PaperFree检测报告简明打印版

相似度：27.85%
编　号：OUC8VGOJHMS03RBV
标　题：毕设论文初稿-王冶
作　者：PaperFree
长　度：21326字符
时　间：2019-05-23 22:17:02
比对库：中国学位论文全文数据库；中国学术期刊数据库；中国重要会议论文全文数据库；英文论文全文数据库；互联网资源；自建比对库

相似资源列表(学术期刊，学位论文，会议论文，英文论文等本地数据库资源)
1. 相似度：1.20%　篇名：《基于深度学习的文本情感分析研究》
来源：《哈尔滨工业大学硕士学位论文》　年份：2016　作者：曹宇慧
2. 相似度：0.66%　篇名：《中文文本分类算法研究》
来源：《南京理工大学硕士论文》　年份：2012　作者：马鹏飞
3. 相似度：0.61%　篇名：《基于深度学习的小儿白内障裂隙图像诊断研究及治疗效果预测》
来源：《西安电子科技大学硕士学位论文》　年份：2017　作者：安莹莹
4. 相似度：0.57%　篇名：《基于语义相关的网络文本情感分类研究》
来源：《广东外语外贸大学硕士学位论文》　年份：2016　作者：王伟
5. 相似度：0.48%　篇名：《基于深度学习卷积神经网络的电影票房预测》
来源：《首都经济贸易大学硕士学位论文》　年份：2017　作者：张雪
6. 相似度：0.42%　篇名：《基于SVM和深度学习的情感分类算法研究》
来源：《重庆邮电大学硕士学位论文》　年份：2016　作者：黄志勇
7. 相似度：0.41%　篇名：《基于句子情感权值合成算法的篇章情感分析》
来源：《太原理工大学硕士学位论文》　年份：2015　作者：邸鹏
8. 相似度：0.37%　篇名：《互联网评论文本情感分析研究》
来源：《山东大学硕士学位论文》　年份：2015　作者：崔连超
9. 相似度：0.37%　篇名：《图像分类中的卷积神经网络方法研究》
来源：《南京邮电大学硕士学位论文》　年份：2016　作者：李明威
10. 相似度：0.36%　篇名：《深度神经网络的快速学习算法》
来源：《嘉应学院学报》　年份：2015　作者：卓维
11. 相似度：0.34%　篇名：《朴素贝叶斯分类器的研究与应用》
来源：《重庆交通大学硕士学位论文》　年份：2010　作者：王国才
12. 相似度：0.34%　篇名：《基于词典和机器学习组合的情感分析》
来源：《西安邮电大学硕士学位论文》　年份：2017　作者：丁蔚
13. 相似度：0.29%　篇名：《基于动态随机卷积神经网络的手写数字识别方法》
来源：《吉林大学硕士学位论文》　年份：2017　作者：刘威
14. 相似度：0.29%　篇名：《基于连续维度型的文本情感强度计算方法研究》
来源：《南昌大学硕士学位论文》　年份：2017　作者：胡佳男
15. 相似度：0.26%　篇名：《基于递归神经网络的微博情感分类研究》
来源：《浙江理工大学硕士学位论文》　年份：2017　作者：孙超红
16. 相似度：0.26%　篇名：《突发事件网络舆情分析与威胁估计方法研究》
来源：《解放军信息工程大学硕士学位论文》　年份：2012　作者：王铁套
17. 相似度：0.25%　篇名：《基于卷积神经网络的短文本分类方法研究》
来源：《西南大学硕士学位论文》　年份：2016　作者：蔡慧苹
18. 相似度：0.25%　篇名：《基于机器学习的专利分类研究》
来源：《上海交通大学硕士论文》　年份：2008　作者：褚晓雷
19. 相似度：0.24%　篇名：《面向不平衡文本数据集分类算法研究》
来源：《东北林业大学硕士学位论文》　年份：2017　作者：姚宇
20. 相似度：0.24%　篇名：《面向话题型微博评论的观点识别及其情感倾向分析研究》
来源：《杭州电子科技大学硕士学位论文》　年份：2015　作者：黄时友

21. 相似度：0.24% 篇名：《流行词语计算机获取模型研究》
来源：《华中师范大学硕士学位论文》 年份：2006 作者：朱薏
22. 相似度：0.24% 篇名：《基于卷积神经网络的植物叶片分类》
来源：《计算机与现代化》 年份：2015 作者：龚丁禧
23. 相似度：0.23% 篇名：《简单语音识别系统的设计实现》
来源：《中国新通信》 年份：2013 作者：任赟
24. 相似度：0.22% 篇名：《文本分类及其相关技术研究》
来源：《北京交通大学博士学位论文》 年份：2007 作者：尚文倩
25. 相似度：0.22% 篇名：《基于多分类器投票集成的半监督情感分类方法研究》
来源：《上海交通大学硕士学位论文》 年份：2015 作者：黄伟
26. 相似度：0.21% 篇名：《基于边界向量预选的支持向量机算法研究》
来源：《哈尔滨工程大学硕士学位论文》 年份：2008 作者：杨显飞
27. 相似度：0.21% 篇名：《基于文本分类的商品评价情感分析》
来源：《计算机应用》 年份：2014 作者：钟将
28. 相似度：0.21% 篇名：《多层级情感分析系统的研究与实现》
来源：《电子科技大学硕士学位论文》 年份：2014 作者：杨彬
29. 相似度：0.20% 篇名：《用支持向量机建模学习过程》
来源：《电子科技》 年份：2012 作者：刘宏义
30. 相似度：0.20% 篇名：《中文微博情感词典的构建研究与应用》
来源：《上海师范大学硕士学位论文》 年份：2017 作者：於伟
31. 相似度：0.20% 篇名：《基于深层卷积神经网络的物体识别研究》
来源：《哈尔滨工业大学硕士学位论文》 年份：2017 作者：黎奉薪
32. 相似度：0.20% 篇名：《电商商品评论情感分析方法及优化研究》
来源：《南昌大学硕士学位论文》 年份：2017 作者：陈颖
33. 相似度：0.19% 篇名：《基于文本分类的商品评价情感分析》
来源：《计算机应用》 年份：2014 作者：钟将
34. 相似度：0.18% 篇名：《支持向量回归增量学习》
来源：《计算机科学》 年份：2017 作者：张一凡
35. 相似度：0.18% 篇名：《做好学习笔记夯实基础》
来源：《中考历史》 年份：2011 作者：张超群
36. 相似度：0.18% 篇名：《基于网络影评文本的关系图谱系统的设计与实现》
来源：《北京邮电大学硕士学位论文》 年份：2017 作者：陈传俊
37. 相似度：0.18% 篇名：《MapReduce框架下的贝叶斯文本分类学习研究》
来源：《山西财经大学硕士学位论文》 年份：2012 作者：卫洁
38. 相似度：0.17% 篇名：《基于AdaBoost-LC的微博垃圾评论识别研究》
来源：《重庆大学硕士学位论文》 年份：2014 作者：黄铃
39. 相似度：0.17% 篇名：《网络文本自动分类器的设计与实现》
来源：《电子科技大学硕士论文》 年份：2011 作者：李琼琼
40. 相似度：0.17% 篇名：《基于SVM的中文文本分类系统实现》
来源：《吉林大学硕士学位论文》 年份：2012 作者：苏红刚
41. 相似度：0.16% 篇名：《统计学习图像去噪方法研究》
来源：《西安电子科技大学硕士论文》 年份：2012 作者：张月圆
42. 相似度：0.16% 篇名：《一种基于深度学习与Labeled-LDA的文本分类方法》
来源：《中山大学硕士学位论文》 年份：2017 作者：庞宇明
43. 相似度：0.15% 篇名：《基于贝叶斯网络的文本分类算法研究》
来源：《中国地质大学硕士学位论文》 年份：2016 作者：王沙沙
44. 相似度：0.14% 篇名：《数据分类算法性能的大规模实验对比分析》
来源：《河南大学硕士学位论文》 年份：2016 作者：刘畅畅
45. 相似度：0.14% 篇名：《基于机器学习的专利分类研究》
来源：《上海交通大学硕士论文》 年份：2008 作者：褚晓雷
46. 相似度：0.14% 篇名：《卷积神经网络的并行化研究》
来源：《郑州大学硕士论文》 年份：2013 作者：凡保磊
47. 相似度：0.13% 篇名：《融合深度学习特征与浅层机器学习特征的中文分词关键技术研究》
来源：《华中师范大学硕士学位论文》 年份：2017 作者：周寅
48. 相似度：0.13% 篇名：《关于酷刑的一些思考》

