

CSDN新首页上线啦，邀请你来立即体验！(http://blog.csdn.net/)

CSDN

博客 (//blog.csdn.net?ref=toolbar) 学院 (//edu.csdn.net?ref=toolbar)

下载 (//download.csdn.net?ref=toolbar) GitChat (//gitbook.cn/?ref=csdn)

更多

0



登录 (https://passport.csdn.net/account/login?ref=toolbar) 注册 (http://passport.csdn.net/account/mobile/register?ref=toolbar&action=mobile) 注册 (http://passport.csdn.net/account/mobile/register?ref=toolbar&action=mobile) 注册 (http://passport.csdn.net/account/mobile/register?ref=toolbar&action=mobile)

GENSIM 使用笔记1 --- 语料和向量空间

原创

2016年12月25日 15:35:37

标签：gensim (http://so.csdn.net/so/search/s.do?q=gensim&t=blog) /



中文 (http://so.csdn.net/so/search/s.do?q=中文&t=blog) /

向量 (http://so.csdn.net/so/search/s.do?q=向量&t=blog) /

序列化 (http://so.csdn.net/so/search/s.do?q=序列化&t=blog) /

教程 (http://so.csdn.net/so/search/s.do?q=教程&t=blog)

1271

GENSIM 使用笔记1 — 语料和向量空间 (http://blog.csdn.net/mebiuw/article/details/53870117)

GENSIM 使用笔记2 — 主题模型和相似性查询 (http://blog.csdn.net/mebiuw/article/details/53870778)

1 本篇说明

加入CSDN，享有精准的内容推荐，与5000万程序员共同成长！



MebiuW (http://blog.csdn.net/mebiuw)

+ 关注

(http://blog.csdn.net/mebiuw)

码云

未开通
(https://gitee.com/mebiuw)

原创
270

粉丝
99

喜欢
0

他的最新文章

更多文章 (http://blog.csdn.net/mebiuw)

Tensorflow 机器翻译NMT笔记 1 快速上手 (http://blog.csdn.net/MebiuW/article/details/77825642)

Leetcode 655. Print Binary Tree 打印二叉树 解题报告 (http://blog.csdn.net/MebiuW/article/details/77200704)

Leetcode 654. Maximum Binary Tree 最大二叉树 解题报告 (http://blog.csdn.net/MebiuW/article/details/77170643)



内容举报



返回顶部



本篇博客来源于GENSIM官方向导文档的第一章 (<http://radimrehurek.com/gensim/tut1.html>)，主要供自己后续的翻阅，并通过分享带给诸位网友一个小小的参照。

从字符串到向量

在这一小节当中，将会讲述如何通过gensim，将一段文本以向量的形式表示。
首先我们看一下我们的基本文档形式：

```
1 documents = [  
2     '拍照反光一直是摄影爱好者较为苦恼的问题',  
3     '尤其是手机这种快速拍照设备的成像效果总是难以令人满意',  
4     '特别是抓拍的珍贵照片',  
5     '遇上反光照片基本作废',  
6     '而索尼最近研发的集成偏振片传感器',  
7     '似乎可以有效的解决拍照反光的问题'  
8 ]
```

和原始教程不一样，这里我不完全参考他的文档，并且换用了中文作为示例，这一点更加贴合我们实际的使用。

在这里，我们简单的表示了下，将每篇文档（这里只是一句话，请根据实际情况替换）表示为了一个字符串，最后用一个list表示所有的文档，也就是我们的语料库了。

随后，我们需要将他进行分词，在这里我是用了jieba中文分词，如果有其他的大家可以自行替换，如果有什么特殊的功能（如停用词等）也可以自行参照修改。

加入CSDN，享受更精准的内容推荐，与5000万程序员共同成长！

```
1 texts = [jieba.lcut(document) for document in documents]
```

(<https://passport.csdn.net/a>)
Leetcode 653. Two Sum IV
BST 两数相加4 解题报告 (<https://passport.csdn.net/a>)
dn.net/MebiuW/article/details/52218249)

Leetcode 652. Find Duplicate
寻找重复子树 解题报告 (<https://passport.csdn.net/a>)
n.net/MebiuW/article/details/52218249)



