复旦大学大数据学院
School of Data Science, Fudan University

NLTK语料库及SRILM简介

刘乾龙

2017年9月27日

NLTK语料库

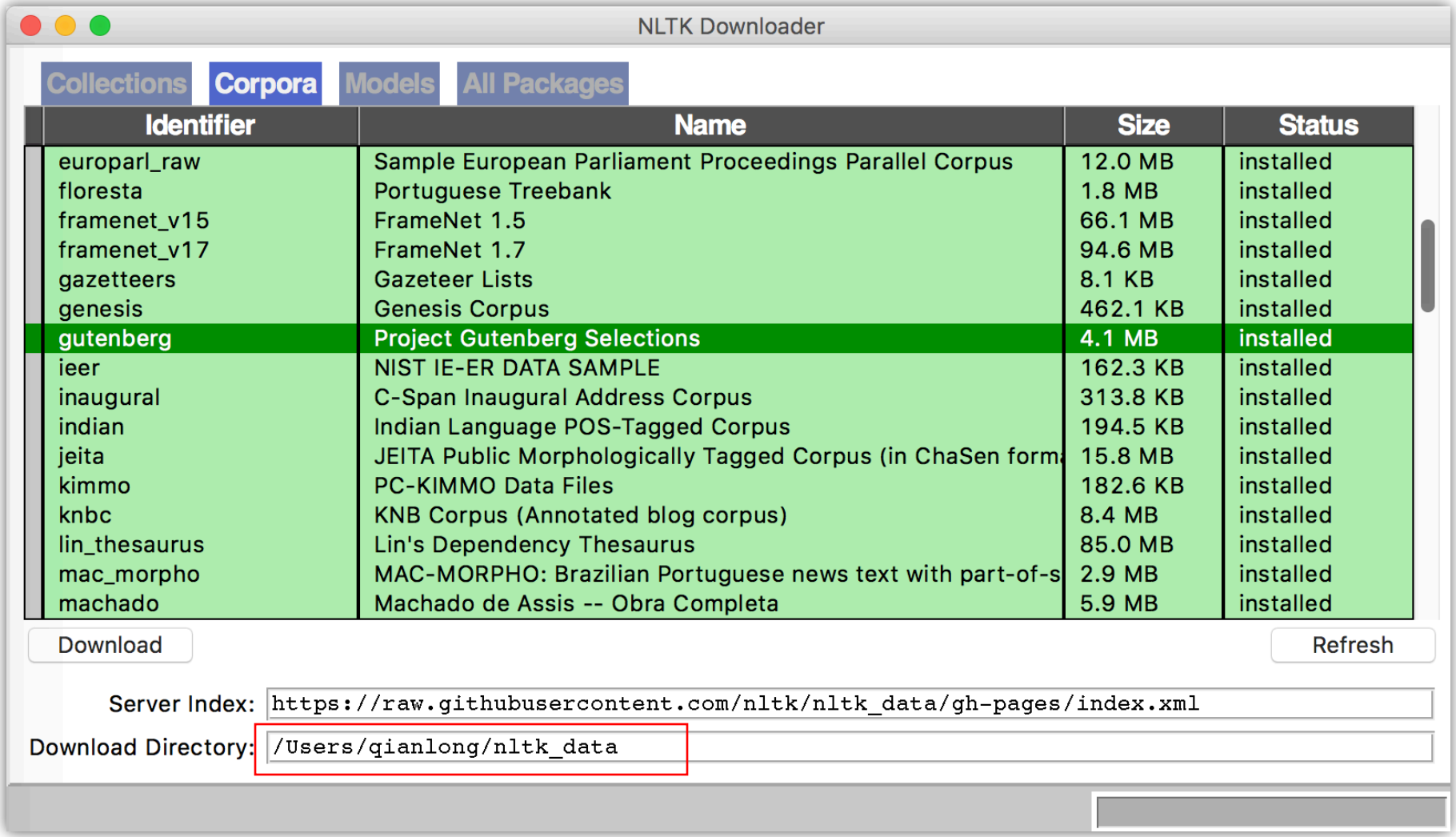
SRILM简介

1. NLTK语料库

- 古腾堡语料库
- 布朗语料库
- 路透社语料库
- 就职演说语料库
- 网络文本和聊天文本
- 其他语言语料
- 文本语料结构
- 加载自己的语料

1. NLTK语料库


```
>>> import nltk
>>> nltk.download()
showing info https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/index.xml
```



