



## 文本复制检测报告单(全文标明引文)

№:ADBD2018R\_2017110615244720181017200038719806711906

检测时间: 2018-10-17 20:00:38

检测文献: 80636129435913428\_田肇阳\_近似新闻合并及正负面评价

作者: 田肇阳

检测范围: 中国学术期刊网络出版总库

中国博士学位论文全文数据库/中国优秀硕士学位论文全文数据库

中国重要会议论文全文数据库

中国重要报纸全文数据库

中国专利全文数据库

互联网资源(包含贴吧等论坛资源)

英文数据库(涵盖期刊、博硕、会议的英文数据以及德国Springer、英国Taylor&Francis 期刊数据库等)

港澳台学术文献库

优先出版文献库

互联网文档资源

图书资源

CNKI大成编客-原创作品库

个人比对库

时间范围: 1900-01-01至2018-10-17

### 检测结果

总文字复制比: 9.7%

跨语言检测结果: 0%

去除引用文献复制比: 9.7%

去除本人已发表文献复制比: 9.7%

单篇最大文字复制比: 3%

重复字数: [ 1066 ]

总字数: [ 10984 ]

单篇最大重复字数: [ 333 ]

总段落数: [ 1 ]

前部重合字数: [ 38 ]

疑似段落最大重合字数: [ 1066 ]

疑似段落数: [ 1 ]

后部重合字数: [ 1028 ]

疑似段落最小重合字数: [ 1066 ]

指标: ☐ 疑似剽窃观点 ☒ 疑似剽窃文字表述 ☐ 疑似自我剽窃

☐ 一稿多投 ☐ 过度引用 ☐ 疑似整体剽窃 ☐ 重复发表

表格: 0

脚注与尾注: 0

(注释: ■ 无问题部分 ■ 文字复制比部分 ■ 引用部分)

#### 1. 80636129435913428\_田肇阳\_近似新闻合并及正负面评价

总字数: 10984

相似文献列表 文字复制比: 9.7% (1066) 疑似剽窃观点 (0)

1	基于语义分析的网络信息采集算法研究与应用 赵佳鹤(导师: 王秀坤) - 《大连理工大学硕士论文》 - 2006-12-05	3.0% (333) 是否引证: 否
2	互联网短文本信息分类关键技术研究 - 豆丁网 - 《互联网文档资源 ( <a href="http://www.docin.com/p-108213576.html">http://www.docin.com/p-108213576.html</a> )》 - 2016	2.8% (313) 是否引证: 否