来源：《现代妇女：理论前沿》 年份：2014 作者：申建梅
49. 相似度：0.13% 篇名：《基于图的半监督学习及其应用研究》
来源：《南京航空航天大学硕士学位论文》 年份：2011 作者：张长帅
50. 相似度：0.13% 篇名：《数学专业英语辅助写作系统》
来源：《吉林大学硕士学位论文》 年份：2017 作者：张爽
51. 相似度：0.13% 篇名：《朴素贝叶斯分类器的集成学习方法研究》
来源：《河北大学硕士学位论文》 年份：2009 作者：郝丽锋
52. 相似度：0.13% 篇名：《中文短文本情感分类方法的研究与实现》
来源：《河北科技大学硕士学位论文》 年份：2016 作者：杨鹏飞
53. 相似度：0.12% 篇名：《基于神经网络和遗传算法的卫星结构参数优化》
来源：《装备制造技术》 年份：2012 作者：韩冲
54. 相似度：0.12% 篇名：《自然语言处理技术综述》
来源：《商情》 年份：2013 作者：妮鲁帕尔·艾山江
55. 相似度：0.12% 篇名：《自然语言处理技术综述》
来源：《商情》 年份：2013 作者：妮鲁帕尔·艾山江
56. 相似度：0.11% 篇名：《基于深度学习的情感词向量生成模型研究》
来源：《北京邮电大学硕士学位论文》 年份：2016 作者：焦晨晨
57. 相似度：0.11% 篇名：《基于微博平台的中文情感分析技术的研究》
来源：《沈阳工业大学硕士学位论文》 年份：2017 作者：葛达明
58. 相似度：0.11% 篇名：《面向舆情监测的主题爬虫设计与分析》
来源：《天津科技大学硕士学位论文》 年份：2014 作者：范瑾
59. 相似度：0.11% 篇名：《主题爬虫系统中的关键技术研究》
来源：《北京邮电大学硕士学位论文》 年份：2017 作者：李灏舟
60. 相似度：0.11% 篇名：《基于贝叶斯方法的可靠性评估研究》
来源：《华中科技大学硕士学位论文》 年份：2007 作者：周献振
61. 相似度：0.10% 篇名：《支持向量机（SVM）主动学习方法研究与应用》
来源：《计算机应用》 年份：2004 作者：张健沛
62. 相似度：0.10% 篇名：《使用进化神经网络进行文本自动分类》
来源：《计算机与现代化》 年份：2011 作者：耿俊成
63. 相似度：0.10% 篇名：《贝叶斯分类算法的研究与应用》
来源：《重庆大学硕士学位论文》 年份：2011 作者：郑默
64. 相似度：0.09% 篇名：《基于卷积神经网络的图像处理与识别算法研究》
来源：《江南大学硕士学位论文》 年份：2017 作者：满凤环
65. 相似度：0.09% 篇名：《在线影评和在线短评对票房收入影响的比较研究》
来源：《北京邮电大学硕士学位论文》 年份：2017 作者：钟碧园
66. 相似度：0.09% 篇名：《基于量化的近似最近邻搜索技术研究》
来源：《中国科学技术大学博士学位论文》 年份：2017 作者：张婷
67. 相似度：0.09% 篇名：《中文短文本的情感分析》
来源：《北京邮电大学硕士学位论文》 年份：2015 作者：袁丁
68. 相似度：0.09% 篇名：《中文微博情感倾向性分析与情感要素抽取方法》
来源：《北京工业大学硕士学位论文》 年份：2015 作者：夏梦南
69. 相似度：0.08% 篇名：《基于集成情感成员模型的文本情感分析方法》
来源：《计算机工程与应用》 年份：2014 作者：朱俭
70. 相似度：0.08% 篇名：《深度学习技术在中文人物关系抽取中的应用研究》
来源：《华东师范大学硕士学位论文》 年份：2017 作者：黄蓓静
71. 相似度：0.07% 篇名：《朴素贝叶斯分类算法研究》
来源：《商情》 年份：2012 作者：余民杰
72. 相似度：0.07% 篇名：《文本分类中基于综合度量特征选择算法的研究》
来源：《华中科技大学硕士学位论文》 年份：2015 作者：李铂鑫
73. 相似度：0.07% 篇名：《中文文本情感倾向分析研究》
来源：《情报资料工作》 年份：2013 作者：马晓玲
74. 相似度：0.07% 篇名：《面向情感搜索的中文语料分析及其分词》
来源：《北京邮电大学硕士学位论文》 年份：2014 作者：刘浩
75. 相似度：0.07% 篇名：《DRIS系统中的中文自动分词模块设计与实现》
来源：《华中科技大学硕士学位论文》 年份：2007 作者：向晖

76. 相似度：0.07% 篇名：《基于统计的汉语意见文本校对系统设计与实现》
来源：《黑龙江大学硕士学位论文》 年份：2014 作者：李柏玲
77. 相似度：0.07% 篇名：《基于支持向量机的医学图像分割》
来源：《兰州大学硕士学位论文》 年份：2010 作者：李涟凤
78. 相似度：0.07% 篇名：《基于约束和特征的结构类零件实体模型重建关键技术研究》
来源：《南京航空航天大学博士学位论文》 年份：2005 作者：吴敏
79. 相似度：0.07% 篇名：《基于朴素贝叶斯的文本分类》
来源：《电脑开发与应用》 年份：2013 作者：菅小艳
80. 相似度：0.07% 篇名：《面向博客的垃圾评论识别方法研究》
来源：《河北大学硕士学位论文》 年份：2011 作者：邓冰娜
81. 相似度：0.07% 篇名：《基于云模型主成分的深度学习算法研究及应用》
来源：《南昌大学硕士学位论文》 年份：2017 作者：陈明威
82. 相似度：0.07% 篇名：《支持向量机多类分类算法的分析与设计》
来源：《扬州大学硕士论文》 年份：2008 作者：马波
83. 相似度：0.07% 篇名：《一种基于粗糙集的朴素贝叶斯分类算法》
来源：《福建电脑》 年份：2013 作者：郑芸芸
84. 相似度：0.06% 篇名：《未来城市架构的神经和网络》
来源：《中国广告》 年份：2013 作者：路甬祥
85. 相似度：0.06% 篇名：《文本分类技术及其在网络信息服务中的应用》
来源：《中国科技信息》 年份：2004 作者：陈辉
86. 相似度：0.06% 篇名：《基于随机游走的交互式图像分割算法研究》
来源：《东北大学硕士学位论文》 年份：2011 作者：程伟
87. 相似度：0.06% 篇名：《中文文本情感分类的研究》
来源：《北京交通大学硕士论文》 年份：2011 作者：曾一平
88. 相似度：0.06% 篇名：《支持向量机预处理对SMO算法的改进研究》
来源：《辽宁工程技术大学硕士学位论文》 年份：2015 作者：王庆菊
89. 相似度：0.06% 篇名：《基于操作码的Python程序防逆转算法研究与实现》
来源：《中国科学技术大学硕士学位论文》 年份：2017 作者：王小强
90. 相似度：0.06% 篇名：《基于集成学习的半监督情感分类方法研究》
来源：《中文信息学报》 年份：2015 作者：高伟
91. 相似度：0.05% 篇名：《深度神经网络的快速学习算法》
来源：《嘉应学院学报》 年份：2015 作者：卓维
92. 相似度：0.05% 篇名：《基于GA和KNN的SVM决策树分类方法研究》
来源：《计算机与数字工程》 年份：2012 作者：陈东莉
93. 相似度：0.05% 篇名：《局部学习支持向量机》
来源：《控制与决策》 年份：2012 作者：陶剑文
94. 相似度：0.05% 篇名：《基于演化超网络的中文文本分类方法》
来源：《江苏大学学报：自然科学版》 年份：2013 作者：王进
95. 相似度：0.05% 篇名：《基于分类器选择的个人信用评估组合模型研究》
来源：《哈尔滨工业大学硕士学位论文》 年份：2015 作者：刘艳芳
96. 相似度：0.05% 篇名：《弱标注文本的情感分类技术研究》
来源：《南京大学硕士学位论文》 年份：2016 作者：许超
97. 相似度：0.05% 篇名：《手机产品垂直搜索引擎的设计与实现》
来源：《西安电子科技大学硕士学位论文》 年份：2012 作者：刘育莲
98. 相似度：0.04% 篇名：《迁移学习支持向量回归机》
来源：《计算机应用》 年份：2015 作者：史荧中
99. 相似度：0.04% 篇名：《社会网络新媒体的信息获取与情感分类关键技术研究及实现》
来源：《河北科技大学硕士学位论文》 年份：2013 作者：刘邵博

相似资源列表(百度文库，豆丁文库，博客，新闻网站等互联网资源)

1. 相似度：0.78% 标题：《《北京青年》高清完整无删减版-免费VIP在线观看-飘零影院》
来源：http://www.zgmhj.com/vod/tv/QbdsaqaoSzHuMH.html
2. 相似度：0.78% 标题：《前一秒爆笑,后一秒爆燃,这部春节档电影,观众们这么形容(转载)_...》
来源：http://bbs.tianya.cn/post-filmtv-618804-1.shtml
3. 相似度：0.74% 标题：《(数据科学学习手札30)朴素贝叶斯分类器的原理详解&Python与R实现》

来源：https://www.cnblogs.com/feffery/p/8954959.html
4. 相似度：0.69% 标题：《决策树应用(一)》
来源：http://www.mamicode.com/info-detail-2401696.html
5. 相似度：0.67% 标题：《机器学习的9个基础概念和10种基本算法总结 - Joyce的笔记 - 博客园》
来源：https://www.cnblogs.com/iter1991/p/5664732.html
6. 相似度：0.67% 标题：《机器学习第6章SVM weiququ》
来源：https://www.cnblogs.com/weiququ/p/9462443.html
7. 相似度：0.64% 标题：《结巴分词5 关键词抽取 老顽童2007》
来源：https://www.cnblogs.com/zhbzz2007/p/6177832.html
8. 相似度：0.49% 标题：《机器学习算法基础概念学习总结 - 文章 - 伯乐在线》
来源：http://blog.jobbole.com/74716/
9. 相似度：0.47% 标题：《gensim-Python 计算 tfidf ,数据较大,报错memory error——CSDN...》
来源：https://ask.csdn.net/questions/249159?locationNum=1
10. 相似度：0.45% 标题：《Java怎么去除文本文件中的停用词_百度知道》
来源：https://zhidao.baidu.com/question/682123421419855692.html
11. 相似度：0.45% 标题：《附近的人 停用了吗_百度知道》
来源：https://zhidao.baidu.com/question/2016724894121259628.html
12. 相似度：0.45% 标题：《什么是停用词,静止词_百度知道》
来源：https://zhidao.baidu.com/question/1673376168939593547.html
13. 相似度：0.44% 标题：《使用TensorFlow和TensorBoard从零开始构建卷积神经网络 - 深度...》
来源：http://www.dataguru.cn/article-12648-1.html
14. 相似度：0.44% 标题：《如何基于TensorFlow使用LSTM和CNN实现时序分类任务》
来源：http://www.mamicode.com/info-detail-2004602.html
15. 相似度：0.39% 标题：《tensorflow学习之等价代码 邪恶的亡灵》
来源：https://www.cnblogs.com/Jerry-PR/p/8043746.html
16. 相似度：0.37% 标题：《飞驰人生观后感_飞驰人生观后感450字》
来源：http://www.51ui.cn/zhishi/4266337/
17. 相似度：0.37% 标题：《免费领取《飞驰人生》电影票!申科上汽大众金乡悦众店福利来..._搜狐》
来源：http://m.sohu.com/a/294819349_175406
18. 相似度：0.37% 标题：《五部超好看大电影,部部吸引人,不容错过!_娱乐频道_东方新闻》
来源：http://mini.eastday.com/a/190513103314841.html
19. 相似度：0.37% 标题：《文本挖掘预处理之TF-IDF - 刘建平Pinard - 博客园》
来源：https://www.cnblogs.com/pinard/p/6693230.html
20. 相似度：0.37% 标题：《FCN训练心得 - 简书》
来源：https://www.jianshu.com/p/f500ad5443dd
21. 相似度：0.33% 标题：《beautifulsoup或是正则表达式匹配问题-CSDN论坛》
来源：https://bbs.csdn.net/topics/391037797
22. 相似度：0.31% 标题：《NLP(TF-IDF)---关键词提取算法实现》
来源：https://www.1data.info/content-527.html
23. 相似度：0.28% 标题：《基于深度学习的问题分类组合模型研究--《华中师范大学》2018年...》
来源：http://cdmd.cnki.com.cn/Article/CDMD-10511-1018244863.htm
24. 相似度：0.26% 标题：《tensorflow 优化器optimizer》
来源：http://www.360doc.com/content/18/0505/10/54605916_751286822.shtml
25. 相似度：0.25% 标题：《Python机器学习笔记:朴素贝叶斯算法 - 战争热诚 - 博客园》
来源：https://www.cnblogs.com/wj-1314/p/10560870.html
26. 相似度：0.24% 标题：《33款可用来抓数据的开源爬虫软件工具_手机搜狐网》
来源：http://www.sohu.com/a/35151316_221016
27. 相似度：0.24% 标题：《【爬虫软件】爬虫软件哪个好用_爬虫工具哪个好_PC6手机下载站》
来源：http://www.pc6.com/wenjian/pachong/
28. 相似度：0.22% 标题：《从梯度下降法、最大间隔法两种角度理解SVM - Allegro - 博客园》
来源：https://www.cnblogs.com/kukri/p/8430377.html
29. 相似度：0.20% 标题：《IT入职必读之《机器学习》第七章》
来源：http://www.sohu.com/a/158376231_824406
30. 相似度：0.19% 标题：《中文分词与词性标注联合模型综述.doc-全文可读》
来源：https://max.book118.com/html/2018/0927/8037112015001125.shtm

