在自然语言处理的实际项目中，通常要使用 大量的语言数据或者语料库。

我们需要思考：

1. 什么是有用的文本语料库和词汇资源，如何使用Python获取它们？
2. 哪些Python结构最适合这项工作
3. 编写Python代码是如何避免重复的工作

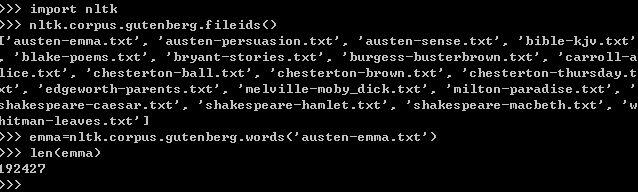
语料库来源

nltk语料库

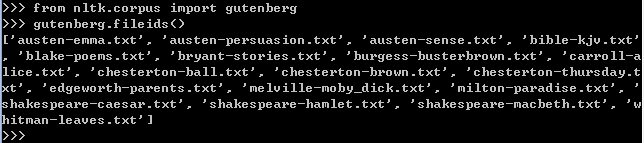
古腾堡语料库<http://gutenberg.org/>

大约有36000本免费电子书

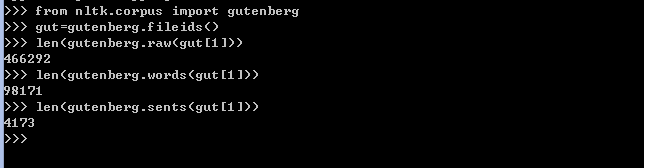
我们加载一下古腾堡语料库的文件标识符，然后选取第一个，简爱



我们也可以使用更熟悉的方式载入语料库



我们可以把一个文本以字符，单词或句子进行分割。



注意raw()函数把空格也计入

现在，我们可以计算一下每个文本的平均词长，句长和单词平均出现次数

from nltk.corpus import gutenberg

for fileid in gutenberg.fileids():

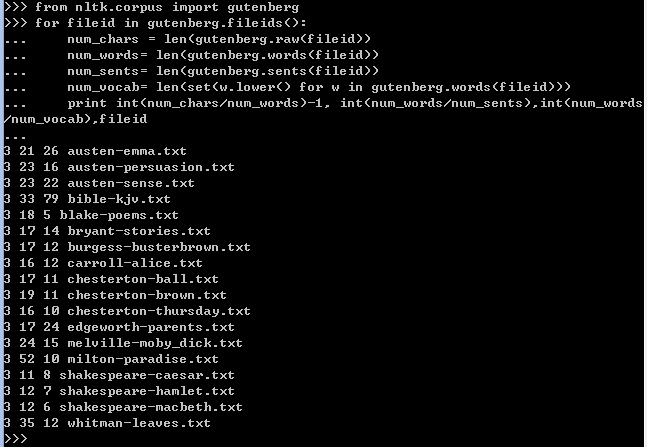
num\_chars = len(gutenberg.raw(fileid))

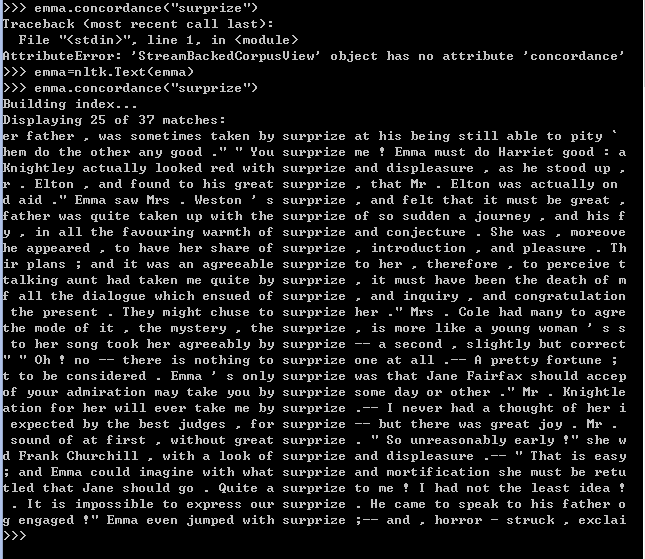
num\_words= len(gutenberg.words(fileid))

num\_sents= len(gutenberg.sents(fileid))

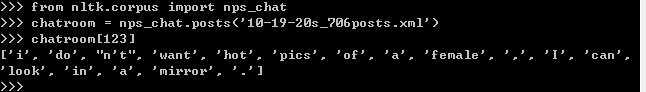
num\_vocab= len(set(w.lower() for w in gutenberg.words(fileid)))

print int(num\_chars/num\_words)-1, 注释：-1是为了去除空格int(num\_words/num\_sents),int(num\_words/num\_vocab),fileid

