


```
5
6 #读取倚天屠龙记文本，进行分词后存在新的文件里
7 fin = open('倚天屠龙记_utf8.txt', 'r')
8 fou = open('倚天屠龙记_segmented.txt', 'w')
9
10 line = fin.readline()
11 while line:
12     newline = jieba.cut(line, cut_all=False) # 精确模式
13     str_out = ''.join(newline).encode('utf-8').replace(' ','').replace('?', '').replace('!', '').replace(' ','') \
14             .replace('(','').replace(')','').replace('(','').replace(')','').replace('(','').replace(')','') \
15             .replace('(','').replace(')','').replace('(','').replace(')','').replace('(','').replace(')','') \
16             .replace('(','').replace(')','').replace('(','').replace(')','').replace('(','').replace(')','') \
17             .replace('(','').replace(')','').replace('(','').replace(')','').replace('(','').replace(')','')
18     print str_out,
19     print >> fou, str_out
20     line = fin.readline()
21 fin.close()
22 fou.close()
23
```

分词后的结果如下，内心里一阵狂喜对不对？！

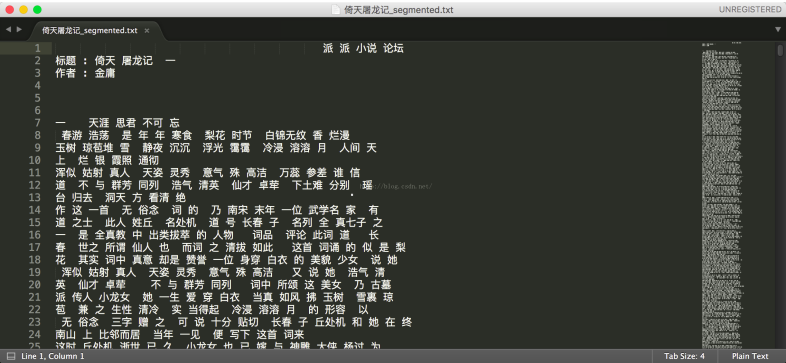
OpenCV和MFC的完美结合 (http://blog.csdn.net/gnehcuoz/article/details/8704012)
1332



内容举报



返回顶部



返回顶部

22

```
nlp_jizhi — -bash — 80x44
Last login: Wed Aug 3 11:25:55 on ttys000
[brycezoudeMacBook-Pro:~ brycezhou$ cd work/PyWork/machine_learning/nlp_jizhi/
[brycezoudeMacBook-Pro:nlp_jizhi brycezhou$ ls
nlp_chinese2utf8.py      倚天屠龙记.txt
nlp_chinese_segmentation.py  倚天屠龙记_model.txt
nlp_jizhi.iml           倚天屠龙记_segmented.txt
word2vec_gensim.py      倚天屠龙记_utf8.txt
[brycezoudeMacBook-Pro:nlp_jizhi brycezhou$ cat word2vec_gensim.py
#coding:utf8
import gensim.models.word2vec as w2v

model_file_name = '倚天屠龙记_model.txt'

...

#模型训练，生成词向量
sentences = w2v.LineSentence('倚天屠龙记_segmented.txt')
model = w2v.Word2Vec(sentences, size=20, window=5, min_count=5, workers=4)
model.save(model_file_name)

...

#使用已经训练好的模型
model = w2v.Word2Vec.load(model_file_name)

print model.similarity('赵敏'.decode('utf-8'), '赵敏'.decode('utf-8'))
print model.similarity('周芷若'.decode('utf-8'), '赵敏'.decode('utf-8'))
print model.similarity('韦一笑'.decode('utf-8'), '赵敏'.decode('utf-8'))
for k in model.similar_by_word('张三丰'.decode('utf-8')):
    print k[0], k[1]

[brycezoudeMacBook-Pro:nlp_jizhi brycezhou$ python word2vec_gensim.py
1.0
0.982774771556
0.80651853522
灭绝师太 0.988478422165
宋青书 0.979941904545
俞莲舟 0.979240238667
宋远桥 0.97885197401
莲舟 0.97605073452
张松溪 0.974860548973
金花婆婆 0.973897099495
纪晓芙 0.969041705132
殷梨亭 0.96669614315
空智 0.964550256729
brycezoudeMacBook-Pro:nlp_jizhi brycezhou$
```

Python源码

计算相似度

与'张三丰'相似的词

4、参考文献

1) 中文分词工具jieba : <http://www.oschina.net/p/jieba/?fromerr=s7MN6pKB> (<http://www.oschina.net/p/jieba/?fromerr=s7MN6pKB>)

2) NLP工具包gensim : <https://radimrehurek.com/gensim/models/word2vec.html> (<https://radimrehurek.com/gensim/models/word2vec.html>)

