**1.目标：构建高校分数信息库**

**2.具体所需数据：**

A）各高校近8年在各省的录取分数线（一般有投档线，最高分，最低分和平均分，其中最重要的是**投档线**）

B）各省近8年的批次线（一本线，二本线等）

C）各省近8年的分段排名数据（1分1段表，5分段表）

D）各高校近8年在各省的各专业的录取分数线（具体到各个专业）

E）各高校的基本信息（如百度百科中的信息，加上其它的特色信息）

**3.目标爬取网站**

A）易高考 http://www.easygaokao.com/

B）中国教育在线 http://gaokao.eol.cn/

C）高考网 http://www.gaokao.com/

D）优志愿 http://www.youzy.cn/

E）高考派 http://www.gaokaopai.com/

F）高考志愿网 http://www.gkzy100.com/

G）完美志愿 http://www.wmzy.com

H）腾讯高考 http://data.edu.qq.com/college/

上述网站目前状况分析：**易高考**是2016年12月份上线的，其数据目前都免费，但是不算全； **中国教育在线**和**高考网**的数据多但是也不全，有些关键数据没有，更新得慢，还不一定准确，但是做了很多年了； **高考派**是近两年来做的不错的网站和app，目前是被好未来收购，其数据目前只能查询前几页（其他需要会员）； **高考志愿网**是一家小公司做的，也是16年才上线，数据还不错，但也不算全，也是只能查一点点，需要会员（298元）才能看； **优志愿**主要是面向江苏省和浙江省，其他省的数据也不多； **完美志愿**数据也一般； **腾讯高考**也一般，更新也慢，完全可以做的比它好。 其他可能还有大大小小的网站，但是知名度不如上面的，肯定数据也一般，有待发现。

**4.信息库构建流程**

A）先完成湖南地区的各分数信息的爬取（之后再爬其他各省）。首先爬取关键信息，湖南近8年的批次线，分段表以及各高校在湖南的录取线。

B）先将数据直接存文件，保存成JSON格式，设计一个保存格式（类似于数据库表），将不同网站的数据统一存储，只存储关键信息。

C）先爬取易高考，中国教育在线，高考网，高考志愿网的湖南部分数据，之后作处理的时候，通过分析同一条数据的一致性来确定该分数是否准确（因为还不能相信网站给出的分数信息是正确的，一般批次线肯定是对的，其他的就不一定了），把有差异的数据标记出来，或者另存文件以作后续的核查。

D）数据处理完后设计数据库表，通JSON格式，存入数据库。

E）开发一个简易的查询系统（仅仅查询），可以是WEB，或者是H5的页面。（暑假时提供给安仁的学弟学妹使用，以供参考）

**5.爬虫方法简介**

以程序的方法通过HTTP向服务器请求一个页面（URL地址），将服务器返回的页面中的关键数据提取出来。一般python获取一个URL的请求页面只有3行代码：

request = urllib2.Request(url, headers = header)  
response = urllib2.urlopen(request, timeout=100)  
html = response.read()

其中URL和header如下：

url = "http://www.gkzy100.com/member/csjbyUni.html"  
header = {'User-Agent': 'Mozilla/5.0'}

请求时，只需输入URL地址和请求头header（模拟一个浏览器）即可。返回的页面中读出来就成了字符串，即一个网页的HTML源代码。之后就在这个HTML字符串中通过分析具体的标签结构，get关键数据，可以用一些库直接操作，例如beautifulSoup，很方便，还有其他的库也可以。

有的网站因为需要切换大量数据，用了AJAX技术，一般使用JSON来传输数据，所以获取AJAX里面的get json数据的URL很关键，如果可以得到这个URL，那返回的就是一段JSON数据，都不需要对HTML结构作分析。

查看JSON数据的URL方法如下：用谷歌浏览器，易高考网站举例。

例如查个地区批次线：

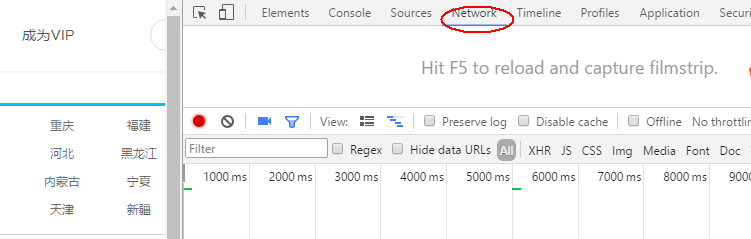
<http://www.easygaokao.com/dataSearch.html?nav_1=sjcx&nav_2=cfsx>



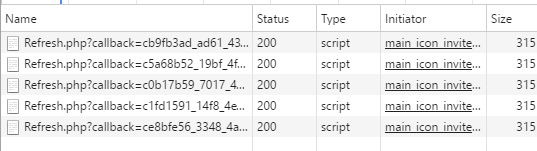
表中的数据是通过AJAX获取，直接右键查看源代码，在html中根本找不到具体的表中的分数内容。