31. 相似度：0.19% 标题：《利用TensorFlow实现卷积神经网络做文本分类 AllenOR灵感的个人...》
来源：https://my.oschina.net/u/3579120/blog/1533584
32. 相似度：0.19% 标题：《聊聊支持向量机的数学原理》
来源：http://www.360doc.com/content/18/0805/10/11935121_775817898.shtml
33. 相似度：0.18% 标题：《机器学习实战之SVM - 笨鸟多学 - 博客园》
来源：https://www.cnblogs.com/zy230530/p/6901277.html
34. 相似度：0.18% 标题：《神奇的贝叶斯定理 - 简书》
来源：https://www.jianshu.com/p/283154606af5
35. 相似度：0.18% 标题：《为什么支持向量的alpha在0和c之间_百度知道》
来源：https://zhidao.baidu.com/question/331447135784946325.html
36. 相似度：0.18% 标题：《一起来读西瓜书:第七章 贝叶斯分类器 - 简书》
来源：https://www.jianshu.com/p/e90ed474e4a5
37. 相似度：0.18% 标题：《python按比例随机分割生成新文本文件》
来源：https://bbs.csdn.net/topics/392017640
38. 相似度：0.17% 标题：《【NLP】CNN文本分类原理及python代码实现 - 程序员大本营》
来源：http://www.pianshen.com/article/1704157006/
39. 相似度：0.16% 标题：《《机器学习》(周志华)西瓜书读书笔记(完结) - Limitlessun - 博客园》
来源：https://www.cnblogs.com/limitlessun/p/8505647.html
40. 相似度：0.16% 标题：《SVM面试题 - 简书》
来源：https://www.jianshu.com/p/fa02098bc220
41. 相似度：0.16% 标题：《基于卷积神经网络和注意力模型的文本情感分析.PDF》
来源：https://max.book118.com/html/2017/0531/110624940.shtm
42. 相似度：0.15% 标题：《分段卷积神经网络在文本情感分析中的应用 - 道客巴巴》
来源：http://www.doc88.com/p-1314934349886.html
43. 相似度：0.14% 标题：《从零到一:IOS平台TensorFlow入门及应用详解(附源码)(一)-云栖社区》
来源：https://m.aliyun.com/yunqi/articles/73722
44. 相似度：0.14% 标题：《TF-IDF基本概念和原理 - 简书》
来源：https://www.jianshu.com/p/9d4df5d39b40
45. 相似度：0.13% 标题：《Only call `sparse_softmax_cross_entropy_with_logits` with named ...》
来源：https://m.oschina.net/question/1422726_2236190
46. 相似度：0.13% 标题：《Python统计汉字频率-CSDN论坛》
来源：https://bbs.csdn.net/topics/390098761
47. 相似度：0.13% 标题：《Python::re 模块 -- 在Python中使用正则表达式 - Now&Fig..._博客园》
来源：https://www.cnblogs.com/now-fighting/p/4495841.html
48. 相似度：0.11% 标题：《朴素贝叶斯 - 简书》
来源：https://www.jianshu.com/p/7f99c2acfa7e
49. 相似度：0.11% 标题：《情感分析-基于深度学习LSTM方法_lidan_新浪博客》
来源：http://blog.sina.com.cn/s/blog_727a704c0102xl21.html
50. 相似度：0.10% 标题：《基于深度学习的情感词向量及文本情感分析的研究_CNKI学问》
来源：http://xuewen.cnki.net/CMFD-1016138384.nh.html
51. 相似度：0.10% 标题：《使用sklearn自带的贝叶斯分类器进行文本分类和参数调优 - 简书》
来源：https://www.jianshu.com/p/0bf2eb488afa
52. 相似度：0.10% 标题：《卷积神经网络为什么能称霸计算机视觉领域?》
来源：http://www.sohu.com/a/229728388_297710
53. 相似度：0.10% 标题：《基于深度学习的文本情感分析研究--《哈尔滨工业大学》2016年硕士...》
来源：http://cdmd.cnki.com.cn/Article/CDMD-10213-1016773836.htm
54. 相似度：0.10% 标题：《TensorFlow训练卷积神经网络中,输入数据必须是什么类型的?-CSDN问答》
来源：https://ask.csdn.net/questions/693102
55. 相似度：0.10% 标题：《SVM支持向量机 - 简书》
来源：https://www.jianshu.com/p/341c5edd85f5
56. 相似度：0.09% 标题：《虚词的词性:副词、介词、连词、助词、拟声词和叹词在线..._喜马拉雅》
来源：https://www.ximalaya.com/youshengshu/10172321/49048195
57. 相似度：0.09% 标题：《【结巴分词资料汇编】结巴中文分词源码分析(2) - 伏草惟存 - 博客园》
来源：https://www.cnblogs.com/baiboy/p/jieba2.html
58. 相似度：0.09% 标题：《分类算法 二(SVM) - KAMINI - 博客园》

来源：https://www.cnblogs.com/kang06/p/9389333.html
59. 相似度：0.08% 标题：《【火炉炼AI】机器学习011-分类模型的评估:准确率,精确率,召回率,...》
来源：https://www.jianshu.com/p/7cb8759b0680
60. 相似度：0.08% 标题：《关于tensorflow里面的tf.contrib.rnn.BasicLSTMCell 中num_units...》
来源：http://www.mamicode.com/info-detail-2507001.html
61. 相似度：0.08% 标题：《使用文本挖掘技术进行小说《圣墟》评论的情感分析——基于python》
来源：http://www.360doc.com/content/17/1112/10/29308714_703081869.shtml
62. 相似度：0.08% 标题：《tf.nn.softmax cross entropy with logits的用法》
来源：https://www.jianshu.com/p/648d791b55b0
63. 相似度：0.08% 标题：《PimaIndiansdiabetes-数据预处理实验（一） - 小婷儿 - 博客园》
来源：https://www.cnblogs.com/xxtalhr/p/10859517.html
64. 相似度：0.07% 标题：《中文分词_百度百科》
来源：https://baike.baidu.com/item/%E4%B8%AD%E6%96%87%E5%88%86%E8%AF%8D
65. 相似度：0.07% 标题：《【我们的四年,2018】学霸宿舍7#108》
来源：http://www.ahpu.edu.cn/fzfzxy/2018/0621/c6267a104285/page.htm
66. 相似度：0.07% 标题：《监督学习最常见的五种算法,你知道几个? | 雷锋网》
来源：https://www.leiphone.com/news/201704/w6SbD8XGrvQ9IQTB.html
67. 相似度：0.07% 标题：《2.3 卷积神经网络-卷积神经网络实战_慕课手记》
来源：http://www.imooc.com/article/263299
68. 相似度：0.07% 标题：《贝叶斯分类器 - SuperZhang828 - 博客园》
来源：https://www.cnblogs.com/super-zhang-828/p/8082500.html
69. 相似度：0.07% 标题：《Python可视化技术在BP神经网络教学中的应用 - 道客巴巴》
来源：http://www.doc88.com/p-8856422351008.html
70. 相似度：0.07% 标题：《朴素贝叶斯的那点事儿》
来源：https://www.jianshu.com/p/bbe85033c1d9
71. 相似度：0.07% 标题：《2018-04-18第三周 svm深入学习+使用线性核函数写出demo+优化1:k折...》
来源：https://www.jianshu.com/p/4c2d20aeb72a
72. 相似度：0.07% 标题：《朴素贝叶斯算法——实现新闻分类(Sklearn实现) - asialee - 博客园》
来源：https://www.cnblogs.com/asialee/p/9417659.html
73. 相似度：0.07% 标题：《如何根据提供的关键字,利用java代码生成一篇文章-CSDN论坛》
来源：https://bbs.csdn.net/topics/391886971?page=1
74. 相似度：0.07% 标题：《scikit-learn(sklearn)支持向量机(SVM)算法类库介绍_键..._新浪博客》
来源：http://blog.sina.com.cn/s/blog_62970c250102xg0g.html
75. 相似度：0.07% 标题：《神经网络(cnn)训练集正确率88%,测试集只有50%,这是为什么-CSDN论坛》
来源：https://bbs.csdn.net/topics/392064652?page=1
76. 相似度：0.07% 标题：《stick-learn朴素贝叶斯的三个常用模型:高斯、多项式、伯努利》
来源：https://www.cnblogs.com/Scorpio989/p/4760281.html
77. 相似度：0.06% 标题：《机器学习入门之达观数据:文本大数据的机器学习自动分类方法》
来源：http://m.zhizuobiao.com/study/study-19052100031/
78. 相似度：0.06% 标题：《结巴分词4--词性标注 - 老顽童2007 - 博客园》
来源：https://www.cnblogs.com/zhbzz2007/p/6165442.html
79. 相似度：0.06% 标题：《神经网络-如何理解最大池化层有几分缩小?——CSDN问答频道》
来源：https://ask.csdn.net/questions/243140
80. 相似度：0.05% 标题：《机器学习有很多关于核函数的说法,核函数的定义和作用是什么?_知乎》
来源：https://www.zhihu.com/question/24627666
81. 相似度：0.05% 标题：《支持向量机(SVM)是什么意思? - 知乎》
来源：https://www.zhihu.com/question/21094489/answer/302111240
82. 相似度：0.05% 标题：《“以梦为马,不负韶华”出自哪里?是什么意思?_百度知道》
来源：https://zhidao.baidu.com/question/181318393910298204.html
83. 相似度：0.05% 标题：《python划分训练集与测试集后如何将他们保存在不同的文本文档中》
来源：https://bbs.csdn.net/topics/392292505
84. 相似度：0.05% 标题：《4. 算法——贝叶斯(69093)》
来源：https://www.cnblogs.com/skyme/p/3564391.html
85. 相似度：0.05% 标题：《直白介绍卷积神经网络(CNN) - 深度学习-炼数成金-Dataguru专业...》

