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个评论 来源：MebiuW的专栏 收藏 我要投稿

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

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
· hdfs
· 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：

```
Python >= 2.6. Tested with versions 2.6, 2.7, 3.3, 3.4 and 3.5. Supp

NumPy >= 1.3. Tested with version 1.9.0, 1.7.1, 1.7.0, 1.6.2, 1.6.1rc

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

```
easy_install -U gensim
或者（这个相对于官网的，我修改过，实测我的没问题）
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

```
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。

```
# coding:utf-8
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+matlotlib

```
reload(sys)
sys.setdefaultencoding( "utf-8" )
from gensim.models import Word2Vec
import logging,gensim,os

class TextLoader(object):
    def __init__(self):
        pass

    def __iter__(self):
        input = open('corpus-seg.txt','r')
        line = str(input.readline())
        counter = 0
        while line!=None and len(line) > 4:
            #print line
            segments = line.split(' ')
            yield  segments
            line = str(input.readline())

sentences = TextLoader()
model = gensim.models.Word2Vec(sentences, workers=8)
model.save('word2vector2.model')
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。

```
# coding:utf-8
import sys
reload(sys)
sys.setdefaultencoding( "utf-8" )
from gensim.models import Word2Vec
import logging,gensim,os

#模型的加载
model = Word2Vec.load('word2vector.model')
#比较两个词语的相似度,越高越好
print('"唐山" 和 "中国" 的相似度:'+ str(model.similarity('唐山','中国')))
print('"中国" 和 "祖国" 的相似度:'+ str(model.similarity('祖国','中国')))
print('"中国" 和 "中国" 的相似度:'+ str(model.similarity('中国','中国')))
#使用一些词语来限定,分为正向和负向的
result = model.most_similar(positive=['中国', '城市'], negative=['学生'
print('同"中国"与"城市"二词接近,但是与"学生"不接近的词有:')
for item in result:
    print('    "'+item[0]+'"  相似度:'+str(item[1]))

```

```
result = model.most_similar(positive=['男人','权利'], negative=['女人']
print('同"男人"和"权利"接近,但是与"女人"不接近的词有:')
for item in result:
    print('   "'+item[0]+'"  相似度:'+str(item[1]))

result = model.most_similar(positive=['女人','法律'], negative=['男人']
print('同"女人"和"法律"接近,但是与"男人"不接近的词有:')
for item in result:
    print('   "'+item[0]+'"  相似度:'+str(item[1]))
#从一堆词里面找到不匹配的
print("老师 学生 上课 校长 , 有哪个是不匹配的? word2vec结果说是:"+model.doe
print("汽车 火车 单车 相机 , 有哪个是不匹配的? word2vec结果说是:"+model.doe
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
#直接查看某个词的向量
print('中国的特征向量是:')
print(model['中国'])
```

这里给出一个我的运行结果：

```
/System/Library/Frameworks/Python.framework/Versions/2.7/bin/python?
"唐山" 和 "中国" 的相似度:0.1720725224
"中国" 和 "祖国" 的相似度:0.456236474841
"中国" 和 "中国" 的相似度:1.0
同"中国"与"城市"二词接近,但是与"学生"不接近的词有:
   "全球"  相似度:0.60819453001
   "亚洲"  相似度:0.588450014591
   "我国"  相似度:0.545840501785
   "世界"  相似度:0.540009200573
   "名城"  相似度:0.518879711628
   "硅谷"  相似度:0.517688155174
   "长三角"  相似度:0.512072384357
   "国内"  相似度:0.511703968048
   "全国"  相似度:0.507433652878
   "国际"  相似度:0.505781650543
同"男人"和"权利"接近,但是与"女人"不接近的词有:
   "权益"  相似度:0.67150759697
   "隐私权"  相似度:0.666741013527
   "选举权"  相似度:0.626420497894
   "财产权"  相似度:0.617758154869
   "利益"  相似度:0.610122740269
   "义务"  相似度:0.608267366886
   "尊严"  相似度:0.605125784874
   "继承权"  相似度:0.603345394135
   "法律"  相似度:0.596215546131
   "优先权"  相似度:0.59428691864
同"女人"和"法律"接近,但是与"男人"不接近的词有:
```

