

资讯 | 安全 | 论坛 | 下载 | 读书 | 程序开发 | 数据库 | 系统 | 网络 | 电子书 | 微信学院 | 站长学院 | QQ | 手机软件 | 考试

频道栏目 | 软件开发 | web前端 | Web开发 | 移动开发 | 综合编程 |

登录 注册



奥迪r8二手



什么丰胸又快



三星s6以旧换新



电脑租赁



收音机全波段



零基础学习手



明天涨停的股



人工智能课程



无网络收

首页 > 程序开发 > Web开发 > Python > 正文

## Python版的Word2Vector -- gensim 学习手札 中文词语相似性度量

2016-08-29 09:36:43

0条评论 来源：MebiuW的专栏

收藏

我要投稿



奥迪r8二手



收音机全波段



电脑租赁



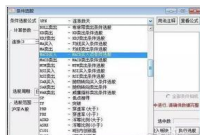
智能家居公司



无锅机顶盒



日语常用语



通达信怎么选



十大装修公司



网络游戏前十



石景山租房网



怎么学习日语



电饭煲蛋糕

## 前言

昨天好不容易试用了一下Google自己提供的Word2Vector的源代码，花了好长时间训练数据，结果发现似乎Python并不能直接使用，于是上网找了一下Python能用的Word2Vector，这么一找，就找到了gensim

gensim（应该要翻墙）：

<http://radimrehurek.com/gensim/models/word2vec.html>



## 电脑租赁



文章

推荐

- 详解DB2中自定义XML存储及其使用环境
- 用Jasperreport计算openingbalance
- TNS-12541，TNS-12560，TNS-00511，TN
- redisclientprotocol实现
- 数据库中表的复杂查询
- sedna加载xml文件
- hdf5
- 在Hbase Endpoint Coprocessor中使用

# 安装

gensim有一些依赖，首先请先确保你安装了这些东西：

```
1 Python >= 2.6. Tested with versions 2.6, 2.7, 3.3, 3.4 and 3.5. Supp?
2
3 NumPy >= 1.3. Tested with version 1.9.0, 1.7.1, 1.7.0, 1.6.2, 1.6.1rc2
4
5 SciPy >= 0.7. Tested with version 0.14.0, 0.12.0, 0.11.0, 0.10.1, 0.9.
```

还有一点特别注意的是，保证你的系统有C的编译器，不然速度会很慢，其实你可以首先编译一下Google官方的C语言版的试试，然后在安装gensim，gensim的word2vector用了官方的代码

根据官网的安装指南，有两种方法可以选择：

使用easy\_install 或者pip，注意这两者可能都需要sudo申请更高的权限

```
1 easy_install -U gensim
2 或者（这个相对于官网的，我修改过，实测我的没问题）
3 pip install --upgrade --ignore-installed six gensim
```

我使用了第二种方式进行的安装，如果这些依赖没有安装的，可以安装python和相关的工具后，直接使用pip或easy\_install安装。

在进行模型训练的时候，如果不安装Cython，无法进行多线程训练，速度很瘦影响，所以接着安装下Cython

```
1 pip install cython
```

1、训练模型：

如果所有安装配置工作都已经做好了，那么可以开始使用gensim了。这里的语料库使用我之前博客里面已经分好词的corpus-seg.txt语料库。这里在完成模型训练后，将他存到一个文件中，这样下次就可以直接使用了。

```
1 # coding:utf-8
2 import sys
```



## 点击排行

- 基于zabbix用Python写一个运维流量气象
- python学习之argparse模块
- 使用Python读取和写入CSV文件
- AttributeError: 'module' object
- XGBoost参数调优完全指南（附Python代
- python各种类型转换-int,str,char,flo
- Python3.x爬虫教程：爬网页、爬图片、
- [python] 安装numpy+scipy+matplotlib

```

3 reload(sys)
4 sys.setdefaultencoding( "utf-8" )
5 from gensim.models import Word2Vec
6 import logging,os
7
8 class TextLoader(object):
9     def __init__(self):
10         pass
11
12     def __iter__(self):
13         input = open('corpus-seg.txt','r')
14         line = str(input.readline())
15         counter = 0
16         while line!=None and len(line) > 4:
17             #print line
18             segments = line.split(' ')
19             yield segments
20             line = str(input.readline())
21
22 sentences = TextLoader()
23 model = gensim.models.Word2Vec(sentences, workers=8)
24 model.save('word2vector2.model')
25 print 'ok'

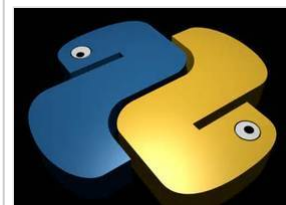
```

