

Newsteinwell

随笔 - 2 文章 - 0 评论 - 3

[博客园](#) [首页](#) [新随笔](#) [联系](#) [订阅](#) [管理](#)

公告

昵称：Newsteinwell
园龄：1年2个月
粉丝：0
关注：0
[+加关注](#)

<	2017年12月						>
日	一	二	三	四	五	六	
26	27	28	29	30	1	2	
3	4	5	6	7	8	9	
10	11	12	13	14	15	16	
17	18	19	20	21	22	23	
24	25	26	27	28	29	30	
31	1	2	3	4	5	6	

搜索

word2vec 构建中文词向量

词向量作为文本的基本结构——词的模型，以其优越的性能，受到自然语言处理领域研究人员的青睐。良好的词向量可以达到语义相近的词在词向量空间里聚集在一起，这对后续的文本分类，文本聚类等等操作提供了便利，本文将详细介绍如何使用word2vec构建中文词向量。

一、中文语料库

本文采用的是搜狗实验室的搜狗新闻语料库，数据链接 <http://www.sogou.com/labs/resource/cs.php>

下载下来的文件名为：news_sohusite_xml.full.tar.gz

二、数据预处理

2.1 解压并查看原始数据

cd 到原始文件目录下，执行解压命令：

```
tar -zxvf news_sohusite_xml.full.tar.gz
```

得到文件 news_sohusite_xml.dat, 用vim打开该文件，

谷歌搜索

常用链接

我的随笔
我的评论
我的参与
最新评论
我的标签

我的标签

python(1)
word2vec(1)
分词(1)
文本替换(1)

随笔分类

道可道，为何道(2)

随笔档案

2016年11月 (2)