相关推荐

gensim使用方法以及例子 (<http://blog.csdn.net/u014595019/article/details/52218249>)

Gensim官方教程翻译（一）——快速入门 (<http://blog.csdn.net/questionfish/article/details/46725475>)

Gensim Word2vec 使用教程 (http://blog.csdn.net/Star_Bob/article/details/47808499)

Gensim实战（一） (<http://blog.csdn.net/u013776640/article/details/42347983>)



内容举报



返回顶部

登录

注册



此时，需要进行了词典的构建，如果需要查看具体的对照信息，也可以print下。具体的方式如下：

```

1 #构造字典 并 保存和加载
2 dictionary = corpora.Dictionary(texts)
3 dictionary.save('mydict.dic')
4 print 'Tokens:Id'
5 print dictionary.token2id
6 new_dictionary = corpora.Dictionary.load('mydict.dic')
7 print(new_dictionary)

```

⋮

那么现在我可以引入一些新的文本，并且通过他生成对应的向量（注意这里保证你的词，已经出现过在之前的语料库之中了，否则多出的这些词是不会统计的）

```

1 #构造新的文本并且获得他的向量
2 new_document = "索尼可以有效解决拍照的问题"
3 new_vector = dictionary.doc2bow(jieba.lcut(new_document))
4 print 'the vector of "%s": (tokenid,frequency)' % new_document
5 print new_vector

```

使用doc2bow这个功能，只会简单的做一些类似于wordcount的东西，并且返回的是一些元组，就是(词 id, 频次)的一个数组，这里需要特别注意下。

最后，我们通过之前的字典和预料，生成一个符合我们格式的语料库

```

1 #生成语料库
2 corpus = [ dictionary.doc2bow(text) for text in texts]

```

加入CSDN，享受更精准的内容推荐，与5000万程序员共同成长！



他的热门文章

Tensorflow #3 使用DNN构造Iris分类器 (http://blog.csdn.net/mebiuw/article/details/53222000)

13286



返回顶部

Python下的自然语言处理利器-LTP语言技术平台 pyltp 学习手札 (http://blog.csdn.net/mebiuw/article/details/52496920)

登录

注册



当我们训练好了一个词典以后，一般希望将其记录到磁盘当中，方便后续使用，而不是每次都单独训练，gensim提供了多种序列化方式，在这里我只选择其中一种进行说明：

```
1 #序列化
2 corpora.MmCorpus.serialize('corpus.mm', corpus)
3 #重新加载预料
4 new_corpus = corpora.MmCorpus('corpus.mm')
5 print(len(new_corpus))
```

如上，就将语料库序列化和反序列化了



原教程，还有一大块是关于如何对接numpy 以及如何节约内存的，这里就不多说了，有需要的自行研究

完整代码

加入CSDN，享受更精准的内容推荐，与5000万程序员共同成长！

登录

注册



📖 12054

(<https://passport.csdn.net/a>

Python版的Word2Vec -- g
札 中文词语相似性度量 V1
sdn.net/mebiuw/article/det

📖 9452

Tensorflow #2 深度学习-RI
NIST手写识别Demo ([http://](http://mebiuw/article/details/5270)
mebiuw/article/details/5270

📖 9070

深度学习（BOT方向）学
quence2Sequence 学习 ([https://](https://mebiuw/article/details/52632847)
net/mebiuw/article/details/52632847)

