

**《人民的名义》**

**豆瓣评论情感分析**

编 撰 人：范为明

审 核 人：

批 准 人：

批准日期：2017-5-20

保密级别：机密

文档版本：0.0.1

**曙光信息产业（北京）有限公司**

目录

[1.数据源的获取 2](#_Toc483417216)

[**1.1利用python进行数据爬取** 2](#_Toc483417217)

[**1.2使用八爪鱼采集器进行数据采集** 3](#_Toc483417218)

[2.文本的处理 3](#_Toc483417219)

[3.文本情感标注 4](#_Toc483417220)

[4.使用LDA模型来获取二类文档类型的关键词 5](#_Toc483417221)

[5.Conclusion 6](#_Toc483417222)

**1.数据源的获取**

所谓巧妇难为无米之炊，数据源的获取自然的成了我们进行文本情感分析的第一步。我们选择在文艺青年聚集地豆瓣网上进行文本评论的获取，我们前后进行了两种方式去爬取数据，一种就是利用Python构建短小精悍的爬虫小工具，另一种就是采用了八爪鱼采集器。

**1.1利用python进行数据爬取**

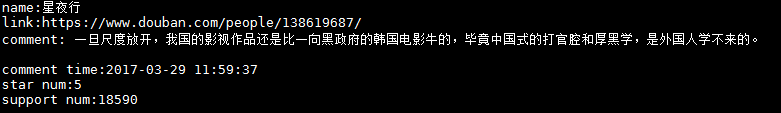
首先我们打开我们想要进行数据爬取的web页面，大家可以感受一下这部热剧的短评页面是这样的画风。



没错，以上就是这部热剧的短评的web页面。找到了想要爬取的页面，接下来的任务就好办多了。我们利用了python里面的BeautifulSoup和urllib2两个模块去解析html里面的东东。解析的核心代码如下，运行以上代码我们大致看一下结果是不是我们所想要的东西。



我们来看一下结果。



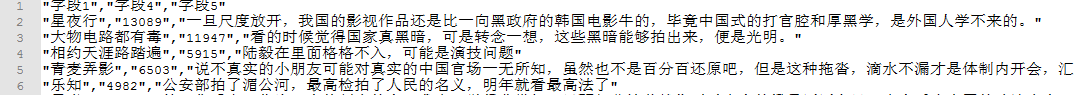
好的，确实是我们想要的东西。其中的comment的内容就是用户的短评。接下来我们接着通过设置user-agent（也就是模拟浏览器动作），设置爬取间隔，设置多用户代理和模拟登陆的操作去批量的爬取我们所需要的数据。但是值得一提的是现在豆瓣的反爬虫机制作的也是越来越好，笔者在进行大批量数据爬取的时候很容易的就会遇到IP被禁的问题，整得笔者痛苦不堪，好吧，我们在这里介绍一下另外一种的数据爬取方式。

**1.2使用八爪鱼采集器进行数据采集**

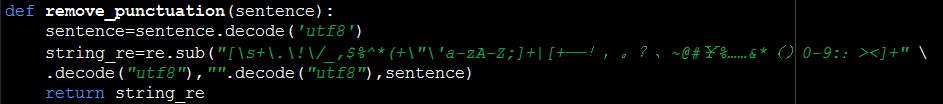
八爪鱼采集器是深圳视界信息技术有限公司研发的一款网页采集软件，具有使用简单，操作简便的特点，甚至可以说是傻瓜式的操作。在进行爬取时候，我们只需要选择给定的爬取模式以及进行爬取的url地址、url地址的变化规律，就可以一键式的进行数据的获取。数据获取可以选择将获取的评论以csv的格式给保存下来。