最新评论

1. Re:word2vec 构建中文词向量
解决了，需要忽略错误的编码才行
with open("news_sohusite_xml.da
t", "rb") as fin: c = 0

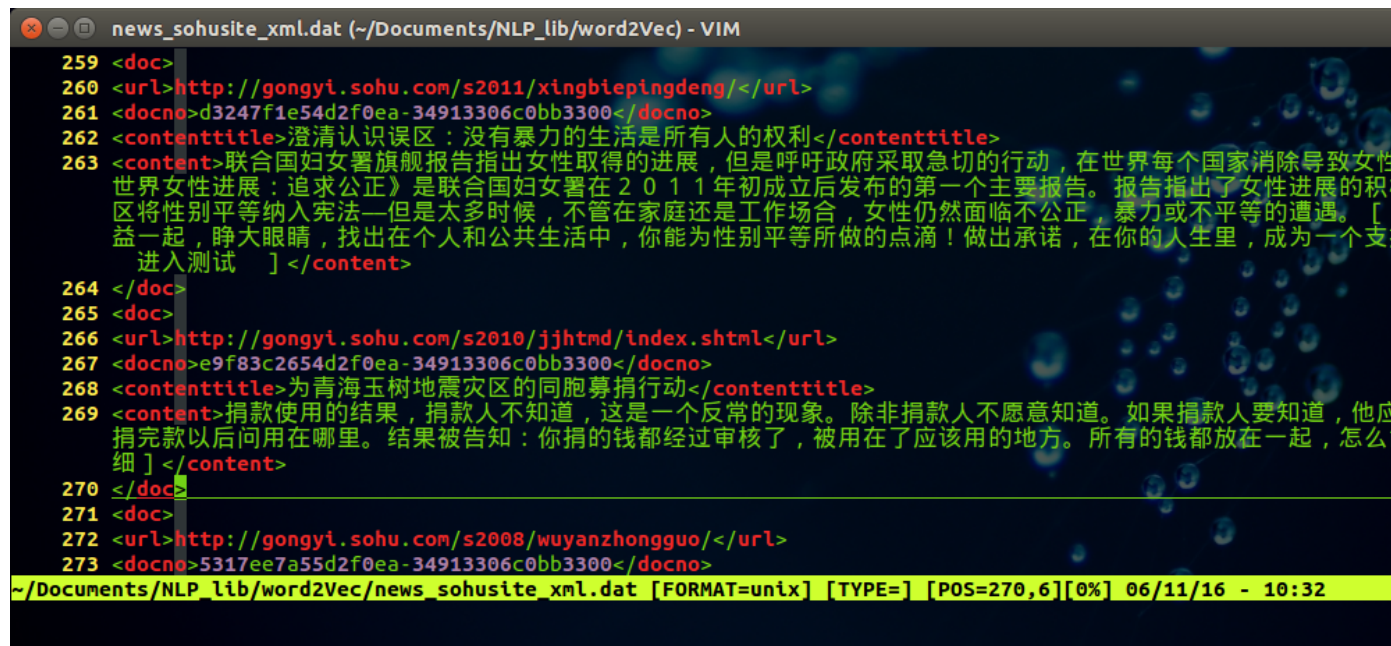
--hain7

2. Re:word2vec 构建中文词向量
我解压news_sohusite_xml.full.tar.
gz后的文件无法识别编码，遇到了
下面的问题：www.itwendao.com/a
rticle/detail/74789.html请问你可以
分.....

--hain7

```
vim news_sohusite_xml.dat
```

得到如下结果：



```
news_sohusite_xml.dat (~/.Documents/NLP_lib/word2Vec) - VIM
259 <doc>
260 <url>http://gongyi.sohu.com/s2011/xingbiepingdeng/</url>
261 <docno>d3247f1e54d2f0ea-34913306c0bb3300</docno>
262 <contenttitle>澄清认识误区：没有暴力的生活是所有人的权利</contenttitle>
263 <content>联合国妇女署旗舰报告指出女性取得的进展，但是呼吁政府采取急切的行动，在世界每个国家消除导致女性
世界女性进展：追求公正》是联合国妇女署在2011年初成立后发布的第一个主要报告。报告指出了女性进展的积极
区将性别平等纳入宪法—但是太多时候，不管在家庭还是工作场合，女性仍然面临不公正，暴力或不平等的遭遇。[
益一起，睁大眼睛，找出在个人和公共生活中，你能性别平等所做的点滴！做出承诺，在你的人生里，成为一个支
进入测试 ]</content>
264 </doc>
265 <doc>
266 <url>http://gongyi.sohu.com/s2010/jjhtmd/index.shtml</url>
267 <docno>e9f83c2654d2f0ea-34913306c0bb3300</docno>
268 <contenttitle>为青海玉树地震灾区的同胞募捐行动</contenttitle>
269 <content>捐款使用的结果，捐款人不知道，这是一个反常的现象。除非捐款人不愿意知道。如果捐款人要知道，他应
捐完款以后问用在哪里。结果被告知：你捐的钱都经过审核了，被用在了应该用的地方。所有的钱都放在一起，怎么
细 ]</content>
270 </doc>
271 <doc>
272 <url>http://gongyi.sohu.com/s2008/wuyanzhongguo/</url>
273 <docno>5317ee7a55d2f0ea-34913306c0bb3300</docno>
~/Documents/NLP_lib/word2Vec/news_sohusite_xml.dat [FORMAT=unix] [TYPE=] [POS=270,6][0%] 06/11/16 - 10:32
```

2.2 取出内容

取出<content> </content> 中的内容,执行如下命令：

```
cat news_tensite_xml.dat | iconv -f gbk -t utf-8 -c | grep "<content>" > corpus.txt
```

