**复旦大学本科生毕业论文（设计）开题报告**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 姓名 | 曲俊杰 | | 学号 | 16307110125 |
| 所在院系 | 软件学院 | | 专业 | 软件工程 |
| 指导教师 | 徐迎晓 | | 职称 | 高级工程师 |
| 校外指导教师  及其所属单位 |  | | 职称 |  |
| 论文（设计）题目 | 基于网络爬虫和数据挖掘技术的新冠疫情可视化分析 | | | |
| **开题报告：建议包含以下内容（可另附页）：**  2020年开年，新冠疫情肆虐全球，给人们带来了空前的恐慌，短短三个月内，数以百万计的人生命受到威胁，亿万居民的生活受到了不同程度的影响，中国作为14亿人口的泱泱大国，充分应用了应对2003年非典疫情时的经验，采用地区强制隔离、分级管控等政策高效抗击疫情，全国上下齐心协力，共同战疫。中国抗击疫情的坚定决心和显著成效给世界人民极大地鼓舞和信心。本次毕业设计将基于网络爬虫、数据挖掘技术，对疫情爆发以来的一系列数据进行汇总，结合一些可视化方法对数据进行可视化呈现，并结合基础的机器学习知识做疫情发展趋势的相关预测。  疫情爆发以来，关于疫情的论文很多，截止二月中旬的疫情文献汇总文献可见：  [1]张龙浩,李柏宏,贾鹏,蒲剑,白蓓,李音,朱培嘉,李雷,曾国军,赵欣,董珊珊,刘梦菡,张楠.新型冠状病毒（SARS-CoV-2）全球研究现状分析[J].生物医学工程学杂志,2020,37(02):236-249.  国内外学者也对疫情数据进行了一些可规划尝试，可供我参考借鉴，学习其中可视化的方法策略：  [2]吴新丽,李云飞,高金.睿思战“疫”——媒体大数据创新实践[J].中国传媒科技,2020(03):11-14.  [3]刘婷,刘颖旭,翟畅.数据可视化在疫情报道中的应用——以新冠肺炎事件报道为例[J].中国传媒科技,2020(03):22-27.  [4]刘勘,周晓峥,周洞汝.数据可视化的研究与发展[J].计算机工程,2002(08):1-2+63.  [5]任磊,杜一,马帅,张小龙,戴国忠.大数据可视分析综述[J].软件学报,2014,25(09):1909-1936.  数据收集中，我们需要用到网络爬虫技术，并且需要了解一定的反反爬虫手段：  [6]刘金红,陆余良.主题网络爬虫研究综述[J].计算机应用研究,2007(10):26-29+47.  [7]孙立伟,何国辉,吴礼发.网络爬虫技术的研究[J].电脑知识与技术,2010,6(15):4112-4115.  [8]余豪士,匡芳君.基于Python的反反爬虫技术分析与应用[J].智能计算机与应用,2018,8(04):112-115.  [9]刘佐. 基于微信公众平台的数据挖掘与可视化研究[D].华北电力大学,2017.  最后我们需要用到基本的机器学习技术，尝试预测分析：  [10]张学工.关于统计学习理论与支持向量机[J].自动化学报,2000(01):36-46.  [11]慕春棣,tsinghua.edu.cn,戴剑彬,叶俊.用于数据挖掘的贝叶斯网络[J].软件学报,2000(05):660-666.  主要研究内容如下：  首先，利用腾讯/百度等互联网门户给的疫情通报数据接口，汇总每日疫情数据，完成全国/世界范围内疫情相关数据的收集工作。利用这些数据和相关接口结合echart技术建立自己的疫情统计页面，方便直观地看到疫情变化。  页面建立后，将数据汇总成表格并进入SQL数据库，使用python绘图技术进行数据的多样化可视化呈现，通过最直观的方式将各地疫情呈现出来。结合基础的机器学习如线性回归、贝叶斯多项式回归、SVM等方式训练参数拟合疫情发展曲线。  此外，本次可视化分析希望对疫情中突出一环——媒体新闻予以适当关注，主要关注微信公众号中的文章，汇总新鲜词汇，侦测大众对于媒体新闻的关注点，并利用jieba分词和wordcloud技术生成热词云图，向在抗疫事业中无私奉献的英雄们致敬。  本次毕业设计中的难点在于：1）反爬虫机制对于数据采集造成的阻碍；2）数据的直观化呈现标准；3）词云图的筛选和生成；4）疫情数据预测的手段和准确性评估  创新点在于：1）数据的时效性强，有收集和分析价值；2）可视化技术同时事热点相结合；3）数据分析融入文章热词人文部分，学科交叉文理交融；4）爬虫技术来收集数据，具体原创性；5）新前端技术的应用服务于本次毕业设计。  预期的成果及形式  1）新冠疫情的相关数据集；  2）每日疫情直观呈现的网页；  3）热词分析系统，自动生成云图；  4）疫情数据可视化图集及预测结果； | | | | |
| **研究进度及具体时间安排（不够写可加行）** | | | | |
| 起止日期 | | 主要研究内容 | | |
| 2020.3.15-2020.4.1 | | 文献阅读和技术准备 | | |
| 2020.4.1-2020.4.10 | | 数据集的收集和构建（爬虫） | | |
| 2020.4.10-2020.4.20 | | 数据分析和可视化挖掘 | | |
| 2020.4.20-2020.5.1 | | 可视化疫情网页构建 | | |
| 2020.5.1-2020.5.10 | | 疫情预测和图集整理 | | |
| 2020.5.10-2020.5.20 | | 论文的撰写和修改 | | |
| **指导教师对课题报告的意见：**  **1．对选题依据、基本思路或技术路线的可行性、创新性的评价；2．存在的主要问题和改进建议。**  （初稿的审核意见如下：本科论文题目用“...研究”有点太大，要具体化到开发一个什么系统的开发。方向可以，但开题报告要细化，讲清楚数据可视化领域前人具体做到什么程度，参考文献要多阅读一些，以便站在前人肩膀上提升深度，另外要注意开发工作量体现得要充足，要比课程project工作量大）  指导教师签名： 年 月 日 | | | | |