维基百科中文语料的获取

2016-08-31 | 驾驭文本

最近做实验需要较大规模的中文语料，听说维基百科的中文数据就比较不错。使用维基百科做训练语料有很多好处：

- 维基百科资源获取非常方便，有Wiki Dump可以直接下载。其他可能需要用爬虫抓取或者付费。
- 维基百科的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。
- 维基百科的质量较高，而且领域广泛。

> 缺点也有：最主要的就是数量较少，相比国内的百度百科、互动百科等，数据量要少一个数量级。

1 下载中文Wiki Dump

中文Wiki Dump的链接是：https://dumps.wikimedia.org/zhwiki/。我需要下载的是zhwiki-latest-pages-articles.xml.bz2。这个压缩包里面存的是标题、正文部分，如果需要其他数据，如页面跳转、历史编辑记录等，可以到目录下找别的下载链接。

2 使用Wikipedia Extractor抽取正文

Wikipedia Extractor是意大利人用Python写的一个维基百科抽取器，使用非常方便。下载之后直接使用这条命令即可完成抽取，运行了大约半小时的时间。

```
$ sudo apt-get install unzip python python-dev python-pip
$ git clone https://github.com/attardi/wikiextractor.git wikiextractor
$ cd wikiextractor
$ sudo python setup.py install
$ ./WikiExtractor.py -b 500M -o extracted zhwiki-latest-pages-articles.xml.bz2
```

参数 -b 500M 表示以 500M 为单位切分文件，默认是1M。由于最后生成的正文文本约1144M，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。

3 繁简转换

维基百科的中文数据是繁简混杂的，里面包含大陆简体、台湾繁体、港澳繁体等多种不同的数据。有时候在一篇文章的不同段落间也会使用不同的繁简字。

解决这个问题最佳的办法应该是直接使用维基百科自身的繁简转换方法URL。不过维基百科网站虽然是开源的，但要把里面的繁简转换功能拆解出来，有一定的难度。

为了方便起见，我直接使用了开源项目opencc。参照安装说明的方法，安装完成之后，使用下面的命令进行繁简转换，整个过程大约需要1分钟：

```
$ sudo apt-get install opencc
$ opencc -i wiki_00 -o zh_wiki_00 -c zht2zhs.ini
```

```
$ opencc -i wiki_01 -o zh_wiki_01 -c zht2zhs.ini
$ opencc -i wiki_02 -o zh_wiki_02 -c zht2zhs.ini
```

命令中的 wiki_00 这个文件是此前使用 Wikipedia Extractor 得到的。

到此为止，已经完成了大部分繁简转换工作。实际上，维基百科使用的繁简转换方法是以词表为准，外加人工修正。人工修正之后的文字是这种格式，多数是为了解决各地术语名称不同的问题：

```
他的主要成就包括Emacs及後來的GNU Emacs，GNU C 編譯器及-{zh-hant:GNU 除錯器;zh-hans:GDB 调试器}-。
```

对付这种可以简单的使用正则表达式来解决。一般简体中文的限定词是 zh-hans 或 zh-cn ，在C#中用以下代码即可完成替换：

```
s = Regex.Replace(s, @"-\{.*?(zh-hans|zh-cn):([^;]*?)(;.*?)?\}-", @"$2")
```

在Python中是这样的：参考资料

```
1  import re
2  p = re.compile(ur'-\{.*?(zh-hans|zh-cn):([^;]*?)(;.*?)?\}-')
3  s = '他的主要成就包括Emacs及後來的GNU Emacs，GNU C 編譯器及-{zh-hant:GNU 除錯器;zh-hans:GDB 调试器}-。
4  print p.sub(ur'\2', s)
```

由于Wikipedia Extractor抽取正文时，会将有特殊标记的外文直接剔除，最后形成类似这样的正文：

西方语言中“数学”（；）一词源自于古希腊语的（）

虽然上面这句话是读不通的，但鉴于这种句子对我要处理的问题影响不大，就暂且忽略了。最后再将 「」『』 这些符号替换成引号，顺便删除空括号，就大功告成了！代码如下：

```
#!/usr/bin/python
# -*- coding: utf-8 -*-
# python exec.py zh_wiki_00
# python exec.py zh_wiki_01
# python exec.py zh_wiki_02
import re
import sys
import codecs

def myfun(input_file):
    p1 = re.compile(ur'-\{.*?(zh-hans|zh-cn):([^;]*?)(;.*?)?\}-')
    p2 = re.compile(ur'[ ( \(][ ， ；。 ？ ！\s]*[ ) \)]')
    p3 = re.compile(ur'[「『]')
    p4 = re.compile(ur'[」』]')
    outfile = codecs.open('std_' + input_file, 'w', 'utf-8')
    with codecs.open(input_file, 'r', 'utf-8') as myfile:
        for line in myfile:
            line = p1.sub(ur'\2', line)
            line = p2.sub(ur'', line)
            line = p3.sub(ur'“', line)
            line = p4.sub(ur'”', line)
            outfile.write(line)
    outfile.close()

if __name__ == '__main__':
    if len(sys.argv) != 2:
        print "Usage: python script.py inputfile"
```

```
28              sys.exit()
29          reload(sys)
30          sys.setdefaultencoding('utf-8')
31          input_file = sys.argv[1]
32          myfun(input_file)
```

4 号外：Python处理超大文件

使用with关键词：

```
1   with open('somefile.txt', 'r') as myfile:
2       for line in myfile:
3           # operation
```

使用fileinput模块：

```
1   import fileinput
2   for line in fileinput.input('myfile'):
3       do_something(line)
```

建立缓存，控制缓冲区大小：

```
1   BUFFER = int(10E6) #10 megabyte buffer
2   myfile = open('somefile.txt', 'r')
3   lines = myfile.readlines(BUFFER)
4   while lines:
5       for line in lines:
```

```
        # operation
    lines = myfile.readlines(BUFFER)
myfile.close()
```

5 使用gensim处理wiki语料

只能说是多学一条路子，目前效果还不太好，比如有很多奇怪的英文字符和丢失一些内容等情况，还有就是获得的text仍需要进行繁简转换：安装指南

```
#!/usr/bin/python
# -*- coding: utf-8 -*-
import sys
import codecs
import logging
from gensim.corpora.wikicorpus import WikiCorpus

if __name__ == '__main__':
    if len(sys.argv) < 3:
        print "Usage: python script.py inputfile outputfile"
        sys.exit()
    reload(sys)
    sys.setdefaultencoding('utf-8')
    logging.basicConfig(format='%(asctime)s : %(levelname)s : %(message)s', level=logging.INFO)
    infile, outfile = sys.argv[1:3]
    output = codecs.open(outfile, 'w', 'utf-8')
    wiki = WikiCorpus(infile, lemmatize=False, dictionary={})
    for text in wiki.get_texts():
        output.write(" ".join(text) + '\n')
    output.close()
```

运行脚本你将得到一份纯文本格式的语料，每行对应一篇文章，去掉了标点符号等内容，并以空格作为连接符。最终抽取出了268740篇文章，命令行：

```
$ python exec.py zhwiki-latest-pages-articles.xml.bz2 wiki.zh.text
$ head -n 1 wiki.zh.text
歐幾里得 西元前三世紀的希臘數學家 現在被認為是幾何之父 ..
```

此文非原创，感谢原作者无私分享。原文链接

#Ubuntu #语料 #维基百科

‹ 如何计算两个文档的相似度 中文维基百科语料上的Word2Vec实验 ›