6 Jan 获取并处理中文维基百科语料

By 苏剑林 | 2017-01-06 | 12665位读者 引用

中文语料库中，质量高而又容易获取的语料库，应该就是维基百科的中文语料了，而且维基百科相当厚道，每个月都把所有条目都打包一次（下载地址在这里：https://dumps.wikimedia.org/zhwiki/），供全世界使用，这才是真正的“取之于民，回馈于民”呀。遗憾的是，由于天朝的无理封锁，中文维基百科的条目到目前只有91万多条，而百度百科、互动百科都有千万条了（英文维基百科也有上千万了）。尽管如此，这并没有阻挡中文维基百科成为几乎是最高质量的中文语料库。（百度百科、互动百科它们只能自己用爬虫爬取，而且不少记录质量相当差，几乎都是互相复制甚至抄袭。）

门槛

尽量下载很容易，但是使用维基百科语料还是有一定门槛的。直接下载下来的维基百科语料是一个带有诸多html和markdown标记的文本压缩包，基本不能直接使用。幸好，已经有热心的高手为我们写好了处理工具，主要有两个：1、Wikipedia Extractor；2、gensim的wikicorpus库。它们都是基于python的。

然而，这两个主流的处理方法都不能让我满意。首先，Wikipedia Extractor提取出来的结果，会去掉{{}}标记的内容，这样会导致下面的情形

> 西方语言中“数学”（；）一词源自于古希腊语的（）

这是因为括号里的词带有{{}}标记，被清空了；而按照网上的教程，直接用gensim.corpora.wikicorpus.WikiCorpus处理，问题更严重，因为它连所有标点都去掉了。对于追求一份高质量语料库的、具有强迫症的笔者来说，这都是不能接受的。因此，自己动手结合gensim，写了一个处理脚本。

代码

```
from gensim.corpora.wikicorpus import extract_pages,filter_wiki
import bz2file
import re
import opencc
from tqdm import tqdm
import codecs

wiki = extract_pages(bz2file.open('zhwiki-latest-pages-articles.xml.bz2'))

def wiki_replace(d):
    s = d[1]
    s = re.sub(':*{\|[\s\S]*?\|}', '', s)
    s = re.sub('<gallery>[\s\S]*?</gallery>', '', s)
    s = re.sub('(.){{([^{}\n]*?\|[^{}\n]*?)}}', '\\1[[\\2]]', s)
    s = filter_wiki(s)
    s = re.sub('\* *\n|\'{2,}', '', s)
    s = re.sub('\n+', '\n', s)
    s = re.sub('\n[:;]|\n +', '\n', s)
    s = re.sub('\n==', '\n\n==', s)
    s = u'【' + d[0] + u'】\n' + s
```

```
    return opencc.convert(s).strip()

i = 0
f = codecs.open('wiki.txt', 'w', encoding='utf-8')
w = tqdm(wiki, desc=u'已获取0篇文章')
for d in w:
    if not re.findall('^[a-zA-Z]+:', d[0]) and d[0] and not re.findall(u'^#', d[
        s = wiki_replace(d)
        f.write(s+'\n\n\n')
        i += 1
        if i % 100 == 0:
            w.set_description(u'已获取%s篇文章'%i)

f.close()
```

注释

可见，代码的主要部分是正则表达式。首先通过bz2file直接不解压来读取下载下来的语料，然后用gensim的extract_pages来提取每个页面，提取后，先处理页面的一些特殊的、非文本的标记，然后将部分有用的{{}}标记替换为[[]]，因为[[]]标记不会被完全清空（具体原理读者得自己测试了），然后用gensim的filter_wiki函数直接清理，接着再处理一下换行的问题，最后通过opencc将繁体转化为简体。

后面的循环中，re.findall('^[a-zA-Z]+:', d[0])这个条件是去掉那些帮助页面，re.findall(u'^#', d[1])这个条件是去掉重定向的页面，最后得到大概就是91.9万个页面。tqdm是用来显示进度的，这个必须有。程序在我的机器上运行了大概40分钟，得到了1.5G左右的纯文本语料。运行时间不重要，因为预处理是一次性的。

值得注意的是，opencc不能用sudo apt-get install opencc来安装，这个默认版本太低，要用源码编译安装，然后pip install opencc安装python接口，这时候在python中调用opencc可能会报“段错误”，这时候要运行

```
cp /usr/lib/libopencc.so.1.0.0 /usr/lib/x86_64-linux-gnu/
```

副产品

上面提到了重定向，重定向意味着两个词具有同样的意思，这样我把中文维基里边所有的重定向都提取出来了，做了个匹配表。也就是说，词表中每一行的两个词都是相同含义的。这算是一个副产品了。

基于中文维基重定向的同义词表：wiki_cn_mapping.7z

转载到请包括本文地址：https://spaces.ac.cn/archives/4176

如果您需要引用本文，请参考：

苏剑林. (2017, Jan 06). 《获取并处理中文维基百科语料》[Blog post]. Retrieved from https://spaces.ac.cn/archives/4176