来源：http://www.dataguru.cn/article-13355-1.html
86. 相似度：0.05% 标题：《inpluslab dataplayer》
来源：https://www.cnblogs.com/inpluslab-dataplayer/
87. 相似度：0.05% 标题：《回首大学生活-【热词推荐-人人网】》
来源：http://share.renren.com/keywords/2167862
88. 相似度：0.05% 标题：《基于改进的卷积神经网络的中文情感分类 - 道客巴巴》
来源：http://www.doc88.com/p-8959604276618.html
89. 相似度：0.04% 标题：《祝福老师的话语_句子大全》
来源：http://www.1juzi.com/new/34602.html
90. 相似度：0.04% 标题：《卷积- 简书》
来源：https://www.jianshu.com/p/d0e55f09fa42
91. 相似度：0.04% 标题：《朴素贝叶斯算法详细总结-电子发烧友网》
来源：http://www.elecfans.com/d/703515.html
92. 相似度：0.04% 标题：《朴素贝叶斯算法下的情感分析——C#编程实现 - 杨睿 - 博客园》
来源：https://www.cnblogs.com/yangruiGB2312/p/5743897.html

全文简明报告

大连海事大学

毕 业 论 文

二○一九年六月

网络文本数据的爬取与分析

专业班级： 软件工程2班

姓 名： 王冶

指导教师： 李楠

信息科学技术学院

摘 要

本文针对互联网中的网络评论数据（以豆瓣影评为例），{ 57% : 利用机器学习算法，分析其包含的个人情感， }从而实现对于评论文本积极、消极情感的判断，以及灌水评论、垃圾评论的识别和剔除，达到舆情分析的初步效果。

第一，使用python对豆瓣网站的电影评论数据进行爬去，进行初步的数据处理并存入MonGoDB数据库中；第二，

提取数据库中的影评文本和对应的评论得分，为每一条评论标注标签；第三，{81% : 使用jieba中文分词工具， }对每一条文本数据进行分词处理、并生成TFIDF特征向量矩阵；第四，使用机器学习中的模型（朴素贝叶斯、支持向量机）以及深度学习中的卷积神经网络进行文本的情感分析。

本次实验基于Linux操作系统，{ 74% : 以python作为开发语言， }使用VsCode编辑器编写程序，{ 57% : 借助于Sklearn机器学习工具包， }以多种方法对文本的情感进行的训练分析，得到了对应的效果。

关键词：网络爬虫；文本分类；情感分析；机器学习；深度学习

I

ABSTRACT

Based on the network comment data in the Internet (taking douban film review as an example), this paper USES machine learning algorithm to analyze the personal emotions contained in it, so as to judge the positive and negative emotions of the comment text, as well as to identify and eliminate comments and garbage comments, and achieve the initial effect of public opinion analysis.

First, use python to crawl the movie review data of douban website, conduct preliminary data processing and store it in MonGoDB database. Second,

Extract the movie review text and corresponding comment score in the database, label each comment;{ 58% : Thirdly, using the Chinese word segmentation tool of jieba, } word segmentation was performed on each piece of text data and

the TFIDF eigenvector matrix was generated; Fourthly, the model in machine learning (naive bayes, support vector machine) and the convolutional neural network in deep learning are used for the emotional analysis of text.

This experiment is based on the Linux operating system. Python is used as the development language, and the VsCode editor is used to write the program. With the help of the Sklearn machine learning kit, the training and analysis of the text's emotion are carried out in various ways, and the ideal effect is obtained.

Keywords:

Web crawlers; Text classification; Emotional analysis; Machine learning; Deep learningIII

目录

网络文本数据的爬取与分析

第1章 绪论

1.1 课题研究的背景及意义

当今社会，在信息化的影响下，每个人都拥有着丰富多彩的网络生活，在享受网络生活的同时，网络也随之带来了许多问题。伴随着大流量app、网站的出现与流行，数以万计的网络数据进入人们视野，而垃圾评论、恶意灌水评论的大量存在混淆了人们的视听，错误引导舆论，甚至严重至歪曲事实真相。

基于这样的现实背景，分析网络文本评论数据、过滤垃圾评论、以及进行舆情把控便具有重大的意义。本文在这样的现实背景下，利用机器学习算法，{ 55%：对评论文本进行了特征提取和情感分析， }从而初步实现对于网络文本数据的正负向情感分析、以及垃圾评论的过滤。

{87%：本文以提高文本情感分析性能为目标， }通过研究机器学习算法（朴素贝叶斯、支持向量机）、深度学习算法（卷积神经网络）并将其应用到中文文本分类这一问题，{ 66%：将有助于提高基于文本情感分析的网络舆情把控、用户评价等分析。 }{85%：因此，本文基于机器学习、深度学习的文本情感分析具有较高的科学研究意义和应用价值。 }

1.2 文本情感分析的研究现状

{ 78%：目前，文本情感分析主要有三类分类方法： }

{89%：基于情感词典的文本情感分析方法； }

基于机器学习的文本情感分析方法

{ 75%：那么随着进年来深度学习技术的不断发展和深入研究， }{ 65%：深度学习在自然语言处理这一领域也广泛应用开来。 }

{ 72%：在本节中，将简单介绍基于情感字典、机器学习、深度学习的文本情感分析的相关技术以及发展现状。 }

1.2.1 基于情感词典的文本情感分析方法

顾名思义，{ 57%：基于词典的文本情感分析方法通常根据人工搭建的情感词典， }利用当前句子中存在的情感词、情感短语的情感加强、反转等规则来判断当前句子的情感类型。就研究现状而言，目前已经实现通过基于搜索引擎的方法、以及每个词语和情感词语之间的相关度等方法实现情感分类，并提高分类效果。不过基于情感词典的文本情感分析方法需要投入大量的人力，将非常耗费人力成本。

1.2.2基于机器学习的文本情感分析方法

{ 62% : 机器学习算法主要分为两类，一种是有监督的机器学习算法，另一种是无监督的机器学习算法。 }

有无监督的区别在于：是否使用人工标注的数据作为数据集进行训练分析，近年来，包括使用朴素贝叶斯、SVM支持向量机、决策树等机器学习算法在文本情感分析问题上面都取得了非常不错的效果，朴素贝叶斯基于贝叶斯公式，将“已知特征求解类别”的问题转化成“已知类别求解特征”和“类别概率”的乘积问题；{85% : SVM基本思想是求解能够正确划分训练数据集并且几何间隔最大的分离超平面。 }{ 63% : 本文将使用这两种机器学习算法进行文本情感分析。 }

1.2.3基于深度学习的文本情感分析方法

{ 72% : 基于深度学习的文本情感分析主要使用的方法有： }{ 57% : CNN（卷积神经网络）、RNN（循环神经网络）。 }{ 65% : 卷积神经网络通过卷积核的卷积运算，提取重要特征， }在经过池化层的进一步处理，{ 62% : 最终在全连接层得出正确的分类结果， }{ 66% : 因此卷积神经网络更适合用于分类的任务。 }而循环神经网络是具有时间顺序的关系，同时属于前馈神经网络。{ 56% : 由于循环神经网络的时间顺序关系， }导致其通道之间具有了前后时间上的联系。所以循环神经网络更适合做上下文的语义分析。目前国内外关于文本情感分析的研究还远远没有达到饱和，由于模型的限制，以及中文数据的不确定性，这也导致了关于中文的情感分析还没有达到英文情感分析的高度。{ 71% : 本文使用的是基于深度学习中卷积神经网络的情感分析方法， }{ 58% : 对中文文本进行了具体的分析。 }

1.3 论文组织结构

本文的内容组织结构如下：

第一章：介绍当前网络生活中的大流量应用所带来的垃圾评论对人们生活的影响，{ 56% : 以及对网络数据进行分析的意义。 }{ 59% : 最后介绍网络文本情感分析主要的研究方法和现阶段的研究现状。 }

{ 58% : 第二章：介绍本文所使用的技术、工具以及相关基础知识。 }包括使用python网络爬虫技术对豆瓣网站进行爬取，同时介绍本文爬虫的具体逻辑；除此之外，还将介绍如何过滤脏数据、冗余数据，{ 72% : 以及正则表达式、BeautifulSoup等工具的使用； }还将介绍自然语言处理中的文本预处理内容——分词、去除停用词；最后，本章着重介绍所使用的机器学习、深度学习算法即朴素贝叶斯、支持向量机、卷积神经网络。

{ 65% : 第三章：详细介绍机器学习算法——朴素贝叶斯， }以及它在文本情感分类中的相关处理过程，{ 69% : 同时详细介绍在使用朴素贝叶斯算法进行文本的情感分类时， }文本预处理的步骤，以及如何使用机器学习工具包sklearn实现朴素贝叶斯算法，并讨论影响文本情感分析的具体因素和遗留的具体问题。

第四章：{ 59% : 详细介绍机器学习算法——支持向量机， }包括支持向量机的数学公式推导、拉格朗日对偶问题的解决，以及SMO算法的数学原理。除此之外还将介绍如何使用sklearn工具包实现支持向量机并应用到中文文本情感分析中。

{ 62% : 第五章：详细介绍卷积神经网络的基本构成， }{ 70% : 以及如何使用Tensorflow构建自己的卷积神经网络， }如何预处理文本来满足神经网络的输入要求，展示使用cnn进行文本的情感分析的具体实验结果。

第六章：主要对比朴素贝叶斯、支持向量机、卷积神经网络的实验效果，以及如何调参优化试验模型，使实验效果达到最优。

1.4 本章小结

{ 64% : 本章主要介绍了文本情感分析的研究背景、以及应用意义。 }{ 57% : 同时也阐述了实现文本情感分析的三种主流方法的原理以及相应的应用现状，主要包括基于情感词典、基于机器学习、基于深度学习的文本情感分析方法。 }{ 67% : 最后，本章给出了本文的内容组织结构。 }{ 56% : 本文将主要使用基于机器学习、深度学习这两种方法实现网络文本的情感分析， }同时还会涉及中文分词，生成词向量、文本预处理等相关内容。

第2章 基础知识及相关技术

网络文本数据的爬取与数据分析涉及到python网络爬虫、机器学习、深度学习、自然语言处理等相关内容，{ 55% : 本文将对网络爬虫技术、自然语言处理中的中文分词， }生成词向量、特征提取，{ 56% : 机器学习中的朴素贝叶斯、支持向量机和深度学习中的卷积神经网络进行详细介绍。 }

2.1 python网络爬虫技术

{100% : 网络爬虫是一个自动提取网页的程序，它为搜索引擎从万维网上下载网页，是搜索引擎的重要组成。 }{ 67% : Python语言中提供了大量的库， }能够帮助我们非常方便的从网络上爬取自几需要的数据。