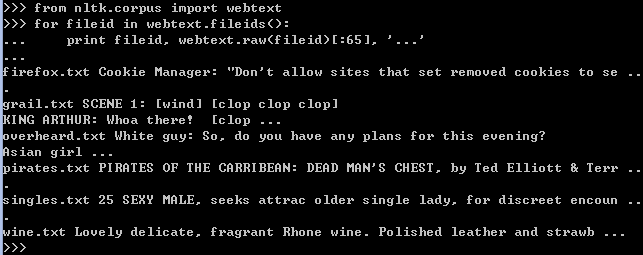


现在，emma不是Text类，所以无法使用concordance函数查看包含某个词的文本，所以必须现在emma转化为Text的成员。

网络和聊天文本

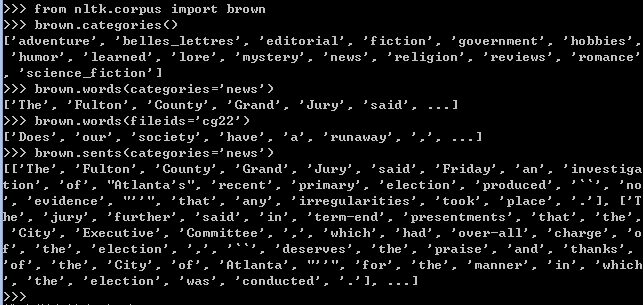


以上是2006年10月19日从20多岁的聊天室中收集的706个贴子的第123个帖子

网络文本集合

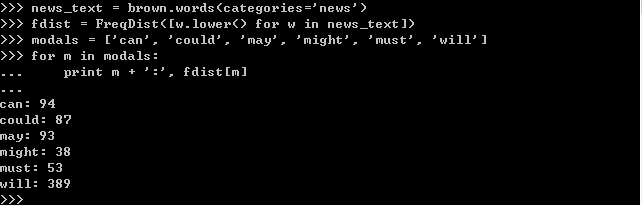
布朗语料库

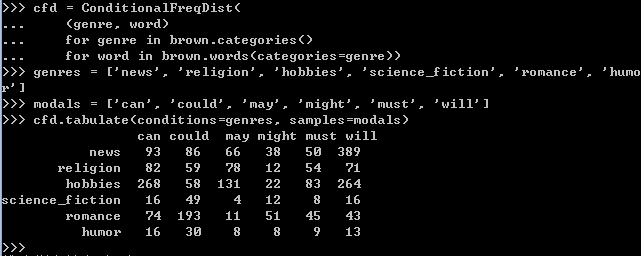
由布朗大学于1961年创建的第一个百万词级英语电子语料库



布朗语料库是一个研究文体之间的系统差异（又叫文体学的语言学研究）的资源。

我们可以看一下新闻类中不同情态动词出现的次数。





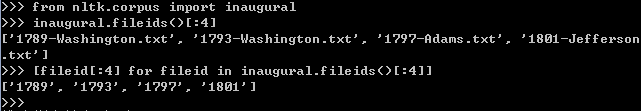
这里我们可以看出，新闻中最常用will，言情小说中最常用could

路透社语料库

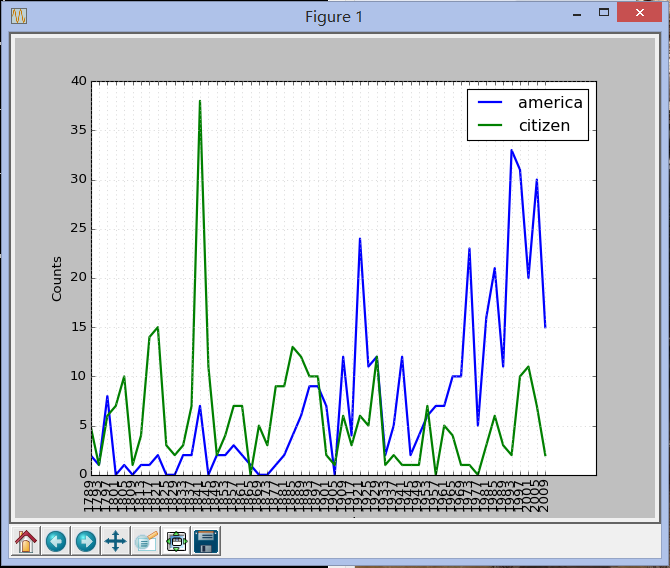
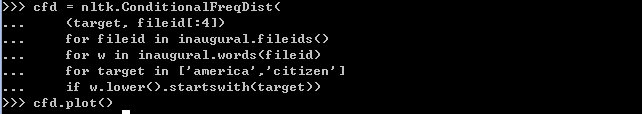
路透社语料库包含10788个新闻文档，共计130万字。这些文档分成90个主题，按照“训练”和“测试”分为两组。

