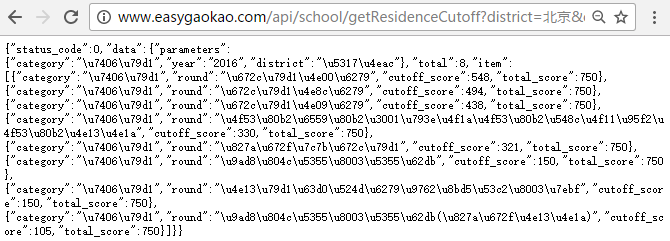
1. 爬取原则：网页中，有用的数据全爬下来，并以一定规则存储。
2. 爬虫爬取的URL对象分两种，一种返回的是一串JSON数据，一种返回的是HTML页面
   1. 返回JSON数据：

此种情况下很好办，返回的是一串由大括号{}括起来的JSON数据，因为考虑保存原生数据，固先不对返回的JSON串做去杂的处理，只要返回了有用数据，统统直接存文件。例如爬取易高考的各省历年批次线，步骤如下：

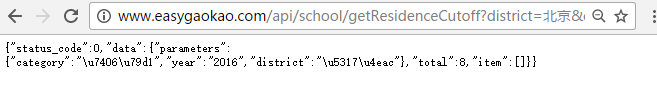
1. 例如，爬虫请求

<http://www.easygaokao.com/api/school/getResidenceCutoff?district=%E5%8C%97%E4%BA%AC&category=%E7%90%86%E7%A7%91&year=2016&ItemCountPerpage=50&page=1>，该URL的返回是一串JSON数据，如下图所示。

1. 判断该JSON数据中包含了所需要的有用信息：



1. 如果返回的JSON中无有用信息，则略过，例如当上述URL最后的参数改为2时，返回的Item里面的数据为空：



1. 存文件，保存为JSON格式（.json），存的时候为了方便后续读取该文件内容，在每个JSON串中间加上隔开符号，例如加上‘，\n’，使得每个JSON占一行，且有逗号区分，下次再读取的时候便可以以此分开不同的JSON串。
2. 爬取过程中，在程序中打印计数器信息，显示当前爬取的url的参数，若出现异常还需记录异常时的url参数信息，以便下一次从该处接着爬，避免重复。
   1. 返回HTML页面：

此种情况下，需对请求后返回的页面作HTML解析，根据HTML标签的属性，定位到目标数据所在的标签内，再获取该数据。需先分析好要获取哪些数据，以便每个页面都存储这些数据。

例如爬取高考网的批次线：

1. 爬虫请求<http://college.gaokao.com/areapoint/a1/s1/y2016/>



1. 查看网页HTML结构，定位上表所在的位置和标签。如下，数据在table里。



1. 确保每个返回的HTML页面中，table的属性和结构都一致，这样就可以存储每个页面的关键信息了。将该页所爬的所有数据构造成一个JSON串，或者其他有结构性的数据结构，方便后续的读取。然后写入文件。