# 一、词云绘制

Python在数据分析中越来越受欢迎，已经达到了统计学家对R的喜爱程度，Python的拥护者们当然不会落后于R，开发了一个个好玩的数据分析工具，下面我们来看看如何使用Python，来读红楼梦，本文是《用Python读红楼梦》系列文章的第一篇，绘制小说中的词云。

首先当然要导入我们需要用到的包，下面import进来的包，都是我们将在接下来的程序中使用到的包，如果大家还没有安装它们，那么尽快安装它们吧。

import jieba

import numpy

import codecs

import pandas

import matplotlib.pyplot as plt

from wordcloud import WordCloud

接着，要分析词频，就要读取我们的《红楼梦》的文本数据，读取文本，我建议使用codecs包，它可以先通过设置文件的编码，对文件进行读入，这样子就不用边读遍转码了，非常实用。

file = codecs.open("D:\\红楼梦.txt", 'r', 'utf-8')

content = file.read()

file.close()

然后，我们就来分词了，中文分词，当然要用大名鼎鼎的jieba包，下面就是分词的方法。

这里我们需要注意两点：

1、为了提高分词的准确度，我们最好寻找我们分词的词库，这里我下载到了红楼梦的分词库，加载如jieba中，然后再进行分词。

2、对于小说中，一个字的词，基本上算是无用的词，或者说是标点符号，因此这里我直接抛弃了。

jieba.load\_userdict('D:\\红楼梦词库.txt');

segments = []

segs = jieba.cut(content)

for seg in segs:

if len(seg)>1:

segments.append(seg);

为了方便统计词频，我们把结果保存在pandas的DataFrame中。

segmentDF = pandas.DataFrame({'segment':segments})

接着我们来移除停用词，停用词包括我们日常的停用词和文言文中的停用词两部分，如下所示：

#移除停用词

stopwords = pandas.read\_csv(

"D:\\StopwordsCN.txt",

encoding='utf8',

index\_col=False,

quoting=3,

sep="\t"

)

segmentDF = segmentDF[~segmentDF.segment.isin(stopwords.stopword)]

wyStopWords = pandas.Series([

  # 42 个文言虚词

  '之', '其', '或', '亦', '方', '于', '即', '皆', '因', '仍', '故',

  '尚', '呢', '了', '的', '着', '一', '不', '乃', '呀', '吗', '咧',

  '啊', '把', '让', '向', '往', '是', '在', '越', '再', '更', '比',

  '很', '偏', '别', '好', '可', '便', '就', '但', '儿',

  # 高频副词

  '又', '也', '都', '要',

  # 高频代词

  '这', '那', '你', '我', '他',

  #高频动词

  '来', '去', '道', '笑', '说',

  #空格

  ' ', ''

]);

segmentDF = segmentDF[~segmentDF.segment.isin(wyStopWords)]

从上面的代码我们可以看到，pandas对数据的处理，真的是非常方便，更加方便的还在下面，我们接着来对词频进行统计。

segStat = segmentDF.groupby(

by=["segment"]

)["segment"].agg({

"计数":numpy.size

}).reset\_index().sort(

columns=["计数"],

ascending=False

);

segStat.head(100)

到这里，我们基本上可以得到词频了，观察一下下面的词频，贾宝玉当然当之无愧是主角，出现次数基本和换行符一致，哈哈，但是出乎意料的是，贾母，竟然是第二，其实这也难怪，红楼梦，哪个场景没有涉及到贾母的？贾宝玉玩完后，都说要到贾母那里请安或者吃饭的，情理之中。有此可见，最佳女配角，非贾母莫属了。



细心的读者可能也发现了，林黛玉出现的次数，竟然还排在了袭人的后面，其实笔者仔细一想，一点也不奇怪。

首先第一个，林黛玉挂得早，挂得早，出现的次数自然就没有那么多了。

还有另外一个原因，不知道看官们注意到不，贾宝玉和女生滚床单，第一个就是袭人（当然秦可卿是不算的，那纯属是贾宝玉自己梦中YY的），因此，你们懂的。如果大家对贾宝玉的私生活感兴趣，可以看这篇文章《[贾宝玉到底和多少人发生过性关系？](https://movie.douban.com/subject/3014183/discussion/24893203/)》

最后要告诉大家的是，黛玉只是林黛玉的昵称，加上妹妹（只是不知道贾宝玉有多少个妹妹咯）和连名带姓的林黛玉的词频，也是超越了袭人的，因此，红楼梦符合广电总局的规定——小三是不能上位的。



  segment    计数

14760      宝玉  3762

0        \r\n  3191

35682      贾母  1272

7738       凤姐  1192

34168      袭人  1134

40972      黛玉  1029

27448     王夫人  1015

13833      如今  1002

35130      说道   978

31820     老太太   974

29301      知道   973

36077      起来   955

14062      姑娘   949

7858       出来   932

4769       众人   872

821        一面   828

13305      太太   825

13686      奶奶   810

10094      只见   791

14774      宝钗   789

2211       两个   771

25441      没有   767

1737       不是   743

1828       不知   702

10940      听见   692

35734      贾琏   689

37195      进来   632

11053      告诉   605

2151       东西   603

16508      平儿   590

...       ...   ...

15894      屋里   286

37827     邢夫人   286

23866     林黛玉   278

15735      尤氏   277

39074      问道   275

28961      看见   271

14016      妹妹   270

1800       不用   265

9373       原来   258

40507      香菱   256

200        一句   255

15032      家里   254

1646       不得   254

248        一声   253

33649      薛蟠   253

14223      媳妇   249

8269       到底   247

36880     这会子   246

37178      进去   246

14033      姊妹   243

8129       别人   240

11753      回去   237

36611      过去   236

22448      明儿   236

21774      方才   233

40871      麝月   233

2446       丫鬟   233

37290      连忙   232

17554      心中   230

14200      婆子   225

为了更加方便地观察数据，我们使用词云的工具，来绘制词云，通过词云，我们可以更加方便简单地查看词频。

#绘画词云

wordcloud = WordCloud(

font\_path='D:\\simhei.ttf',

background\_color="black"

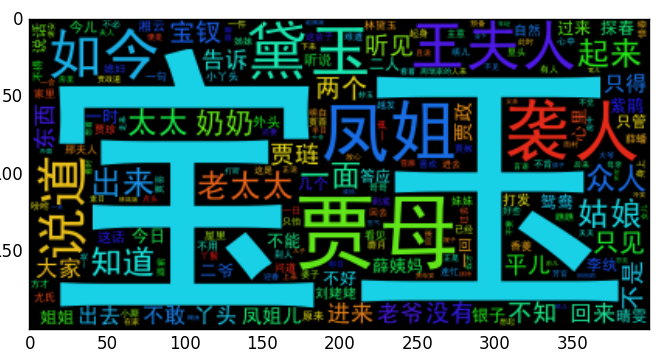
)

plt.figure(num=None, figsize=(1000, 600), dpi=800, facecolor='w', edgecolor='k')

wordcloud = wordcloud.fit\_words(segStat.head(1000).itertuples(index=False))

plt.imshow(wordcloud)

plt.close()



好了，这个就是用Python来看红楼梦的第一篇文章，如果你觉得这个系列的文章值得期待，欢迎关注笔者微博@大数据分析实战，欲知后续如何分析，请听下回分解。