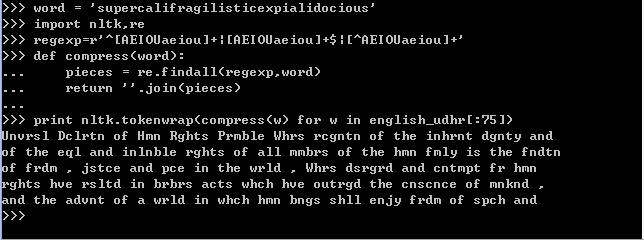
提取字符块

我们可以使用findall函数来提取字符块

|  |  |
| --- | --- |
| |  | | --- | | **>>> word = 'supercalifragilisticexpialidocious' >>> re.findall(r'[aeiou]', word) ['u', 'e', 'a', 'i', 'a', 'i', 'i', 'i', 'e', 'i', 'a', 'i', 'o', 'i', 'o', 'u'] >>> len(re.findall(r'[aeiou]', word)) 16** | |

word = 'supercalifragilisticexpialidocious'

英语中的词，去掉元音异能看懂，所以我们对人权宣言去掉单词中间部分的元音。



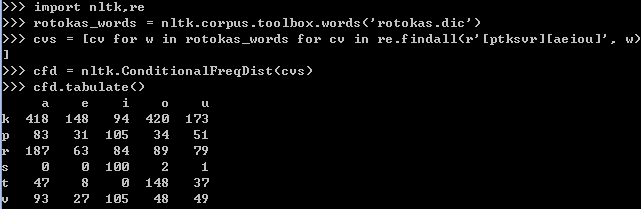
结合正则与条件频率

rotokas\_words = nltk.corpus.toolbox.words('rotokas.dic')

cvs = [cv for w in rotokas\_words for cv in re.findall(r'[ptksvr][aeiou]', w)]

cfd = nltk.ConditionalFreqDist(cvs)

cfd.tabulate()



使用nltk.Index()可以用来建立索引



搜索已分词文本

<>用来标记标志符的边界，尖括号之间的所有空白都被忽略（只能在findall函数中使用）

（）表示要选取的内容

