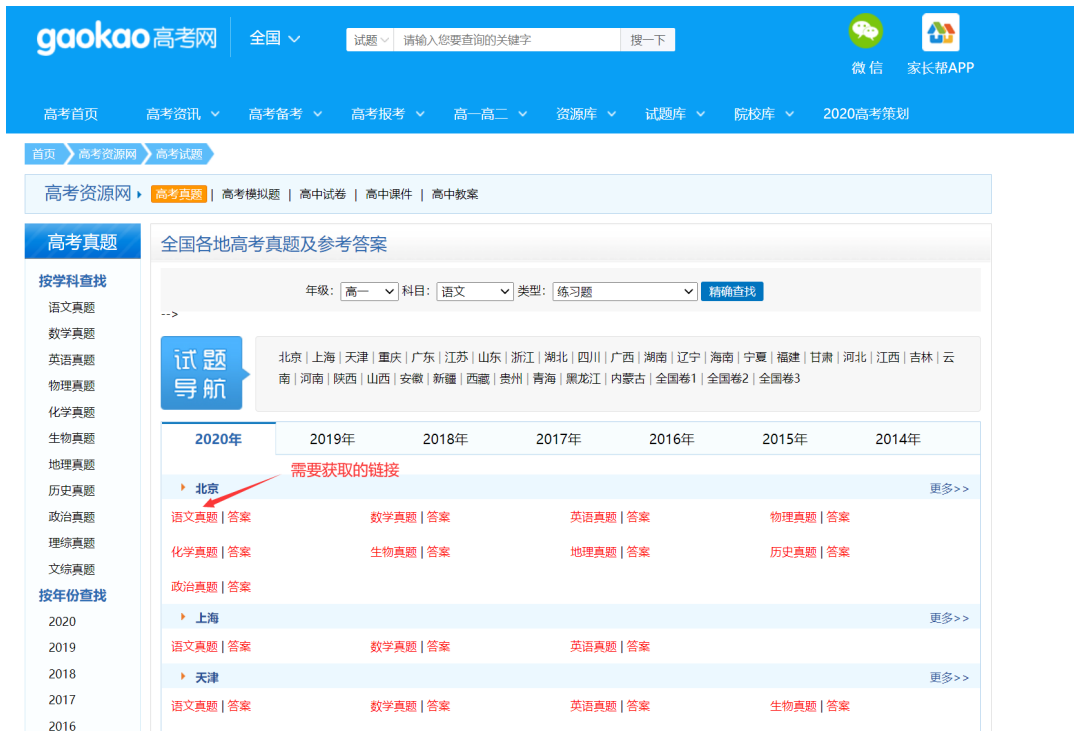


README

大致思路

1. 选取目录页作为爬取起始页



观察Xpath寻找规律：

```
<div align="center">...</div>
<div class="stMain">...</div>
<div class="hr_10">...</div>
<div class="tabBox">
  <ul class="tab_tit_gkst">...</ul>
  <div class="tab_con_gkst">
    <!--2020年-->
    <div style="display: block;">
      <table width="100%" border="0" cellspacing="0" cellpadding="0">
        <tbody>
          <tr>...</tr>
          <tr class="tag_con_st">
            <td width="25%">
              <a href="http://www.gaokao.com/e/20190610/5cfe331dd49fb.shtml"
                target="_blank" title="2020年北京高考语文真题" style="color:#F00">语
                文真题</a> == $0
              " | "
              <a href="http://www.gaokao.com/e/20190418/5cb8599d02310.shtml"
                target="_blank" title="2020年北京高考语文真题及答案" style="color:#F0
                0">答案</a>
            </td>
            <td width="25%">...</td>
            <td width="25%">...</td>
            <td width="25%">...</td>
          </tr>
          <tr class="tag_con_st">...</tr>
          <tr class="tag_con_st">...</tr>
          <tr class="tag_con_st">...</tr>
          <tr class="tag_con_st">...</tr>
        </tbody>
      </table>
    </div>
  </div>
</div>
... orderD.widbox820.ztMain.right div.tabBox div.tab_con_gkst div table tbody tr.tag_con_st td a ...
```

```
<tr class="tag_con_st">
  <td width="25%">...</td>
  <td width="25%">
    <a href="http://www.gaokao.com/e/20200710/5f07d8c4d55cc.shtml"
      target="_blank" title="2020年北京高考数学真题" style="color:#F00">数
      /html/body/div[5]/div[1]/div[5]/div[5]/div/div[1]/table/tbody/tr[2]/td[2]/a[1]
    <a href="http://www.gaokao.com/e/20200710/5f07e2232e720.shtml"
      target="_blank" title="2020年北京高考数学真题及答案" style="color:#F0
      0">答案</a> == $0
  </td>
</tr>
<tr class="tag_con_st">...</tr>
<tr class="tag_con_st">...</tr>
<tr class="tag_con_st">...</tr>
<tr class="tag_con_st">...</tr>
```

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。



