

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

使用python进行简单的分词与词云

原创

2017年08月22日 10:45:44

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

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

193

目标：

1. 导入一个文本文件
2. 使用jieba对文本进行分词
3. 使用wordcloud包绘制词云

环境：

Python 3.6.0 |Anaconda 4.3.1 (64-bit)



weixin_3506...

(//my.csdn.net/)

(//write.blog.csdn.net/postedit/activity?ref=toolbar)

ref=toolbar)source=csdnblog

番番要吃肉 (http://blog.csdn.net/xiexf189)

+ 关注

(http://blog.csdn.net/xiexf189)

码云

未开通
(https://gitee.com/xiexf189)

原创
4

粉丝
4

喜欢
0

未开通
(https://gitee.com/xiexf189)

他的最新文章

更多文章 (http://blog.csdn.net/xiexf189)

Python数据分析练习：北京、广州PM 2.5空气质量分析（2）(http://blog.csdn.net/xiexf189/article/details/77368583)

Python数据分析练习：北京、广州PM 2.5空气质量分析（1）(http://blog.csdn.net/xiexf189/article/details/77367504)

Python-sklearn 机器学习的第一个样例（7）(http://blog.csdn.net/xiexf189/article/details/77477283)

立即体验



¥30.00/件



内容举报



返回顶部

工具：

jupyter notebook

从网上下载了一篇小说《老九门》，以下对这篇小说进行分词，并绘制词云图。

分词使用最流行的分词包jieba，参考：<https://github.com/fxsjy/jieba> (<https://github.com/fxsjy/jieba>)

词云使用wordcloud包，参考：https://github.com/amueller/word_cloud (https://github.com/amueller/word_cloud)

这两个包都不是anaconda自带的，需要按官网的步骤安装。

In [1]:

```
import wordcloud as wc
import jieba
import matplotlib.pyplot as plt
from scipy.misc import imread

%matplotlib inline

plt.rc('figure', figsize=(15, 15))
```

首先读取文件，保存到一个字符串对象中。

In [2]:

```
all_text = open(file='老九门.txt', encoding='utf-8').read()
```

查看一下字符串的内容，发现其中有很多多余的字符：'\n'、'\u3000'。

In [3]:

```
all_text
```

Out[3]:

[cle/details/72598976](#))

Python-sklearn机器学习的第
(6) (<http://blog.csdn.net/x>
[cle/details/72598910](#))

Python-sklearn机器学习的第
(5) (<http://blog.csdn.net/x>
[cle/details/72560725](#))

相关推荐

Python 使用itchat 对微信好友数据进行简单分析 (<http://blog.csdn.net/vcx08/article/details/72126189>)

Python 使用pdb进行简单调试 (<http://blog.csdn.net/cay22/article/details/8938677>)

使用Python进行简单的验证码识别 (<http://blog.csdn.net/ethantang520/article/details/52102284>)

简单数据预测—使用Python训练回归模型并进行预测（转自蓝鲸网站分析博客） (<http://blog.csdn.net/u011089412/article/details/66967993>)



广告 ¥30.00/件



内容举报



返回顶部

'\ufe0f《盗墓笔记》中，一段与二月红有关的故事。《老九门》壹：二月红①丝帐许久没有换过了。半夜入不了眠，睁开眼睛，便看到床边垂下的帐面，在月光下看着有一死暗淡。原来可是丝丝的带着光亮，好像最白的银拉出来的丝一般。果然再好的东西，也总是由好往坏了去。以往一过立秋，... <以下省略>

在分词之前先把这些多余字符剔除掉。

In [4]:

```
all_text = all_text.replace('\n', ' ')
all_text = all_text.replace('\u3000', ' ')
```

下面先尝试做一次分词，把所有分词用空格分开，输出看一下分词的结果：

In [6]:

```
seg_list = jieba.cut(all_text, cut_all=False)
words = ' '
for seg in seg_list:
    words = words + seg + ' '
print(words)
```

Out[6]:

《盗墓笔记》中，一段与二月红有关的故事。《老九门》壹：二月红①丝帐许久没有换过了。她半夜入不了眠，睁开眼睛，便看到床边垂下的帐面，在月光下看着有一死暗淡。原来可是丝丝的带着光亮，好像最白的银拉出来的丝一般。果然再好的东西，也总是由好往坏了去。以往一过立秋，她就会亲自拆下这块帐头，亲自去漂洗，她知道这东西的脾气，得小心伺候着，一寸一寸地过水。如今不让她下床，这东西没人伺候了，倒也显得越来越不值当被这么细心对待起来。也许，下一个立秋的时候，才有人敢动这个东西，但那个人，必然不是自己了。中午大夫和他说的这些话，虽然是在屋外，但是她还是听到了几分，自己的病，不知道还有多少日子可熬。她舒了口气，胸中的那丝痛楚似乎好了一些。多少日... <以下省略>

