《大数据导论》作业说明

2022 年秋季学期

主讲教师：厦门大学信息学院　林子雨　副教授　ziyulin@xmu.edu.cn

一、　作业题目

网页数据采集。

二、　作业目的

运用 Python 语言编写网页爬虫获取网页数据。

三、　作业性质

必做。作为评定期末总成绩的依据之一，占期末总成绩的 10%。

四、　作业考核方法

作业成绩评定方法如下：

- 不按时交作业、所提交的作业无法打开或抄袭他人作业：零分
- 作业评分范围：0-100 分

温馨提示：作业必须自己独立完成（所有作业全部要求自己独立完成，没有采用团队合作的形式），不得抄袭他人作业，不得直接拷贝厦门大学数据库实验室网站上提供的案例，否则，期末总成绩不及格。

五、　提交日期与方式

1、必须在 2022 年 12 月 17 日（周六）0 时到 24 时之间提交，不要在其他时间提交。

2、提交的内容为压缩文件 RAR 文件，最后把压缩包文件发送到助教刘浩然同学邮箱：1609126475@qq.com（如果邮件太大，可以使用 QQ 邮箱超大附件功能发送）；

3、文件名命名为“姓名学号.rar”，例如“王小明 23020191152890.rar”；

4、文件夹中应该包含实验报告 WORD 文档、工程文件、采集到的数据文件以及其他有必要提交的文档，使得老师可以根据这些信息在老师电脑上可以重现实验内容。

六、　作业工具和环境要求

（1）　必须在 Windows 系统下完成作业，必须使用 Python 语言，不能使用其他语言。没有学习过 Python 语言的同学，可以找资料自学，1 天就可以学会基本语法。本作业只需要使用 Python 的基本语法知识。

（2）　相关软件的版本要求如下：

- Python3.X
- Windows7、Windows10 及以上
- PyCharm

七、　作业内容和要求

（1）任务 1：从网络上寻找一个网站 A，参考《（自学材料）第 3 章　网络数据采集.pdf》3.3 节、3.4 节、3.5 节、3.6 节的知识，编写 Python 程序，爬取网页数据，保存到一个文本文件中。提交作业时，在“姓名学号”文件夹下面，新建一个名称为“任务 1”的文件夹，这个文件夹下要包含一个实验报告 WORD 文档、代码文件（.py 格式）和采集到的数据文件，WORD 文档里面把具体任务描述清楚，包括采集网页的地址、网页格式描述、要采集什么样的数据、编写程序的基本思路。

（2）任务 2：针对同一个网站 A，使用 Scrapy 框架（参考《（自学材料）第 3 章　网络数据采集.pdf》3.7 节知识）爬取与任务 1 相同的数据，需要使用 PyCharm 编写 Python 程序。提交作业时，在“姓名学号”文件夹下面，新建一个名称为“任务 2”的文件夹，这个文件夹下要包含一个实验报告 WORD 文档、工程文件（需要把 PyCharm 的工程文件打成压缩

包提交，里面包含了代码和采集到的数据文件），WORD 文档里面把具体任务描述清楚，包括采集网页的地址、网页格式描述、要采集什么样的数据、编写程序的基本思路。

八、 优秀作业奖励

为了鼓励同学高质量地完成作业，一方面，优秀的作业可以获得较高的成绩，另一方面，可以获得荣誉和现金奖励。将从所有作业中，**评选出 3 名优秀奖（每人奖励 500 元）**。优秀作业评选的要素包括程序质量和实验报告质量。

九、 附录 1:教师介绍

林子雨(1978－),男,博士,厦门大学计算机科学系 副教授,主要研究领域为数据库,数据仓库,数据挖掘,大数据

主讲课程：大数据处理技术

办公地点：厦门大学海韵园科研 2 号楼

E-mail: ziyulin@xmu.edu.cn

林子雨，男，1978 年出生，博士（毕业于北京大学），现为厦门大学计算机科学系副教授，曾任厦门大学信息科学与技术学院院长助理、晋江市发展和改革局副局长。中国计算机学会数据库专业委员会执行委员，中国计算机学会信息系统专业委员会执行委员，荣获“2017 中国大数据创新百人”称号，入选“2021 年高校计算机专业优秀教师奖励计划”。中国高校首个“数字教师”提出者和建设者，厦门大学数据库实验室负责人，厦门大学云计算与大数据研究中心主要建设者和骨干成员，2013 年度、2017 年度和 2020 年度厦门大学奖教金获得者，荣获 2022 年福建省高等教育教学成果奖特等奖和 2018 年福建省教学成果二等奖，主讲的《大数据技术原理与应用》课程荣获“2018 国家精品在线开放课程”。主要研究方向为数据库、数据仓库、数据挖掘、大数据、云计算和物联网，并以第一作者身份在《软件学报》《计算机学报》和《计算机研究与发展》等国家重点期刊以及国际学术会议上发表多篇学术论文。作为项目负责人主持的科研项目包括 1 项国家自然科学青年基金项目(No.61303004)、1 项福建省自然科学青年基金项目(No.2013J05099)和 1 项中央高校基本科研业务费项目(No.2011121049)；作为课题负责人主持的教学项目包括 1 项福建省教改课题和 1 项教育部产学合作育人项目。同时，作为课题负责人完成了国家发改委城市信息化重大课题、国家物联网重大应用示范工程区域试点泉州市工作方案、2015 泉州市互联网经济调研等课题。中国高校首个“数字教师”提出者和建设者，2009 年至今，“数字教师”大平台累计向网络免费发布超过 500 万字高价值的研究和教学资料，累计网络访问量超过 1800 万次。打造了中国高校大数据教学知名品牌，编著出版了中国高校第一本系统介绍大数据知识的专业教材《大数据技术原理与应用》，并成为京东、当当网等网店畅销书籍；建设了国内高校首个大数据课程公共服务平台，为教师教学和学生学习大数据课程提供全方位、一站式服务，单年访问量超过 400 万次，累计访问量超过 1800 万次，成为全国高校大数据教学知名品牌。