1.1 古腾堡语料库

- 古腾堡项目：超过54000本免费电子图书
  - www.gutenberg.org
  - www.gutenberg.org/robot/harvest

```
>>> import nltk
>>> books = nltk.corpus.gutenberg.fileids()
>>> books
['austen-emma.txt', 'austen-persuasion.txt', 'austen-sense.txt', 'bible-kjv.txt'
, 'blake-poems.txt', 'bryant-stories.txt', 'burgess-busterbrown.txt', 'carroll-a
lice.txt', 'chesterton-ball.txt', 'chesterton-brown.txt', 'chesterton-thursday.t
xt', 'edgeworth-parents.txt', 'melville-moby_dick.txt', 'milton-paradise.txt', '
shakespeare-caesar.txt', 'shakespeare-hamlet.txt', 'shakespeare-macbeth.txt', 'w
hitman-leaves.txt']
>>> len(books)
18
>>> emma = nltk.corpus.gutenberg.words("austen-emma.txt")
>>> len(emma)
192427
>>> emma[0:20]
['[', 'Emma', 'by', 'Jane', 'Austen', '1816', ']', 'VOLUME', 'I', 'CHAPTER', 'I'
, 'Emma', 'Woodhouse', ',', 'handsome', ',', 'clever', ',', 'and', 'rich']
```

```
>>> from nltk.corpus import gutenberg
>>> emma = gutenberg.words("austen-emma.txt")
>>> len(emma)
192427
```

1.1 古腾堡语料库

- 以单词为单位获取文本
- 以句子为单位获取文本
- 以纯文本的形式获取文本

```
>>> emma_raw = gutenberg.raw("austen-emma.txt")
>>> emma_sent = gutenberg.sents("austen-emma.txt")
>>> emma_raw[0:30]
'[Emma by Jane Austen 1816]\n\nVO'
>>> emma_sent[3:5]
[['Emma', 'Woodhouse', ',', 'handsome', ',', 'clever', ',', 'and', 'rich', ',',
'with', 'a', 'comfortable', 'home', 'and', 'happy', 'disposition', ',', 'seemed'
, 'to', 'unite', 'some', 'of', 'the', 'best', 'blessings', 'of', 'existence', ';
', 'and', 'had', 'lived', 'nearly', 'twenty', '-', 'one', 'years', 'in', 'the',
'world', 'with', 'very', 'little', 'to', 'distress', 'or', 'vex', 'her', '.'], [
'She', 'was', 'the', 'youngest', 'of', 'the', 'two', 'daughters', 'of', 'a', 'mo
st', 'affectionate', ',', 'indulgent', 'father', ';', 'and', 'had', ',', 'in', '
consequence', 'of', 'her', 'sister', "'", 's', 'marriage', ',', 'been', 'mistres
s', 'of', 'his', 'house', 'from', 'a', 'very', 'early', 'period', '.']]
```

1.2 布朗语料库

- 1961年由布朗大学创建，是第一个百万词语级别的英语语料库;
- 平衡语料库，包含15个类别的文本;

ID	File	Genre	Description
A16	ca16	news	Chicago Tribune: *Society Reportage*
B02	cb02	editorial	Christian Science Monitor: *Editorials*
C17	cc17	reviews	Time Magazine: *Reviews*
D12	cd12	religion	Underwood: *Probing the Ethics of Realtors*
E36	ce36	hobbies	Norling: *Renting a Car in Europe*
F25	cf25	lore	Boroff: *Jewish Teenage Culture*
G22	cg22	belles_lettres	Reiner: *Coping with Runaway Technology*
H15	ch15	government	US Office of Civil and Defence Mobilization: *The Family Fallout Shelter*
J17	cj19	learned	Mosteller: *Probability with Statistical Applications*
K04	ck04	fiction	W.E.B. Du Bois: *Worlds of Color*
L13	cl13	mystery	Hitchens: *Footsteps in the Night*
M01	cm01	science_fiction	Heinlein: *Stranger in a Strange Land*
N14	cn15	adventure	Field: *Rattlesnake Ridge*
P12	cp12	romance	Callaghan: *A Passion in Rome*
R06	cr06	humor	Thurber: *The Future, If Any, of Comedy*

