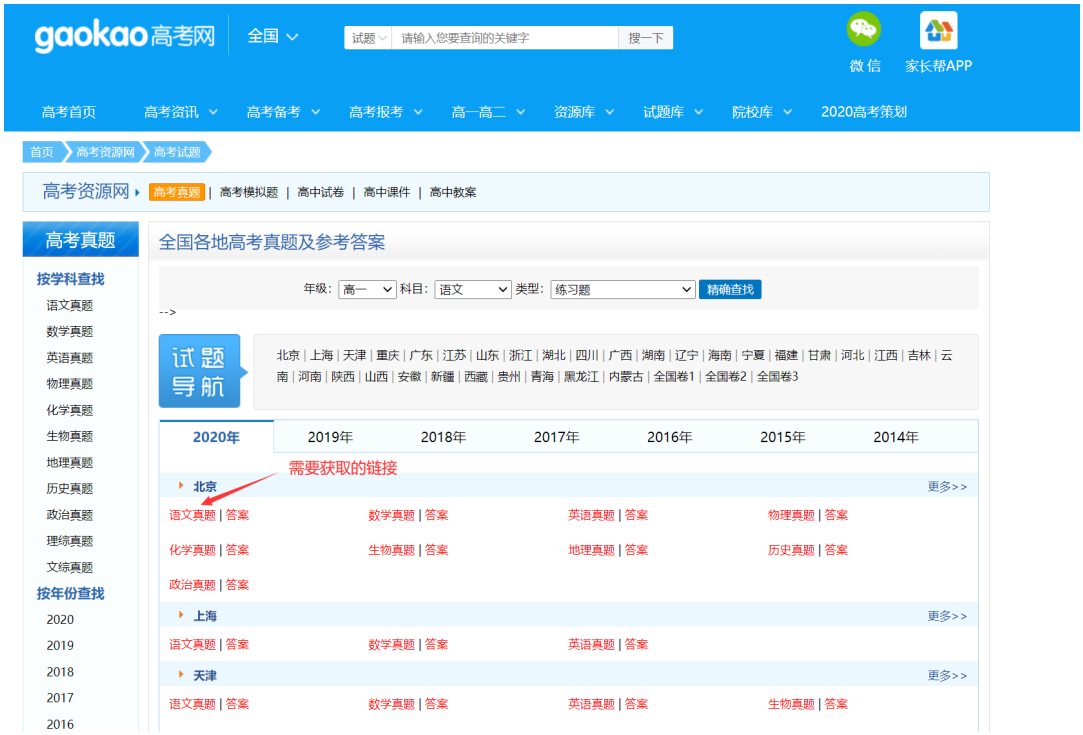
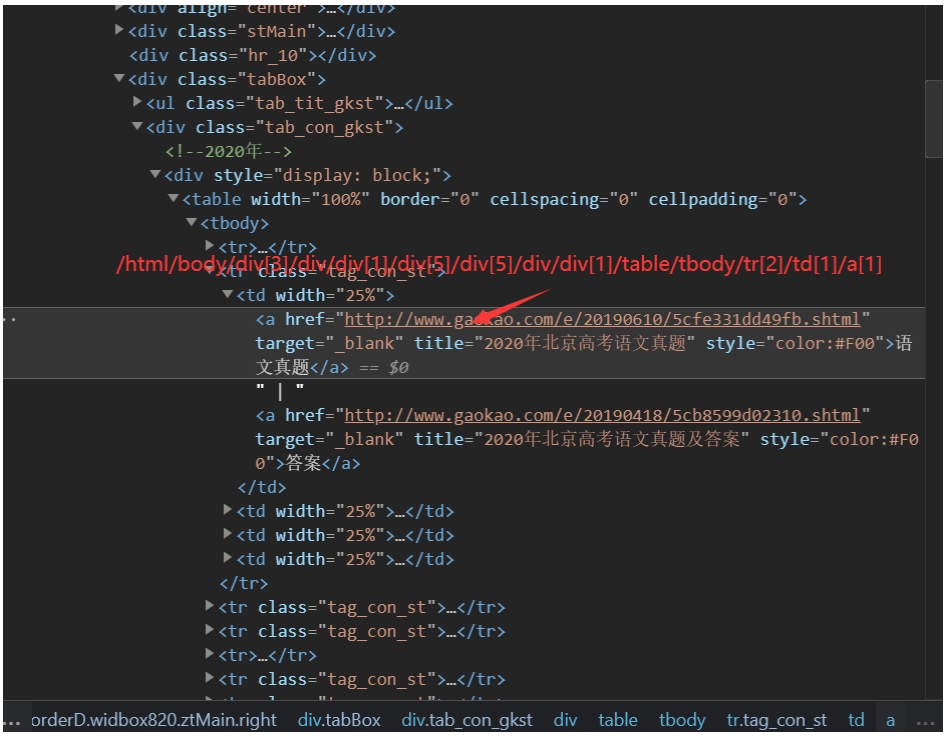
README

大致思路

1. 选取目录页作为爬取起始页

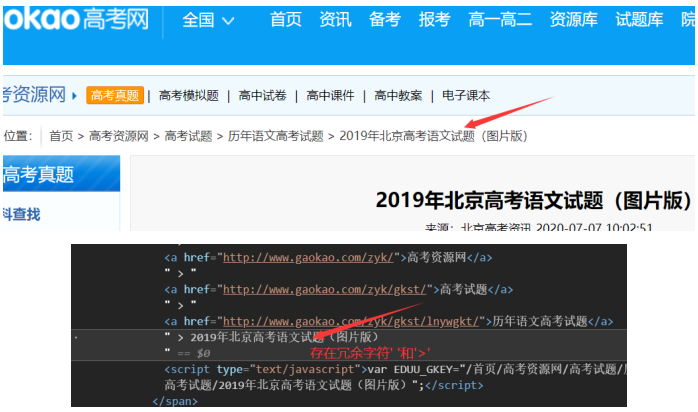


观察Xpath寻找规律：



2. 由于此页面链接Xpath中含有tbody标签导致获取列表为空，将页面body下源码存为txt文件来读取。
3. 获取足够的套卷首页链接后，循环下载30套试卷
    - 从试卷首页开始，获取全部试卷列表

    - 获取标题，通过比对地区列表与科目列表来命名下载的文件



    - 判断是word格式还是图片格式，如果是word格式标题中会含有‘word’字符串。word格式则下载.doc文件，图片格式则根据试卷列表下载.jpg文件。