得到文件名为corpus.txt的文件，可以通过vim 打开

```
vim corpus.txt
```

得到如下效果：

3. Re:word2vec 构建中文词向量

很好的分享，学习了

--BlanKen

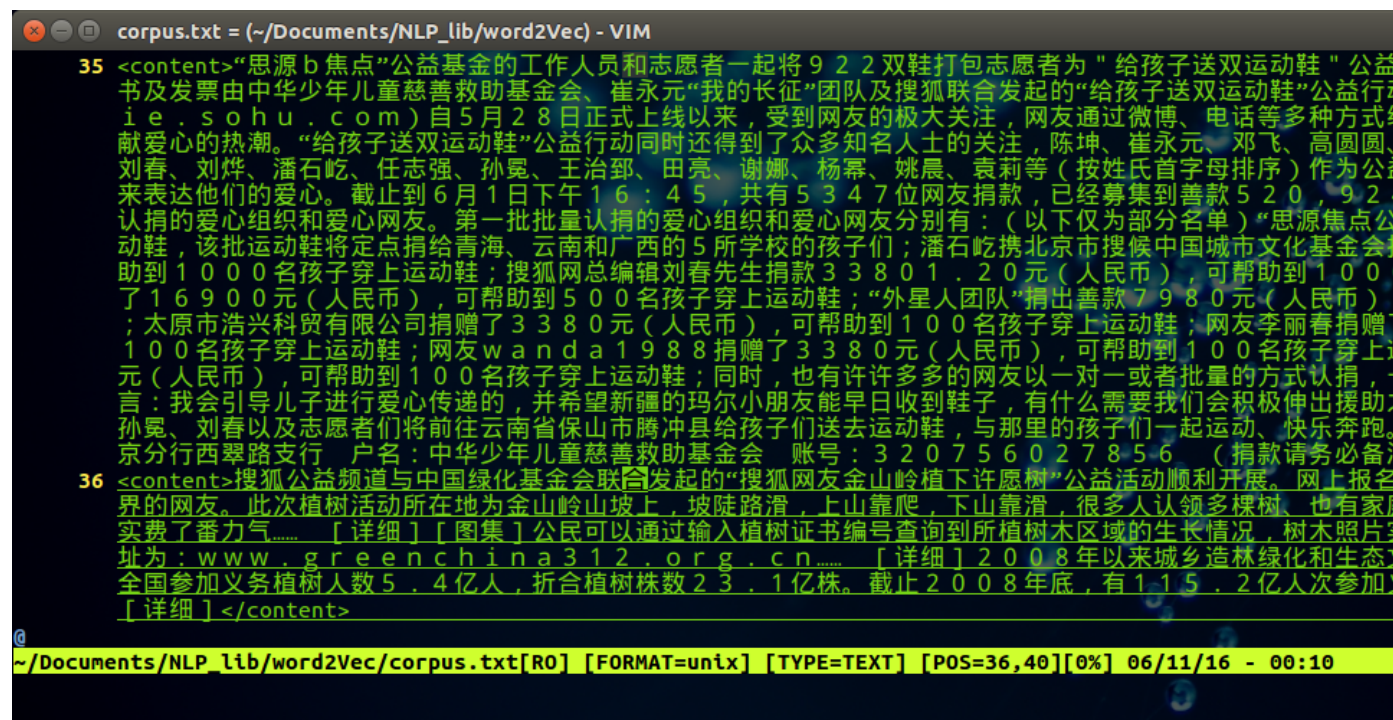
阅读排行榜

1. word2vec 构建中文词向量(13768)

2. 替换句子中的多个不同的词——python 实现(484)

评论排行榜

1. word2vec 构建中文词向量(3)



2.3 分词

注意，送给word2vec的文件是需要分词的，分词可以采用jieba分词实现，安装jieba 分词

```
pip install jieba
```

对原始文本内容进行分词，python 程序如下：

```
1 ##!/usr/bin/env python
2 ## coding=utf-8
3 import jieba
4
5 filePath='corpus.txt'
6 fileSegWordDonePath ='corpusSegDone.txt'
```

```
7 # read the file by line
8 fileTrainRead = []
9 #fileTestRead = []
10 with open(filePath) as fileTrainRaw:
11     for line in fileTrainRaw:
12         fileTrainRead.append(line)
13
14
15 # define this function to print a list with Chinese
16 def PrintListChinese(list):
17     for i in range(len(list)):
18         print list[i],
19 # segment word with jieba
20 fileTrainSeg=[]
21 for i in range(len(fileTrainRead)):
22     fileTrainSeg.append([' '.join(list(jieba.cut(fileTrainRead[i][9:-11], cut_all=False))))])
23     if i % 100 == 0 :
24         print i
25
26 # to test the segment result
27 #PrintListChinese(fileTrainSeg[10])
28
29 # save the result
30 with open(fileSegWordDonePath, 'wb') as fW:
31     for i in range(len(fileTrainSeg)):
32         fW.write(fileTrainSeg[i][0].encode('utf-8'))
33         fW.write('\n')
```