- 一次获得一个文件的内容
- 同时获得某一类文件的内容
- 同时获得多个文件或多个类别的内容

```
>>> brown.words(fileids=["ca01"])
[u'The', u'Fulton', u'County', u'Grand', u'Jury', ...]
>>> brown.words(categories="news")
[u'The', u'Fulton', u'County', u'Grand', u'Jury', ...]
>>> brown.sents(categories=["news","editorial","reviews"])
[[u'The', u'Fulton', u'County', u'Grand', u'Jury', u'said', u'Friday', u'an', u'
investigation', u'of', u"Atlanta's", u'recent', u'primary', u'election', u'produ
ced', u'``', u'no', u'evidence', u"''", u'that', u'any', u'irregularities', u'to
ok', u'place', u'.'], [u'The', u'jury', u'further', u'said', u'in', u'term-end',
 u'presentments', u'that', u'the', u'City', u'Executive', u'Committee', u',', u'
which', u'had', u'over-all', u'charge', u'of', u'the', u'election', u',', u'``',
 u'deserves', u'the', u'praise', u'and', u'thanks', u'of', u'the', u'City', u'of
', u'Atlanta', u"''", u'for', u'the', u'manner', u'in', u'which', u'the', u'elec
tion', u'was', u'conducted', u'.'], ...]
```

Reuters Corpus: http://disi.unitn.it/moschitti/corpora.htm

- RCV1：Reuters Corpus, Volume 1
  - 英语，810 000篇新闻报道，对新闻进行人工分类；
- RCV2：Reuters Corpus, Volume 2
  - 13语言，487 000篇新闻报道，非平行语料库；

1.3 路透社语料库

- 路透社语料库：10788个新闻文档，130万词;
- 按照内容分成90个类别，同时分为“训练集”(7769)和“测试集”(3019);
- 不同类别之间会有重叠;

```
>>> from nltk.corpus import reuters
>>> reuters_fileids = reuters.fileids()
>>> len(reuters_fileids)
10788
>>> reuters_categories = reuters.categories()
>>> len(reuters_categories)
90
>>> reuters_fileids[0:5]
['test/14826', 'test/14828', 'test/14829', 'test/14832', 'test/14833']
```

- 获得一个文件或多个文件所属类别；
- 获得一个类别或多个类别包含的文件；

```
>>> reuters.categories(["training/9880"])
['money-fx']
>>> reuters.categories("training/9865")
['barley', 'corn', 'grain', 'wheat']
>>> reuters.categories(["training/9865","training/9880"])
['barley', 'corn', 'grain', 'money-fx', 'wheat']
>>> reuters.fileids("barley")[0:5]
['test/15618', 'test/15649', 'test/15676', 'test/15728', 'test/15871']
>>> reuters.fileids(["barley","corn"])[0:5]
['test/14832', 'test/14858', 'test/15033', 'test/15043', 'test/15106']
```