```
    "劳动法"  相似度:0.652353703976
    "司法"  相似度:0.652238130569
    "婚姻法"  相似度:0.631354928017
    "民法"  相似度:0.624598622322
    "法规"  相似度:0.623348236084
    "刑法"  相似度:0.611774325371
    "国际法"  相似度:0.608191132545
    "诉讼"  相似度:0.607495307922
    "ＲＥＡＣＨ"  相似度:0.599701464176
    "强制力"  相似度:0.597045660019
老师 学生 上课 校长 ， 有哪个是不匹配的? word2vec结果说是:上课
汽车 火车 单车 相机 ， 有哪个是不匹配的? word2vec结果说是:相机
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
中国的特征向量是:
[-0.08299727 -3.58397388 -0.55335367  1.4152931   3.94189262 -2.03232
  1.31824613 -1.75067747 -1.66100371 -1.70273054 -3.47409034  2.70463
 -0.87696695 -2.53364205 -2.12181163 -7.60758495 -0.6421982   2.91871
  1.38164878 -0.05457138  1.02129567  1.64029694  0.21894537 -0.82295
  3.30296516 -0.65931851  1.39501953  0.71423614  2.0213325   2.97903
  1.46234405 -0.30748805  2.45258284 -0.51123774 -1.84140313 -0.92091
 -4.28990364  4.0552578  -2.01020265  0.85769647 -4.6681509  -2.88254
 -1.80714786  0.52874494  3.31922817  0.43049669 -3.03839922 -1.20092
  2.75143361  0.99246925  0.41537657 -0.78819919  1.28469515  0.12056
 -4.54702759 -1.36031103  0.35673267 -0.36477017 -3.63630986 -0.21103
  2.16747832 -0.47925043 -0.63043374 -2.25911093 -1.47486925  4.23800
 -0.22334123  3.2125628   0.91901672  0.66508955 -2.80306172  3.42943
  2.26001453  5.24837303 -4.0164156  -3.28324246  4.40493822 -0.14068
 -4.31880903  1.98531461  0.2576215  -2.69446373  0.59171939 -0.48250
 -0.67274201  1.96152794 -2.83031917  0.54468328  2.57930231 -1.44152
 -0.61808151  1.03311574 -3.48526216 -2.35903311 -3.9816277  -0.93071
  2.77195001  1.8912288  -3.45096016  4.93347549]

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

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xlrd——中文文档

Python抓取中文网页

python实现支持unicode中文的AC自动机

python中文decode和encode转码

python之中文字符串的处理方法介绍

django-registration 0.8 中文文档[

python中文分词

Python中文编码问题

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

图文推荐

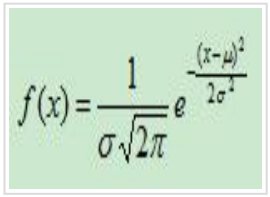

电脑租赁


无锅机顶盒

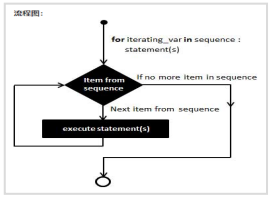

三星s6以旧换新


图形工作站

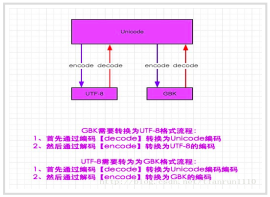

三星小note



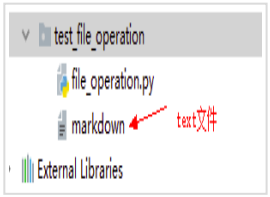
Python数据可视化正态



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