从分词结果里可以发现，有一些固定词语，例如“盗墓笔记”、“老九门”、“二月红”、“张大佛爷”、“齐铁嘴”等书名、人名被分开了。在这篇小说的环境下，这些才成为固定词语，而默认的分词策略根据通常的认识来分词的。



电脑硬件学习



Python机器学习



机



创



电子
¥30.00/件



发际线高



术前
咬肌肥大

橡树湾 Python机器学习 电脑硬件学习
整容的费用 让鼻翼缩小 补牙的危害
矫正团购 机器学习python 联合办公
开后眼角 超强注意力 it培训机构排

他的热门文章

Python数据分析练习：北京、广州PM2.5 内容举报
空气质量分析（1）(<http://blog.csdn.net/xiexf189/article/details/77367504>)

826



返回顶部

Python-sklearn机器学习的第一个样例
（6）(<http://blog.csdn.net/xiexf189/article/details/72598910>)

针对这个情况，jieba有一个“用户词典”的机制，把用户认为应该成为整词的，放到词典里，在分词的时候遇到它们，就会当作一个整体。

在目录下新建一个文本文件dict.txt，输入以下词语：

老九门

二月红



盗墓笔记

张大佛爷

解九爷



解九



狗五

以下为jieba引入这个词典文件，再做一次分词：

In [7]:

```
jieba.load_userdict('dict.txt')
all_seg = jieba.cut(all_text, cut_all=False)
all_word = ' '
for seg in all_seg:
    all_word = all_word + seg + ' '
print(all_word)
```

📖 737

Python-sklearn机器学习的第
(3) (<http://blog.csdn.net/xiexf189/article/details/72528755>)

📖 718

Python-sklearn机器学习的第
(2) (<http://blog.csdn.net/xiexf189/article/details/72528667>)

📖 589

Python-sklearn 机器学习的第
(1) (<http://blog.csdn.net/xiexf189/article/details/72518860>)

📖 497



广告

¥30.00/件



内容举报



返回顶部

《盗墓笔记》中，一段与二月红有关的故事。《老九门》壹：二月红①丝帐许久没有换过了。她半夜入不了眠，睁开眼睛，便看到床边垂下的帐面，在月光下看着有一死暗淡。原来可是丝丝的带着光亮，好像最白的银拉出来的丝一般。果然再好的东西，也总是由好往坏了去。以往一过立秋，她就会亲自拆下这块帐头，亲自去漂洗，她知道这东西的脾气，得小心伺候着，一寸一寸地过水。如今不让她下床，这东西没人伺候了，倒也显得越来越不值当被这么细心对待起来。也许，下一个立秋的时候，才有人敢动这个东西，但那个人，必然不是自己了。中午大夫和他说的这些话，虽然是在屋外，但是她还是听到了几分，自己的病，不知道还有多少日子可熬。她舒了口气，胸中的那丝痛楚似乎好了一些。多少日子了？

. 0. ...<以下省略>

从这个结果来看，就不会再把人名、书名分开了。

以下开始制作词云。

制作词云，使用的是wordcloud包，由两个参数需要特别注意，一个是字体，一个是背景图片。字体好理解，就不解释了。背景图片，是词云显示的背景形状。这里选用了—个心形图案。



以下是词云制作过程：

In [8]:



内容举报

返回顶部

```
# 引入字体
font=r"C:\WINDOWS\Fonts\simhei.ttf"

#读取背景图片,生成矩阵
color_mask = imread("love.jpg")

# 生成词云对象,设置参数
cloud = wc.WordCloud( font_path=font,#设置字体
                        background_color="black", #背景颜色
                        max_words=2000,# 词云显示的最大词数
                        mask=color_mask,#设置背景图片
                        max_font_size=100, #字体最大值
                        random_state=42)

# 绘制词云图
mywc = cloud.generate(all_word)
```



In [9]:

```
plt.imshow(mywc)
```

Out[9]:

<matplotlib.image.AxesImage at 0x1ecef5e588>



内容举报



返回顶部



 内容举报

 TOP

返回顶部

```
<wordcloud.wordcloud.WordCloud at 0x1ece4b9bc88>
```

以上就是使用python进行分词，并绘制词云图的简单操作。对于jieba和wordcloud的更高级的使用方法，还需要进一步研究和学习。



发表你的评论

(http://my.csdn.net/weixin_35068028)



相关文章推荐



Python 使用itchat 对微信好友数据进行简单分析 (<http://blog.csdn.net/vcx08/article/details/...>)

人生苦短，我用Python！Python 热度一直很高，我感觉这就是得益于拥有大量的包资源,极大的方便了开发人员的需求。最近在一个微信公众号上看到一个调用微信 API 可以对微信好友进行简单数据...