这里的文件加载用了自己的代码，当然也可以使用自带的Line Sentence,之所以贴出上面的代码是因为，如果你的文件格式比较特殊可以参照上面的代码进行处理。

```

1 # coding:utf-8
2 import sys
3 reload(sys)
4 sys.setdefaultencoding( "utf-8" )
5 from gensim.models import Word2Vec
6 import logging,os
7
8 #模型的加载
9 model = Word2Vec.load('word2vector.model')
10 #比较两个词语的相似度,越高越好
11 print('"唐山" 和 "中国" 的相似度:'+ str(model.similarity('唐山','中国')))
12 print('"中国" 和 "祖国" 的相似度:'+ str(model.similarity('祖国','中国')))
13 print('"中国" 和 "中国" 的相似度:'+ str(model.similarity('中国','中国')))
14 #使用一些词语来限定,分为正向和负向的
15 result = model.most_similar(positive=['中国', '城市'], negative=['学生'])
16 print('同"中国"与"城市"二词接近,但是与"学生"不接近的词有:')
17 for item in result:
18     print('    '+item[0]+'    相似度:'+str(item[1]))
19

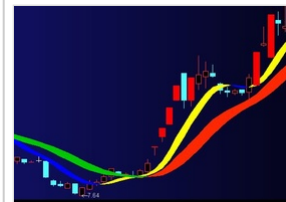
```



python课程



android音频



通达信怎么选股



手板模型



```

20 result = model.most_similar(positive=['男人','权利'], negative=['女人'])
21 print('同"男人"和"权利"接近,但是与"女人"不接近的词有:')
22 for item in result:
23     print('    '+item[0]+'    相似度:'+str(item[1]))
24
25 result = model.most_similar(positive=['女人','法律'], negative=['男人'])
26 print('同"女人"和"法律"接近,但是与"男人"不接近的词有:')
27 for item in result:
28     print('    '+item[0]+'    相似度:'+str(item[1]))
29 #从一堆词里面找到不匹配的
30 print("老师 学生 上课 校长 , 有哪个是不匹配的? word2vec结果说是:"+model.doe
31 print("汽车 火车 单车 相机 , 有哪个是不匹配的? word2vec结果说是:"+model.doe
32 print("大米 白色 蓝色 绿色 红色 , 有哪个是不匹配的? word2vec结果说是:"+model
33 #直接查看某个词的向量
34 print('中国的特征向量是:')
35 print(model['中国'])

```

这里给出一个我的运行结果：

```

1 /System/Library/Frameworks/Python.framework/Versions/2.7/bin/python?.
2 "唐山" 和 "中国" 的相似度:0.1720725224
3 "中国" 和 "祖国" 的相似度:0.456236474841
4 "中国" 和 "中国" 的相似度:1.0
5 同"中国"与"城市"二词接近,但是与"学生"不接近的词有:
6     "全球" 相似度:0.60819453001
7     "亚洲" 相似度:0.588450014591
8     "我国" 相似度:0.545840501785
9     "世界" 相似度:0.540009200573
10    "名城" 相似度:0.518879711628
11    "硅谷" 相似度:0.517688155174
12    "长三角" 相似度:0.512072384357
13    "国内" 相似度:0.511703968048
14    "全国" 相似度:0.507433652878
15    "国际" 相似度:0.505781650543
16 同"男人"和"权利"接近,但是与"女人"不接近的词有:
17    "权益" 相似度:0.67150759697
18    "隐私权" 相似度:0.666741013527
19    "选举权" 相似度:0.626420497894
20    "财产权" 相似度:0.617758154869
21    "利益" 相似度:0.610122740269
22    "义务" 相似度:0.608267366886
23    "尊严" 相似度:0.605125784874
24    "继承权" 相似度:0.603345394135
25    "法律" 相似度:0.596215546131
26    "优先权" 相似度:0.59428691864
27 同"女人"和"法律"接近,但是与"男人"不接近的词有:

```



```

28     "劳动法" 相似度:0.652353703976
29     "司法" 相似度:0.652238130569
30     "婚姻法" 相似度:0.631354928017
31     "民法" 相似度:0.624598622322
32     "法规" 相似度:0.623348236084
33     "刑法" 相似度:0.611774325371
34     "国际法" 相似度:0.608191132545
35     "诉讼" 相似度:0.607495307922
36     "R E A C H" 相似度:0.599701464176
37     "强制力" 相似度:0.597045660019
38 老师 学生 上课 校长 , 有哪个是不匹配的? word2vec结果说是:上课
39 汽车 火车 单车 相机 , 有哪个是不匹配的? word2vec结果说是:相机
40 大米 白色 蓝色 绿色 红色 , 有哪个是不匹配的? word2vec结果说是:大米
41 中国的特征向量是:
42 [-0.08299727 -3.58397388 -0.55335367 1.4152931 3.94189262 -2.03232
43 1.31824613 -1.75067747 -1.66100371 -1.70273054 -3.47409034 2.70463
44 -0.87696695 -2.53364205 -2.12181163 -7.60758495 -0.6421982 2.91871
45 1.38164878 -0.05457138 1.02129567 1.64029694 0.21894537 -0.82295
46 3.30296516 -0.65931851 1.39501953 0.71423614 2.0213325 2.97903
47 1.46234405 -0.30748805 2.45258284 -0.51123774 -1.84140313 -0.92091
48 -4.28990364 4.0552578 -2.01020265 0.85769647 -4.6681509 -2.88254
49 -1.80714786 0.52874494 3.31922817 0.43049669 -3.03839922 -1.20092
50 2.75143361 0.99246925 0.41537657 -0.78819919 1.28469515 0.12056
51 -4.54702759 -1.36031103 0.35673267 -0.36477017 -3.63630986 -0.21103
52 2.16747832 -0.47925043 -0.63043374 -2.25911093 -1.47486925 4.23806
53 -0.22334123 3.2125628 0.91901672 0.66508955 -2.80306172 3.42943
54 2.26001453 5.24837303 -4.0164156 -3.28324246 4.40493822 -0.14068
55 -4.31880903 1.98531461 0.2576215 -2.69446373 0.59171939 -0.48256
56 -0.67274201 1.96152794 -2.83031917 0.54468328 2.57930231 -1.44152
57 -0.61808151 1.03311574 -3.48526216 -2.35903311 -3.9816277 -0.93071
58 2.77195001 1.8912288 -3.45096016 4.93347549]
59
60 Process finished with exit code 0

