采集网站数据

1、爬取高校的信息

链接：<https://yz.chsi.com.cn/sch/>

需要爬取以下6个信息+高校的代码：



这里的高校缺少了高校的代码。在第2个采集中有。

我建好了数据库的表头，详见文件：“高校信息库”

1. 爬取招生目录信息

链接：<https://yz.chsi.com.cn/zsml/>

先简单介绍一下硕士专业目录的构成：在考研报名的时候，首先会选择一个学校，学校的下面，会有学院，学院下面又有不同的专业，专业下面有研究方向，每个研究方向有相应的“考试方式”“学习方式”“拟招生人数”“考试范围”“跨专业”“备注”。

需要采集的就是：每个高校下面的学院信息、专业、方向，以及这个方向所属的门类、学科类别、学习方式、指导教师、拟招生人数、考试范围、是否跨专业、备注。考试范围可能会有多个，需要分别保存。我建好了一个数据库的表头，详见文件“2019招生目录”

即：

**地区 学校 学院 门类 学科类别 专业 研究方向** 学习方式 指导教师 拟招生人数 考试范围 跨专业 备注

上面的蓝色加粗，表示包括了“数字代码+名称”两项，保存的时候，分开保存。