版权声明：本文为博主原创文章，转载请注明出处

内容举报
返回顶部



发表评论

http://my.csdn.net/weixin_35068028

ytong82 (ytong82) 4小时前

2楼

(/ytong82)拾的代码调试不通。。。 回复

u014697424 (/u014697424) 2017-08-01 16:39

1楼

(/u014697424)问一下先分词再去掉标点符号和先去掉标点符号再分词有没有区别呢？ 回复 1条回复

相关文章推荐

中文维基百科语料上的Word2Vec实验 (<http://blog.csdn.net/yangyangrenren/article/details/...>)

此文主要参考52nlp-中英文维基百科语料上的Word2Vec实验，按照上面的步骤来做的，略有改动，因此不完全是转载的。这里，为了方便大家可以更快地运行gensim中的word2vec模型，我提供了w...

yangyangrenren (<http://blog.csdn.net/yangyangrenren>) 2017年02月22日 12:10

5023

Delphi7高级应用开发随书源码 (<http://download.csdn.net/download/chenx...>)

<http://download.csdn.net/download/chenx...>

2003年04月30日 00:00 676KB

下载

2017年前端报告：程序员薪酬上涨70%！

前端程序员的薪酬曝光，2017年，平均上涨70%，月薪20的人最为常见！以下为详细数据.....

(http://www.baidu.com/cb.php?c=lgF_pyfqHmknj0dP1f0lZ0qnfK9ujYzP1nYPH0k0Aw-5Hc3rHnYnHb0TAq15HfLPWRznb0T1d9rAmvryDkuhPbPyPBnHR40AwY5HDdnHndPj6vP1f0lgF_5y9YlZ0lQzq-uZR8mLPbUB48ugfElAqspynElvNBnHqdlAdxTvqdThP-5yF_UvTkn0KzujYk0AFV5H00TZcqn0KdpyfqHRLPjnvnfKEpyfqHc4rj6kP0KWpyfqP1cvrHnz0AqLUWYs0ZK45HcsP6KWThnqPjR4rHm)

word2vec词向量训练及中文文本相似度计算 (<http://blog.csdn.net/Eastmount/article/details/...>)

本文是讲述如何使用word2vec的基础教程，文章比较基础，希望对你有帮助！ 官网C语言下载地址：<http://word2vec.googlecode.com/svn/trunk/Word2vec...>

Eastmount (<http://blog.csdn.net/Eastmount>) 2016年02月18日 00:35

46215

用word2vec 跑搜狗SogouCS语料 - 大小4G | 6.8 亿词长 | 57万词汇 (<http://blog.csdn.net/kev...>)

[训练] \$ time ./word2vec -train /data/sogou/sohunews_segmented_1line.txt -output /data/sogou/vectors_s...

kevinew (<http://blog.csdn.net/kevinew>) 2013年09月06日 20:51

13807



利用 word2vec 训练的字向量进行中文分词 (<http://blog.csdn.net/peghoty/article/details/171...>)

内容举报

返回顶部



最近针对之前发表的一篇博文《Deep Learning 在中文分词和词性标注任务中的应用》中的算法做了一个实现，感觉效果还不错。本文主要是将我在程序实现过程中的一些数学细节整理出来，借此优化一下自己的...

 peghoty (<http://blog.csdn.net/peghoty>) 2013年12月04日 18:28  23618





一学就会的 WordPress 实战课

认真学习完这个系列课程之后，会深入了解 WordPress 的使用和开发，并掌握基本的 WordPress 的开发能力，后续可以根据需要开发适合自己的主题、插件，打造最个性的 WordPress 站点。

(http://www.baidu.com/cb.php?c=lgF_pyfqnHmknjTYPHf0IZ0qnfK9ujYzP1f4Pjnd0Aw-5Hc4nj6vPjm0TAq15Hf4rjn1n1b0T1Ysnj7bnjTsuAf3nhN9uWfv0AwY5HDdnHndPj6vP1m0lgF_5y9YIZ0lQzqMpgwBUvqoQhP8QvIGIAPCmgfEmvq_lyd8Q1N9nHmvnj7hnHPWnjFhPAD1Pyn4uW99ujqdlAdxTvqdThP-5HDznHN9mhcEuskzujYk0AFV5H00TZcq0KdpfyqnHRLPjnvnfKEpyfqnHnsnj0YnsKWpyfqP1cvrHnz0AqLUWYs0ZK45HcsP6KWThnqPHmLPWc)



word2vec中文相似词计算和聚类的使用说明及c语言源码 (<http://blog.csdn.net/Eastmount/ar...>)

word2vec使用说明及源码介绍 1.下载地址 2.中文语料 3.参数介绍 4.计算相似词语 5.三个词预测语义语法关系 6.关键词聚类 -train Result_Country.txt 表示的是...