3	<u>搜索引擎中中文WEB文本自动分类研究</u> 刘宏伟(导师: 孟小华) - 《暨南大学硕士论文》 - 2007-04-01	2.8% (303) 是否引证: 否
4	中文分词技术 - 深之JohnChen的专栏 - 博客频道 - CSDN.NET - 《网络 ( <a href="http://blog.csdn.net/byxdaz/article/details/5815677">http://blog.csdn.net/byxdaz/article/details/5815677</a> ) 》 - 2013	2.6% (286) 是否引证: 否
5	中文分词原理 - xiaomin1991222的专栏 - 博客频道 - CSDN.NET - 《网络 ( <a href="http://blog.csdn.net/xiaomin1991222/article/details/50981853">http://blog.csdn.net/xiaomin1991222/article/details/50981853</a> ) 》 - 2017	2.6% (286) 是否引证: 否
6	中文分词技术(中文分词原理) - u010384318的专栏 - 博客频道 - CSDN.NET - 《网络 ( <a href="http://blog.csdn.net/wbgxx333/article/details/11178693">http://blog.csdn.net/wbgxx333/article/details/11178693</a> ) 》 - 2013	2.6% (286) 是否引证: 否
7	文本相似性检测----中文分词技术 - Johline的博客 - CSDN博客 - 《网络 ( <a href="http://blog.csdn.net/johline/article/details/59109811">http://blog.csdn.net/johline/article/details/59109811</a> ) 》 - 2017	2.6% (281) 是否引证: 否
8	互联网舆情信息获取与分析研究 - 豆丁网 - 《互联网文档资源 ( <a href="http://www.docin.com/p-669365673.html">http://www.docin.com/p-669365673.html</a> ) 》 - 2017	2.5% (280) 是否引证: 否
9	《跨语言文本相似性检测》第一周一前期调研 - Johline的博客 - CSDN博客 - 《网络 ( <a href="http://blog.csdn.net/johline/article/details/59111894">http://blog.csdn.net/johline/article/details/59111894</a> ) 》 - 2017	2.4% (269) 是否引证: 否
10	互联网媒体信息热点主动发现关键技术研究 - 豆丁网 - 《互联网文档资源 ( <a href="http://www.docin.com/p-669364954.html">http://www.docin.com/p-669364954.html</a> ) 》 - 2017	2.4% (263) 是否引证: 否
11	互联网媒体信息热点主动发现关键技术研究 - 豆丁网 - 《互联网文档资源 ( <a href="http://www.docin.com/p-122549484.html">http://www.docin.com/p-122549484.html</a> ) 》 - 2013	2.4% (263) 是否引证: 否
12	<u>中文分词技术及其应用初探</u> 余战秋 - 《电脑知识与技术》 - 2004-11-27	2.3% (253) 是否引证: 否
13	<u>民航发动机健康管理数据库设计与故障诊断</u> 张煜(导师: 李书明;黄燕晓) - 《中国民航大学硕士论文》 - 2016-06-30	2.3% (251) 是否引证: 否
14	关于现代中文分词技术的综述 - 《互联网文档资源 ( <a href="http://wenku.baidu.com/view/bf11ce4be518964bcf847cff.html">http://wenku.baidu.com/view/bf11ce4be518964bcf847cff.html</a> ) 》 - 2017	2.3% (251) 是否引证: 否
15	信息检索论文 - 《互联网文档资源 ( <a href="http://wenku.baidu.com/view/708f660390c69ec3d5bb753c.html">http://wenku.baidu.com/view/708f660390c69ec3d5bb753c.html</a> ) 》 - 2017	2.2% (242) 是否引证: 否
16	<u>英文自动问答系统中数值型问句的理解研究</u> 赵龙(导师: 刘亚清;黄威) - 《大连海事大学硕士论文》 - 2016-05-01	2.1% (234) 是否引证: 否
17	<u>社交网络舆情传播监督管理系统的设计与实现</u> 郑罡(导师: 阚忠良;姚德明) - 《黑龙江大学硕士论文》 - 2015-10-20	1.9% (212) 是否引证: 否
18	<u>基于用户兴趣度的网络信息过滤模型研究</u> 王翠平(导师: 刘培玉) - 《山东师范大学硕士论文》 - 2007-04-27	1.9% (212) 是否引证: 否
19	中文分词算法 - llandrj的博客 - CSDN博客 - 《网络 ( <a href="http://blog.csdn.net/llandrj/article/details/49412141">http://blog.csdn.net/llandrj/article/details/49412141</a> ) 》 - 2017	1.9% (212) 是否引证: 否
20	[PDF][精品文]基于用户兴趣度的网络信息过滤模型研究 - 豆丁网 - 《互联网文档资源 ( <a href="http://www.docin.com/p-475602698.html">http://www.docin.com/p-475602698.html</a> ) 》 - 2017	1.9% (212) 是否引证: 否
21	[PDF][精品文]基于用户兴趣度的网络信息过滤模型研究 - 豆丁网 - 《互联网文档资源 ( <a href="http://www.docin.com/p-475602698.html">http://www.docin.com/p-475602698.html</a> ) 》 - 2016	1.9% (212) 是否引证: 否
22	<u>基于条件随机场和空间推理的地理编码方法</u> 周海(导师: 李宏伟) - 《解放军信息工程大学硕士论文》 - 2015-04-20	1.8% (196) 是否引证: 否
23	<u>论四头双导程蜗杆车削挂轮的选配</u> 梁宗斌; - 《现代商贸工业》 - 2017-08-05	1.5% (165) 是否引证: 否
24	<u>基于Hadoop平台分布式SVM分类研究</u> 蔡鑫怡; - 《电脑迷》 - 2018-06-21	1.4% (155) 是否引证: 否
25	<u>基于经济普查大数据的上海“三新”经济发展态势研究</u>	1.4% (154)

	杭敬;苑立波;张志远; - 《统计科学与实践》 - 2016-11-25	是否引证: 否
26	<u>吕苏语口语标注语料的自动分词方法研究</u> 于重重;操镭;尹蔚彬;张泽宇;郑雅; - 《计算机应用研究》 - 2016-07-15 1	1.3% (147) 是否引证: 否
27	<u>媒体情绪能够影响投资者情绪吗——基于新兴市场门槛效应的研究</u> 黄宏斌;刘树海;赵富强; - 《山西财经大学学报》 - 2017-10-30 1	0.9% (103) 是否引证: 否
28	<u>汉语篇章连贯性自动分析方法研究</u> 王小虎(导师: 钟茂生) - 《华东交通大学硕士论文》 - 2015-06-30	0.8% (92) 是否引证: 否
29	<u>面向篇章的省略恢复及其在机械设计中的应用</u> 万棋顺(导师: 赵克) - 《西安电子科技大学硕士论文》 - 2008-01-01	0.7% (75) 是否引证: 否
30	<u>熔融金属红外热像测温精度的研究</u> 高悦(导师: 马翠红;李北丹) - 《华北理工大学硕士论文》 - 2016-12-05	0.6% (66) 是否引证: 否
31	<u>金泽大厦建筑工程项目绿色施工管理研究</u> 胡通文(导师: 蔡为民;刘美秀) - 《天津工业大学硕士论文》 - 2018-01-26	0.5% (60) 是否引证: 否
32	<u>习近平:各级党政机关和领导干部要学会通过网络走群众路线</u> - 《共产党员》 - 2016-05-03	0.3% (38) 是否引证: 否
33	<u>对网络直播乱象说“不”</u> 曹振国; - 《求学》 - 2017-02-15	0.3% (37) 是否引证: 否

## 原文内容

近似新闻合并与正负面评价

清华大学附属中学田肇阳

指导教师: 牛建伟张予瑶

2018年7月

摘要

在全面建设小康社会步伐的大背景下, 人民的生活水平大幅提高。在当下的信息化时代当中, 互联网正在成为一个日趋发展的平台和日渐重要的媒介。互联网不是有百利而无一害的, 在具有丰富信息的同时具有许多不良信息。网络空间是亿万民众共同的精神家园。网络空间天朗气清、生态良好, 符合人民利益。因此, 建立一个负面信息过滤系统迫在眉睫。

