',
 '1929-Hoover.txt',
 '1933-Roosevelt.txt',
 '1937-Roosevelt.txt',
```

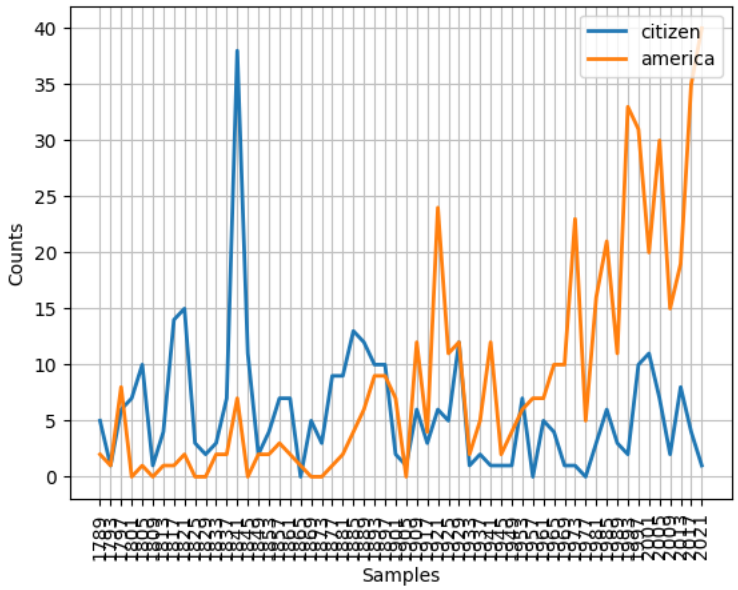
```
        '1941-Roosevelt.txt',
        '1945-Roosevelt.txt',
        '1949-Truman.txt',
        '1953-Eisenhower.txt',
        '1957-Eisenhower.txt',
        '1961-Kennedy.txt',
        '1965-Johnson.txt',
        '1969-Nixon.txt',
        '1973-Nixon.txt',
        '1977-Carter.txt',
        '1981-Reagan.txt',
        '1985-Reagan.txt',
        '1989-Bush.txt',
        '1993-Clinton.txt',
        '1997-Clinton.txt',
        '2001-Bush.txt',
        '2005-Bush.txt',
        '2009-Obama.txt',
        '2013-Obama.txt',
```

```
[fileid[:4]
for fileid in inaugural.fileids()]
```

```
     ['1789',
      '1793',
      '1797',
      '1801',
      '1805',
      '1809',
      '1813',
      '1817',
      '1821',
      '1825',
      '1829',
      '1833',
      '1837',
      '1841',
      '1845',
      '1849',
      '1853',
      '1857',
      '1861',
      '1865',
      '1869',
      '1873',
      '1877',
      '1881',
      '1885',
      '1889',
      '1893',
      '1897',
      '1901',
      '1905',
      '1909',
      '1913',
      '1917',
      '1921',
      '1925',
      '1929',
      '1933',
      '1937',
      '1941',
      '1945',
      '1949',
      '1953',
      '1957',
      '1961',
      '1965',
      '1969',
      '1973',
      '1977',
      '1981',
      '1985',
      '1989',
      '1993',
      '1997',
      '2001',
      '2005',
      '2009',
      '2013',
      '2017',
```

```
cfd = nltk.ConditionalFreqDist(
(target, fileid[:4])
for fileid in inaugural.fileids()
for w in inaugural.words(fileid)
for target in ['america','citizen']
if w.lower().startswith(target)
)
```

```
cfd.plot()
```




```
<Axes: xlabel='Samples', ylabel='Counts'>
```