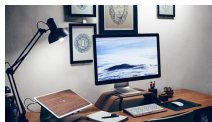
vcx08 (<http://blog.csdn.net/vcx08>) 2017年05月15日 12:14 697

Python 使用pdb进行简单调试 (<http://blog.csdn.net/cay22/article/details/8938677>)

Python 使用pdb进行简单调试 pdb 是 python 自带的一个包，为 python 程序提供了一种交互的源代码调试功能，主要特性包括设置断点、单步调试、进入函数调试、查看当前代码...



cay22 (<http://blog.csdn.net/cay22>) 2013年05月17日 10:56 2309



广告

票选结果：Python再上天，微软要求全员学Python？

宇宙语言Python荣登年度排行榜，吴恩达，微软纷纷为它站台，Python这么牛逼的原因是....



广告

¥30.00/件



内容举报



返回顶部

(http://www.baidu.com/cb.php?c=lgF_pyfqHmknjnvPjc0IZ0qnfK9ujYzP1nYPH0k0Aw-5Hc3rHnYnHb0TAq15HfLPWRznjb0T1Y4nH01uAD1PWbYrHT4PjnL0AwY5HDdnHfzrHDLPHb0IgF_5y9YIZ0IQzq-uZR8mLPbUB48ugfEIAqspynETZ-YpAq8nWqdlAdxTvqdThP-5yF_UvTkn0KzujYk0AFV5H00TZcqN0KdpyfqNHRLPjnvnfKEpyfqHc4rj6kP0KWpyfqP1cvrHnz0AqLUWYs0ZK45HcsP6KWThnqPWcvP6)

使用Python进行简单的验证码识别 (<http://blog.csdn.net/ethantang520/article/details/52102...>)



0、环境 系统：Windows7 旗舰版 64位 Python：2.7.12 Pycharm: profession 2016.2 1、资源集合...



ethantang520 (<http://blog.csdn.net/ethantang520>) 2016年08月03日 11:57 513

简单数据预测—使用Python训练回归模型并进行预测（转自蓝鲸网站分析博客）(<http://blog.c...>)

使用Python训练回归模型并进行预测 2016年9月2日 By 蓝鲸 1 Comment 回归分析是一种常见的统计方法，用于确定不同变量间的相互关系。在Excel中可以通过数据分析菜单中的回归功...



u011089412 (<http://blog.csdn.net/u011089412>) 2017年03月27日 09:28 3267

【学习笔记】使用Python对文件进行简单操作 (<http://blog.csdn.net/sealvor/article/details/5...>)

文件处理中常用的Python代码1



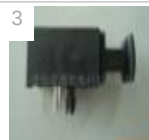
sealvor (<http://blog.csdn.net/sealvor>) 2016年08月26日 05:45 162



议价
供应特价销售 制造销售Fu"-50u"镀金 8P8C



18.00/片
DD001 蓝牙LED灯带模块 蓝牙4.0模块



1.50/个
供应TOTX147PL,光纤连接器



广告

¥30.00/件





内容举报




返回顶部

在Python中使用正则表达式同时匹配邮箱和电话并进行简单的分类 (<http://blog.csdn.net/tdmy...>)

在Python使用正则表达式需要使用re(regular exprssion)模块，使用正则表达式的难点就在于如何写好p=re.compile ('r' 正则表达式') 的内容。下面是在Python中使用...

 tdmyl (<http://blog.csdn.net/tdmyl>) 2013年09月04日 14:44  2410

使用Python进行简单的图像处理 (<http://blog.csdn.net/aoyingge/article/details/51679265>)

1、 环境搭建 python安装好后（我的版本是2.7），需要在安装Numpy和opencv（注意对应的版本号）1）numpy的安装：在<http://www.lfd.uci.edu/~gohl...>

 aoyingge (<http://blog.csdn.net/aoyingge>) 2016年06月15日 10:43  2889

  **c与python程序通过openMPI进行并行通信的简单例子** (<http://download.csdn.net/detail/...>)
 (<http://download.csdn.net/detail/...>) 2012年03月21日 10:38 3KB 



python中文分词，使用结巴分词对python进行分词 (<http://blog.csdn.net/yangjiyue0520/article/details/77477283>)

在采集美女站时,需要对关键词进行分词,最终采用的是python的结巴分词方法. 中文分词是中文文本处理的一个基础性工作，结巴分词利用进行中文分词。其基本实现原理有三点：基于Trie树结构实现...