```

目前这个手札只是介绍几本的安装和使用，更多的工作将会在后续博客中写入。

## 使用可能遇到的问题：

ValueError: numpy.dtype has the wrong size, try recompiling :

<http://stackoverflow.com/questions/17709641/valueerror-numpy-dtype-has-the-wrong-size-try-recompiling>

## 参考资料

- 1、官方教程：<http://radimrehurek.com/gensim/models/word2vec.html>

	<a href="#">doxygen</a>		<a href="#">flash 动画制作</a>
<a href="#">什么丰胸又快又好</a>	<a href="#">4k电视是什么意思</a>	<a href="#">python 爬虫</a>	<a href="#">python线程同步</a>
	<a href="#">WebRTC</a>		<a href="#">python线程锁</a>

点击复制链接 与好友分享!

[回本站首页](#)

相关TAG标签

[相似性](#)

[手札](#)

[中文](#)

[词语](#)

上一篇：[python爬虫（中）--数据建模与保存（入库）](#)

下一篇：[Python“隐藏”特性](#)

### 相关文章

[python新手开发必碰到的问题：encod](#)

[python之中文字符串的处理方法介绍](#)

[Python之xlrd——中文文档](#)

[django-registration 0.8 中文文档\[](#)

[Python抓取中文网页](#)

[python中文分词](#)

[python实现支持unicode中文的AC自动机](#)

[Python中文编码问题](#)

[python中文decode和encode转码](#)

[Python基础之--注释与语句结束符“;”](#)

热门专题推荐

[python](#)

[div+css](#)

[css教程](#)

[html5](#)

[html教程](#)

[jquery](#)

[Android SDK](#)

[php](#)

[mysql](#)

[oracle](#)



# 百度公益频道全新上线

## 图文推荐



电脑租赁



无锅机顶盒



三星s6以旧换新



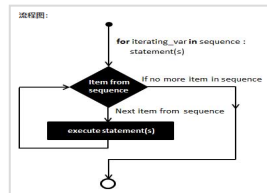
图形工作站



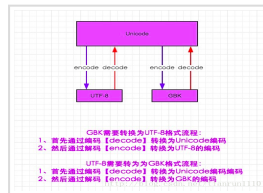
三星小note

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

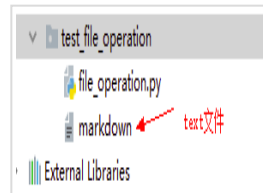
Python数据可视化正态



python是如何定义并使



python字符编码与转码



python文件操作步骤讲

[登录](#)

来说两句吧...

还没有评论，快来抢沙发吧！

红黑联盟正在使用畅言

[订单管理系统](#)
[tensor](#)
[moto新款手机](#)
[免费云主机](#)
[语言理解](#)
[多屏显示](#)
[python学习路线](#)
[电脑学习基础](#)
[招商加盟词语](#)
[电动车电池修复方法](#)
[开发一个app多少钱](#)
[歌华机顶盒](#)
[ppt制作](#)
[奥迪r8二手](#)
[石景山租房网](#)



The banner features a blue background with a diagonal line pattern. On the left, there is an icon of a blue graduation cap and a rolled-up diploma, followed by the text '证书 + 能力' in a stylized, glowing font. The main text in the center is in large, bold, yellow characters: '安全工程师 软件工程师 网站工程师 网络工程师 电脑工程师'. Below this, in smaller pink characters, it says '为新手量身定做的课程, 让菜鸟快速变身高手 正规公司助您腾飞'. On the right side, there is a red button with white text that says '立即加入'. Above the button, in pink characters, it says '不断增加新科目'.

证书 + 能力

安全工程师 软件工程师 网站工程师 网络工程师 电脑工程师

为新手量身定做的课程, 让菜鸟快速变身高手 正规公司助您腾飞

不断增加新科目

立即加入

---

[关于我们](#) | [联系我们](#) | [广告服务](#) | [投资合作](#) | [版权申明](#) | [在线帮助](#) | [网站地图](#) | [作品发布](#) | [Vip技术培训](#) | [举报中心](#)

版权所有: 红黑联盟--致力于做实用的IT技术学习网站