2.1.1 豆瓣爬虫逻辑

本文使用python-requests模块对豆瓣电影网站进行了爬取，共爬到电影评论数据4万条左右，数据内容主要包括电影的文本评论、对应评论的官方打分等相关内容。

本文需要情感丰富的文本评论数据作为数据集进行训练分析，而评论文本又不宜过长，故此选择了豆瓣电影的电影短评数据作为本次实验的训练集数据，网站数据如下图所示：

矩形框内便是需要爬去的电影短评数据，箭头指向的便是这条评论的得分数据，后期根据这个数据对标签进行人工标注。

具体的爬虫逻辑如下：

对于爬虫爬取下来的网络文本数据，要进行了特殊字符以及英文字符的处理，{ 74%：这里可以使用正则表达式、BeautifulSoup等工具， }同时还要根据评论的得分情况，对其进行人工标注，1,2标注为0，表示为消极情感。4,{ 60%：5分标注为1，表示为正向情感。 }

2.2 文本预处理技术

2.2.1去除脏数据

由于爬虫爬取下来的网络数据包含大量的HTML标签、表情、符号、以及包括很多英文字符，这些字符的存在将大大降低情感分析的精准度以及效率。{ 56%：本文使用Python中自带的Re模块即正则表达式模块对英文字符、表情、符号进行剔除， }{ 73%：同时使用BeautifulSoup去除HTML标签。 }

2.2.2中文分词处理

英文较中文来讲，由于空格的存在，天然不需要分词这一处理。而中文则不同。目前中文分词的工具包包括jieba、THULAC以及北大开源工具包pkuseg。目前分词工具实现的原理基本是基于规则、统计、语义、理解这四种方式实现。{ 76%：而本文使用jieba正是基于统计的分词方法， }{ 68%：jieba分词工具包共有三种模式： }{81%：精确模式、全模式、搜索引擎模式。 }而这三种模式具有不同的特点和优势，{ 68%：精确模式能够将句子精确地切分开，特别适合做文本分析。 }{ 56%：全模式对于句子词语扫描速度很快，但却无法解决歧义问题。 }搜索引擎模式，顾名思义更适合用于搜索引擎分词。本文的分词采取的是“精确模式”，这样利于文本的情感分析。

2.2.3停用词处理

停用词是一些不会影响当前句子感情的词或字，在做自然语言处理的时候，更多的采取将这些词过滤掉，从而提高后续工作的效率。目前对于停用词的处理，大都是基于停用词表，通过遍历的方式去除停用词。本文选用了哈工大停用词表，对豆瓣影评分词后的数据进行了停用词过滤。本文以5000条影评数据为例，对比了在未去除停用词和去除停用词之后的特征数，如下图所示：

可以看到再去除停用词后，特征数由26748下降到26517，当数据量达到几十万条乃至几十G的时候，去除停用词便显得尤为重要。

2.2.4平凡词、独特词处理

平凡词指的是在众多文本数据中均出现的词语，例如“电影”，由于出现的频数过高，其所代表的情感意义也就不是很突出，过于平凡。独特词指的是，在众多文本数据中某个词语仅出现很少的次数，也就是仅在少数文本数据中存在，那么它所代表也只是少数的特点，并不具备说服力。为了降低特征矩阵的特征数，进而提高中文情感分析的准确度和效率，本文采用统计词语在文件中出现的次数frequence、词语的文件占比率rate两项数据作为参考指标，设置阈值，将词语的文件占比率rate超过0.8，词语在文件中出现的次数frequence少于10次的词语全部剔除，极大的降低了特征数。如下图所示：

可见，在设置阈值，剔除平凡词和独特词之后，特征数成功下降到原来的十分之一。

2.3 机器学习算法、深度学习算法

2.3.1朴素贝叶斯

{95%：贝叶斯决策论是概率框架下实施决策的基本方法， }{ 66%：对于分类任务来说，当所有相关概率在已知的理想情况下，贝叶斯决策理论将根据这些相关概率和误判损失进行最优分类， }从而得出最优的分类类别标记。

而基于贝叶斯公式，我们知道：

（2.1）

——类别的先验概率

——样本A对于类别B的条件概率

——归一化证据因子

——后验概率

这其中，我们把P(B)成为“类别”的先验概率；{ 64% : 而p(A|B)是样本A对于类别B的条件概率；P(A)称为用于归一化的证据因子， }那么对于给定的样本A，p(A)与类别标记没有任何关系，{ 57% : 那么问题：“在已知特征样本A的情况下， }求解该样本为B的概率？”就将转换成如何根据训练数据估计先验概率P(B)、条件概率P(A|B)的乘积问题。

{ 65% : 类先验概率P(B)表示的是各类样本占总样本的比例， }反映了子样本的数量情况。{ 67% : 根据大数定律可知，当训练集的样本数量充足时，各个样本满足独立同分布时，P(B)即可用各类样本出现的频率来表示。 }

{ 66% : 对类条件概率P(A|B)而言，他涉及的是有关于所有特征样本A的联合概率， }{ 59% : 那么当各个特征相互独立的时候， }{ 56% : 类条件概率便可转换为各个子特征属性的条件概率乘积 }

{ 56% : 那么在贝叶斯分类器的基础上， }{ 62% : 当我们假设各个特征属性相互独立时，那么每个特征属性都将对分类的结果产生影响， }{ 70% : 这时我们将贝叶斯分类器称为朴素贝叶斯分类器。 }{ 56% : “朴素”即代表特征之间独立假设的成立。 }

那么由上述条件可知，朴素贝叶斯的数学公式应用到特征分类领域为：

（2.2）

{ 68% : 由于各个特征属性之间相互独立， }那么条件概率便可以写成各个子特征的条件概率之积，而先验概率仍然保持不变。{ 68% : 那么朴素贝叶斯分类器的训练过程便是，基于训练集来估计类别的先验概率P(B)，并且估计每一个特征属性的条件概率。 }

{ 76% : 朴素贝叶斯分类器共有三种模型， }{ 64% : 分别是多项式模型、伯努利模型、高斯模型。 }三个模型各有特点，也又所区别。其中多项式模型的特征为单词、特征值为该类单词出现的词频占百分比，{85% : 并且在多项式朴素贝叶斯分类器中， }特征向量多为离散型向量，应用于文本分类；{ 57% : 伯努利模型以文本为特征，特征值为布尔型数据， }标为0或者；高斯模型中，{ 70% : 特征向量是连续性变量，并且假定所有特征的取值是符合高斯分布的。 }高斯模型适用于连续性变量预测。

2.3.2 支持向量机（SVM）

支持向量机算法模型在1995年被提出之后，得到了迅速发展，{ 60% : 并在解决小样本、非线性和高维的模式识别问题中，均取得非常不错的效果。 }{ 56% : 支持向量机根据其使用的核函数可分为： }{ 68% : 线性、多项式、高斯、拉普拉斯、Sigmoid类型的SVM。 }本文就情感分析问题上主要使用的是线性支持向量机，故此下文将详细的介绍线性支持向量机以及其数学推导过程。

{ 79% : 支持向量机是一种有监督的学习算法。 }在二维平面上，散落着很多数据点，假设数据点仅有两类，那么我们可以找到一条直线对其进行分割，{82% : 使不同类的数据点位于直线的两侧。 }同理在三维的空间中我们仍然可以找到一个面，将数据点分割开来，继而将维度上升至n维，那么也必定能够找到n-1的对象将n维中的数据分为不同的类别。这个n-1维的对象称为分隔超平面。在分割的过程中，{ 69% : 离分隔超平面最近的点叫作支持向量。 }在实际应用中，{ 65% : 人们通常希望找到最优的分隔超平面， }所谓最优，就是指分隔超平面两侧的支持向量间的距离最大，当满足这个条件时，{ 68% : 我们把它称为最大分类间隔超平面。 }

通过数学建模可知，在二维的情况下，分隔超平面的线性方程为：

（2.3）

——分隔超平面的法向量

——分隔超平面方程的截距量

{ 75% : 那么支持向量到分隔超平面的距离为： }

(2.4)

同时，如下图所示：

对于支持向量来说，其满足的线性方程为：

(2.5)

{ 56% : 因此异类支持向量之间的距离为： }

(2.6)

——分隔超平面法向量的绝对值

那么求解最大分类间隔的超平面即为求解两个异类支持向量之间的最大距离，也就是求解W的最小值。为了方便研究W的最值问题以及方便求导，将求解W的最小值转化成½|W2|的最小值，即MIN(½|W2|)。{ 55% : 同时对于分隔超平面每侧的支持向量， }均满足下面的关系式。

(2.7)

{ 61% : 那么对于支持向量机的基本数学关系式为： }

(2.8)

在约束条件：

(2.9)

在后续的求解过程中，{ 66% : 由于问题本身就是一个凸二次规划问题， }故此可以使用拉格朗日对偶问题的求解思路进行解决。首先设置拉格朗日α因子，并规定α大于等于0。那么得到的拉格朗日函数为：

(2.10)

{ 63% : 根据拉格朗日函数式，对W、α求偏导： }

(2.11)

(2.12)

进而推导拉格朗日函数式为：

(2.13)

同时问题转换成：

(2.14)

其应该满足的条件为：

(2.15)

则问题由原来的拉格朗日函数式问题转换成求解合适α，使关系式取得最大值，那么求解α也就是smo算法问题，smo算法的数学推导如下：

计算误差：

(2.16)

计算上下界：

(2.17)

计算η

(2.18)

更新αj

(2.19)

修正αj

(2.20)

更新αi

(2.21)

更新b1和b2：

(2.22)

根据b1、b2更新

(2.23)

{100%：SMO算法的工作原理是：每次循环中选择两个alpha进行优化处理。 }{100%：一旦找到了一对合适的alpha，那么就增大其中一个同时减小另一个。 }{98%：这里所谓的“合适”就是指两个alpha必须符合以下两个条件，条件之一就是两个alpha必须要在间隔边界之外，而且第二个条件则是这两个alpha还没有进进行过区间化处理或者不在边界上。 }

2.3.3 卷积神经网络（CNN）

2.3.3.1 传统的神经网络的基本结构、优缺点

传统的神经网络是根据生物的神经网络系统的特点，模仿着生物通过神经网络对现实世界的判断，从而实现算法对于现实问题的交互反应。对于生物神经网络系统而言，神经元与神经元之间的联系依靠神经元分泌的化学物质，上一个神经元分泌的化学物质刺激下一个神经元，并最终激活该神经元，实现神经元之间的信息互通。

{ 55%：传统神经网络的基本结构是：输入层，隐含层（通常包含多层），输出层，而每一层都是由若干的神经元构成的。 }输入层作为接收数据的基础，接收数据之后传递至下一层神经元。在传递的过程中，{ 56%：神经元之间通过激活函数传递信息即权重，从而激活下一层的神经元， }并这样一直传递下去，最终在输出层得到结果，回归或者分类等等。但是对于传统的神经网络，由于每层的神经元都需要进行传递信息和运算，将导致运算规模较大、运算速度较慢，从而并不适合本文的中文情感分析流程。