借助谷歌的开发者工具（F12）

点击network（它用来查看一次页面请求中所有从服务器传来的文件）

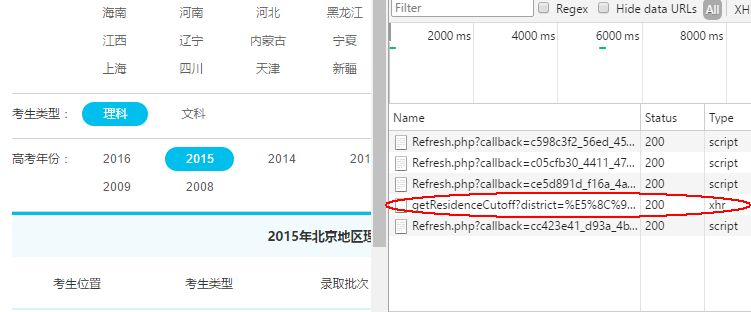


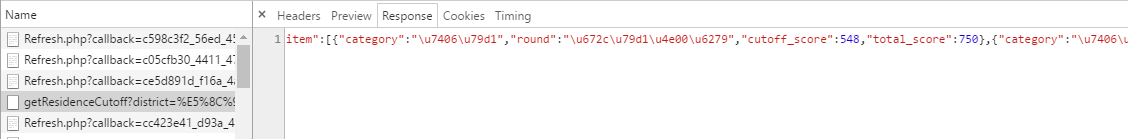
看到底下的refresh，先不管，它是自动刷新了什么鬼。



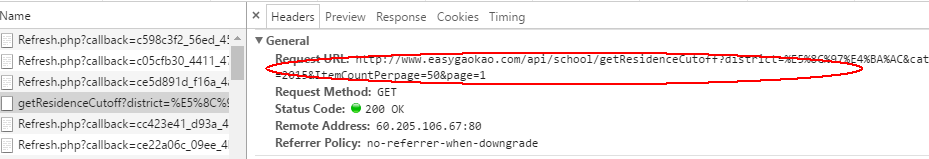
点击一下左边网页中的任何按钮，例如点一下2015年，换个年份查询。

右边出来一个特殊的URL请求。点开一看，就是它。

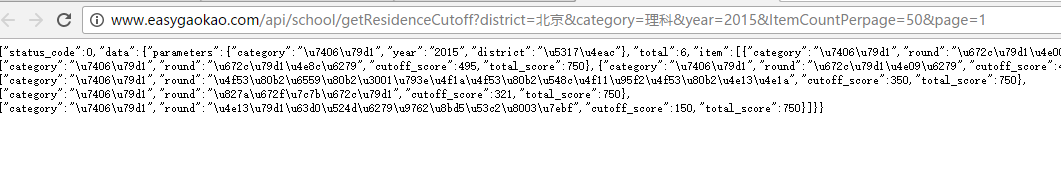




再看看他的具体的http请求怎么写

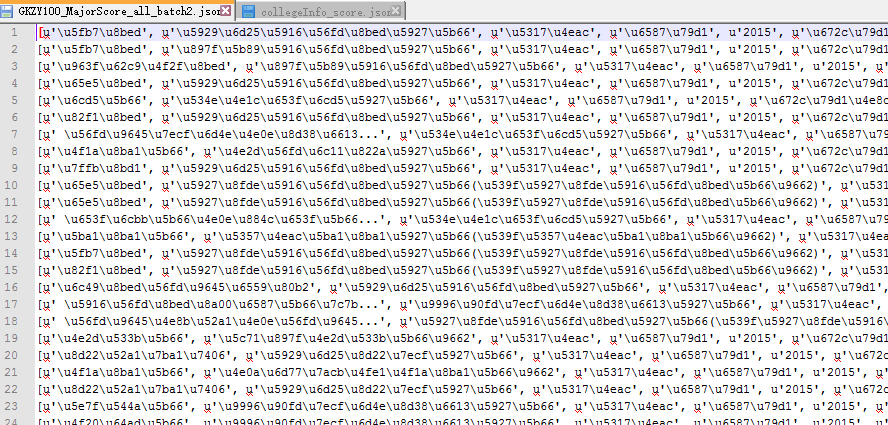


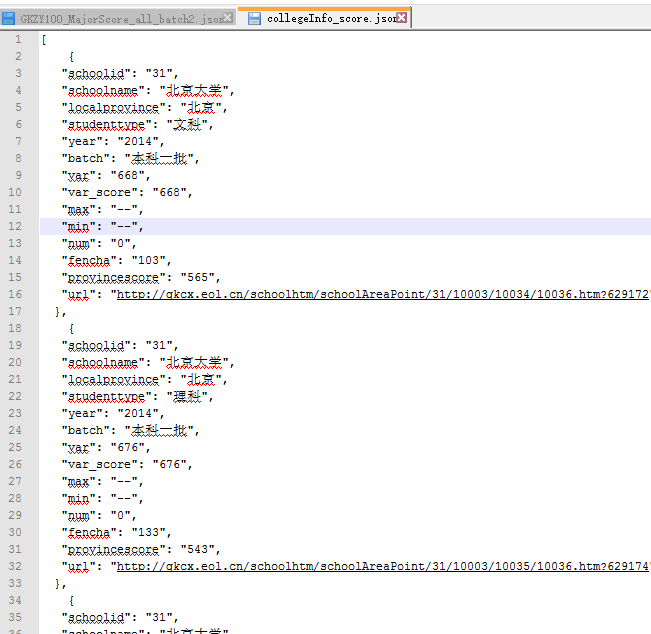
复制该URL直接在浏览器中打开



就是这个URL了，通过改变参数值，就可获取所有的数据了。这个爬下来可以直接存储了。

当时存成这样了，没做什么处理，返回的JSON是什么结构就存什么结构了。





Python网络爬虫案例，百度一大把，都还可以。

附一个python基础教程，和爬取高考志愿网的分数线的代码示例，比较粗鲁。