**2.文本的处理**

获取完评论之后，因为原始的文本数据因为无意义的字符，格式不统一等问题，需要进行文本清洗工作，以获得相对干净的文本数据。

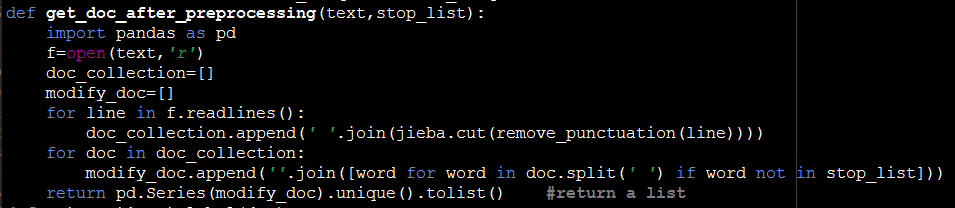


我们通过过滤特殊字符，去除标点，来获得干净的文本数据

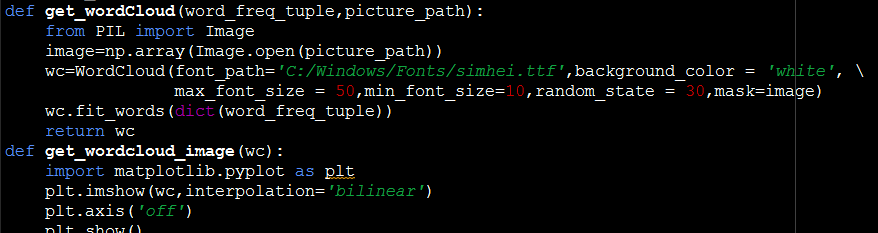


接下来，我们需要进行分词。这里我们选择了结巴分词来进行中文文本的分词。

我们将分词之后的文本转换成为python里面的list形式。



到此为止，我们算是完成了对文本的初步清理工作；这个时候，我们希望看一眼清洗完之后的评论文本的高频词呈现出怎样的可视化形式，我们利用wordclound和Image两个模块去构建可定制化的词云可视化工具。



我们可以选择我们喜欢的背景图片来充当词云的背景，最后我们构建出来的词云是这个样子的，大家来感受一下下。

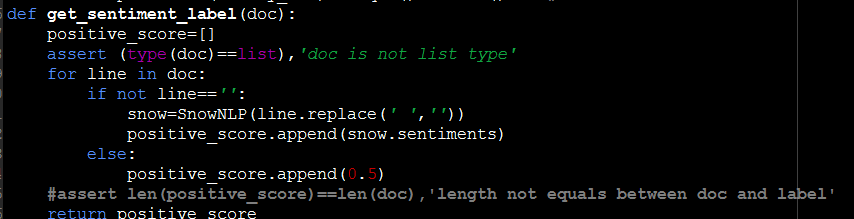


**3.文本情感标注**

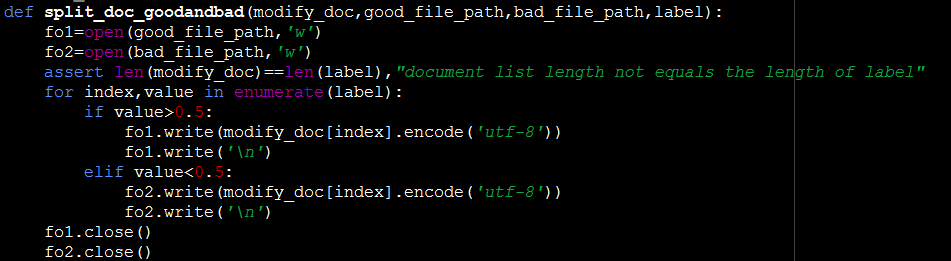
在文本的情感标注模块，我们使用了python下面的snownlp模块。SnowNlp是一个simplified Chinese Text Processing的模块，模块的github的地址为<https://github.com/isnowfy/snownlp>

SnowNLP是一个python写的类库，可以方便的处理中文文本内容，是受到了TextBlob的启发而写的，由于现在大部分的自然语言处理库基本都是针对英文的，于是写了一个方便处理中文的类库，并且和TextBlob不同的是，这里没有用NLTK，所有的算法都是自己实现的，并且自带了一些训练好的字典。注意本程序都是处理的unicode编码，所以使用时请自行decode成unicode。