卷积神经网络（CNN）针对传统神经网络的缺点很好的做了弥补和优化，{ 65%：例如如果我们去识别一张“猫”图片， }我们不需要分析每一个像素点之后，才能得出这是一只猫的结论，如果查看眼睛、耳朵等部位也可以得出同样的结论。那么这样无论速度、还是准确率都将大大的提升。{ 61%：卷积神经网络的正是通过卷积运算减少特征， }从而实现通过局部特征进行分析，得出结论。

{ 61%：2.3.3.2 卷积神经网络基本结构： }

{ 58%：卷积神经网络的基本构成为卷积层、池化层、dropout层，全连接层。 }下面将分别进行介绍：

{ 58%：卷积层中的卷积核在数字信号处理领域被称为滤波器，主要的种类有高斯滤波器等。 }而卷积层的作用就是类似于特征选择器，{ 57%：通过卷积核与原特征矩阵进行卷积运算，从而提取原特征矩阵中的关键特征元素。 }这样也就解决了传统神经网络的参数、特征太多问题。同时卷积运算的公式如下：

（2.24）

{ 56%：其中n为原矩阵维度，f为卷积核大小， }p为填充大小，s为步长。在这里p即填充大小，填充的目的在于保证经过卷积运算之后得到的矩阵大小仍然和原输入矩阵大小相同。S作为步长，规定了卷积运算时横向、纵向移动的长度。

池化层可以理解为在卷积层经过卷积运算的基础上，将运算后得到的数据输入到池化层，而池化层通过既定的运算规则，进一步的提取重要的特征，{ 58%：从而降低运算的复杂性，提高准确率。 }{ 65%：池化的具体方式有两种，一种是取最大值MaxPooling，另一种是取平均值Avgpooling。 }前者是根据步长，去相应矩阵中的最大值进行合并，后者是根据矩阵元素的平均值，作为新的代表元素进行合并。

{ 56%：Dropout层是研究人员为了防止在训练模型、测试数据的时候出现“过拟合”现象而采用的技巧， }通俗来讲就是通过设置概率阈值，{ 55%：每次训练的过程中按照一定的概率丢弃相应的神经网络，是训练的网络减少， }通过实验验证，这种做法对于降低过拟合很有帮助。

{ 60%：全连接层作为卷积神经网络的最后一层，也就是输出层， }在这一层上，之前经过卷积、池化后的特征矩阵将在全连接层上做最后的运算，最终得到我们需要的结果。{ 66%：以上便是卷积神经网络的基本构成， }在后文中，{ 59%：将会介绍在进行中文情感分析的过程中， }如何搭建具体的卷积神经网络，{ 73%：以及使用神经网络进行情感分析。 }

2.4 本章小结

在本章中，第一部分首先讲述了python网络爬虫技术，以及本文爬取数据的具体逻辑，之后对于爬到的电影影评数据进行了预处理，去掉了其中的英文、字符、表情、标点，这样整个数据集中便只剩下了由纯中文构成的中文影评文本。那么为了文本向量化，{ 60%：本章又讲述了使用jieba中文分词工具对预处理过后的文本进行分词， }从而得到分词后的评论文本。而对于分词后的评论文本，还讲述了如何使用去除停用词，也就是对于感情色彩并不重要的词或字。从而得到最终的数据集。{ 55%：第二部分本章具体的介绍了机器学习算法、深度学习算法。 }{ 59%：主要包括机器学习算法朴素贝叶斯、支持向量机、深度学习卷积神经网络。 }

第3章 朴素贝叶斯算法在中文文本情感分析中的应用

3.1 贝叶斯定理与朴素贝叶斯

根据统计学相关知识可知，{ 59% : 贝叶斯定理是由英国数学家托马斯-贝叶斯提出， }当时由于数据量过于庞大，而计算量过于繁重，{ 61% : 贝叶斯定理在当时并没有引起人们的注意， }如今，计算机的出现不但加快了数据的计算速度，同时也扩大了数据的运算规模，而贝叶斯定理也深入的应用到各个领域，在机器学习领域，朴素贝叶斯算法也成功的应用到文本分类、垃圾评论过滤、以及预测分析的问题上。

{ 58% : 机器学习中的朴素贝叶斯算法， }{ 57% : 是在贝叶斯定理的基础上，提出相关假设， }并将贝叶斯公式成功进行了应用。在原贝叶斯定理的基础上，如果假设问题的各个特征相互独立且发生的概率互不影响，此时，贝叶斯定理也就精确为朴素贝叶斯。

3.1.1 贝叶斯公式

根据统计数的相关知识，可知贝叶斯公式为：

（3.1）

{ 62% : P(A|B)、P(B|A)均表示条件概率， }而对应的含义分别为：

P(A|B)：在事件B发生的条件下，事件A发生的概率

P(B|A)：在事件A发生的条件下，事件B发生的概率

其中，{ 66% : P(A|B)称作后验概率，P(A)称为先验概率，P(B|A)/P(B)称为可能性函数，又称做影响因子。 }那么由以上含义，条件概率便有了新的阐释方式：

条件概率 = 可能性函数（影响因子） * 先验概率

3.1.2 朴素贝叶斯在文本情感分析中的具体应用

而在具体的文本情感分析、垃圾邮件分类等的问题中，{ 69% : 假设问题的各个特征之间互不影响、相互独立， }那么此时贝叶斯公式便发生适当转化，也就是朴素贝叶斯公式。

在文本情感分析的过程中，将分词后的句子的每个词作为该句子的特征，并且这些特征出现的可能性均相互独立。那么此时问题便转化成：

{ 55% : “在已知特征的情况下，求解对应情感类别的概率” ， }公式表达如下：

P(类别|特征)

那么根据贝叶斯公式可知：已经提到特征是不止一个的，那么对于一句中文文本如何使用朴素贝叶斯公式求解呢？

{ 73% : 并且已经假设各个特征之间相互独立， }且互不影响，{ 56% : 那么此时由朴素贝叶斯公式可知： }

（3.2）

可见，{ 62% : 在特征特别多的文本情感分类问题上， }朴素贝叶斯仍然适用。

3.2 朴素贝叶斯中文文本情感分析处理流程

基于机器学习算法朴素贝叶斯的中文文本情感分析处理流程主要包括：中文文本预处理、停用词处理、自定义词典、中文分词、TFIDF特征提取、中文词性标注、词频统计、训练数据、验证数据。下面将详细介绍。

3.2.1 中文文本预处理

由于数据集在爬取的过程中，难免出现夹杂冗余数据的情况，而冗余数据主要包括：标点、符号、表情、英文等。这些词的存在，不仅加大了算法处理数据的负担，{ 63% : 同时也大大降低了文本情感分析的准确率， }所以在进行数据训练之前必须过滤掉这些脏数据。本文使用的方法是利用python的正则模块re，匹配英文、标点符号、表情，并去除掉这些数据。

3.2.2 停用词处理

{100% : 停用词是指在信息检索中，为节省存储空间和提高搜索效率，在处理自然语言数据（或文本）之前或之后会自动过滤掉某些字或词，这些字或词即被称为Stop Words（停用词）。 }可见，停用词在文本的情感分析上面，并不具备很好的情感倾向，故此完全可以去除掉停用词，这样也就降低了这些词对于情感分析的影响，{ 74% : 从而提高中文本情感分析的准确性。 }以下面的句子为例：

{98% : “结局很出乎意料，韩式幽默和腾式幽默的碰撞，前半段很搞笑后半段很煽情。 }{98% : 打败自己的只有自己，很燃看完竟然泪流满面，韩寒给我们青春的回忆” }

在这句中，停用词包括“的”、“和”可见这些词或字并没有对这句文本的情感正负性产生任何影响，因此还是应该去掉，从而降低算法训练数据的压力。

3.2.3 中文分词

由于中文使用标点符号，分隔句子，{ 59% : 这种特点便加大了词与词之间的联系， }为了分析中文文本的情感，必须将句子分隔成不同的词语，{ 69% : 这样每个词语作为该文本的特征， }而整个句子也被分割成了多个词语，这样一个中文句子便被分割成了若干个词语。例如该句：

{98% : “结局很出乎意料，韩式幽默和腾式幽默的碰撞，前半段很搞笑后半段很煽情。 }{92% : 打败自己的只有自己，很燃看完竟然泪流满面，沈腾和韩寒给我们青春的回忆” }

进过中文分词处理后：

“结局 很 出乎意料 韩式 幽默 和 腾式 幽默 的 碰撞 前半段 很 搞笑 后半段 很 煽情 打败 自己 的 只有 自己 很燃 看 完 竟然 泪流满面 韩寒 沈 腾 给 我们 青春 的 回忆”

只有经过分词处理后，一个中文文本的句子才可以转换成向量矩阵进行运算，{ 59% : 从而对中文的情感分析进行进一步处理。 }

3.2.4 自定义词典

自定义词典的原因主要是：无论多么好的分词工具，也不可能将自己的数据集中的中文文本全部准确无误的分隔成词，比如从上文我们可知，“韩寒”作为具体的人称被分词工具成功的发现，但是“沈腾”却并没有被发现并提取出来，这样也就造成了分词的误差，而太多的分词误差便会导致整个文本情感分析的准确性，因此我们必须自己根据数据集的情况手动添加词语进入自定义词典，并在分词工具分词前加载进去。这样便能够使分词按照我们的意愿进行，也防止了一些词汇被误分隔，一些词汇没有被发现的问题的出现。具体例子如下：

未定义词典：

{ 77% : ” 打败 自己 的 只有 自己 很燃 看 完 竟然 泪流满面 韩寒 沈 腾 给 我们 青春 的 回忆” }

自定义词典后：

{ 78% : “打败 自己 的 只有 自己 很燃 看 完 竟然 泪流满面 韩寒 沈腾 给 我们 青春 的 回忆” }

可见，沈腾作为名词被识别出来了。

3.2.5 TF-IDF特征提取

当中文文本被分词处理后，得到了若干词组成的词向量，{ 68% : 为了更好的分析该文本的情感倾向， }需要进一步提取文本的特征。{ 59% : 如果把所有分词后得到的词汇都当做特征， }那么在大文本数据量的背景下，特征维度过大，导致算法的运算时间过长，{ 60% : 同时也不利于情感分析准确率的提高。 }而如果只考虑词频，按照词频的高低选取固定数目的词作为该句文本的特征，那么便会出现如下情况：

“我今天非常不开心，因为我的书包丢了”这句话经过分词后，得到如下：

“我 今天 非常 不 开心 ， 因为 我 的 书包 丢 了” ，可知“我”这个字出现了两次，是整个句子出现频率最高的词，不过如果只选择我作为该句话的特征，也就是仅考虑词频高低的情况下，{ 56% : 文本的情感分析将会受到大大影响。 }