在使用reuters.categories()是出现了错误，返回值为空，原因不详

就职演说语料库



我们可以看一下American和citizen随时间推移的使用情况。

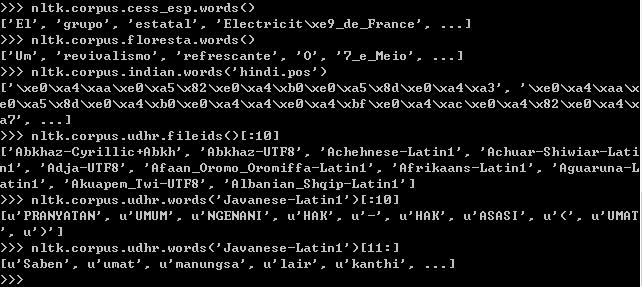


标注文本语料库

|  |  |  |
| --- | --- | --- |
| Corpus | Compiler | Contents |
| Brown Corpus | Francis, Kucera | 15 genres, 1.15M words, tagged, categorized |
| CESS Treebanks | CLiC-UB | 1M words, tagged and parsed (Catalan, Spanish) |
| Chat-80 Data Files | Pereira & Warren | World Geographic Database |
| CMU Pronouncing Dictionary | CMU | 127k entries |
| CoNLL 2000 Chunking Data | CoNLL | 270k words, tagged and chunked |
| CoNLL 2002 Named Entity | CoNLL | 700k words, pos- and named-entity-tagged (Dutch, Spanish) |
| CoNLL 2007 Dependency Treebanks (sel) | CoNLL | 150k words, dependency parsed (Basque, Catalan) |
| Dependency Treebank | Narad | Dependency parsed version of Penn Treebank sample |
| FrameNet | Fillmore, Baker et al | 10k word senses, 170k manually annotated sentences |
| Floresta Treebank | Diana Santos et al | 9k sentences, tagged and parsed (Portuguese) |
| Gazetteer Lists | Various | Lists of cities and countries |
| Genesis Corpus | Misc web sources | 6 texts, 200k words, 6 languages |
| Gutenberg (selections) | Hart, Newby, et al | 18 texts, 2M words |
| Inaugural Address Corpus | CSpan | US Presidential Inaugural Addresses (1789-present) |
| Indian POS-Tagged Corpus | Kumaran et al | 60k words, tagged (Bangla, Hindi, Marathi, Telugu) |
| MacMorpho Corpus | NILC, USP, Brazil | 1M words, tagged (Brazilian Portuguese) |
| Movie Reviews | Pang, Lee | 2k movie reviews with sentiment polarity classification |
| Names Corpus | Kantrowitz, Ross | 8k male and female names |
| NIST 1999 Info Extr (selections) | Garofolo | 63k words, newswire and named-entity SGML markup |
| Nombank | Meyers | 115k propositions, 1400 noun frames |
| NPS Chat Corpus | Forsyth, Martell | 10k IM chat posts, POS-tagged and dialogue-act tagged |
| Open Multilingual WordNet | Bond et al | 15 languages, aligned to English WordNet |
| PP Attachment Corpus | Ratnaparkhi | 28k prepositional phrases, tagged as noun or verb modifiers |
| Proposition Bank | Palmer | 113k propositions, 3300 verb frames |
| Question Classification | Li, Roth | 6k questions, categorized |
| Reuters Corpus | Reuters | 1.3M words, 10k news documents, categorized |
| Roget's Thesaurus | Project Gutenberg | 200k words, formatted text |
| RTE Textual Entailment | Dagan et al | 8k sentence pairs, categorized |
| SEMCOR | Rus, Mihalcea | 880k words, part-of-speech and sense tagged |
| Senseval 2 Corpus | Pedersen | 600k words, part-of-speech and sense tagged |
| SentiWordNet | Esuli, Sebastiani | sentiment scores for 145k WordNet synonym sets |
| Shakespeare texts (selections) | Bosak | 8 books in XML format |
| State of the Union Corpus | CSPAN | 485k words, formatted text |
| Stopwords Corpus | Porter et al | 2,400 stopwords for 11 languages |
| Swadesh Corpus | Wiktionary | comparative wordlists in 24 languages |
| Switchboard Corpus (selections) | LDC | 36 phonecalls, transcribed, parsed |
| Univ Decl of Human Rights | United Nations | 480k words, 300+ languages |
| Penn Treebank (selections) | LDC | 40k words, tagged and parsed |
| TIMIT Corpus (selections) | NIST/LDC | audio files and transcripts for 16 speakers |
| VerbNet 2.1 | Palmer et al | 5k verbs, hierarchically organized, linked to WordNet |
| Wordlist Corpus | OpenOffice.org et al | 960k words and 20k affixes for 8 languages |
| WordNet 3.0 (English) | Miller, Fellbaum | 145k synonym sets |