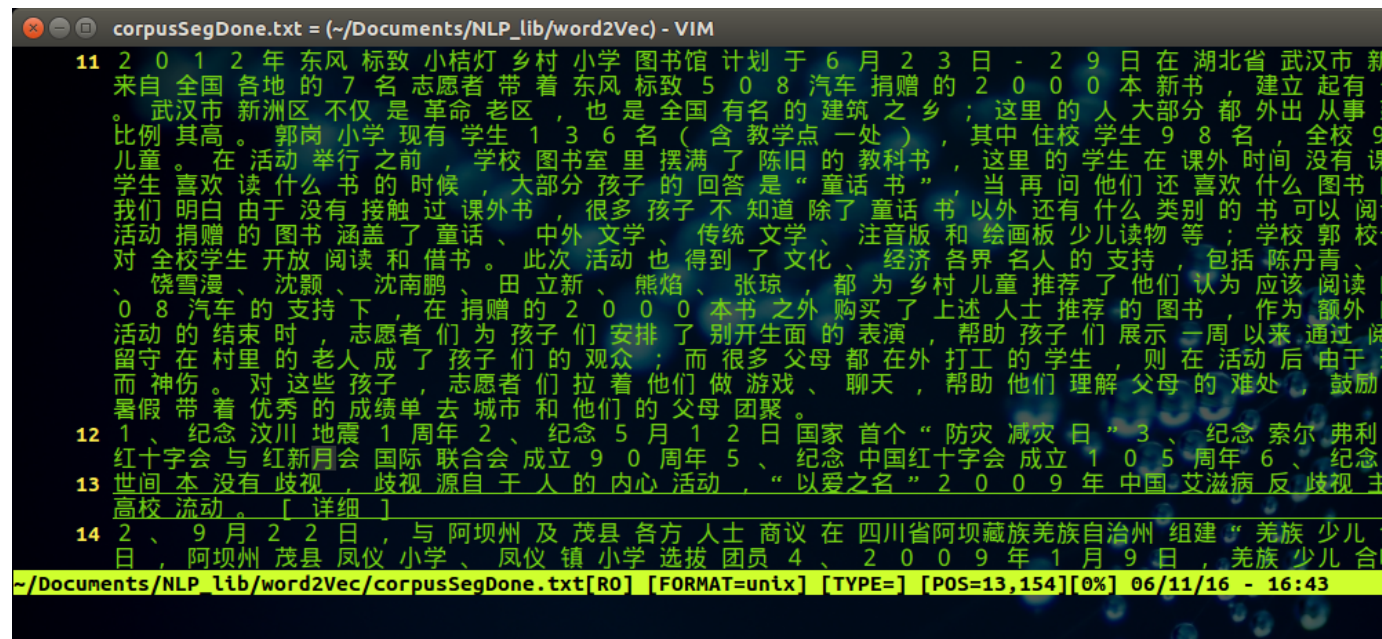


可以得到文件名为 corpusSegDone.txt 的文件，需要注意的是，对于读入文件的每一行，使用结巴分词的时候并不是从0到结尾的全部都进行分词，而是对[9:-11]分词 (如行22中所示: fileTrainRead[i][9:-11])，这样可以去掉每行（一篇新闻稿）起始的<content> 和结尾的</content>。

同样的，可以通过vim 打开分词之后的文件，执行命令：

```
vim corpusSegDone.txt
```

得到如下图所示的结果：



三、构建词向量

3.1 安装word2vec

```
pip install word2vec
```

3.2 构建词向量

执行以下程序：

```
import word2vec
word2vec.word2vec('corpusSegDone.txt', 'corpusWord2Vec.bin', size=300, verbose=True)
```