{100% : TF-IDF是Term Frequency - Inverse Document Frequency的缩写，即“词频-逆文本频率”。 }{ 63% : 它由两部分组成，TF和IDF。TF指的是词频，也就是一个词在当前文本中出现的次数， }{ 73% : 而IDF指的是一个词在所有文本中出现的次数，如果一个词在所有文本中出现的次数很高，那么它的IDF值将会很低， }{ 58% : 而如果在所有文本中出现的次数不多，那么其IDF值将会很高。 }{ 58% : 而TF-IDF的主要思想也就是，如果一个词在当前文本中出现的次数很高，即TF很高，且其在所有文本中出现的次数很低，也就是IDF很高，那么我们认为这个词具备很好的分类能力， }故此将其提取出来。

本文使用机器学习工具包sklearn中的TFIDFVectorizer实现文本的向量化，并且根据每个词的TFIDF特征值，提取相关词作为当前语句的特征，从而实现降低词向量维度，同时提高准确率。

3.2.6 中文词性标注

词性标注是指在分词的过程中，根据词性表将分好的词语标注词性，目前比较权威的汉语词性表是ICTCLAS汉语词性表，它主要将词语的词性归类为：

* 实词：名词、动词、形容词、状态词、区别词、数词、量词、代词

* 虚词：{93% : 副词、介词、连词、助词、拟声词、叹词。 }

{ 58% : 本文在ICTCLAS汉语词性表的基础上， }{ 73% : 利用jieba分词工具，在分词的同时， }也对词语的词性进行了标注，{ 60% : 更好的区分了动词、形容词、副词等等。 }具体词性标注效果如下(仅列举部分)：

表演：v 台词： n 没： v 刻意： v 营造: v 笑点: n 燃点 : n

3.2.7 词频统计

对于词频的统计，在TFIDF特征值处理的时候，已经计算过了，此处不再赘述。

3.2.8 训练数据、验证数据

完成上述步骤之后，{ 55% : 便可以利用机器学习朴素贝叶斯算法，训练数据， }当模型训练完毕后，在测试集上测试，最终得到相关的指标。本文采用两种方式实现朴素贝叶斯训练数据。下文将详细介绍，与此同时本文还选择了准确率、召回率、以及F-score、学习曲线作为模型的评判标准。由混淆矩阵可知：

True Positive(真正，TP)：将正类预测为正类数

True Negative(真负，TN)：将负类预测为负类数

False Positive(假正，FP)：将负类预测为正类数误报 (Type I error)

False Negative(假负，FN)：将正类预测为负类数→漏报 (Type II error)

准确率指得是经过模型训练、分析、预测得到的分类正确的信息数与总信息数的比值，{ 68% : 准确率越高，该分类器的效果就越好。 }准确率计算公式如下：

（3.3）

召回率指的是经过模型训练、分析、预测得到的正确的分类数量与总共的正确数量的比值。召回率计算公式如下：

（3.4）

score是指当准确率、召回率发生冲突时，对二者进行综合的考虑得到的指标。F-score的具体计算公式如下：

（3.5）

其中α为调和参数。

{ 55% : 本文先后采用了自实现朴素贝叶斯算法进行分类训练， }预测分析和使用机器学习工具库sklearn朴素贝叶斯分类模块进行分类，具体的分类效果以及评判标准如下：

{ 64% : 使用python编程语言编写朴素贝叶斯算法， }并在训练集上面训练模型，当模型训练完毕后，进行测试。本文以9893条豆瓣影评进行训练，2000条豆瓣影评进行测试，得到的最终结果如下：

可以看到，{ 58% : 训练完的模型在测试集上面的准确率达到了87%左右。 }不过自实现的算法在大数据量的情况下，算法运行的时间将会大大增加，{ 62% : 并且最终的准确率也大大降低。 }这体现了自实现朴素贝叶斯算法的不完备性。

{ 72% : 使用sklearn中的朴素贝叶斯分类器模块，进行分类。 }{100% : sklearn是一个Python第三方提供的非常强力的机器学习库，它包含了从数据预处理到训练模型的各个方面。 }{100% : 在实战使用scikit-learn中可以极大的节省我们编写代码的时间以及减少我们的代码量，使我们有更多的精力去分析数据分布，调整模型和修改超参。 }在上文的论述中，我们了解到朴素贝叶斯共有三种模型，{ 57% : 分别是多项式模型，高斯模型，伯努利模型， }而这三种模型都有着不同的应用领域。{ 69% : 多项式模型，更多的用于特征是离散的情况； }高斯模型更多的处理连续特征的情况；{89% : 与多项式模型一样，伯努利模型适用于离散特征的情况，所不同的是，伯努利模型中每个特征的取值只能是1和0。 }而对于中文情感分析问题，{ 56% : 本文分别使用了三种模型对其分析效果进行了对比。 }主要对比如下：

准确率、召回率、F-Score对比：

多项式模型：

高斯模型：

伯努利模型：

通过对比可以看到，{ 61% : 无论是准确率、召回率、还是F-Score指标， }都是多项式模型更适合做情感分析。

通过学习曲线，对比拟合程度

多项式模型：

高斯模型：

伯努利模型：

通过上述三图分析可知：

{ 59% : 高斯模型在训练集逐渐增大的情况下， }训练集的出错概率也在逐渐增大，而测试集准确率也并没有随着数据的增大有所改善。伯努利模型和多项式模型，测试集、训练集的学习曲线均有着相同的趋势。二者均随着数据量的增加而不断收敛至一个标准值，且由图可以判断二者均出现了欠拟合问题，不过就准确率、召回率的高低比较而言，{ 56% : 多项式模型具有更好的实验效果。 }

3.3 本章小结

在本章中，{ 60% : 具体介绍了朴素贝叶斯算法如何应用到中文文本的情感分析中来， }同时在3.2小节中，{ 57% : 详细的介绍了朴素贝叶斯情感分析的处理流程。 }{ 60% : 主要包括文本预处理、特征提取、训练数据、测试数据等。 }并在最后分别展示了朴素贝叶斯的三种模型高斯模型、伯努利模型、多项式模式分别用于情感分析的效果对比，主要包括准确率、召回率、F-Score等。除此之外，还分别绘制了学习曲线从而查看不同的模型的拟合程度。同时本章也对比了不同数量的数据集对于实验效果的影响，可以看到在增大数据集之后明显实验效果会改善。最后本章介绍了sklearn工具包，{ 57% : 以及使用其实现朴素贝叶斯算法等。 }

第4章 支持向量机算法在中文文本情感分析中的应用

机器学习算法支持向量机在中文文本情感分析中的主要流程为：文本预处理、中文分词、向量化、特征提取、svm训练数据、测试数据等。而文本预处理、中文分词、向量化、特征提取这些部分的内容均与朴素贝叶斯中文文本分类的步骤相同，在这里便不再重复赘述。本文使用机器学习工具包sklearn实现svm分类模型的训练，并在测试集上进行了测试，也取得不错的效果。

4.1 支持向量机·核函数

对于支持向量机的基本定义、最大间隔、支持向量以及拉格朗日对偶问题的使用等已经在上文详细介绍过，在此不做过多赘述。而对于支持向量机而言，{ 58% : 所使用的核函数不同，得到的分类效果也不尽相同。 }

在之前讨论的支持向量、以及分隔超平面的相关内容，我们首先是在假设数据点线性可分的基础上进行推导的。{ 69% : 那么如果数据点并不是线性可分的呢， }例如异或问题：

对于这种问题我们不可能找到一条超平面或者直线将数据点分割成两个类别，那么为了解决这类问题，我们引入映射函数将原样本空间中数据点映射至更高维的特征空间中，{ 57% : 使数据点在高维的特征空间中是可分的。 }例如下图，异或问题便得到了解决：

而在上文中使用的映射函数被称为核函数，可想而知在使用svm支持向量机算法做文本的情感分析的时候，{87% : 我们希望样本在相应的特征空间中线性可分，那么特征空间的好坏对支持向量机的性能至关重要，}而核函数的选择也是隐性的选择了特征空间，{ 58% : 故此适当的选择核函数对于文本的情感分析至关重要。 }{ 70% : SVM支持向量机的常用核函数有线性核，多项式核，高斯核，拉普拉斯核以及Sigmoid核。 }下面详细介绍这些核函数：

线性核函数：

线性核的表达式如下：

（4.1）

{ 59% : 线性核函数主要处理于线性可分的数据点样本， }{ 56% : 也就是说如果数据点样本是线性可分的， }那么输入线性核函数后形成的特征空间与原样本空间是完全一致的。

多项式核函数：

多项式核的表达式如下：

（4.2）

{ 55% : 多项式核函数能够实现将原样本空间映射成高维的特征空间， }从而实现原样本点不可分转换成在特征空间内可以分割的情况，对于多项式核函数来说，{ 55% : 它的参数更多，也比线性核函数更加复杂， }运算的规模更大，速度更慢。

高斯核函数：

高斯核函数的表达式如下:

(4.3)

{ 66% : 高斯核函数也可以将原样本空间映射到高维的特征空间， }{ 61% : 不过与多项式核函数相比，高

斯核函数的参数更少，}并且实用性更强，无论大样本数据还是小样本数据都可以取得非常不错的效果。

4.2 不同核函数的SVM文本情感分析对比

本文共使用了三种核函数对svm中文情感分析进行了实验，并取得了不同的效果。

4.2.1 线性核函数实验效果：

注：以下实验均使用豆瓣影评为数据集，其中训练集5000条，测试集1200条

在机器学习工具包sklearn中，{ 59% : 设置参数,kernel = 'linear'，即为线性核函数。 }使用线性核函数时的情感分类效果如下：

可以看到，{ 57% : 分类的准确率达到了74.08%， }召回率达到了71%左右。

4.2.2 多项式核函数实验效果：

在同样的数据集下如果和函数采用多项式核函数的话，准确率下降到了0.48，召回率达到了100%，{ 56% : 但这并不能说明分类的效果很好， }因为召回率反应了该分类模型对正向情感分类的效果非常好，但是准确率不足一半又恰恰说明了该分类模型对负向也就是消极的情感并不具备很好的辨识能力。可见多项式核函数并不适合此次中文文本情感分类。

4.2.3 高斯核函数实验效果：

由上图所示，高斯模型在同样的数据集下的分类效果和多项式核函数模型相同，故此可以推断，{ 57% : 在5000训练集，1200测试集的情况下， }线性核函数的分类效果更加优越。当然，在后文的调参优化、增大数据集操作之后，高斯、多项式核函数也会展现出改善的效果。

4.3 本章小结

{ 58% : 本章主要介绍了支持向量机（SVM）基本结构， }在4.1中本章主要介绍了支持向量机的核函数以及核函数的用途，同时讲述了svm的主要核函数，{ 64% : 包括线性核函数、高斯核函数、多项式核函数等。 }在下一节中，本文使用sklearn实现不同核函数的支持向量机的中文文本的感情分析，最终得到不同的效果，可以从上文的图片得知，只有线性核达到了本次实验的基本要求。