因为SnowNlp模块的情感标注毕竟存在所局限的地方，所以目前最好的方法就是使用snownlp模块结合人工标注，对文本评论进行情感标注，积极情感为1，消极类的情感为0，中间是一些连续变化的实数。 越靠近实数1说明该文本越偏向于积极的情感指向。



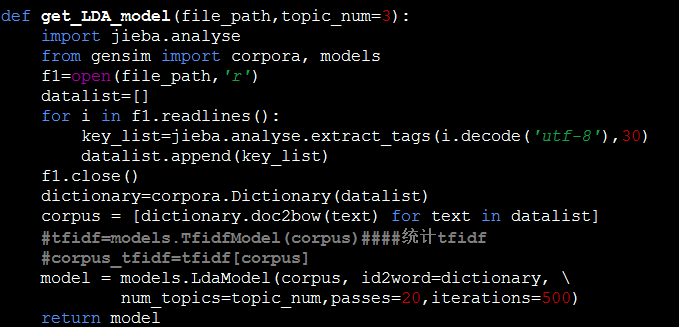
以上代码是利用SnowNlp来进行获取文本评论情感标注。然后我们根据文本标注值的大小来进行文本评论的二分类。Sentiment label超过0.5的评论将其划分为积极类文本，而label的得分低于0.5的部分被划分为消极类的情感。



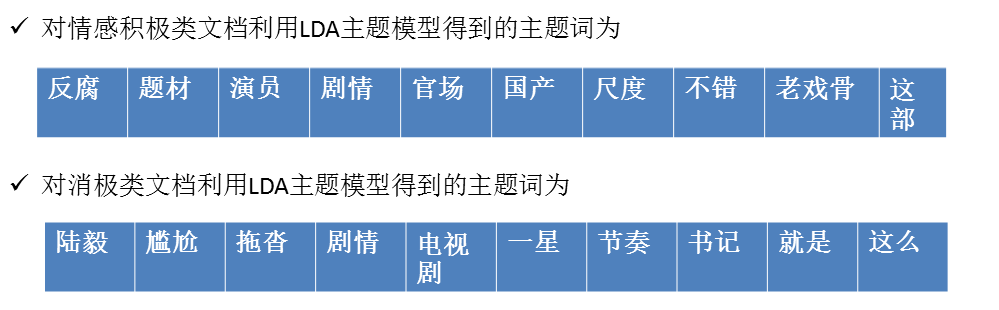
**4.使用LDA模型来获取二类文档类型的关键词**

LDA（Latent Dirichlet Allocation）是一种文档主题生成模型，也称为一个三层贝叶斯概率模型，包含词、主题和文档三层结构。所谓生成模型，就是说，我们认为一篇文章的每个词都是通过“以一定概率选择了某个主题，并从这个主题中以一定概率选择某个词语”这样一个过程得到。文档到主题服从多项式分布，主题到词服从多项式分布。

LDA是一种非监督机器学习技术，可以用来识别大规模文档集（document collection）或语料库（corpus）中潜藏的主题信息。它采用了词袋（bag of words）的方法，这种方法将每一篇文档视为一个词频向量，从而将文本信息转化为了易于建模的数字信息。但是词袋方法没有考虑词与词之间的顺序，这简化了问题的复杂性，同时也为模型的改进提供了契机。每一篇文档代表了一些主题所构成的一个概率分布，而每一个主题又代表了很多单词所构成的一个概率分布。



以上代码采用gensim模块去基于语料库训去练LDA主题模型，最后我们可以得到消极类文档的主题词、积极类文档的主题词。



**5.Conclusion**

到此为止，我们可以根据两类评论来获取观众的主流情感指向。

1. 观众在宏观层面上去赞扬了这部电视剧的题材、剧情
2. 观众对这部剧的反腐性质的题材非常的有兴趣，并且对这部国产剧产生了极大的兴趣
3. 这部剧有一些不错的演员、以及一些老戏骨非常的抓人眼球。
4. 同样剧情的拖沓和节奏也是观众吐槽较多的点
5. 陆毅这个高频词反常的出现在了消极类文档的主题词当中，可以看出来，观众对陆毅在这部电视剧里的表现不敢恭维啊
6. ………………..