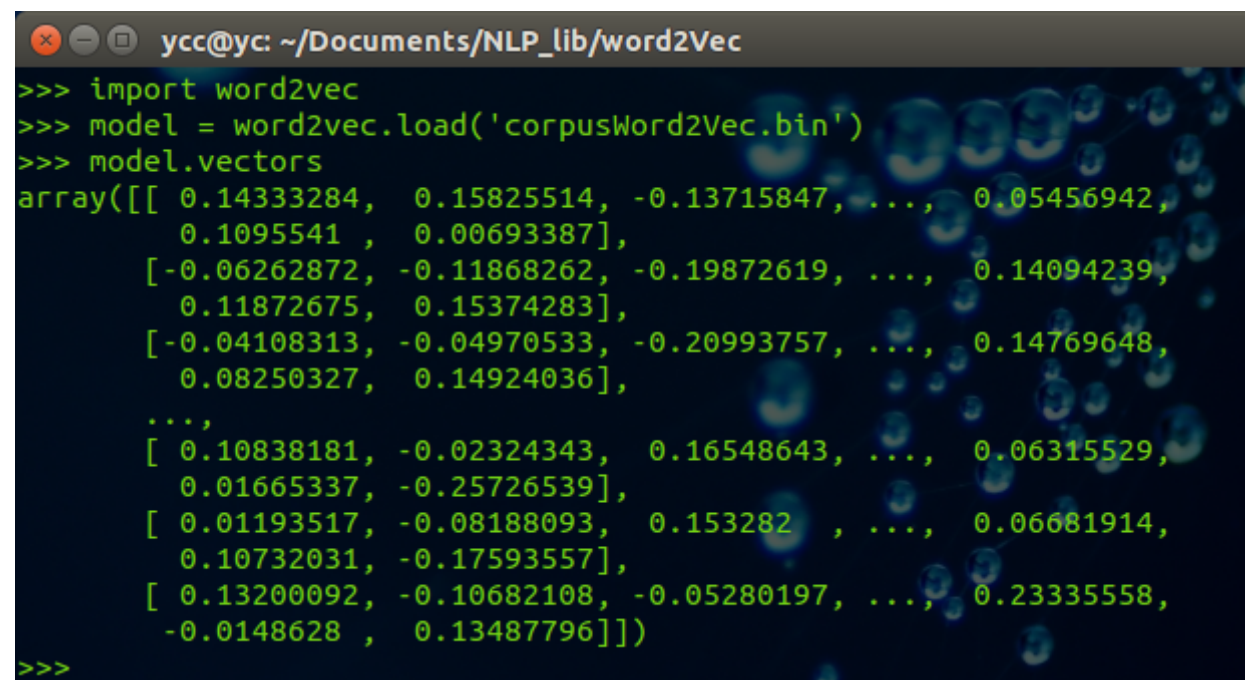

即可构建词向量，得到结果放在文件名为 corpusWord2Vec.bin的文件中。可以通过设定size 的大小来指定词向量的维数。用vim打开生成的二进制文件会出现乱码，目前不知道解决方法。

3.3 显示并使用词向量

3.3.1 查看词向量

```
import word2vec
model = word2vec.load('corpusWord2Vec.bin')
print (model.vectors)
```