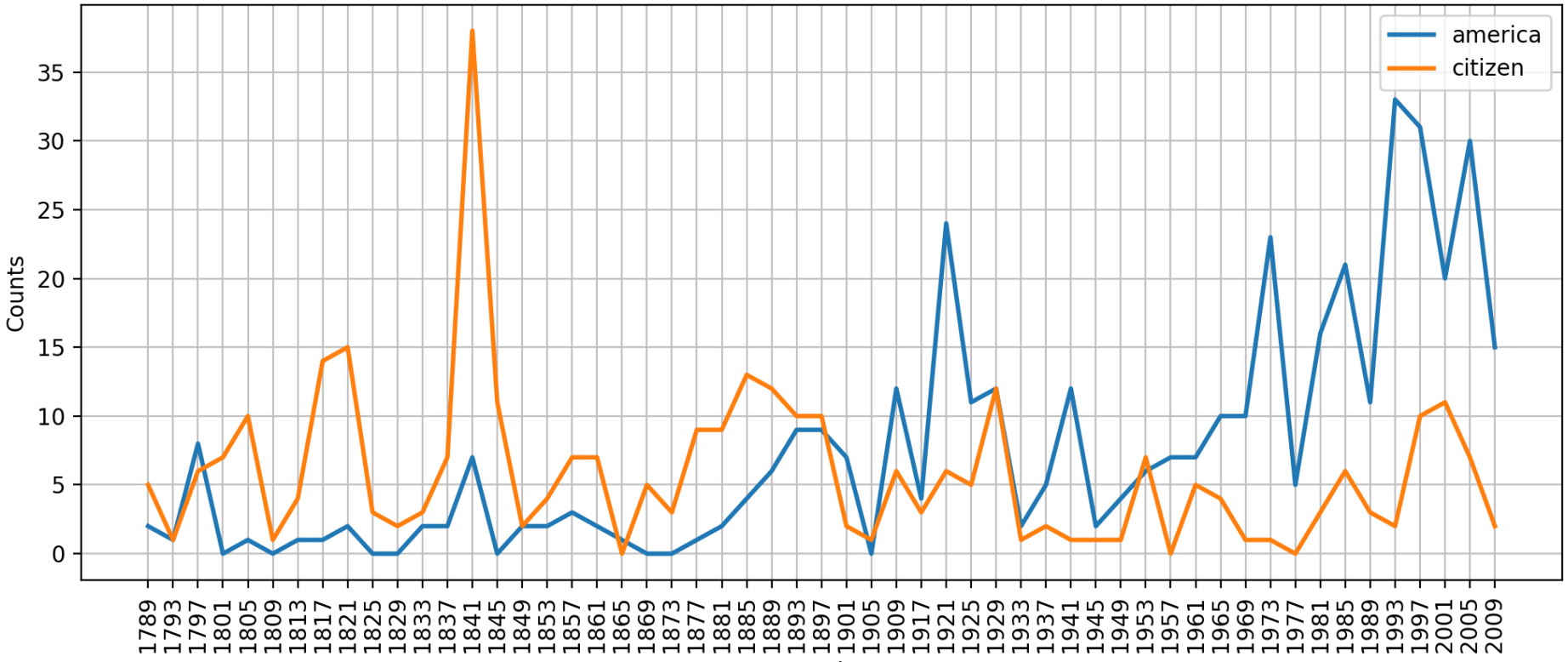
1.4 就职演说语料库

- 美国历史上56届总统的就职演说；

```
>>> from nltk.corpus import inaugural
>>> address = inaugural.fileids()
>>> address[0:5]
['1789-Washington.txt', '1793-Washington.txt', '1797-Adams.txt', '1801-Jefferson.t
xt', '1805-Jefferson.txt']
>>> len(address)
56
```

- “America”“citizen”两个词的使用随着时间的变化是如何变化的？

```
>>> 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()
```



- 包含：Firefox交流论坛、《加勒比海盗》电影剧本、个人广告、葡萄酒评论等。

```
>>> from nltk.corpus import webtext
>>> 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 ...
```