简要的来说, 本文完成了以下几个任务: 文章获取, 即利用URLLIB模块从百度爬取搜索结果的网址; 寻找关键词、对文本语义有影响的词, 即将截取网站内容并利用JIEBA模块进行分词; 试图复现他人的工作, 利用WORD CLOUD模块提取文章特征, 从而进行词的判别; 利用TENSORFLOW模块人工分类一部分数据以后将剩余数据进行判别。本文以关键词为江歌案为例开展研究, 但又不局限于这一个关键词, 本文提出了一个文章负面信息过滤的方法。

对于第一个部分, 本文首先设定一定的超时时间, 当超出这个超时时间时程序便会获得下一个网址而不再停留。然后需要设定总共爬取的页数。再往下要切换当前的网页链接为下一页的网页链接; 若下一页为空, 也就是说到达了预设地方的终止, 则退出程序。紧接着本文判断是否爬取完毕所有的网址, 如果爬取完毕就退出程序。接下来本文需要将网页链接上的内容存储到本地。本文读取记事本里的每一个网址, 然后访问并读取其中的内容, 利用本文给出的符号去掉一切的非中文字符和空格, 最后连同每一个网址一起存储在本地。

本文利用TENSORFLOW工具进行判别。本文将正面和负面的文本分别存储在不同的文档里, 然后进行读取。首先本文进行全判别, 将所有的训练集作为测试集进行判别, 输出最终判别得到的准确率, 以求得到较高的判别准确率和模型精准度。然后本文利用三分之一训练集三分之二测试集进行部分判别, 以求得存在一些实际应用的价值和效果。本文设定输入和输出的文本占位符, 然后设置一个函数用以跟踪判别的准确率。然后制作一个嵌入层。再为每个文本创建一个卷积和池化层用来缩小大小。最后进行判别并不断修正数据。

本文将神经网络的优点, 即对于文章本身而非词有较好的把握, 和关键词库的优点, 即对于文字这一载体有较好的应用结合起来。研究计划做的项目是上网搜索类似的新闻, 然后进行正面和负面的评价, 最后将评价得到的结果输出给用户。本文通过TENSORFLOW模块, 先通过训练集进行机器学习, 然后通过测试集进行判别。

本文所设计的程序有多种用途，可以放置在搜索引擎上过滤负面信息，过滤有损于国家和社会安全的信息，以维护正常的社会秩序。同时也可以提供给家长来监控孩子浏览的信息，确保他们不会观察到负面的信息，以确保孩子的健康发展。

关键词：

卷积神经网络，二分类，TENSORFLOW，URLLIB，JIEBA，WORDCLOUD

目录

1. 前言

1.1 文献综述

1.2 研究背景和目的

2. 网址爬取

2.1 模型假设

2.2 符号说明

3. 网站内容获得

4. JIEBA分词

4.1 JIEBA模块介绍

4.2 JIEBA模块应用

5. WORDCLOUD词云制作

5.1 WORDCLOUD模块介绍

5.2 WORDCLOUD词云制作

6. TENSORFLOW神经网络判别

6.1 TENSORFLOW模块介绍

6.2 神经网络介绍

6.3 训练模型函数分析

6.4 测试模型函数分析

7. 结论

8. 致谢

9. 参考文献

10. 附录

1. 前言

1.1 文献综述

在《文本情感分析》[1]中，作者提出了一个基于JIEBA分词模块得到的具有词汇正负面评价的功能的软件，但是这个软件当文本中没有关键词可索引的时候则无能为力，它不能判别这样的文本。同时，它依靠关键词进行判别，而缺少对文本语义的把握。

在姜新猛[2]的研究中，作者利用卷积神经网络达到了利用分类功能中的卷积层和池化层对于图片进行分类。他通过池化层则以缩小数据大小，实现了对整体进行更加深入的分析从而得到抽象程度更高的特征。作者系统介绍了神经网络的研究历史，目前神经网络方法现状，以及如何利用TENSORFLOW模块实现之。作者重点研究了卷积神经网络结构中的卷积层和池化层，并且搭建了实验平台，阐释TENSORFLOW的工作原理及框架结构。

在王银利[3]的研究中，作者提出了负面文字过滤技术的一个操作方法，利用了判别式文本分类算法实现其目的。作者对于目前因特网上出现的负面信息的出现方式进行了规则库的设置和规则的设计。作者提出了一种基于启发式规则和文本分类算法相结合的多级信息过滤模型，介绍了基于启发式规则的信息过滤模型，对于目前因特网上出现的负面信息的出现方式进行了规则库的设置和规则的设计。然后比较了规则导向的信息过滤和文字信息导向的信息过滤算法。综合两种方法的长处和不足，作者提出了一种基于启发式规则和判别式分类算法相结合的多级信息过滤模型。

在石锋[4]、李灏舟[5]、LUO XI[6]和TAOHONG ZHANG[7]等诸位学者的研究中，作者提出了一种基于实体关系抽取的文本分类方法。作者抽取了模型的一些特征，然后通过这些特征进行分类和判别。作者提出了一种新的方

