

毕业设计(论文)任务书

（适用于工科类、理科类专业）

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 课题名称 | | | 一个面向特定领域的微博数据采集和分析系统 | | | | | | | | | | | |  | | | |
| 副 标 题 | | |  | | | | | | | | | | | |  | | | |
| 学院（系） | | | 软件学院 | | | | | | | | | | | |  | | | |
| 专 业 | | | 软件工程 | | | | | | | | | | | |  | | | |
| 学生姓名 | | | 陈薇伊 | | | | | 学 号 | | 1252874 | | | | |  | | | |
|  | | |  | | | | |  | |  | | | | |  | | | |
| 毕业设计(论文)起讫时间： | | | | | | | | | | | | | | | | | | |
| 自 | 2016 | 年 | |  | 月 |  | 日至 | | 2016 | | 年 |  | 月 |  | | 日共 | 16 | 周 | |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 指导教师签名 |  | 2016年 |  | 月 |  | 日 |
| 教学院长(系主任)签名 |  | 2016年 |  | 月 |  | 日 |

一、毕业设计（论文）的课题背景

随着网络的迅猛发展，互联网正在逐渐地改变人们的生活交流方式。从简单的电子邮件，到如今各种社交网络平台，新的社交方式正逐渐深入到人们的日常生活中。其中，微博作为一个基于用户关系信息分享、传播以及获取的平台，以其即时分享的时效性和随意性跻身成为最热门的社交网络平台之一。

2009年8月，新浪微博作为中国门户网站中的第一个微博网站进入人们的视野。据2015年微博发布的第三季度财报中显示，截止2015年9月30日，微博月活跃用户数(MAU)已经达到2.12亿人，较上年同期增长48%。以其庞大的用户群、特有的用户关系结构，新浪微博显示出巨大的影响力，其覆盖的行业领域也在不断扩大，在人们的社交生活中也扮演着越来越重要的角色，已经成为信息传播的新势力。因此，如何有效地收集新浪微博中我们所关心的有效数据并加以分析具有重大的意义。但在新浪微博数据采集过程中会出现登陆模式等问题，难以运用普通网络爬虫的抓取模式，所以首选使用新浪微博提供的官网API进行数据采集，在数据量不够的情况下，再选择使用网络爬虫以扩大数据量。基于新浪微博API对Python具有良好的支持，系统使用Python进行开发，以提高开发效率，保证较强的可读性、可移植性、文件操作简易性。

在采集到新浪微博的数据后，将数据进行处理和存储。根据数据的性质，选用高效率、高存储量、低成本的NoSQL数据库。MongoDB作为一个面向文档的数据库，可以仅仅使用一条数据表现复杂的层次关系。同时，在MongoDB中，数据库的访问和数据的修改的速度极快，还提供了索引、聚合等独特的功能。

在储存数据的同时，需要对数据进行一定的文本分析，即自然语言处理。分词、信息检索、信息分类、关键词抽取等，都属于自然语言处理的范畴。其中，NLTK（Natural Language Toolkit）包含了它包含了词性标识符、分类、分块、解析、语义解释等多模块，是一个基于Python的自然语言工具包。在文本分析中，首先要进行的是中文分析。在当前基于Python的中文分词开源库中，Jieba分词组件扩展性好，提供了精确模式、全模式和搜索引擎模式等三种分词模式，同时支持繁体分词和自定义词典。在已分词的基础上，通过主题建模进行主题词提取。通过TF-IDF可以找出文档集中便于搜索的主题词，其主要思想是：如果某个词或短语在一篇文章中出现的频率TF高，并且在其他文章中很少出现，则认为此词或者短语具有很好的类别区分能力，适合用来分类。同时，通过LDA（Latent Dirichlet Allocation），三级层次贝叶斯模型，进行文字隐含主题建模，可以对一抓取的数据提取潜在的主题，并利用相关主题进行进一步的数据抽取可以使抽取的数据覆盖面更广。其中，Gensim是一个基于Python的开源主题模型工具包，它可以有效地自动抽取予以主题，其中包括了LSA（Latent Semantic Analysis）,LDA和RP（Random Projections）等主题模型的算法。

最后，对经过上述处理的数据进行数据分析。由于微博作为社交平台，所采集到的数据主要涉及用户的生活，对各方面的评价以及社会舆论等，所以在数据分析阶段可以进行情绪分析。通过有效的分析，可以确定某一个状态或评论的情绪是正面，负面或者是中性的。

本课题便是在这个背景下提出，旨在设计并实现一个有效的面向特定领域微博数据收集分析系统。

二、毕业设计（论文）的技术参数（研究内容）

1. 熟悉和掌握使用新浪微博API
2. 熟练和掌握python
3. 研究NoSQL,熟悉和掌握MongoDB
4. 研究主题模型
5. 研究并进行情绪分析

三、毕业设计（论文）应完成的具体工作

1. 使用新浪微博API对新浪微博进行数据抽取。
2. 搭建MongoDB数据库，将从新浪微博抽取到的数据存储到MongoDB中。
3. 使用主题模型，从所提取到的数据中提取相关主题，进行二次数据抽取，扩大相关数据的数据量，为数据分析奠定数据基础。
4. 进行数据分析
5. 搭建微博数据收集分析系统，生成报告，撰写论文

四、毕业设计（论文）进度安排

|  |  |  |
| --- | --- | --- |
| 序 号 | 设计（论文）各阶段名称 | 时间安排（教学周） |
| 1 | 了解、学习相关知识 | 第1至2周 |
| 2 | 整体架构设计和准备工作 | 第3至4周 |
| 3 | 开始抽取数据并存入数据库 | 第5周 |
| 4 | 研究主题模型 | 第6周 |
| 5 | 研究分析数据模型 | 第7周 |
| 6 | 编写代码并实现原型系统 | 第8至9周 |
| 7 | 对系统进行评审和测试，撰写相应文档 | 第10至12周 |
| 8 | 工作总结，撰写学位论文 | 第13至15周 |
| 9 | 后续工作及答辩准备 | 第16周 |
| 10 |  |  |

同组学生姓名：无。

五、应收集的资料及主要参考文献

[1]廉捷, 周欣, 曹伟,等. 新浪微博数据挖掘方案[J]. 清华大学学报:自然科学版, 2011(10):1300-1305.

[2] Chodorow, Kristina. MongoDB: the definitive guide[M]. " O'Reilly Media, Inc.", 2013.

[3]Han J, Haihong E, Le G, et al. Survey on NoSQL database[C]// Pervasive Computing and Applications (ICPCA), 2011 6th International Conference on. IEEE, 2011:363-366.

[4] Ramos, Juan. "Using tf-idf to determine word relevance in document queries." Proceedings of the first instructional conference on machine learning. 2003.

[5] Lutz, Mark. Learning python[M]. " O'Reilly Media, Inc.", 2013.

[6] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation[J]. Journal of Machine Learning Research, 2003, 3:993-1022.

[7] Bird, Steven, Ewan Klein, and Edward Loper. Natural language processing with Python[M]. " O'Reilly Media, Inc.", 2009.

[8] Pang B, Lee L. Opinion Mining and Sentiment Analysis[J]. Foundations and Trends in Information Retrieval，2008， 2(1–2):1-135.

[9] Pak, Alexander, and Patrick Paroubek. Twitter as a Corpus for Sentiment Analysis and Opinion Mining[C]// Conference on International Language Resources & Evaluation. 2010: 1320-1326

[10] Tan, Songbo, and J. Zhang. An empirical study of sentiment analysis for chinese documents[J]. *Expert Systems with Applications* 34.4(2008):2622-2629.