可以得到如下结果：

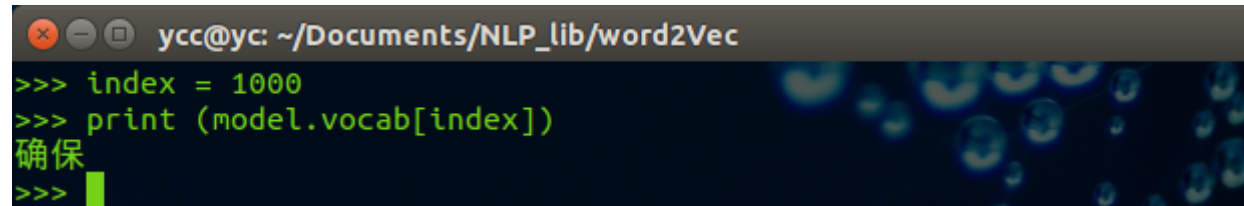


```
ycc@yc: ~/Documents/NLP_lib/word2Vec
>>> import word2vec
>>> model = word2vec.load('corpusWord2Vec.bin')
>>> model.vectors
array([[ 0.14333284,  0.15825514, -0.13715847, ...,  0.05456942,
         0.1095541,  0.00693387],
       [-0.06262872, -0.11868262, -0.19872619, ...,  0.14094239,
         0.11872675,  0.15374283],
       [-0.04108313, -0.04970533, -0.20993757, ...,  0.14769648,
         0.08250327,  0.14924036],
       ...,
       [ 0.10838181, -0.02324343,  0.16548643, ...,  0.06315529,
         0.01665337, -0.25726539],
       [ 0.01193517, -0.08188093,  0.153282, ...,  0.06681914,
         0.10732031, -0.17593557],
       [ 0.13200092, -0.10682108, -0.05280197, ...,  0.23335558,
        -0.0148628,  0.13487796]])
>>>
```

3.3.2 查看词表中的词

```
import word2vec
model = word2vec.load('corpusWord2Vec.bin')
index = 1000
print (model.vocab[index])
```

得到结果如下：



```
ycc@yc: ~/Documents/NLP_lib/word2Vec
>>> index = 1000
>>> print (model.vocab[index])
确保
>>>
```

可以得到词表中第1000个词为 确保。

3.3.3 显示空间距离相近的词

一个好的词向量可以实现词义相近的一组词在词向量空间中也是接近的，可以通过显示词向量空间中相近的一组词并判断它们语义是否相近来评价词向量构建的好坏。代码如下：

```
import word2vec
model = word2vec.load('corpusWord2Vec.bin')
indexes = model.cosine(u'加拿大')
for index in indexes[0]:
    print (model.vocab[index])
```

得到的结果如下：

```
ycc@yc: ~/Documents/NLP_lib/word2Vec
>>> indexes = model.cosine(u'加拿大')
>>> for index in indexes[0]:
...     print (model.vocab[index])
...
澳大利亚
新西兰
澳洲
新加坡
英国
美国
奥地利
法国
新南威尔士州
马来西亚
>>>
```

可以修改希望查找的中文词，例子如下：

```
ycc@yc: ~/Documents/NLP_lib/word2Vec
>>> indexes = model.cosine(u'清华大学')
>>> for index in indexes[0]:
...     print (model.vocab[index])
...
北京大学
武汉大学
上海交通大学
南京大学
上海交大
香港浸会大学
台湾大学
中国科技大学
浙江大学
山东大学
>>>
```



```
ycc@yc: ~/Documents/NLP_lib/word2Vec
>>> indexes = model.cosine(u'爱因斯坦')
>>> for index in indexes[0]:
...     print (model.vocab[index])
...
相对论
亚里士多德
量子力学
黑格尔
李银河
苏格拉底
马克思
假说
概率论
物理学
>>>
```

四、二维空间中显示词向量

将词向量采用PCA进行降维，得到二维的词向量，并打印出来，代码如下：

```
1 #!/usr/bin/env python
2 # coding=utf-8
3 import numpy as np
4 import matplotlib
5 import matplotlib.pyplot as plt
6
7 from sklearn.decomposition import PCA
8 import word2vec
9 # load the word2vec model
10 model = word2vec.load('corpusWord2Vec.bin')
11 rawWordVec=model.vectors
12
13 # reduce the dimension of word vector
14 X_reduced = PCA(n_components=2).fit_transform(rawWordVec)
15
```