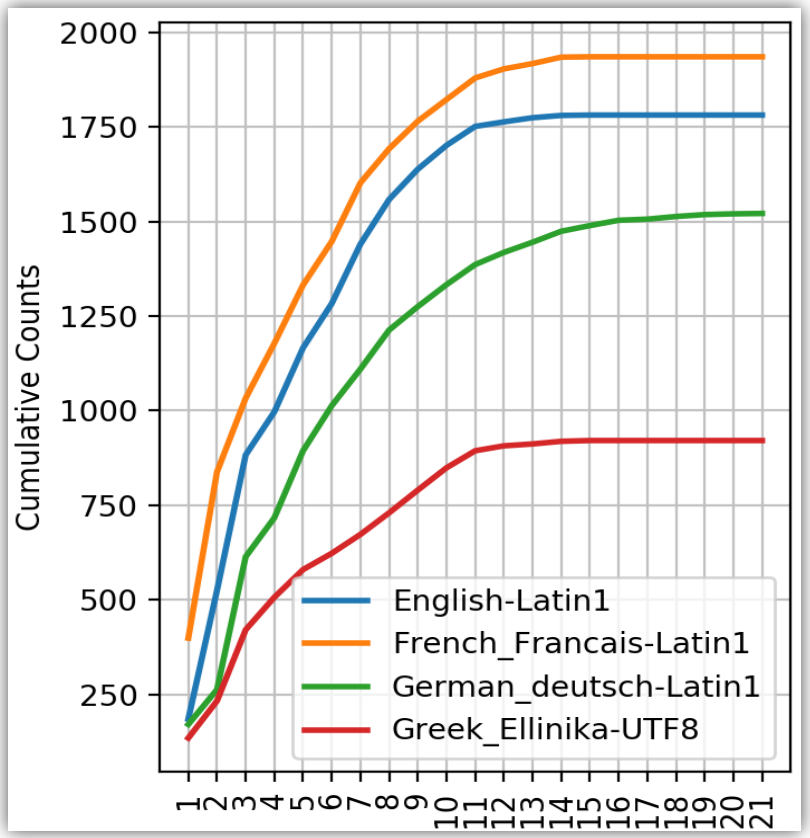
1.5 网络文本和聊天文本

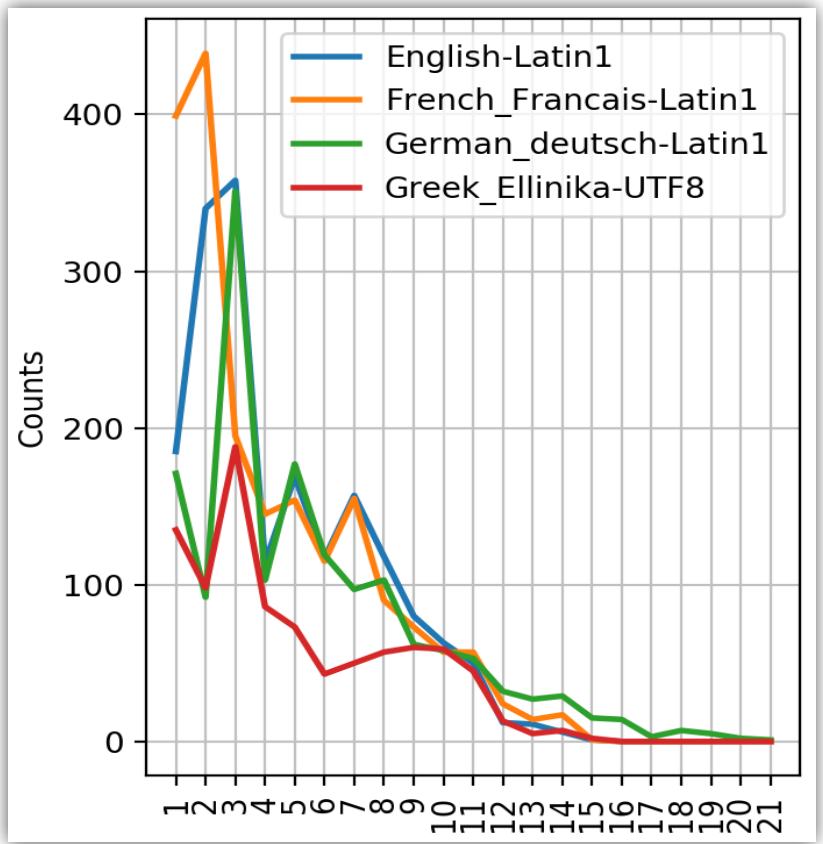
- 即时消息聊天会话语料库，最初由美国海军学院为研究自动检测互联网虐童而收集的。
- 以年龄阶段划分网络聊天室(teens, 20s, 30s, 40s, a generic adults)
- “10-19-20s_706posts.xml”:表示2006年10月19日从20s聊天室收集而来的706个帖子。

```
>>> from nltk.corpus import nps_chat
>>> nps_chat.fileids()
['10-19-20s_706posts.xml', '10-19-30s_705posts.xml', '10-19-40s_686posts.xml', '10
-19-adults_706posts.xml', '10-24-40s_706posts.xml', '10-26-teens_706posts.xml', '1
1-06-adults_706posts.xml', '11-08-20s_705posts.xml', '11-08-40s_706posts.xml', '11
-08-adults_705posts.xml', '11-08-teens_706posts.xml', '11-09-20s_706posts.xml', '1
1-09-40s_706posts.xml', '11-09-adults_706posts.xml', '11-09-teens_706posts.xml']
>>> chatroom = nps_chat.posts("10-19-20s_706posts.xml")
>>> chatroom[123]
['i', 'do', "n't", 'want', 'hot', 'pics', 'of', 'a', 'female', ',', 'I', 'can', 'l
ook', 'in', 'a', 'mirror', '.']
```