第5章 卷积神经网络算法在中文文本情感分析中的应用

在第三,第四章的内容中,{ 58% : 主要讲述的是基于机器学习的基本算法对中文文本的情感进行分析, }{ 72% : 主要的算法包括朴素贝叶斯,支持向量机。 }本章将使用非传统方法来对中文文本的进行分析，也就是使用卷积神经网络对现有数据集——豆瓣电影评论进行分析。下面将详细介绍具体流程。

5.1 文本预处理

{ 60% : 前文中无论是使用朴素贝叶斯、还是使用支持向量机算法， }{ 77% : 都需要对数据集进行文本预处理， }{ 64% : 而使用的深度学习中的卷积神经网络进行分析时， }同样是需要进行文本预处理的，不过此处的文本预处理与前文并不一致，{ 58% : 在本章中，将使用两种方式实现卷积神经网络， }一种是基于字粒度，而另外一种是基于词粒度。当然实现的方法不同，对应的文本预处理也就不一致。

5.1.1 基于字粒度的文本预处理

之所以是基于字粒度，{ 56% : 原因是不对影评数据集进行分词处理， }而是将训练集中所有评论的数据逐字进行统计，提取出不重复的所有字，作为字符表，用于后期生成每个影评数据的向量。具体的做法如下：

读取MongoDB数据库中的影评数据，同时剔除重复数据，{ 63% : 不能将重复的数据作为训练集， }这样将导致数据测试的准确率下降。利用正则表达式、BeautifulSoup等工具剔除文本中掺杂的标点符号、英文数字。

将处理好的文本数据按照相应比例分为训练集、测试集、验证集写入不同的文件，用于后期的训练验证和测试。

遍历训练集数据，统计每个字出现的频率，使用Counter类过滤重复汉字，并按照频率高低进行排序，生成字典的形式，{ 59% : 也就是key为汉字，value为频率。 }写入文件中，形成全部的字符表。为后期的训练时生成向量做准备。

以上便是基于字粒度的文本预处理流程，最主要的目的就通过预处理得到全部的不重复字符表，以作为后期向量生成的标准和参考。

5.1.2 基于词粒度的文本预处理

基于词粒度的文本预处理，便是不在以字符作为输入单位，{ 62% : 而是将影评数据集使用jieba工具进行

分词， }然后统计全部的不重复词作为词汇表。

读取MongoDB数据库中的影评数据，同时剔除重复数据，{ 63% : 不能将重复的数据作为训练集， }这样将导致数据测试的准确率下降。利用正则表达式、BeautifulSoup等工具剔除文本中掺杂的标点符号、英文数字。

将处理好的文本数据按照相应比例分为训练集、测试集、验证集写入不同的文件，用于后期的训练验证和测试。

遍历训练集数据，使用结巴工具分词，{ 56% : 统计分词后的词语出现的频率， }使用Counter类过滤重复词语，并按照频率高低进行排序，生成字典的形式，{ 59% : 也就是key为词语，value为频率。 }写入文件中，形成全部的词汇表。为后期的训练时生成向量做准备。

词粒度的做法更能很好的体现评论语句与词汇的联系性，分词后得到的词汇表的词汇量

要远远大于基于字符粒度的总量。也就是说在使用分词后的训练数据时，神经网络的维度要变得更大。

5.2 中文文本情感分析时卷积神经网络的结构

{ 62% : 在第二章时，本文讲述了基本的卷积神经网络的构成，主要包括输入层、卷积层、池化层、全连接层、输出层等。 }那么在本章中，将会详细的介绍，{ 56% : 进行中文文本情感的分析时，如何搭建合适的网络结构。 }

词向量嵌入层

词向量嵌入层作为是本文自定义的网络的第一层，其作用便是将通过文本预处理得到的词汇表，{ 56% : 将每句影评映射到相应维度的词向量。 }由于影评文本的长度不一，那么得到的词向量的形状便不一样，这样的话不利于之后的数据批处理，故此可以采用设置阈值，同时使用keras中的特殊标记^PAD^，自动的将文本的长度扩大到制定阈值，这样也就保证了每个文本的映射的词向量长度都是固定的，对数据的批处理阶段会更方便。主要代码如下：

```
embedding = tf.get_variable('embedding', [self.config.vocab_size, self.config.embedding_dim]) embedding_inputs = tf.nn.embedding_lookup(embedding, self.input_x)
```

卷积层、池化层

在上文中，我们已经提到过卷积层的作用相当于特征提取，而卷积层所做的运算主要是卷积运算。通过卷积运算，选取更能代表数据样本的特征。在本文中，设定卷积核数目为256个，同时指定卷积核大小为5。{ 57% : 同时为了减少参数，防止过拟合现象的出现， }池化层的设置也必不可少。具体设置代码如下：

```
# CNN layer 卷积层 conv = tf.{ 72% : layers.conv1d(embedding_inputs, }self.config.num_filters, self.config.kernel_size, name='conv') # global max pooling layer 最大池化层 gmp = tf.reduce_max(conv, reduction_indices=[1], name='gmp')
```

全连接层、Dropout层

{ 55% : 上文也提到了全连接层、Dropout层的具体作用， }分别是最后的运算和丢弃部分神经元防止过拟合现象。本章中我们设定全连接层神经元的个数为128个，设定分类器，并在参数上赋值为积极、消极两种类别。{ 68% : 同时设置dropout比例， }最后链接relu激活函数。主要代码如下：

```
# dense表示全连接层 在这里不指定激活函数，也就是使用默认的激活函数即为线性激活，后面接dropout以及relu激活 # 指定全连接层神经元128个 fc = tf.{88% : layers.dense(gmp, self. }config.hidden_dim, name='fc1') # print(self.keep_prob) fc = tf.contrib.layers.dropout(fc, self.keep_prob) fc = tf.nn.relu(fc) # 分类器 self.{ 73% : logits = tf.layers. }{ 69% : dense(fc, self.config. }num_classes, name='fc2') self.y_pred_cls = tf.argmax(tf.nn.softmax(self.logits), 1) # 预测类别
```

定义损失函数、设置准确率函数

```
{100% : with tf.name_scope("optimize"): } # 损失函数，交叉熵 #添加L2正则化 reg=tf.contrib.layers.apply_regularization(tf.contrib.layers.l2_regularizer(1e-4), tf.{ 57% : trainable_variables()) cross_entropy = tf. }{98% : nn.softmax_cross_entropy_with_logits(logits=self. }{96% : logits, labels=self. }input_y) self.loss = tf.{80% : reduce_mean(cross_entropy)+reg self. }{ 56% : loss = tf.reduce_mean(cross_entropy) # 优化器 self. }{ 70% : optim = tf.train. }{ 73% : AdamOptimizer(learning_rate=self. }config.learning_rate).{ 76% : minimize(self.loss) with tf. }{ 67% : name_scope("accuracy"): # 准确率 correct_pred = tf. }{100% : equal(tf.argmax(self. }{ 72% :
```

input_y, 1), self. }y_pred_cls) self.{85% : acc = tf.reduce_mean(tf. }{ 73% : cast(correct_pred, tf.float32)) }

{87% : 损失函数是对模型所造成误差的度量， }而度量值越小说明模型的误差越小，也体现模型对于当前分类问题的适应性。为了度量误差的大小，我们使用交叉熵损失函数来对模型的误差程度进行评估。{ 67% : 定义了准确率函数，目的是在训练阶段、测试阶段跟踪模型的性能。 }

5.3 卷积神经网络训练、验证、测试结果

在上一节中，{ 57% : 具体介绍了怎样搭建卷积神经网络， }{ 56% : 从而用于中文文本的情感辨析中。 }本节在已有网络的基础上，对训练集、验证集数据进行了训练和交叉验证，之后在测试集进行了测试，得到的具体分类指标如下：

数据训练阶段

从图中可以得出结论，{ 60% : 随着不断地迭代，训练集的损失越来越小， }训练集的准确率越来越高，可见模型在训练集上的学习能力逐渐增强；但可以看到，验证集的准确率虽然也在一直提高，但并没有达到训练集的准确率，除此之外，验证集的损失也一直在0.45左右。

测试阶段：

{ 63% : 可以看到，模型在测试集上的准确率为76.44%， }并且损失率也达到了0.61，结合验证集的验证指标，可以得出结论：经过训练集训练的模型并不被具备很好的泛化能力，{ 61% : 并且模型有可能出现了过拟合问题。 }

{ 72% : Tensorboard 曲线： }

Accuracy曲线：

Loss曲线：

虽然模型的训练准确率总体趋势为逐渐增大，损失率总体趋势为逐渐减少，不过准确率的波动很大，说明模型的稳定性和泛华性并不是很好。1

5.4 本章小结

在本章中，{ 62% : 主要介绍了如何根据数据集搭建卷积神经网络的网络模型结构， }以及设置那些超参数会对实验效果产生影响。同时通过使用tensorflow成功搭建词向量嵌入层、卷积层、池化层、全连接层，并对4万余条数据进行了训练、验证、测试，得到了相应的实验结果。同时根据tensorboard训练准确率、损失率图表，可以得出当前神经网络模型还存在待优化的可能，也出现了过拟合的问题有待解决。

致谢

致 谢

行文至此，{ 64% : 预示着我的大学生涯也即将结束。 }{ 71% : 回首四年的大学时光，往事依然历历在目， }这四年里有过大一的无知、大二的贪玩、也有大三大四的努力与沉淀。可无论怎样，这四年都是属于自己不平凡的四年，都是若干年后回味起来仍然津津有味的四年。

若要用一个词进行总结，那必定是“感谢”一词。

感谢这四年陪伴自己的好友。忘不了与你们一同走过的那些时光，因为它是我人生最宝贵的经历。

感谢于我传道受业解惑的老师们，您们的谆谆教诲我将铭记于心，学生后悔未能珍惜课堂时光，深知为时已晚，却也无力弥补，{ 63% : 只能在此祝福老师们身体健康， }桃李天下。

感谢我的父母，是你们给了我充足的空间去成长，给了我足够的权利去选择。作为您们的孩子，四年时光，伴您左右少之又少，无以回报，但父亲您对我的要求，母亲您对我的教诲，漫漫人生，孩子绝不会忘。

感谢19岁到23岁的自己，是你的年少无知、懵懵懂懂让我明白了原来吃亏是福，是你的敢于尝试，不愿低人一等的倔强给了我如今的自信，感谢你！

挥别过去，放眼未来，更加艰巨的任务还需要自己去完成，希望自己仍然保有19岁、20岁的满腔热血，21岁、22岁的毅力与坚持，学好本事，放低身姿，{ 58% : 以梦为马，不负韶华，成为更好的自己。 }