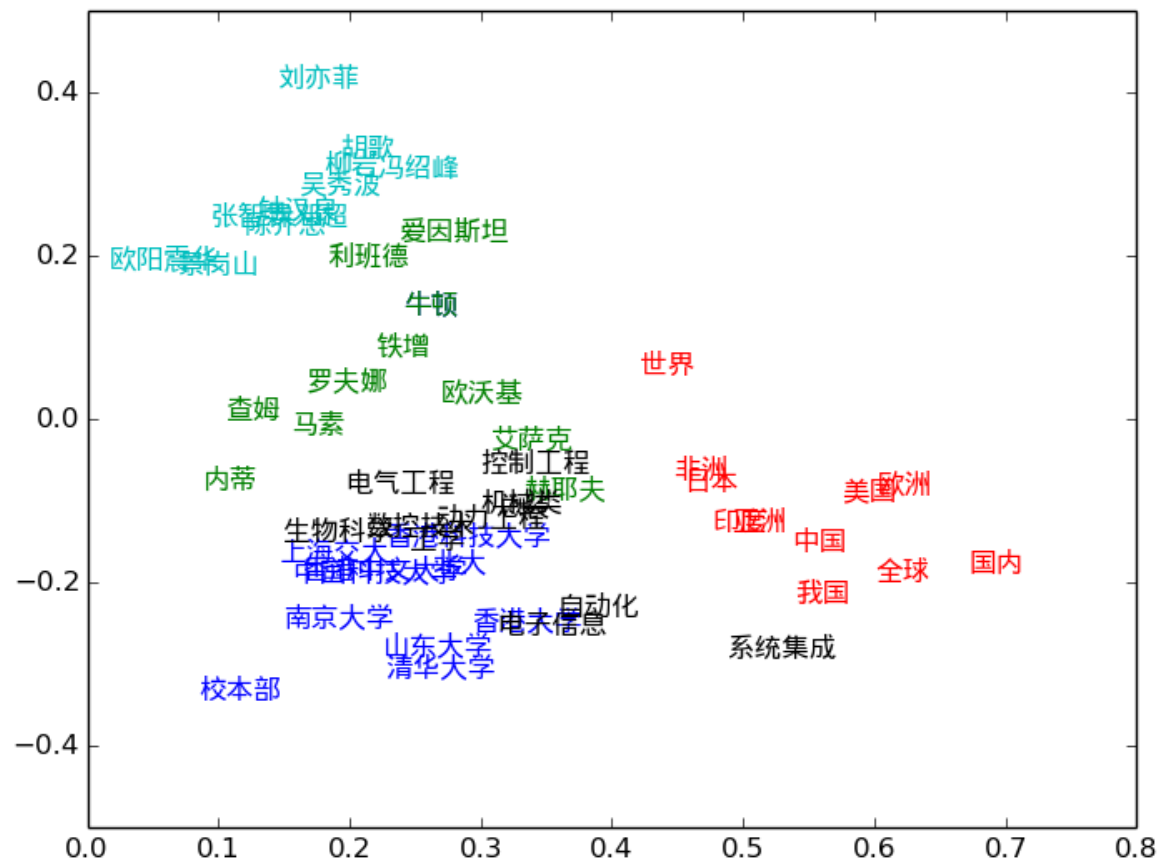
```
16 # show some word(center word) and it's similar words
17 index1,metrics1 = model.cosine(u'中国')
18 index2,metrics2 = model.cosine(u'清华')
19 index3,metrics3 = model.cosine(u'牛顿')
20 index4,metrics4 = model.cosine(u'自动化')
21 index5,metrics5 = model.cosine(u'刘亦菲')
22
23 # add the index of center word
24 index01=np.where(model.vocab==u'中国')
25 index02=np.where(model.vocab==u'清华')
26 index03=np.where(model.vocab==u'牛顿')
27 index04=np.where(model.vocab==u'自动化')
28 index05=np.where(model.vocab==u'刘亦菲')
29
30 index1=np.append(index1,index01)
31 index2=np.append(index2,index03)
32 index3=np.append(index3,index03)
33 index4=np.append(index4,index04)
34 index5=np.append(index5,index05)
35
36 # plot the result
37 zhfont = matplotlib.font_manager.FontProperties(fname='/usr/share/fonts/truetype/wqy/wqy-
microhei.ttc')
38 fig = plt.figure()
39 ax = fig.add_subplot(111)
40
41 for i in index1:
42     ax.text(X_reduced[i][0],X_reduced[i][1], model.vocab[i],
fontproperties=zhfont,color='r')
43
44 for i in index2:
45     ax.text(X_reduced[i][0],X_reduced[i][1], model.vocab[i],
fontproperties=zhfont,color='b')
46
47 for i in index3:
48     ax.text(X_reduced[i][0],X_reduced[i][1], model.vocab[i],
fontproperties=zhfont,color='g')
```

```
49
50 for i in index4:
51     ax.text(X_reduced[i][0],X_reduced[i][1], model.vocab[i],
fontproperties=zhfont,color='k')
52
53 for i in index5:
54     ax.text(X_reduced[i][0],X_reduced[i][1], model.vocab[i],
fontproperties=zhfont,color='c')
55
56 ax.axis([0,0.8,-0.5,0.5])
57 plt.show()
```