 yangjiyue0520 (<http://blog.csdn.net/yangjiyue0520>) 2017年11月04日 14:53  33

Python使用jieba分词并用weka进行文本分类 (<http://blog.csdn.net/guohuiJI/article/details/77477283>)

一、安装pycharm 二、安装Python 三、在Python下安装pip，如下图所示，pip安装成功 四、在python下安装jieba：如下图所示，jieba安装成功：...

 guohuiJI (<http://blog.csdn.net/guohuiJI>) 2017年06月16日 14:41  270





内容举报


返回顶部


Python利用结巴分词进行中文分词 (http://blog.csdn.net/jiahui_zhu/article/details/50165771)

利用结巴分词进行中文分词，选择全模式，建立词倒排索引，并实现一般多词查询和短语查询 # -*- coding: utf-8 -*- import jieba "" Created on 2015-...

 jiahui_zhu (http://blog.csdn.net/jiahui_zhu) 2015年12月03日 20:24 3133


使用Python+jieba和java+庖丁分词在Spark集群上进行中文分词统计 (<http://blog.csdn.net/sh...>)

写在前边的话： 本篇博客也是在做豆瓣电影数据的分析过程中，需要对影评信息和剧情摘要信息进行分析而写的一篇博客 以前学习Hadoop时，感觉做中文分词也没那么...

 shuyun123456789 (<http://blog.csdn.net/shuyun123456789>) 2016年11月15日 19:14 582

利用Python对文本文件进行简单的处理 (<http://blog.csdn.net/studyhard232/article/details/6...>)

在诸多软件压缩包中或是项目压缩包中都会存在一个readme.txt文件，其中的内容无非是对软件的简单介绍和注意事项。但是在该文本文件中，内容没有分段分行，是非常冗杂地混在一起。当然处理手段多种多样，而...

 studyhard232 (<http://blog.csdn.net/studyhard232>) 2017年03月28日 23:21 895

python requests 自动管理 cookie 。 get后进行post发送数据 - - - 最简单的刷票 (<http://b...>)

Request URL: <http://musicman.migu.cn/activity/ccontent/voteWorks.do> Request Method: POST Sta...

 ipqhjjbj (<http://blog.csdn.net/ipqhjjbj>) 2013年10月20日 23:49 3868

Python3网络爬虫(一)：利用urllib进行简单的网页抓取 (<http://blog.csdn.net/u013383813/arti...>)



运行平台：Windows Python版本：Python3.x IDE：Sublime text3 转载 原作者和出处：<http://blog.csdn.net/c406495762...>



内容举报



返回顶部

 u013383813 (<http://blog.csdn.net/u013383813>) 2017年07月01日 19:10  509

利用Python进行数据分析(4) NumPy基础: ndarray简单介绍 (<http://blog.csdn.net/IAlexander...>)

一、NumPy 是什么 NumPy 是 Python 科学计算的基础包，它专为进行严格的数字处理而产生。在之前的随笔里已有更加详细的介绍，这里不再赘述。利用 Python 进行数据分析（一）简单介...




IAlexanderI (<http://blog.csdn.net/IAlexanderI>) 2017年11月23日 11:39  92



Python3 对网页进行简单爬虫操作 (http://blog.csdn.net/yuan_csdn1/article/details/7855331...)

序一直想好好学习一下Python爬虫，之前断断续续的把Python基础学了一下，悲剧的是学的没有忘的快。只能再次拿出来滤了一遍，趁热打铁，借鉴众多大神的爬虫案例，加入Python网络爬虫的学习大军~~...




yuan_csdn1 (http://blog.csdn.net/yuan_csdn1) 2017年11月16日 17:10  83

python进行文档抽取与解析的简单实现 (<http://blog.csdn.net/gugugujiawei/article/details/42...>)

python进行文档抽取与解析的简单实现




gugugujiawei (<http://blog.csdn.net/gugugujiawei>) 2015年01月18日 13:36  1476

linux下用python进行opencv开发----简单的图片操作 (<http://blog.csdn.net/topgun38/article/...>)

初学opencv做的例子程序，保存一下。之所以选择用python，是因为python上手快，开发快。#!/usr/bin/python2 # coding: utf-8 import cv2...



topgun38 (<http://blog.csdn.net/topgun38>) 2013年08月30日 11:02  9844



内容举报



返回顶部

Python连接数据库并进行简单操作整理 (<http://blog.csdn.net/u010159842/article/details/469...>)

下载安装MySQLdb 如果已经安装了easy_install插件，那么就好说了，你想装什么库或是包，只需使用easy_install + 库，就可以了。但是遇到了这个问题： ...



u010159842 (<http://blog.csdn.net/u010159842>) 2015年07月20日 09:30 375



0



内容举报



返回顶部