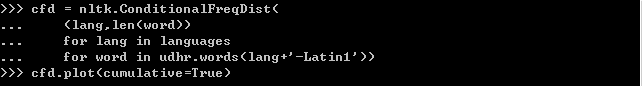
其他语料库

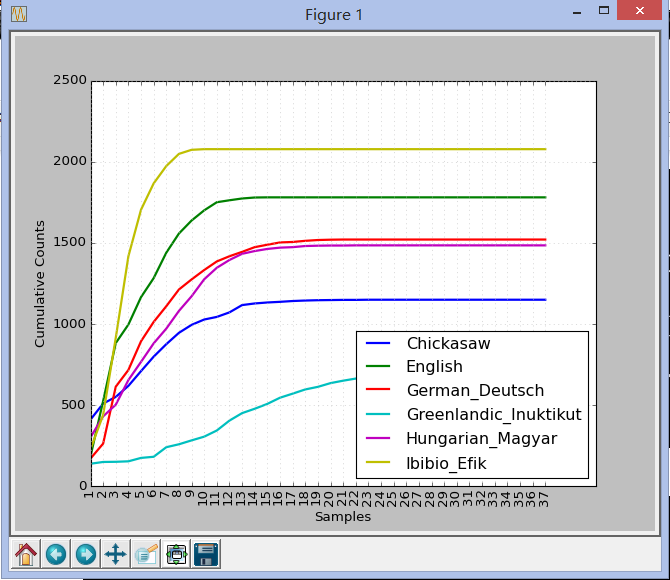
nltk包含多国语言语料库，需要学习字符编码处理。



udhr是世界人权宣言，fileids包括了所有文件所使用的字符编码信息。

我们可以看看不同语言的词长不同





如何处理中文？中文是UFT8，上面的做法并不适用

**Table 1.3**:

Basic Corpus Functionality defined in NLTK: more documentation can be found using help(nltk.corpus.reader) and by reading the online Corpus HOWTO at http://nltk.org/howto.

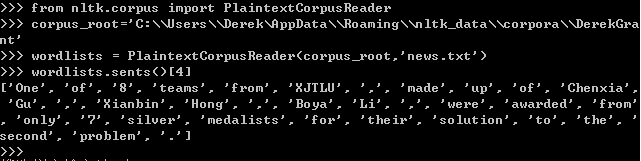
|  |  |
| --- | --- |
| Example | Description |
| fileids() | the files of the corpus |
| fileids([categories]) | the files of the corpus corresponding to these categories |
| categories() | the categories of the corpus |
| categories([fileids]) | the categories of the corpus corresponding to these files |
| raw() | the raw content of the corpus |
| raw(fileids=[f1,f2,f3]) | the raw content of the specified files |
| raw(categories=[c1,c2]) | the raw content of the specified categories |
| words() | the words of the whole corpus |
| words(fileids=[f1,f2,f3]) | the words of the specified fileids |
| words(categories=[c1,c2]) | the words of the specified categories |
| sents() | the sentences of the whole corpus |
| sents(fileids=[f1,f2,f3]) | the sentences of the specified fileids |
| sents(categories=[c1,c2]) | the sentences of the specified categories |
| abspath(fileid) | the location of the given file on disk |
| encoding(fileid) | the encoding of the file (if known) |
| open(fileid) | open a stream for reading the given corpus file |
| root | if the path to the root of locally installed corpus |
| readme() | the contents of the README file of the corpus |

<http://www.nltk.org/book/ch02.html>

载入你自己的语料库

我把news.txt放在C:\Users\Derek\AppData\Roaming\nltk\_data\corpora\DerekGrant中

按一下方法载入这篇新闻，并提取其中的一个句子。



这节好长，总算完了