中文的显示需要做特殊处理，详见代码 line: 37

下图是执行结果：



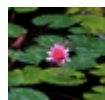
主要参考

<http://blog.csdn.net/zhaoxinfa/article/details/11069485>

<http://nbviewer.jupyter.org/github/danielrfg/word2vec/blob/master/examples/word2vec.ipynb>

分类: [道可道，为何道](#)

标签: [word2vec](#), [分词](#)

[好文要顶](#)[关注我](#)[收藏该文](#)[Newsteinwell](#)[关注 - 0](#)[粉丝 - 0](#)[+加关注](#)

0

0

» 下一篇: [替换句子中的多个不同的词——python 实现](#)

posted @ 2016-11-06 19:27 Newsteinwell 阅读(13768) 评论(3) 编辑 收藏

评论列表

#1楼 2017-09-30 13:13 BlanKen

很好的分享，学习了

支持(0) 反对(0)

#2楼 2017-10-21 20:05 hain7

我解压news_sohusite_xml.full.tar.gz后的文件无法识别编码，遇到了下面的问题：

www.itwendao.com/article/detail/74789.html

请问你可以分享一下你的解压后的文件么？

支持(0) 反对(0)

#3楼 2017-10-21 20:13 hain7

解决了，需要忽略错误的编码才行

```
1 with open("news_sohusite_xml.dat", "rb") as fin:
2     c = 0
3     for x in fin:
4         y = x.decode("GBK", 'ignore').encode('utf-8')
5         print(y)
6         c += 1
```



```
7 |         if c> 10: break
```

支持(0) 反对(0)

[刷新评论](#) [刷新页面](#) [返回顶部](#)

注册用户登录后才能发表评论，请 [登录](#) 或 [注册](#)，[访问网站首页](#)。

【推荐】50万行VC++源码: 大型组态工控、电力仿真CAD与GIS源码库

【促销】腾讯云技术升级10大核心产品年终让利

【推荐】高性能云服务器2折起，0.73元/日节省80%运维成本

【新闻】H3 BPM体验平台全面上线



葡萄城报表
千万种报表 同一种选择

在线设计 报表
数据价值即刻体现

立即了解

最新IT新闻:

- 永辉超市：未来12个月内 腾讯无继续增持计划
- 拓宽AR的娱乐价值，Snapchat推出首款桌面应用「Lens Studio」
- 腾讯发现者揭秘：怎么应对TensorFlow的安全风险，修复有多难
- 怼周鸿祎92女生再发声：我愿意跟周教主面对面对话

- 「中国特斯拉」蔚来ES8的「汽车重新定义」和它的iPhone时刻

» 更多新闻...



最新知识库文章:

- 以操作系统的角度述说线程与进程
- 软件测试转型之路
- 门内门外看招聘
- 大道至简，职场上做人做事做管理
- 关于编程，你的练习是不是有效的？

» 更多知识库文章...

Copyright ©2017 Newsteinwell