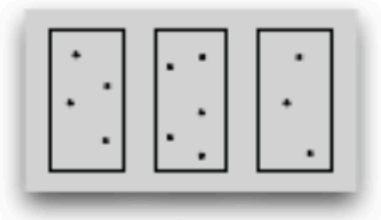
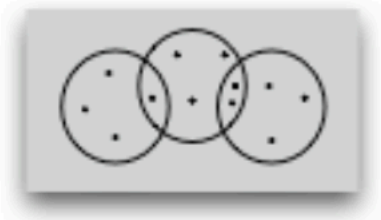
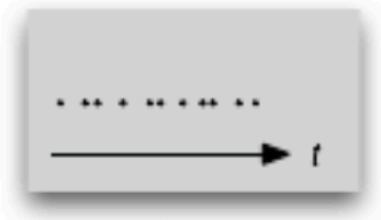
1.6 其他语言语料库


```
>>> import nltk
>>> from nltk.corpus import udhr
>>> languages = ['Greek_Ellinika-UTF8','English-Latin1','German_deutsch-Latin1',
'French_Francais-Latin1']
>>> cfd = nltk.ConditionalFreqDist(
...     (lang, len(word))
...     for lang in languages
...     for word in udhr.words(lang))
>>> cfd.plot(cumulative=True)
>>> cfd.plot()
```





常见结构：
孤立、分类、重叠、时序

孤立	分类	重叠	时序
			
古腾堡语料库	布朗语料库	路透社语料库	就职演说

1.7 其他语言语料库

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

1.8 加载自己的语料库

- 第一个参数：自己语料的目录
- 第二个参数：
  - 包含文件名称的list
  - 正则表达式匹配文件名称

```
>>> from nltk.corpus import PlaintextCorpusReader
>>> corpus_root = "/Users/qianlong/Desktop/my_corpus/"
>>> my_corpus = PlaintextCorpusReader(corpus_root, ".*")
>>> my_corpus.fileids()
['text-1.txt', 'text-2.txt', 'text-3.txt']
>>> text1 = my_corpus.words("text-1.txt")
>>> text1[:5]
['Note', 'Important', ':', 'Our', 'inline']
```

- CNN：
  - 来自于CNN的90000新闻报道；
  - 对应380 000个问题；
- Daily Mail
  - 来自Daily Mail的197 000新闻报道；
  - 对应897 000个问题；

- 平均每一篇文章对应4个问题；
- 每一个问题是缺失了一个单词或词组的一句话，而对应的答案可以在相应的文章中找到；

http://cs.nyu.edu/~kcho/DMQA/

NLTK语料库

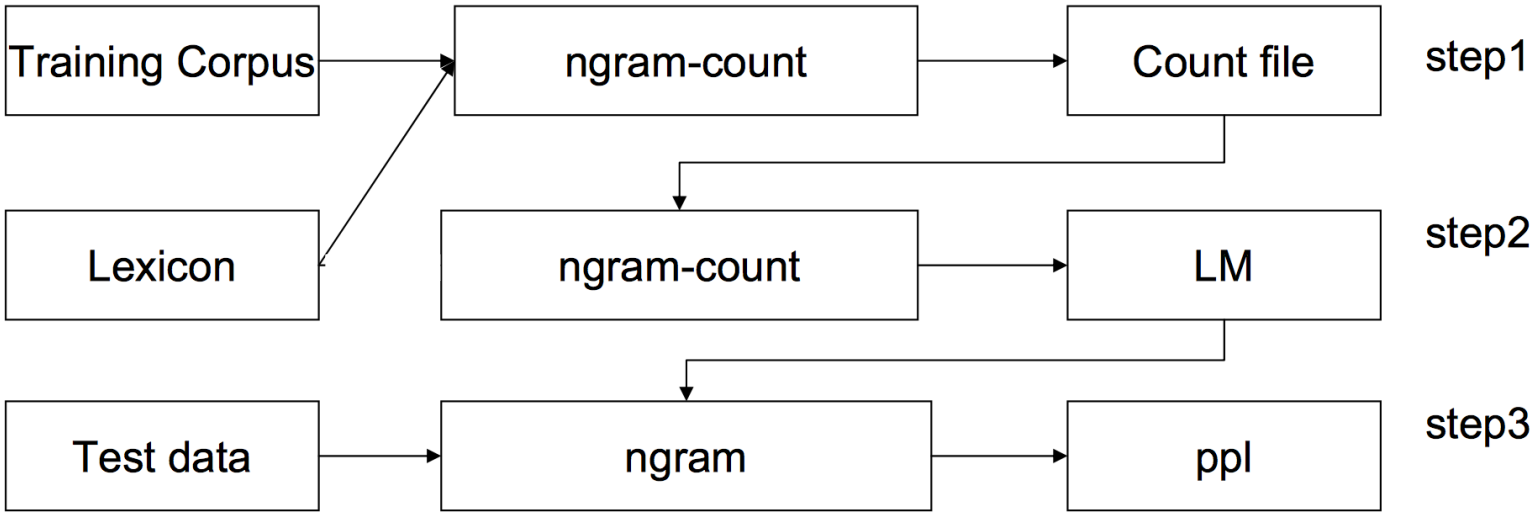
SRILM简介

- 诞生于1995年，有Stanford Research Institute（SRI）实验室负责开发维护；
- 主要应用于统计语言模型，包括语音识别、统计标注以及机器翻译；