法表示文本特征并利用python对主题爬虫系统进行了测试实验，实现了分类器的准确率在90%以上。作者提出了对英文文本进行分类的普遍方法，但是对于中文仍然缺少适当的分类方法的提出。

## 1.2 研究背景和目的

本研究可以分为两个部分：一是使用JIEBA模块以及WORDCLOUD模块进行的关键词、词频分析研究，试图通过一节文章中的常用词来判别测定文本的思想内涵；二是基于TENSORFLOW模块的学习-测试程序。首先进行全判别，即将全部测试集进行回判，然后进行三分之一训练数据三分之二测试数据的判别。

研究将神经网络的优点，即对于文章本身而非词有较好的把握，和关键词库的优点，即对于文字这一载体有较好的应用结合起来。研究计划做的项目是上网搜索类似的新闻，然后进行正面和负面的评价，最后将评价得到的结果输出给用户。本文通过TENSORFLOW模块，先通过训练集进行机器学习，然后通过测试集进行判别。本文的流程如图1：

图1：工作流程图

### 1. 网址爬取

#### 2.1 设计思路

本文的第一部分是网址爬取。本文首先设计了读写和存储。利用PYTHON自带的open函数，可以建立一个文件，这个文件的每一行都用于存储一个独立的网址。

本文然后进行网址的爬取。本文的思路是将程序伪装成一个浏览器进行获取。本文首先设定一个crawler函数，这个函数的-k部分是必须的，作为一个搜索关键词，-t后接一个整数，作为超时时间，如果不加设定则默认设定为60秒，当超出这个超时时间时程序便会获得下一个网址而不再停留。-p后接一个整数，是要爬取的总页数，如果不加设定则默认为5页。

首先本文设定了一个网址函数，可以作为浏览的地址，然后设定一个超时时间函数，然后设置总共爬取的页数。本文设定了一个变量用以存放当前的网址，然后通过解析地址得到下一页按钮所对应的下一页的网址，进而运行这个网址。本程序的一个难点是不知道百度的控制页码的方式，因此本文通过解析下一页按钮的方式来获取下一页的网址。

再往下要切换当前的网页链接为下一页的网页链接；若下一页为空，也就是说到达了预设地方的终止，则退出程序。若不为空，则爬取当前链接所指页面的内容，保存到html中，然后从当前html中解析出搜索结果的链接并保存。本文需要对爬取到的链接进行去重。再接着本文需要判断是否爬取完毕所有的网址，如果爬取完毕就退出程序。

由于百度具有反爬虫系统，给出的搜索链接都是临时链接，因此本文需要找到临时链接所指向的真实链接。本文设计的内容是伪装成一个浏览器，通过访问一定时间以后自动跳转，休眠一定时间。

本文接下来访问这个临时地址，休眠后再次获取地址，获得真实地址。读取到重定向的网址后保存到链接中，然后将链接输出，存储在out.txt文档中。

#### 2.2 运行示例

以江歌案为例，程序运行截图如图2所示，本文获取到的网址如附录1所示。

图2：获取网址

### 2. 数据收集

本文继续对网址内容进行摘录。接下来本文将网页链接上的内容存储到本地。本文期望读取记事本里的每一个网址，然后访问并读取其中的内容，利用本文给出的符号去掉一切的非中文字符和空格，最后连同每一个网址一起存储在本地。按照网址出现的次序进行out加数字命名文件名。本文列举了多种非中文字符，包括符号和数字。运行结果如图2-3所示：

图2：程序运行截图

图3：程序运行结果

## 3. JIEBA分词

### 4.1 JIEBA模块介绍

对于中文来说，词是最小的能够独立活动的有意义的语言成分。汉语是以字位单位，不像西方语言，词与词之间没有空格等的标志指示词的边界。分词问题为中文文本处理的基础性工作，对于本文的中文信息处理起到关键作



用。

对于中文分词来说，其存在诸多的难点。

——分词规范不明确，词的定义仍然存在模糊。

——未登录词问题，即某些新词尚未收录入词库。

——歧义切分问题，多义组合切分歧义等。

中文本身就存在大量的歧义语句，甚至有些语句结合上下文依然无法判别。如以下这个例子：

结婚的和尚未结婚的，可以分成一下两种形式。

结婚 / 的 / 和 / 尚未 / 结婚 / 的

结婚 / 的 / 和尚 / 未 / 结婚 / 的

未登录词存在两种情况，一是已有的词表中没有收录的词，二是已有的训练语料中未曾出现过的词。据学界普遍认为对于大规模真实文本来说，未登录词对于分词的精度的影响远超歧义切分。一些网络新词，自造词通常都属于这些词。

汉语分词方法主要有以下几种：

——基于字典、词库匹配的分词方法（基于规则）

——基于词频度统计的分词方法（基于统计）

——基于知识理解的分词方法。

基于字符串匹配分词，机械分词算法。将待分的字符串与一个充分大的机器词典中的词条进行匹配。分为正向匹配和逆向匹配；最大长度匹配和最小长度匹配；单纯分词和分词与标注过程相结合的一体化方法。所以常用的有：正向最大匹配，逆向最大匹配，最少切分法。实际应用中，将机械分词作为初分手段，利用语言信息提高切分准确率。优先识别具有明显特征的词，以这些词为断点，将原字符串分为较小字符串再机械匹配，以减少匹配错误率，或将分词与词类标注结合。

