20190321 作业

1、 进入“2017 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/2017/index.html）。以此为起点，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”。

输出格式：

代码+名字\n

\t 代码+名字\n

……

\t…\t 区划代码+城乡分类名称+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；

输出格式：111\t 数量\n112\t 数量\n，即一行显示一分类

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；

输出格式：

内蒙古自治区

字\t 数量\n 字\t 数量\n，即一行显示一个字

河南省

字\t 数量\n 字\t 数量\n，即一行显示一个字

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

7、 作业提交截止时间：2019-03-31 23:59 前。