

NLP学习 笔记

Contents

1	语料库	2
1.1	wiki简体中文语料	2
1.1.1	获取	2
1.1.2	预处理	2
1.1.3	优点	3
1.2	复旦语料库	3
1.3	搜狗语料库精简版	3

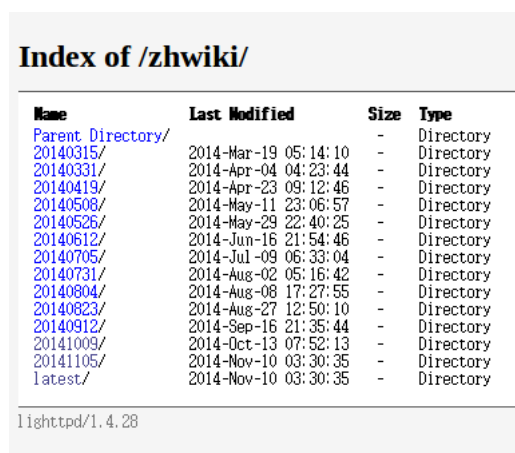
1 语料库

自然语言处理显然少不了使用语料库,这里使用的语料库有kiki简体中文语料,复旦大学,搜狗.下面重点介绍一下这三个语料库的获取和预处理.

1.1 wiki简体中文语料

1.1.1 获取

wiki语料获取非常方便,链接是<http://download.wikipedia.com/zhwiki/>,在该目录下面(见图1)可以找到wiki近期所有的中文语料库,我们使用的20141105目录下面的`zhwiki-20141105-pages-articles-multistream.xml.bz2`文件,这个压缩包有1.1G,里面存放的是标题和正文部分,如果需要其他数据,如页面跳转或历史编辑记录等,可以在当前页面上找对应的链接.



Name	Last Modified	Size	Type
Parent Directory/		-	Directory
20140315/	2014-Mar-19 05:14:10	-	Directory
20140331/	2014-Apr-04 04:23:44	-	Directory
20140419/	2014-Apr-23 09:12:46	-	Directory
20140508/	2014-May-11 23:06:57	-	Directory
20140526/	2014-May-29 22:40:25	-	Directory
20140612/	2014-Jun-16 21:54:46	-	Directory
20140705/	2014-Jul-09 06:33:04	-	Directory
20140731/	2014-Aug-02 05:16:42	-	Directory
20140804/	2014-Aug-08 17:27:55	-	Directory
20140823/	2014-Aug-27 12:50:10	-	Directory
20140912/	2014-Sep-16 21:35:44	-	Directory
20141009/	2014-Oct-13 07:52:13	-	Directory
20141105/	2014-Nov-10 03:30:35	-	Directory
latest/	2014-Nov-10 03:30:35	-	Directory

lighttpd/1.4.28

Figure 1: wiki下载目录

1.1.2 预处理

对wiki中文语料库的预处理有两步.

1. 抽取正文文本

```
1 hjy@hgy:yourworkspace$ bzcat zhwiki-latest-pages-articles.xml.bz2 | python WikiExtractor.py -b1000M -o extracted >output.txt
```

上面命令中**`bzcat`**是将**`.bz2`**文件中的**`.xml`**文件解压并将里面文本显示到终端,这里用管道|将该文本作为下个python程序的输入,**`-b1000m`**表示以1000M为单位切分文件,默认是500K,由于最后生成的正文文本不到600M,把参数设置的大一些可以保证最后的抽取结果全部存在一个文件里.执行命令后,获得两个文件,一个是**`extracted/AA`**目录下的**`wiki_00`**文件,里面的文本格式如下

```
1 <doc id="*" url="*" title="TITLE">
2 TITLE
3
4 TEXT
5 </doc>
```

另一个是`output.txt`,它是对`wiki_00`中所有的id和title进行了提取,每一对占一行.

2. 繁简转换

wiki中文语料库中是简繁混杂的,里面包含大陆简体、台湾繁体、港澳繁体等多种不同的数据。这里使用开源项目`openccc`进行处理,其在ubuntu上的安装使用如下命令.

```
1 hjy@h jy: yourworkspace$ sudo apt-get install openccc
```

接着就可以使用该命令对`wiki_00`和`output.txt`进行繁简转换.

```
1 hjy@h jy: yourworkspace$ openccc -i wiki_00 -o wiki_chs -c
zht2zhs.ini
```

最终我们获得`wiki_chs`和`wiki_output_chs.txt`两个文件.

1.1.3 优点

优点有如下三点:

1. 资源获取非常方便,相比之下,其他很多语料都需要用爬虫抓取,或者付费获得。
2. 文档解析有非常多的成熟工具,直接使用开源工具即可完成正文的提取。
3. 维基百科的质量较高,而且领域广泛

1.2 复旦语料库

复旦语料库的下载地址<http://www.nlpir.org/?action-viewnews-itemid-103>

1.3 搜狗语料库精简版

搜狗语料库的下载地址<http://www.sogou.com/labs/dl/c.html>