相邻的字同时出现的次数越多，越有可能构成一个词语，对语料中的字组频度进行统计，基于词的频度统计的分词方法是一种全切分方法。JIEBA是基于统计的分词方法，JIEBA分词采用了动态规划查找最大概率路径，找出基于词频的最大切分组合。[8]

该方法主要基于句法、语法分析，并结合语义分析，通过对上下文内容所提供信息的分析对词进行定界，它通常包括三个部分：分词子系统、句法语义子系统、总控部分。在总控部分的协调下，分词子系统可以获得有关词、句子等的句法和语义信息来对分词歧义进行判断。这类方法试图让机器具有人类的理解能力，需要使用大量的语言知识和信息。由于汉语语言知识的笼统、复杂性，难以将各种语言信息组织成机器可直接读取的形式。因此目前基于知识的分词系统还处在试验阶段。[8]

JIEBA中文分词模块支持三种分词模式：

A. 精确模式，试图将句子最精确地切开，适合文本分析；

B. 全模式，把句子中所有的可以成词的词语都扫描出来，速度非常快，但是不能解决歧义；

C. 搜索引擎模式，在精确模式的基础上，对长词再次切分，提高召回率，适合用于搜索引擎。

全模式是这样的：我/ 来到/ 北京/ 清华/ 清华大学/ 华大/ 大学。

而精确模式则是这样的：我/ 来到/ 北京/ 清华大学。[8]

JIEBA主要功能包括：分词；添加自定义词典，即开发者可以指定自己自定义的词典，以便包含 JIEBA 词库里没有的词；关键词提取；词性标注；并行分词等。

JIEBA分词的算法策略是基于前缀词典实现高效的词图扫描，生成句子中汉字所有可能成词情况所构成的有向无环图。此后采用了动态规划查找最大概率路径，找出基于词频的最大切分组合。

JIEBA还可以添加自定义词典。目前，网络迅速发展，出现了许多词典中没有但实际上十分常用的词语，在新闻报道中也频频出现（如：“厉害了”，“惊”），这些词语的使用意与原意并不相同，此时需要建立新词典。虽然JIEBA有新词识别能力，但是自行添加新词可以保证更高的正确率，用户也可以通过建立新词典来防止歧义。

JIEBA还可以去停用词。去停用词的意思是有一个文件存放要改的文章，一个文件存放停用表，然后和停用表里的词比较，相同者则删除之，最后把结果存放在一个文件中。停用词可以有效地解决得到关键词无效的问题。

## 1.2 JIEBA词频分析

对于获取得到的文章，本文利用JIEBA模块进行词频分析，得到以下图表数据。

表1：文章1词频统计数据

江歌 1 千佐子 1  
刘鑫 1 受害者 1  
陈世峰 1 母亲 1  
妈妈 1 真相 1  
女儿 1 案卷 1  
江秋莲 1 自己 1  
被害案 1 案件 1  
死刑 1 卷案 1  
日本 1 一个 1  
搜狐 1 没有 1

表2：文章2词频统计数据

江歌 30 帮助 4  
11 22 社会 4  
2017 18 中野 4  
百科 15 请愿 4  
陈世峰 13 签名 4  
日本 13 发布会 4  
母亲 11 公民 4  
遇害案 10 软弱无力 4  
刘鑫 10 无知 4  
引用 10 谴责 4  
日期 10 道德 4  
编辑 9 以为 4  
留日 9 凶手 4  
女生 9 历史 3  
词条 8 知识 3  
前男友 8 合作 3  
12 8 法律 3  
案件 8 池袋 3  
东京 7 解读 3  
嫌疑人 7 事件 3  
2016 6 案发 3  
留学生 6 警方 3  
20 6 10 3  
犯罪 6 江秋莲 3  
24 5 上诉 3  
中国 5 属地 3  
管辖 5 质疑 3  
室友 5 央广网 3  
现场 5 告破 3  
14 5 2018 3  
江歌案 4

图4：文章2词频统计数据

图5: 文章2词频统计数据

## 5 WORLDCLOUD词云

## 5.1 WORDCLOUD模块介绍

WORDCLOUD工具及其制成品主要服务于“快速阅读”模块。它不仅可以提取关键词,也可以直接提取句子。最重要的是,WORDCLOUD模块的成品可以是一张图片,背景的形状和字体大小颜色设计都可以根据用户的喜好设定。

这样一来,阅读新闻从文字变成了图片,可视感更强,更容易得到用户的喜爱。词云以词语为基本单位,可以更加直观和艺术的展示文本

WORDCLOUD库把词云当作一个WORDCLOUD对象,wordcloud.WordCloud()命令代表一个文本对应的词云,因此可以根据文本中词语出现的频率等参数绘制词云。本文以WORDCLOUD对象为基础,通过最基本的步骤,配置参数,加载文本,输出文件。w.to\_file(".jpg")命令代表输出词云。WORDCLOUD词云制作有以下几个基本步骤:

——配置对象函数

——加载词云文本

——输出词云文件

本文需要设置以下参数:

——分隔:以空格分隔单词

——统计:单词出现次数并过滤

——字体:根据统计配置字号

——布局:颜色环境尺寸

## 5.2 WORDCLOUD词云制作

本文利用WORDCLOUD模块进行词云制作,获得了以下直观的词频大小关系。

设计思路如下:本文首先将UTF-8的源文件编码格式转换为ANSI编码格式,然后本文获取当前文件路径,读取文本alice.txt在包文件的example目录下,然后设置背景图片,设定背景颜色、词云显示的最大词数为2000和字体最大值。然后生成词云,输入全部文本,从背景图片生成颜色值。通过绘制图片,绘制词云和保存词云来实现。

对于每篇文章制作的词云结果如下图所示。从这些图可以明显的看出,两位主角在文章中出现的词频较高,而其他人物出现的词频相对较低。同时对于相对支持刘鑫的第三张图片而言,这篇文章的用词相对较为客观,而在其他文章则出现了大量“女儿”,“签名”等的煽动性词汇。

图6: 文章1词云

图7: 文章2词云

图8: 文章3词云

图9: 文章4词云

## 6 TENSORFLOW神经网络分析

## 6.1 TENSORFLOW模块介绍

本文利用TENSORFLOW工具进行判别。本文将正面和负面的文本分别存储在不同的文档里,然后进行读取。首先本文进行全判别,将所有的训练集作为测试集进行判别,输出最终判别得到的准确率,以求得到较高的判别准确率和模型精准度。然后本文利用三分之一训练集三分之二测试集进行部分判别,以求得存在一些实际应用的价值和效果。本文设定输入和输出的文本占位符,然后设置一个函数用以跟踪判别的准确率。然后制作一个嵌入层。再为每个文本创建一个卷积和池化层用来缩小大小。最后进行判别并不断修正数据。

假设一个模型的因变量y和自变量x满足这样的关系,本文称其为线性模型。 $y=kx+b$

其中k, b为模型参数。当这样的模型仅有一个输入的时候,自变量和因变量代表的点的集合在坐标系上表现为一条直线。类似的如果有n个输入, x和y就可以在n+1维空间中组成一个超平面。本文称每个自变量到每个因变量的映射为一个线性变换。由于线性变换无论如何组合表达式必然为线性,因此本文通过线性变换,在单层神经网络之上,全连接神经网络本文也可以表示其对于自变量的数据保留程度。作为一种较为简单的算法,该算法适用于本文所研究的题目。



损失函数可以度量神经网络模型的训练的效果和目标。大多数情况下利用损失函数可以解决神经网络中的分类问题是。一种普遍的利用神经网络获得多分类问题解的方法是确定 $n$ 个输出节点，从而获取 $n$ 维矩阵，这矩阵中每个输出节点表征一个类目。理想情况如下：如果某个样本判别为第 $k$ 类，那么该样本对于该类别对应的这一节点的输出值，也就是维度值，应为1，同时其他任一节点的输出均为0。本文了解到常用的损失函数之一是交叉熵，其用来描述距离在两个概率分布中。本文利用交叉熵来刻画真实输出向量和理想向量的差距。

在深度学习下，“学习”的意思是本文预期基于训练得到的模型对未知类别的条目作出判别。但是存在这样的问题，即模型在训练集上的结果有时不能表征它在测试集上的结果。一个学生在某一类型的题做的非常熟练，然而在这个学生遇到新的问题的时候可能会无所适从。这就在训练中表现为过拟合，即训练后的模型能够近乎完美的描述细节，甚至各个训练数据中的白噪声的部分而忽略了应当学习训练数据中整体的方向。对于学生来说，表现为通过背下来每一道题目而非题目背后的方法来在考试时获得较高的成绩。

图10：神经网络训练情况示例

图10表达了训练模型的各种不同情况。在第一幅图中，明显看出一次线性拟合不符合点集分布，而没有达到更高维度的曲线，因此不能描述点集的分布。在第二幅图中相对符合拟合的趋势，一方面不太关注训练集中的白噪声，另一方面可以良好的描述数据的普遍情况。第三幅图描述的是过拟合模型了，显然这幅图能够完美的划分各个数据点，然而明显可以看出这样的拟合模型不能够良好的对测试集作分划，因为它过度拟合了训练集中的噪声而忽略了模型的内在联系。为了减少这样的问题，正则化是一种普遍使用的方法。正则化的整体思想就是将描述复杂程度的指标加入到损失函数中，以实现模型精准和复杂的一个平衡。

## 6.2 神经网络介绍

BP神经网络利用本体的高阶特征对自己进行数值化，用这种方式将高阶数据实现降维的一种神经网络。这种神经网络是一种无监督学习的特点检测神经网络，即它能够利用较少的的基本特点通过连接层获取更高阶的不易发现的特征。

卷积神经网络是一种基于视觉神经的工作方式和机理发明的神经网络。它们的组织方式都是基于逐层的节点。在这之中每一个节点均代表一个神经元。在BP神经网络中，各相邻层间的节点均有边连接。在卷积神经网络中，相邻两层之间不是全部相连的，而仅有部分节点连接。用这种BP神经网络处理文字的主要挑战是全连接层的参数过多，同时抑郁引发过拟合的情况。

