数据获取设计

一、数据爬取

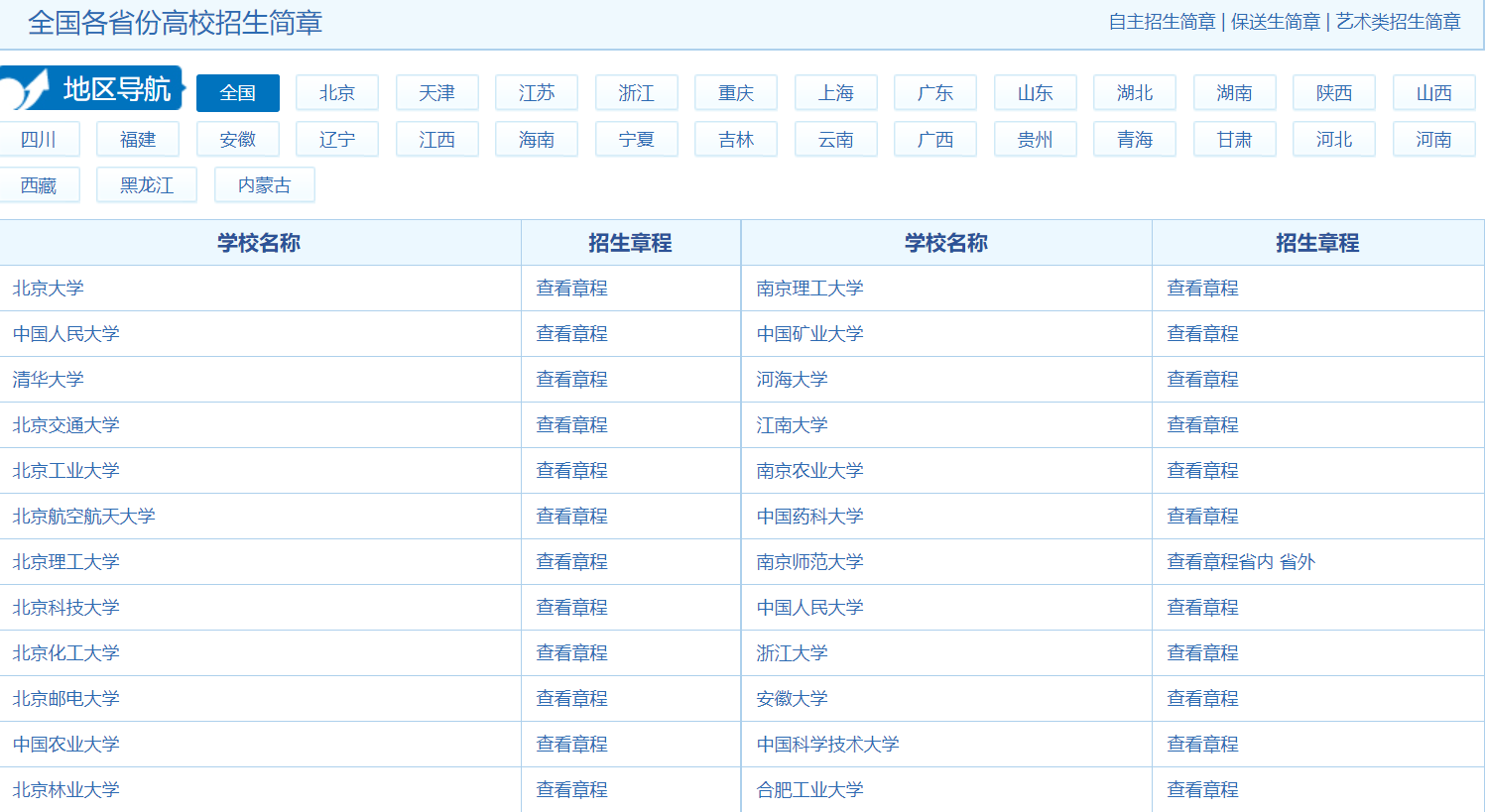
1、各省历年分数线；一本大学历年分数线；二本大学历年分数线；三本大学历年分数线

http://www.gaokao.com/baokao/lqfsx/gsfsx/

对爬到的数据进行预处理：提取出高校名称，省份，高校类型，以及该院校历年分数线文件路径等信息，以高校+省份为主键，保存到院校信息数据表。

2、各学校招生简章

http://www.gaokao.com/baokao/yxdq/zsjz/



爬取数据时需要逐个进入学校官网进行爬取，得到每个学校的高校隶属，高校所在地，院士数量，博士点数量，硕士点数量，通讯地址，联系电话，电子邮箱，学校网址，等信息，根据高校名称+省份的主键，存入数据表的对应表项中。

根据主键将院校简介文件路径，招生简章文件路径，专业目录文件路径也抓取下来存入院校信息数据表。

3、各种专业排名

http://www.gaokao.com/baokao/zydq/xzzy/



进入到各个专业的排名，需提取出专业的名称作为排名信息数据库中一个表的表头，获取排名表中高校和其对应的省内名次，全国名次，学科评级等，按降序写入专业排名表。

4、各种大学排名

http://www.gaokao.com/dxpm/zgdxpm/



进入到中国762所大学的自然科学与社会科学排名，获取排名表中高校和其对应的省内名次，全国名次，学科评级等，按降序写入自然科学排名表、社会科学排名表。

5、各大学专业目录(还可以在本界面进行分数线查询)

<http://college.gaokao.com/schlist/>

点进学校界面，专业介绍中是专业目录





根据高校名称+省份的主键，将专业目录文件路径抓取下来存入院校信息数据表。

二、数据转入

得到爬取到的各类信息表，可由管理员在管理端上传文件或直接上传到数据库，服务器接收到文件时按照主键和对应关系存入数据库。

三、数据更新

每隔两个月对数据源进行一次重新爬取，并根据爬取结果更新数据库。在每年5月、6月、7月每半个月进行一次数据更新。