

《大数据计算机基础》2022 秋季学期作业 3



本次作业基于链接：http://www.moe.gov.cn/jyb_sjzl/moe_560/2020/，是教育部教育统计数据。

本次作业要求如下：

- 1、 下载 2000 年到 2020 年各地基本情况的数据，要求通过网络爬虫技术，自动下载爬取网页，并分析网页获得各省从小学到高中的各年级学生人数包括毕业生人数，最终形成数据文件名为 students.txt，第一列名称为省份，第二列为年份，第三列为年级，第四列为人数。年级规范为小 1，小 2 一直到小 5、小 6，初中分别为初 1、初 2，高中与之类似；对于毕业人数，列入年级列，分为：小毕、初毕、高毕；
- 2、 通过搜索引擎获得 2000、2010、2020 年的各省人口普查数据，并假设每个省人口增长率均等（不同省不同增长率），计算出 2000-2020 年各省每年的人口总数，作为 students.txt 数据的第 5 列，名为总人口；
- 3、 将 students.txt 数据导入到 Excel 形成 students.xlsx，并形成统计图，分别是：各省各年小学一年级人数图、各省各年小学一年级人数与总人口比例图，各省高中毕业生人数与总人口比例图、各省各年小学毕业生人数与高中人数比例图；
- 4、 利用上述数据，并结合其他可能数据，探索其他用途。