📖 8467



内容举报



返回顶部

(https://passport.csdn.net/a



```

1  #coding:utf-8
2  import gensim
3  import jieba
4  from gensim import corpora
5  documents = [
6      '拍照反光一直是摄影爱好者较为苦恼的问题',
7      '尤其是手机这种快速拍照设备的成像效果总是难以令人满意',
8      '特别是抓拍的珍贵照片',
9      '遇上反光照片基本作废',
10     '而索尼最近研发的集成偏振片传感器',
11     '似乎可以有效的解决拍照反光的问题'
12 ]
13 texts = [jieba.lcut(document) for document in documents]
14 #构造字典 并 保存和加载
15 dictionary = corpora.Dictionary(texts)
16 dictionary.save('mydict.dic')
17 print 'Tokens:Id'
18 print dictionary.token2id
19 new_dictionary = corpora.Dictionary.load('mydict.dic')
20 print(new_dictionary)
21
22 #构造新的文本并且获得他的向量
23 new_document = "索尼可以有效解决拍照的问题，佳能就不可以"
24 new_vector = dictionary.doc2bow(jieba.lcut(new_document))
25 print 'the vector of "%s": ( tokenid,frequency)' % new_document
26 print new_vector
27
28 #生成语料库
29 corpus = [ dictionary.doc2bow(text) for text in texts]
30 #序列化
31 corpora.MmCorpus.serialize('corpus.mm', corpus)
32 #重新加载预料

```

加入CSDN，享受更精准的内容推荐，与5000万程序员共同成长！
#重新加载预料

⚠
内容举报

⬆
TOP
返回顶部


登录

注册

×

```
33 new_corpus = corpora.MmCorpus('corpus.mm')
34 print(len(new_corpus))
```

Reference

 <http://radimrehurek.com/gensim/tut1.html> (<http://radimrehurek.com/gensim/tut1.html>)

0



版权声明：本文为博主原创文章，未经博主允许不得转载。



相关文章推荐

gensim使用方法以及例子 (<http://blog.csdn.net/u014595019/article/details/52218249>)

gensim是一个python的自然语言处理库，能够将文档根据TF-IDF, LDA, LSI 等模型转化成向量模式，以便进行进一步的处理。此外，gensim还实现了word2vec功能，能够将单词转...



u014595019 (<http://blog.csdn.net/u014595019>) 2016年08月16日 10:58 12622

加入CSDN，享受更精准的内容推荐，与5000万程序员共同成长！

Gensim官方教程翻译（一）——快速入门 (<http://blog.csdn.net/questionfish/article/details/...>)

[登录](#)[注册](#)[内容举报](#)[返回顶部](#)<https://passport.csdn.net/a>

为了方便自己学习，翻译了官方的教程，原文：<http://radimrehurek.com/gensim/tutorial.html>。本教程按照一系列的实例组织，用以突出gensim的各种特征。本教程...



questionfish (<http://blog.csdn.net/questionfish>) 2015年07月02日 13:41 11377



程序员想转管理有捷径吗？一位老前辈给我指了这条路！靠谱吗？

做程序员5年了收获蛮多，但是最近【中兴跳楼事件】发生后，我在想如果我到了40岁，会被辞退吗...



(http://www.baidu.com/cb.php?c=lgF_pyfqHmknjT3P160lZ0qnfK9ujYzP1nsrjDz0Aw-5Hc3rHnYnHb0TAq15HfLPWRznjb0T1Y3PAnznjkn1m4ny7BP1f30AwY5HDdnHn3njc1PHn0lgF_5y9YlZ0lQzqBTLn8mLPbUB48ugfEUiqYULKuZNXuq99UHqdlAdxTvqdThP-5yF_UvTkn0KzujYk0AFV5H00TZcqn0KdpyfqHRLPjnvnfKEpyfqHc4rj6kP0KWpyfqP1cvrHnz0AqLUWYs0ZK45HcsP6KWThnqnWb4rj6)

5Hc3rHnYnHb0TAq15HfLPWRznjb0T1Y3PAnznjkn1m4ny7BP1f30AwY5HDdnHn3njc1PHn0lgF_5y9YlZ0lQzqBTLn8mLPbUB48ugfEUiqYULKuZNXuq99UHqdlAdxTvqdThP-

5yF_UvTkn0KzujYk0AFV5H00TZcqn0KdpyfqHRLPjnvnfKEpyfqHc4rj6kP0KWpyfqP1cvrHnz0AqLUWYs0ZK45HcsP6KWThnqnWb4rj6)



Gensim Word2vec 使用教程 (http://blog.csdn.net/Star_Bob/article/details/47808499)



本文主要基于Radim Rehurek的Word2vec Tutorial.**准备输入**Gensim的word2vec的输入是句子的序列. 每个句子是一个单词列表代码块例如：>>> # impor...