卷积神经网络如图11所示，主要由以下四层组成：输入层、卷积层、池化层、全连接层。输入层是整个神经网络的输入。卷积层设置的目的是把卷积神经网络各个小部分通过更细致地分析，以期获得出更高维度的表征。过滤器是卷积层结构中最重要的组成部分。池化层的目的是减小矩阵的规模以降低最终全连接层中的参数个数，因而带来这样的有点，即加速计算过程以及减少过拟合的产生。

图11：卷积神经网络结构图

## 6.3 训练模型函数分析

深度学习神经网络往往有过多的Hyperparameter需要调优，优化算法、学习率、卷积核尺寸等很多参数都需要不断调整，使用命令行参数是非常方便的。本文调用TENSORFLOW自带的app.flags实现。本文通过调用python的argparse包，调用函数parser.parse\_known\_args()解析命令行参数。代码运行后得到的FLAGS是一个结构体。本文首先定义一个tf.app.flags对象，调用自带的DEFINE\_string, DEFINE\_boolean, DEFINE\_integer, DEFINE\_float设置不同类型的命令行参数及其默认值。

然后本文利用data\_helpers中的load\_test\_data\_and\_labels函数进行训练数据的加载。这个加载首先进行读取正面和负面的文档，然后进行分词，然后通过遍历每一个标签，将正面和负面的词汇成为词汇对。

然后本文建立一个词汇表，并且将数据进行洗牌，通过生成随机序列。然后本文将数据按训练train和测试dev分块。

训练主函数方面，本文首先进行卷积池化网络导入，确定二分类的目标，然后将上定义的filter\_sizes导入，按逗号分割，获取正则化项的数目。然后本文定义优化器，保存损失函数和准确率的参数、训练数据测试数据。然后本文获取当前时间，打印每一次的判别率，判准率和损失率。

最后本文定义测试函数，保留全部神经元，获取当前的batch数据和Session与global\_step值。每FLAGS.evaluate\_every次，即每100次，执行一次测试。每checkpoint\_every次执行一次保存模型。定义模型保存路径，保存c

checkpoint文件。图12是本文的程序运行截图：

图12：TENSORFLOW训练模型程序截图

可以看到，本文的判准率呈现一个不断提升的趋势。由于判准率能够达到1，存在一定的过拟合的倾向。

#### 6.4 测试模型函数分析

测试方面，本文依然首先定义一个tf.app.flags对象，调用自带的DEFINE\_string, DEFINE\_boolean, DEFINE\_integer, DEFINE\_float设置不同类型的命令行参数及其默认值。然后本文继续进行函数的加载。这个加载首先进行读取正面和负面的文档，然后进行分词，然后通过遍历每一个标签，将正面和负面的词汇成为词汇对。

评价方面，首先，本文将数据匹配到词汇表中。然后加载保存的图像和恢复变量。然后获取每一个图像的占位符，设置我们想要判别的传感器，对于判别的每一个阶段产生批，收集预测。

最后打印准确率，将判别的结果存储在一个csv表格中。在对最终文章的判别过程中，本文实现了67%的判准率，说明本文的程序可以较好的应用于本文所要解决的问题，存在较高的可信度。

图13：最终结果截图

#### 7 结论

本文通过使用JIEBA模块以及WORDCLOUD模块进行的关键词、词频分析研究，试图通过一节文章中的常见用词来判别测定文本的思想内涵；还使用基于TENSORFLOW模块的学习-测试程序。本文首先进行全判别，即将全部测试集进行回判，然后进行三分之一训练数据三分之二测试数据的判别。

本文首先基于URLLIB模块进行百度爬虫，获取了网址。通过伪装成浏览器的方法模拟浏览网页，通过临时地址进而获得每一个链接真实网址。此后本文访问网站获取网页内容，并对于多余的符号进行去除处理。

在词频分析方面，本文利用了直观的词云法和详实的数据法，实现了定性与定量的结合。同时通过读者阅读这样的词云和关键词，读者能够联想到文章所要讲解的事物，可以实现对文章主要内容拥有一个迅速且准确的了解，满足本文设定的对于快速获取信息的研究目标。

在正负面判别方面，本文通过KNN算法和卷积神经网络算法，利用了卷积层，池化层等各层次的应用，实现了负面信息过滤的要求，同时也实现了对于文本内容和价值倾向的二分类，达到了67%的准确率。这初步满足了本文所预期实现的对于文章信息正负面二分类和价值取向二分类的目标。

整体而言，本文简要介绍了分词软件、词云制作、神经网络，介绍了基本的原理概念以及它们在PYTHON平台下的实现方式。本文基于近似新闻合并及正负面评价重点制作了词云和词频分析，为接下来的研究奠定了定性分析的基础。本文此后利用了深度学习核心知识，通过卷积神经网络结构，制作了近似新闻合并及正负面评价的程序并且应用于实际运行，对准确率进行了测试，获得了一定的判准率，表明程序具有一定的可信度及效度。

同时，这样的准确率是仍然存在提升空间的，具有进一步研究的必要。一部分原因是基于过拟合，而过拟合的产生原因是本文的使用方法导致的。本文制作的内容应用了基本的卷积神经网络，如果使用更加复杂的判别方式可以系统的提升本文的判准率，对本文设计的程序加以改进。

