继续了解 NLTK

《用 Python 玩转数据》　　by Dazhuang@NJU

NLTK 包括获取语料库、字符串处理、搭配发现、词性标注、机器学习、分块解析、语义解释、指标评测、概率与估计等多项语言任务，在处理时非常方便，例如要载入并去掉停用词可用类似如下几行简单代码就可以完成：

```
from nltk.corpus import stopwords
stopwords = stopwords.words('english')
… if words not in stopwords…
```

再以布朗语料库中的一个经典的例子来了解 NLTK 和条件频率分布的功能。

布朗语料库中有不同类别的文本，每种类别文本中包含多个词，例如想要获得新闻文体中所有的词可用 words()函数获得：

```
>>> from nltk.corpus import brown
>>> brown.words(categories = 'news')
['The', 'Fulton', 'County', 'Grand', 'Jury', 'said', ...]
>>> brown.words(fileids = 'ca16')
['Romantic', 'news', 'concerns', 'Mrs.', 'Joan', ...]
```

参数“categories = 'news'”表明文体类别为 news，“fileids = 'ca16'”表明文件标识是 ca16 的文件，具体分类可参照 http://icame.uib.no/brown/bcm-los.html 上完整的列表。

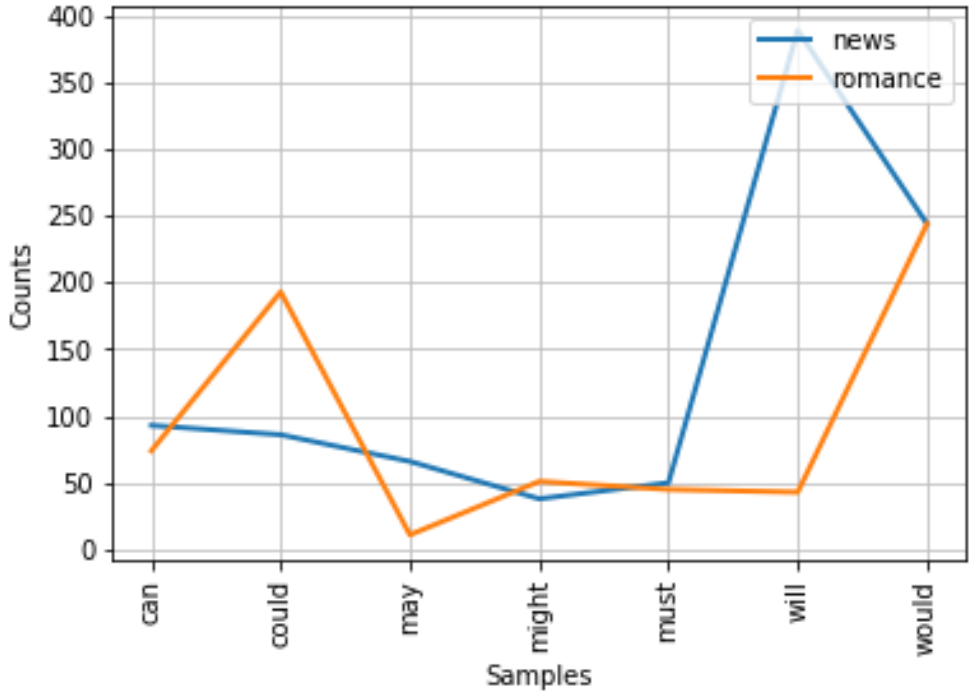
情态动词在文本很常用，不同文体中情态动词是否有不同的使用规律，以下程序用来比较情态动词 can、could、may、might、must、will、would 在新闻(news)和浪漫(romance)这两种不同文体中的频率，依据不同的“条件”即文体计算每个类别的频率分布，tabulate()方法用来为条件概率分布绘制分布表，其 conditions 参数指定要显示的条件这里是文体，默认为所有条件，samples 参数指定要显示的样本这里是情态动词。

```
import nltk
from nltk.corpus import brown

cfd = nltk.ConditionalFreqDist((genre, word)
                for genre in brown.categories()
                for word in brown.words(categories = genre))
genres = ['news', 'romance']
modals = ['can', 'could', 'may', 'might', 'must', 'will', 'would']
cfd.tabulate(conditions = genres, samples = modals)
cfd.plot(conditions = genres, samples = modals)
```

输出结果为：

	can	could	may	might	must	will	would
news	93	86	66	38	50	389	244
romance	74	193	11	51	45	43	244


400
350
300
250
200
150
100
50
0
Counts
news
romance
can
could
may
might
must
will
would
Samples