Star_Bob (http://blog.csdn.net/Star_Bob) 2015年08月20日 15:26 25674

Gensim实战（一）(<http://blog.csdn.net/u013776640/article/details/42347983>)

作为自然语言处理爱好者，大家都应该听说过或使用过大名鼎鼎的Gensim吧，这个一款具备多种功能的神器，为了深入了解该工具的使用方法，本人将使用该工具进行一系列实战。 该系列博客共分为以下...



u013776640 (<http://blog.csdn.net/u013776640>) 2015年01月02日 23:24 7386

python 环境下gensim中的word2vec的使用笔记 (<http://blog.csdn.net/philosophyatmath/article/details/53870117>)
加入CSDN，享受更精准的内容推荐，与5000万程序员共同成长！

(<https://passport.csdn.net/a>)



内容举报



返回顶部

登录

注册



centos 7, python2.7, gensim (0.13.1)语料 : <http://211.136.8.18/files/10940000015A9F94/mattmahoney.net/dc...>



philosophyatmath (<http://blog.csdn.net/philosophyatmath>) 2016年08月29日 16:57 17055



2.30/条
厂家定做cat6 七类跳
线ftp/stp BC 纯铜屏蔽



18.00/包
【批发】塑料吸盘定位
片25*25 不干胶自粘式



460.00/箱
PHILIPS飞利浦正品网
线SWA6310/93-305米

Gensim官方教程翻译（二）——语料库与向量空间（Corpora and Vector Spaces）(<http://b...>)



本文内容：如何利用gensim将文本信息转换为分析用的语料库，以及如何读取/存储语料库。...



questionfish (<http://blog.csdn.net/questionfish>) 2015年07月03日 11:02 7383



gensim 实践篇 (<http://blog.csdn.net/zhangxb35/article/details/73333633>)



继上篇文章了解了一些模型的基本原理以后，这里来讲讲怎么用 gensim，主要参考官方网站的 gensim: Tutorials，这篇博文也只是简单记下一点笔记。主要有三块内容，先讲怎么把文档表示成向量...



zhangxb35 (<http://blog.csdn.net/zhangxb35>) 2017年06月16日 16:21 1016

gensim学习笔记（一）- Vector space model (http://blog.csdn.net/John_xyz/article/details...)

gensim是基于python的自然语言处理库，可以自动的从文档中提取特征，语义信息等等。包括向量空间模型，word2vec, LSI, LDA, 转换之类的操作，非常方便。下面总结一些其基本用法，具...



John_xyz (http://blog.csdn.net/John_xyz) 2017年01月25日 23:02 999

Python 文本挖掘：使用gensim进行文本相似度计算 (<http://blog.csdn.net/chencheng126/art...>)

加入CSDN，享受更精准的内容推荐，与5000万程序员共同成长！

(<https://passport.csdn.net/a>)



内容举报





返回顶部

登录


注册





在文本处理中，比如商品评论挖掘，有时需要了解每个评论分别和商品的描述之间的相似度，以此衡量评论的客观性。评论和商品描述的相似度越高，说明评论的用语比较官方，不带太多感情色彩，比较注重描述商品的...


 chencheng126 (<http://blog.csdn.net/chencheng126>) 2015年11月27日 14:38  21633



[python] 使用scikit-learn工具计算文本TF-IDF值 (<http://blog.csdn.net/Eastmount/article/de...>)

 在文本聚类、文本分类或者比较两个文档相似程度过程中，可能会涉及到TF-IDF值的计算。这里主要讲述基于Python的机器学习模块和开源工具：scikit-learn。文章包括：一.Scikit-lea...

 Eastmount (<http://blog.csdn.net/Eastmount>) 2016年08月08日 16:46  17335



Python脚本中写日志的问题 (<http://blog.csdn.net/Alex1syyl/article/details/51511354>)

 可能会有很多人用Python脚本进行测试等工作，在挂机测试的过程中，随时都可能会出现错误，因此，写日志的功能必不可少。那么，日志的路径怎么写呢，很简单，把当前的目录添加到文件中即可对吧，但是如果这个文...