#### 8 致谢

将近一年的的课题研究即将告一段落了，从未感觉到时间过得如此之快。蓦然回首，有太多难以忘怀的时刻。

在英才计划研究报告结题论文即将完成之际，本文作者要特别感谢牛建伟教授和李青峰老师的的热情关怀和悉心指导。在本文作者研究和论文撰写的过程中，教授和老师都倾注了大量的心血和汗水，无论是在论文的选题、构思和资料的收集方面，还是在论文的研究方法以及成文定稿方面，本文作者都得到了教授和老师悉心细致的教诲和无私的帮助，特别是牛教授广博的学识、深厚的学术素养、严谨的治学精神和一丝不苟的工作作风，以及李老师解答问题的耐心、对于本文作者的爱心，都使本文作者终生受益，本文作者在此表示真诚地感激和诚挚的谢意。

在论文的写作过程中，本文作者也得到了同学、朋友、老师和学校的宝贵建议，在此一并致以诚挚的谢意。本文作者还要感谢家人，谢谢其默默的付出，是你们基于了我力量，让我乘风破浪，勇往直前。

最后，向各位在本文作者研究中给予过我一切帮助的人表示衷心的感谢。

#### 9 参考文献

- [1]文本情感分析[EB/OL].[2018-5-26]. [https://blog.csdn.net/qq\\_22765745/article/details/70947728](https://blog.csdn.net/qq_22765745/article/details/70947728).
- [2] 姜新猛.基于Tensorflow的卷积神经网络的应用研究[D].武汉:华中师范大学计算机学院, 2017.
- [3] 王银利.基于启发式规则和文本分类的信息过滤技术[D].北京:北京交通大学, 2007.

[4] 石锋.面向中文新闻文本的实体关系抽取研究[D].哈尔滨:哈尔滨工业大学, 2016.

[5] 李灏舟.主题爬虫系统中的关键技术研究[D].北京:北京邮电大学, 2016.

[6] Luo Xin.An improved text classifier based on random forest algorithm - [J].Advances in Engineering Research, 2017(150).

[7] Taohong Zhang, Cunfang Li, Nuan Cao, Rui Ma, Shaohua Zhang, and Nan Ma. Abstracts of the Third International Conference of Pioneering Computer Scientists[C].Changsha, China: Engineers and Educators, 2017.

[8] <http://blog.csdn.net/flysky1991/article/details/73948971>自然语言处理入门 (4) ——中文分词原理及分词工具介绍

**跨语言检测结果:** 0%



原文内容	相似内容来源
------	--------

## 指 标

### 疑似剽窃文字表述

1. 基于字符串匹配分词，机械分词算法。将待分的字符串与一个充分大的机器词典中的词条进行匹配。分为正向匹配和逆向匹配；最大长度匹配和最小长度匹配；单纯分词和分词与标注过程相结合的一体化方法。所以常用的有：正向最大匹配，逆向最大匹配，最少切分法。实际应用中，将机械分词作为初分手段，利用语言信息提高切分准确率。优先识别具有明显特征的词，以这些词为断点，将原字符串分为较小字符串再机械匹配，以减少匹配错误率，
2. 方法主要基于句法、语法分析，并结合语义分析，通过对上下文内容所提供信息的分析对词进行定界，它通常包括三个部分：分词子系统、句法语义子系统、总控部分。在总控部分的协调下，分词子系统可以获得有关词、句子等的句法和语义信息来对分词歧义进行判断。这类方法试图让机器具有人类的理解能力，需要使用大量的语言知识和信息。由于汉语语言知识的笼统、复杂性，难以将各种语言信息组织成机器可直接读取的形式。
3. 支持三种分词模式：A. 精确模式，试图将句子最精确地切开，适合文本分析；B. 全模式，把句子中所有的可以成词的词语都扫描出来，
4. 分词等。 JIEBA分词的算法策略是基于前缀词典实现高效的词图扫描，生成句子中汉字所有可能成词情况所构成的有向无环图。此后采用了动态规划查找最大概率路径，找出基于词频的最大切分组合。
5. 特别感谢牛建伟教授和李青峰老师的的热情关怀和悉心指导。在本文作者研究和论文撰写的过程中，教授和老师都倾注了大量的心血和汗水，无论是在论文的选题、构思和资料的收集方面，还是在论文的研究方法以及成文定稿方面，本文作者都得到了教授和老师悉心细致的教诲和无私的帮助，特别是牛教授广博的学识、深厚的学术素养、严谨的治学精神和一丝不苟的工作作风，

**说明：**

1. 仅可用于检测期刊编辑部来稿，不得用于其他用途。
2. 总文字复制比：被检测文章总重合字数在总字数中所占的比例。
3. 去除引用文献复制比：去除系统识别为引用的文献后，计算出来的重合字数在总字数中所占的比例。
4. 去除本人已发表文献复制比：去除作者本人已发表文献后，计算出来的重合字数在总字数中所占的比例。
5. 指标是由系统根据《学术期刊论文不端行为的界定标准》自动生成的。
6. 红色文字表示文字复制部分；绿色文字表示引用部分。
7. 本报告单仅对您所选择比对资源范围内检测结果负责。
8. Email: [amlc@cnki.net](mailto:amlc@cnki.net)  <http://e.weibo.com/u/3194559873>  [http://t.qq.com/CNKI\\_kycx](http://t.qq.com/CNKI_kycx)  
<http://check.cnki.net/>