 Eastmount (<http://blog.csdn.net/Eastmount>) 2016年02月20日 01:53  4144



Google开源的Deep-Learning项目word2vec处理中文 (<http://blog.csdn.net/jdbc/article/deta...>)

推荐word2Vec，说的非常强大、有意思。故找了篇文章看，分享下。全文转自<http://www.cnblogs.com/wowarsenal/p/3293586.html> google最近...

 jdbc (<http://blog.csdn.net/jdbc>) 2015年10月29日 15:59  1485

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

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出词与词之间的...

 churximi (<http://blog.csdn.net/churximi>) 2016年05月21日 20:57  25189





Delphi7高级应用开发随书源码 (<http://download.csdn.net/download/chenx...>)

(<http://download.csdn.net/download/chenx...>) 2003年04月30日 00:00 676KB [下载](#)



【python gensim使用】word2vec词向量处理中文语料 (<http://blog.csdn.net/BigBzheng/art...>)

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出词...

 BigBzheng (<http://blog.csdn.net/BigBzheng>) 2016年08月17日 22:30  1330

TensorFlow实战11：实现Word2Vec (<http://blog.csdn.net/Felaim/article/details/70160667>)

1.Word2Vec简介Word2Vec也称Word Embeddings，中文比较常见的叫法是“词向量”或者是“词嵌入”。通俗的说就是把单词进行编码，变成数字的形式让计算机知道那个单词的代号。哈哈...


 Felaim (<http://blog.csdn.net/Felaim>) 2017年04月13日 18:15  1810

 内容举报

 返回顶部

 22




 [Fidant \(http://blog.csdn.net/mylove0414\)](http://blog.csdn.net/mylove0414) 2017年04月13日 10:13 评论444

tensorflow实例：实现word2vec语言模型 (http://blog.csdn.net/mylove0414/article/details/6...

本文算是对上一篇博文大白话讲解word2vec到底在做什么基于tensorflow的技术实现吧。...

 [mylove0414 \(http://blog.csdn.net/mylove0414\)](http://blog.csdn.net/mylove0414) 2017年04月08日 23:41 评论3205




Delphi7高级应用开发随书源码 (http://download.csdn.net/download/chenx...

<http://download.csdn.net/download/chenx...> 2003年04月30日 00:00 676KB [下载](#)

用Word2vec训练中文wiki，构造词向量并做词聚类 (http://blog.csdn.net/accumulate_zhang...

I利用word2vec训练中文wiki,构造词向量,并搞搞词聚类。

 [accumulate_zhang \(http://blog.csdn.net/accumulate_zhang\)](http://blog.csdn.net/accumulate_zhang) 2016年09月25日 14:57 评论5574




Delphi7高级应用开发随书源码 (http://download.csdn.net/download/chenx...

<http://download.csdn.net/download/chenx...> 2003年04月30日 00:00 676KB [下载](#)


word2vec实战：获取和预处理中文维基百科(Wikipedia)语料库，并训练成word2vec模型 (http...

前言自然语言处理有很多方法，最近很流行的是谷歌开源项目word2vec，详见谷歌官网：官网链接。其主要理论由Tomas Mikolov大神团队的2篇论文组成：Efficient Estimation ...

 [qq_32166627 \(http://blog.csdn.net/qq_32166627\)](http://blog.csdn.net/qq_32166627) 2017年04月01日 10:59 评论2227

使用word2vec训练wiki中文语料 (http://blog.csdn.net/sparkexpert/article/details/68921780)

实验环境：Ubuntu + eclipse + python3.5 首先（1）下载最新中文wiki语料库：wget https://dumps.wikimedia.org/zhwiki/la...

 [sparkexpert \(http://blog.csdn.net/sparkexpert\)](http://blog.csdn.net/sparkexpert) 2017年03月31日 09:47 评论1352


用word2vec 测试腾讯新闻语料 (一) (http://blog.csdn.net/kevinew/article/details/11064859)

有的效果还可以，有的不行，可能是数据太少了的原因。 效果比较好的词：1. Enter word or sentence (EXIT to break): 足球 Word: 足球 Posit...

 [kevinew \(http://blog.csdn.net/kevinew\)](http://blog.csdn.net/kevinew) 2013年09月04日 16:02 评论11032

THCHS-30：一个免费的中文语料库 (http://blog.csdn.net/sut_wj/article/details/70662181)

本文主要介绍了一个免费的开源中文语音识别数据库，附带的一些资源也做出了说明，例如语典，LM，和一些训练方法...

 [sut_wj \(http://blog.csdn.net/sut_wj\)](http://blog.csdn.net/sut_wj) 2017年04月24日 20:51 评论2404

word2vec使用指导 (http://blog.csdn.net/zhoub1668/article/details/24314769)

 内容举报

 返回顶部

 内容举报

 返回顶部

word2vec是一个将单词转换成向量形式的工具。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，来表示文本语义上的相似度。 一、理论概述（主要来源于[!\[\]\(34b4f260a8587d2e97eeaee361cc357b_img.jpg\) zhoubl668 \(<http://blog.csdn.net/zhoub668>\) 2014年04月22日 16:34 !\[\]\(b5f3742814ad7ea0f0989480e393a386_img.jpg\) 131511](http://l...</p></div><div data-bbox=)



22



内容举报



返回顶部