20200416 作业

1、 进入“2020 年统计用区划代码和城乡划分代码网页”（http://www.stats.gov.cn/tjsj/tjbz/tjyqhdmhcxhfdm/2020/index.html）。以此为起点，自动爬取各省市自治区相应统计单位一直到最基层为止；

2、 爬取后的输出格式如下图所示(没有到最基层，以此类推)。每一级统计单位相对于上一级用\t 缩进，最高一级为省市自治区。每个统计单位前为统计用区划代码。最高一级的统计用区划代码为该区域前两位，后面依次补零。最

110000000000北京市
	110100000000市辖区
		110101000000东城区
		110102000000西城区
		110105000000朝阳区
		110106000000丰台区
		110107000000石景山区
		110108000000海淀区

基层有“城乡分类代码”，将该代码放到名称之后，形如：“110112005001 天赐良园社区居委会 111”，其中的 111 为“城乡分类代码”。

3、 将上述内容按格式保存到文件“学号_StatData.txt”之中；

4、 “城乡分类代码”含义如下：“111 表示主城区，112 表示城乡结合区，121 表示镇中心区，122 表示镇乡结合区，123 表示特殊区域，210 表示乡中心区，220 表示村庄”。分别统计各分类最基层统计单位数量格式如下：

	主城区（111）	城乡结合去（112）	镇中心区（121）	镇乡结合区（122）	特殊区域（123）	乡中心区（210）	村庄（220）
北京市							
天津市							
……							

省市名称顺序自定，间距自定或采用 Tab。

5、 分别针对“内蒙古自治区”和“河南省”含有“村委会”的最基层统计单位，统计去除“村委会”后，最常用字前

100 个，观察其异同，输出按字的频率又高到低顺序输出；

6、 根据文后附属的姓氏排行，统计带有不同姓氏的地名数量。注意：仅统计第一个字，仅统计最低两个层次；输出按文后给出的姓氏顺序，格式为每行前一个字符串为姓氏，中间以 Tab 隔开，后面为该形式的地名数。

7、 第 4、5、6 两小题的数据输出到“学号_ComputingData.txt”之中；

8、 建议分为多个独立程序运行。

9、 搜索学习：“我们分析了 67 万个村名，找到了中国地名的秘密”。

附：

2018 年中国姓氏前 100 名：

01 李 02 王 03 张 04 刘 05 陈 06 杨 07 赵 08 黄 09 周 10 吴

11 徐 12 孙 13 胡 14 朱 15 高 16 林 17 何 18 郭 19 马 20 罗

21 梁 22 宋 23 郑 24 谢 25 韩 26 唐 27 冯 28 于 29 董 30 萧

31 程 32 曹 33 袁 34 邓 35 许 36 傅 37 沈 38 曾 39 彭 40 吕

41 苏 42 卢 43 蒋 44 蔡 45 贾 46 丁 47 魏 48 薛 49 叶 50 阎

51 余 52 潘 53 杜 54 戴 55 夏 56 钟 57 汪 58 田 59 任 60 姜

61 范 62 方 63 石 64 姚 65 谭 66 廖 67 邹 68 熊 69 金 70 陆

71 郝 72 孔 73 白 74 崔 75 康 76 毛 77 邱 78 秦 79 江 80 史

81 顾 82 侯 83 邵 84 孟 85 龙 86 万 87 段 88 漕 89 钱 90 汤

91 尹 92 黎 93 易 94 常 95 武 96 乔 97 贺 98 赖 99 龚 100 文