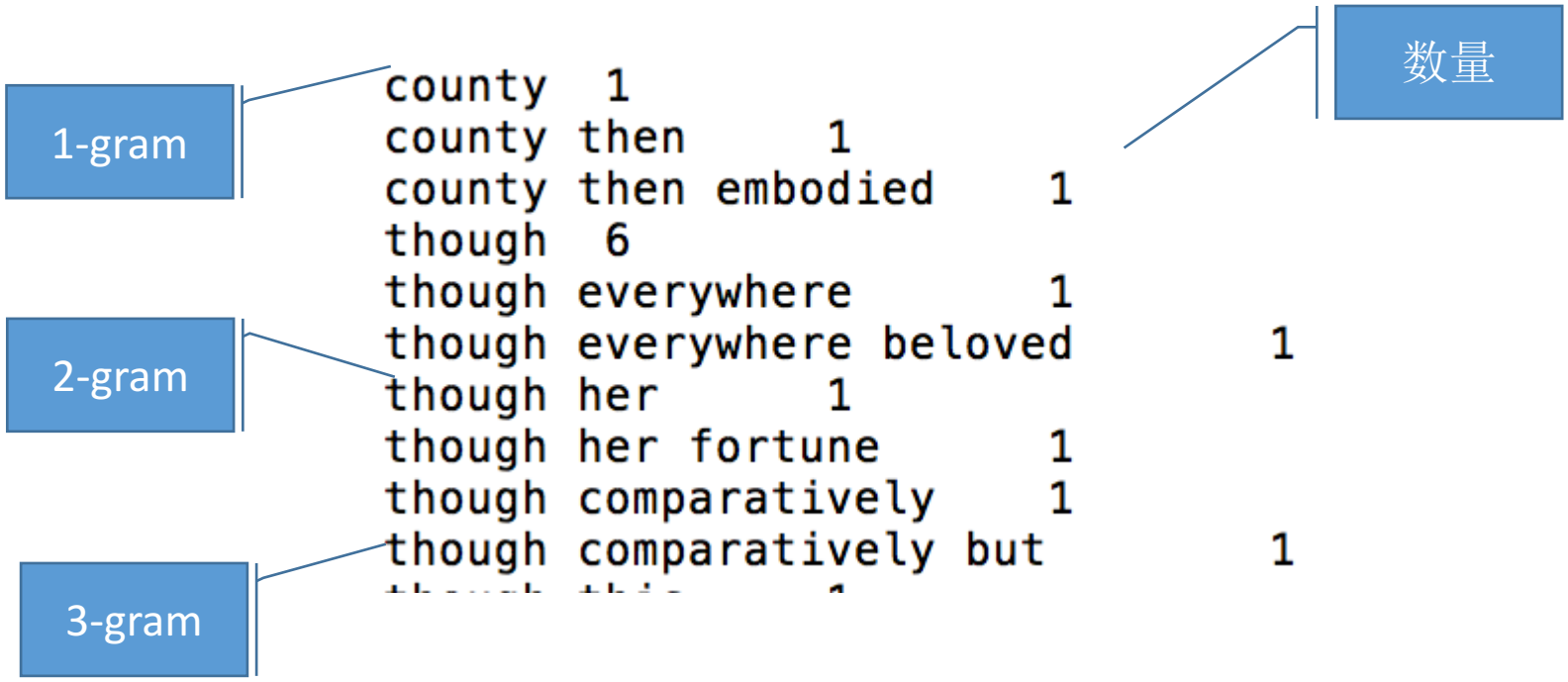
- SRILM的主要目标是支持语言模型的估计和评估。
  - Estimate：估计是从训练数据中，使用最大似然估计和相应的平滑算法得到一个语言模型，ngram-count工具实现；
  - Evaluate：评估是利用测试集计算模型困惑度，ngram工具实现。