 Alex1syyl (<http://blog.csdn.net/Alex1syyl>) 2016年05月26日 22:59  751

文本主题模型之潜在语义索引(LSI) (<http://blog.csdn.net/suv1234/article/details/72851262>)

在文本挖掘中，主题模型是比较特殊的一块，它的思想不同于我们常用的机器学习算法，因此这里我们需要专门来总结文本主题模型的算法。本文关注于潜在语义索引算法(LSI)的原理。 1. 文本主题模型的问题特点...

 suv1234 (<http://blog.csdn.net/suv1234>) 2017年06月03日 12:03  233

NLP02-Gensim语料与向量空间 (<http://blog.csdn.net/ld326/article/details/78353338>)

摘要：对Gensim的语料与向量空间的官方文档的学习，对相关内容进行记录与翻译，并实践操作进行记录。gensim使用文档：《Corpora and Vector Space》来源：<https://github.com/ld326/gensim> 加入CSDN，享受更精准的内容推荐，与5000万程序员共同成长！

(<https://passport.csdn.net/a>)



内容举报




返回顶部

登录


注册



 ld326 (<http://blog.csdn.net/ld326>) 2017年10月26日 14:20 121

gensim文档-语料库与向量空间 (<http://blog.csdn.net/w5310335/article/details/49514815>)

gensim文档-语料库与向量空间 import logging logging.basicConfig(format='%(asctime)s : %(le...

 w5310335 (<http://blog.csdn.net/w5310335>) 2015年10月30日 17:52 394

0

文本分析--Gensim向量空间 (<http://blog.csdn.net/kevinelstri/article/details/70145681>)


-*- coding: utf-8 -*- import gensim "" Tutorial 1: Corpora and Vector Spaces "" import logginglog...

 kevinelstri (<http://blog.csdn.net/kevinelstri>) 2017年04月12日 17:51 609




线性代数笔记（1）：向量空间与子空间 (<http://blog.csdn.net/u010480899/article/details/556...>)

一、向量空间的定义：A vector space V over a field F consists of a set on which two operations (called addition...

 u010480899 (<http://blog.csdn.net/u010480899>) 2017年02月18日 20:53 997

【python gensim使用】word2vec词向量处理英文语料 (<http://blog.csdn.net/jdbc/article/de...>)

word2vec是google的一个开源工具，能够根据输入的词的集合计算出词与词之间的距离。它将term转换成向量形式，可以把对文本内容的处理简化为向量空间中的向量运算，计算出向量空间上的相似度，来表...

 jdbc (<http://blog.csdn.net/jdbc>) 2016年11月26日 19:05 3723

【python gensim使用】word2vec词向量处理英文语料 (<http://blog.csdn.net/churximi/articl...>)

加入CSDN，享受更精准的内容推荐，与5000万程序员共同成长！

(<https://passport.csdn.net/a>)



内容举报




返回顶部

登录


注册



word2vec介绍word2vec官网：<https://code.google.com/p/word2vec/> word2vec是google的一个开源工具，能够根据输入的词的计算算出词与词之间的...


 churximi (<http://blog.csdn.net/churximi>) 2016年05月21日 20:36 11211


在unity向量空间内绘制几何(1):通过将极坐标转换为直角坐标，绘制阿基米德螺线，对数螺线与...

 极坐标内的每个点都有两个参数： r 与 θ 。 r 为此点到极点（中心点）的距离， θ 为此点到极点的线段与极轴（类似 x 轴）的夹角。很多几何图形公式都可以用极坐标简洁的表示，例如：阿...

 liu_if_else (http://blog.csdn.net/liu_if_else) 2016年05月20日 00:38 2509

1. 向量空间的定义 (http://blog.csdn.net/my_live_123/article/details/76600162)

 上的加法和标量的乘法具有的性质： 1. 加法具有交换性和结合性，并且具有单位元；每个元素都有加法逆元。 2. 标量乘法具有结合性，1乘向量不改变该向量； 3. 分配性质将...

 my_live_123 (http://blog.csdn.net/my_live_123) 2017年08月03日 00:04 193

(<https://passport.csdn.net/a>



内容举报



返回顶部

加入CSDN，享受更精准的内容推荐，与5000万程序员共同成长！

登录

注册

