





1、 代码实现

([http://blog.csdn.net/sinat\\_26917383/article/details/51620019](http://blog.csdn.net/sinat_26917383/article/details/51620019))  
📖 20812

持数种单词相似度任务:  
相似词+相似系数 (model.most\_similar)、model.doesnt\_match、model.similarity (两两相似)

```
1 model.most_similar(positive=['woman', 'king'], negative=['man'], topn=1)
2 [('queen', 0.50882536)]
3
4 model.doesnt_match("breakfast cereal dinner lunch".split())
5 'cereal'
6
7 model.similarity('woman', 'man')
8 .73723527
```

2、 词向量

通过以下方式来得到单词的向量:

```
1 model['computer'] # raw NumPy vector of a word
2 array([-0.00449447, -0.00310097, 0.02421786, ...], dtype=float32)
```

案例一：800万微信语料训练

来源于：【不可思议的Word2Vec】2.训练好的模型 (<http://spaces.ac.cn/archives/4304/>)

训练语料	微信公众号的文章，多领域，属于中文平衡语料
语料数量	800万篇，总词数达到650亿
模型词数	共352196词，基本是中文词，包含常见英文词
模型结构	Skip-Gram + Huffman Softmax
向量维度	256维
分词工具	结巴分词，加入了有50万词条的词典，关闭了新词发现
训练工具	Gensim的Word2Vec，服务器训练了7天
其他情况	窗口大小为10，最小词频是64，迭代了10次

[http://blog.csdn.net/sinat\\_26917383](http://blog.csdn.net/sinat_26917383)

训练过程：


```
1 import gensim, logging
2 logging.basicConfig(format='%(asctime)s : %(levelname)s : %(message)s', level=logging.INFO)
3
4 import pymongo
5 import hashlib
6
7 db = pymongo.MongoClient("172.16.0.101").weixin.text_articles_words
8 md5 = lambda s: hashlib.md5(s).hexdigest()
9
10 class sentences:
11     def __iter__(self):
12         texts_set = set()
13         for a in db.find(no_cursor_timeout=True):
14             if md5(a["text"].encode("utf-8")) in texts_set:
15                 continue
16             else:
17                 texts_set.add(md5(a["text"].encode("utf-8")))
18                 yield a["words"]
19
20 print u'最终计算了%s篇文章'%len(texts_set)
21
22 word2vec = gensim.models.word2vec.Word2Vec(sentences(), size=256, window=10, min_count=64, sg=1, hs=1, iter=10, wv
23 word2vec.save("word2vec_wx")
```

这里引入hashlib.md5是为了对文章进行去重（本来1000万篇文章，去重后得到800万），而这个步骤不是必要的。

参考于：

基于python的gensim word2vec训练词向量 (<http://blog.csdn.net/lk7688535/article/details/52798735>)  
Gensim Word2vec 使用教程 ([http://blog.csdn.net/Star\\_Bob/article/details/47808499](http://blog.csdn.net/Star_Bob/article/details/47808499))  
官方教程：<http://radimrehurek.com/gensim/models/word2vec.html>  
(<http://radimrehurek.com/gensim/models/word2vec.html>)

版权声明：本文为博主原创文章，转载请注明来源“素质云博客”，谢谢合作！！微信公众号：素质云笔记


 发表你的评论

([http://my.csdn.net/weixin\\_35068028](http://my.csdn.net/weixin_35068028))

相关文章推荐

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

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 11235

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 25268



霸气！重磅改革！吴恩达说：女儿识字后就教她学Python！

Python的火爆最近越来越挡不住了，连身边多年工作经验的朋友都开始学Python了！他是这么说的....

(http://www.baidu.com/cb.php?c=lgF\_pyfqHmknjnvPjc0IZ0qnFK9ujYzP1ndPWb10Aw-5Hc3rHnYnHb0TAq15HfLPWRznb0T1dhuHD4PH7hPA7BPHIWPhmv0AwY5HDdnHn1rHDsPH60lgF\_5y9YIZ0IQzq-uZR8mLPbUB48ugfEIAqspynETZ-YpAq8nWqdlAdxTvqdThP-5yF\_UvTkn0KzujYk0AFV5H00TZcqn0KdpyfqHRLPjnvnfKEpyfqHc4rj6kP0KWpyfqP1cvrHnz0AqLUWYs0ZK45HcsP6KWThnqH01P6)

Gensim入门教程 (http://blog.csdn.net/Together\_CZ/article/details/7288505)

转自: http://www.cnblogs.com/iloveai/p/gensim\_tutorial.html What is Gensim? Gensim是一款开源的第三方Py...

Together\_CZ (http://blog.csdn.net/Together\_CZ) 2017年06月06日 22:08 348

【gensim中文教程】开始使用gensim (http://blog.csdn.net/DuinoDu/article/details/766186...)

原文链接介绍了基本概念，以及理解和使用gensim的基本元素，并提供了一个简单的例子。...

DuinoDu (http://blog.csdn.net/DuinoDu) 2017年08月03日 14:08 416

gensim 教程 -Part1 (http://blog.csdn.net/Thinking\_boy1992/article/details/53285592)

本文翻译自 Gensim使用Python的标准日志模型，在不同的优先级中来记录各种东西；为了激活日志，运行：>>> import logging >>> logging.basicConfig...

Thinking\_boy1992 (http://blog.csdn.net/Thinking\_boy1992) 2016年12月04日 09:31 324



程序员跨越式成长指南

完成第一次跨越，你会成为具有一技之长的开发者，月薪可能翻上几番；完成第二次跨越，你将成为拥有局部优势或行业优势的专业人士，获得个人内在价值的有效提升和外在收入的大幅跃迁.....