三个主要功能：

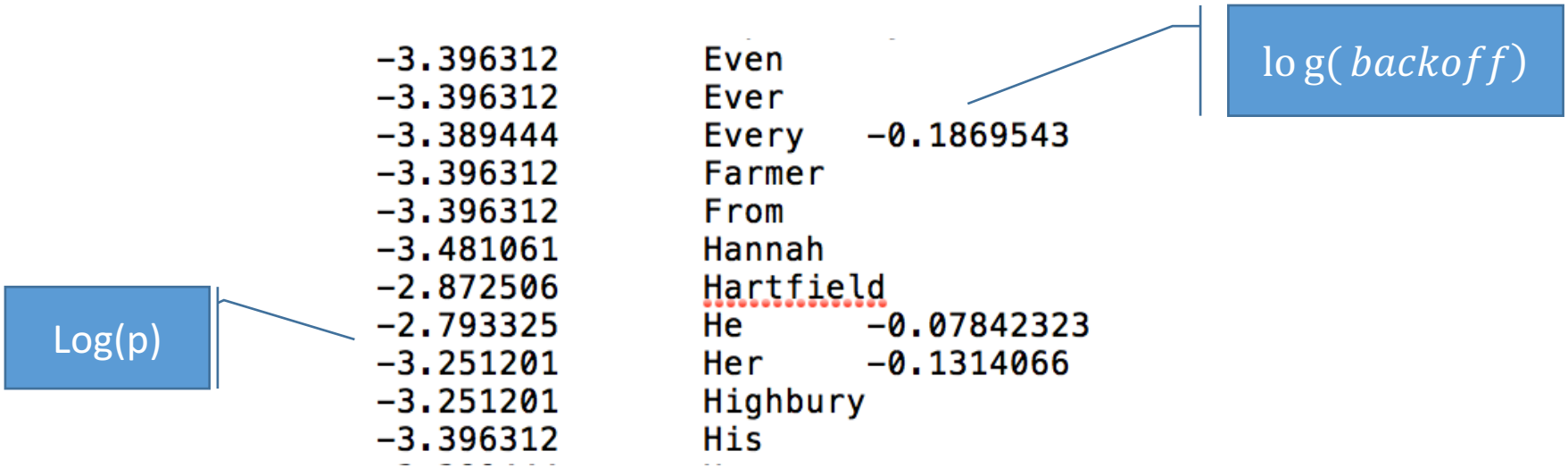
- 从训练集中产生包含每个ngram的计数文件；
- 使用上述的计数文件产生模型；
- 使用上述的模型计算测试集的困惑度；



- ./ngram-count -text yourTextFile -write countFileSaveTo
  - -order 3
  - -no-sos -no-eos



- ./ngram-count -read yourCountFile –lm modelSaveTo -interpolate -kndiscount
  - -addsmooth 0



- ./ngram -lm yourModelFile -ppl yourTestFile

```
~/Downloads/srilm-1.7.2/bin/macosx ./ngram -lm ~/Desktop/LM/brown.lm -ppl ~/De
sktop/LM/test.txt
file /Users/qianlong/Desktop/LM/test.txt: 3869 sentences, 72214 words, 43174 OOVs
0 zeroprobs, logprob= -114493.3 ppl= 3013.622 ppl1= 8762.095
```

```
~/Downloads/srilm-1.7.2/bin/macosx ./ngram-count -help
Usage of command "./ngram-count"
 -version:                  print version information
 -order:                    max ngram order
                Default value: 3
 -varprune:                 pruning threshold for variable order ngrams
                Default value: 0
 -debug:                    debugging level for LM
                Default value: 0
 -recompute:                recompute lower-order counts by summation
 -sort:                     sort ngrams output
 -write-order:              output ngram counts order
                Default value: 0
 -tag:                      file tag to use in messages
 -text:                     text file to read
 -text-has-weights:         text file contains count weights
 -no-sos:                   don't insert start-of-sentence tokens
 -no-eos:                   don't insert end-of-sentence tokens
 -read:                     counts file to read
 -intersect:                intersect counts with this file
 -read-with-mincounts:      apply minimum counts when reading counts file
 -read-google:              Google counts directory to read
 -write:                    counts file to write
 -write1:                   1gram counts file to write
 -write2:                   2gram counts file to write
```

Thank you!