(http://www.baidu.com/cb.php?c=lgF\_pyfqHmknjzrjD0IZ0qnFK9ujYzP1f4PjnY0Aw-5Hc4nj6vPjm0TAq15Hf4rjn1n1b0T1YynvcsmlTduHKBNWRvPAF-0AwY5HDdnHn1rHDsPH60lgF\_5y9YIZ0IQzqMpgwBUvqoQhP8QvGIAPCmgfEmvq\_lyd8Q1R4uWc4uHf3uAckPHRkPWN9PhcsmW9huWqdlAdxTvqdThP-5HDknWFBmhkEusKzujYk0AFV5H00TZcqn0KdpyfqHRLPjnvnfKEpyfqHnsnj0YnsKWpyfqP1cvrHnz0AqLUWYs0ZK45HcsP6KWThnqP1T1r0)

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 12392

gensim做主题模型 (http://blog.csdn.net/luzhonghe1991/article/details/52196026)


作为python的一个库，gensim给了文本主题模型足够的方便，像他自己的介绍一样，topic modelling for humans 具体的tutorial可以参看他的官方...

luzhonghe1991 (http://blog.csdn.net/luzhonghe1991) 2016年08月12日 22:55 914



**基于gensim的Doc2Vec简析 (http://blog.csdn.net/lenbow/article/details/52120230)**

摘要：本文主要描述了一种文章向量（doc2vec）表示及其训练的相关内容，并列出了相关例子。两位大牛Quoc Le 和 Tomas Mikolov（搞出Word2vec的家伙）在2014年的《Distr...

 lenbow (http://blog.csdn.net/lenbow) 2016年08月04日 16:10 24069


**GENSIM 使用笔记2 — 主题模型和相似性查询 (http://blog.csdn.net/MebiuW/article/details/5...**

GENSIM 使用笔记1 — 语料和向量空间 GENSIM 使用笔记2 — 主题模型和相似性查询 在上一个笔记当中，使用gensim针对中文预料创建了字典和语料库，在这一章节中，主要讲下如何创建相...

 MebiuW (http://blog.csdn.net/MebiuW) 2016年12月25日 17:04 1976


**gensim做主题模型 (http://blog.csdn.net/whzhcahzh/article/details/17528261)**

作为python的一个库，gensim给了文本主题模型足够的方便，像他自己的介绍一样，topic modelling for humans 具体的tutorial可以参看他的官方网页，当然是全英...

 whzhcahzh (http://blog.csdn.net/whzhcahzh) 2013年12月24日 15:28 17212


**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 7315


**词向量之加载word2vec和glove (http://blog.csdn.net/u010041824/article/details/70832295)**

1 Google用word2vec预训练了300维的新闻语料的词向量googlenews-vecctors-negative300.bin，解压后3.39个G。 可以用gensim加载进来，...

 u010041824 (http://blog.csdn.net/u010041824) 2017年04月26日 20:57 3861


**Gensim Word2vec 使用教程 (http://blog.csdn.net/qq\_36330643/article/details/78702223)**

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...

 qq\_36330643 (http://blog.csdn.net/qq\_36330643) 2017年12月03日 16:17 29


**DefaultKeyedVector和KeyedVector用法 (http://blog.csdn.net/Qidi\_Huang/article/details/5...**

【用法示例】 在 Android Framework 源码中经常可以看到使用 DefaultKeyedVector 类型的容器。举个例子，在 AudioManagerBas...

 Qidi\_Huang (http://blog.csdn.net/Qidi\_Huang) 2016年09月22日 08:03 3495



**gensim实现python对word2vec的训练和计算 (http://blog.csdn.net/qdhy199148/article/deta...**

词向量（word2vec）原始的代码是C写的，python也有对应的版本，被集成在一个非常牛逼的框架gensim中。 我在自己的开源语义网络项目graph-mind（其实是我自己写的小玩具）中使用了这...

 qdhy199148 (http://blog.csdn.net/qdhy199148) 2016年06月27日 12:29 8021



**Gensim官方教程翻译（三）——主题与转换（Topics and Transformations） (http://bloq.cs...**

gensim官方教程翻译。本篇主要介绍了gensim提供的各种空间向量模型转换方法及其使用。...

 questionfish (<http://blog.csdn.net/questionfish>) 2015年07月03日 15:38  5161

**Android的底层库libutils介绍 (<http://blog.csdn.net/yu741677868yu/article/details/50589292>)**

Android的底层库libutils介绍 第一部分 libutils概述 libutils是Android的底层库，这个库以C++实现，它提供的API也是C++的。Android的层次的...

 yu741677868yu (<http://blog.csdn.net/yu741677868yu>) 2016年01月26日 18:58  679



**Android 安全攻防（一）：SEAndroid的编译 (<http://blog.csdn.net/yiyaaixuexi/article/detail...>)**

SEAndroid的编译SEAndroid概述SEAndroid（Security-Enhanced Android），是将原本运用在Linux操作系统上的MAC强制存取管控套件SELinux，移植到...

 yiyaaixuexi (<http://blog.csdn.net/yiyaaixuexi>) 2012年12月19日 11:00  31621



**gensim中使用word2vec (<http://blog.csdn.net/a1368783069/article/details/52025764>)**

训练语料由于自己在csdn的上传空间不够，暂时将语料放在百度云上 链接: <https://pan.baidu.com/s/1qYKRXOo> 密码: 4psr 文件名是 text8 或者在参考文章...

 a1368783069 (<http://blog.csdn.net/a1368783069>) 2016年07月25日 17:34  6133

**python 环境下gensim中的word2vec的使用笔记 (<http://blog.csdn.net/philosophyatmath/ar...>)**

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  16943



5

