数据库系统实验三

网页数据抓取与分析

数据科学与计算机学院　17大数据与人工智能

17341015　陈鸿峥

下面会详细阐述本次实验的完整流程。

一、 读取任务清单

这对于python来说易如反掌，我甚至没有调用任何库就从json文件中将我所需要处理的学校序号读取了进来。

```
task = eval(open("tasks.json","r").read())["17341015"]
task.sort()
```

即通过eval函数直接将输入文本处理为字典，并通过我的学号进行索引，同时对读取出来的列表进行排序。

二、 读取学校名称

接下来要将序号与对应的学校中文名称进行对应，通过读取“普通高等大学名单.xls”，进行分析即可。这里采用了pandas包，同样在几行之内可以将学校名称提取出来。这里需要注意学校序号并不和Excel的行号一一对应，故需要先与字典中的每一项序号进行匹配，创建一个mask数组后，再读取需要的行/学校。

```
import pandas as pd

df = pd.read_excel("普通高等大学名单.xls",header=2)
mask = [True if item in task else False for item in df["序号"]]
school_name = [school for school in df["学校名称"][mask]]
```

三、 获取学校信息网站

这里通过教育部官方平台阳光招考网[1]获取学校的官方信息，由于学校序号同样不是一一对应的，故这里需要进行一次模拟搜索。通过观察搜索网页的url地址可以发现，通过在yxmc后填写学校中文名称即可实现筛选学校的功能。

[1] https://gaokao.chsi.com.cn

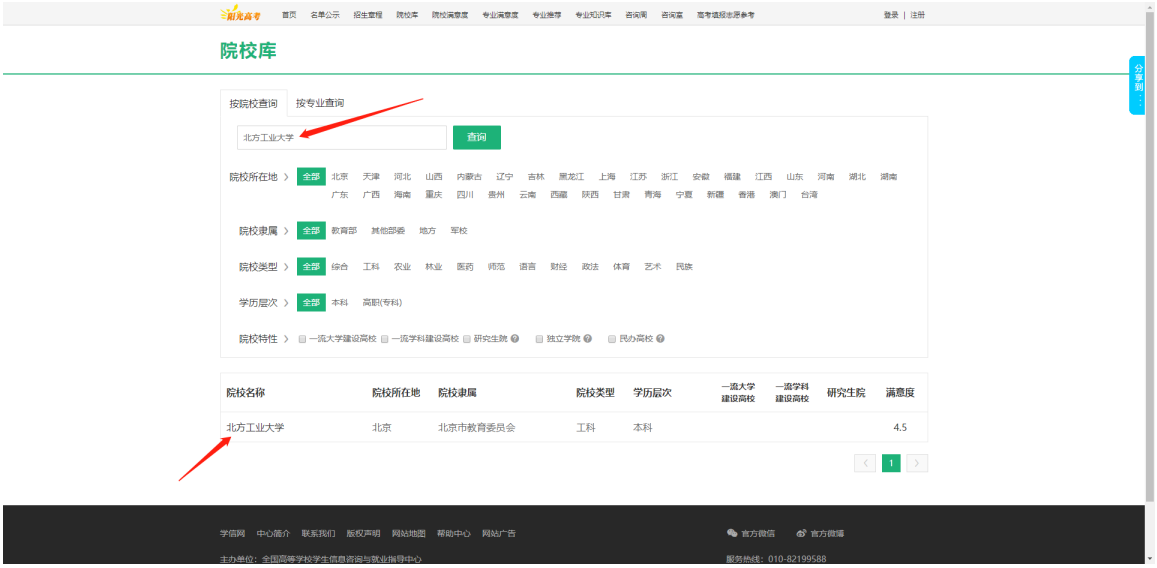


图 1: 阳光招考网的搜索方式与搜索结果

这里用request包实现网页的下载，然后通过BeautifulSoup对html源码进行解析，并找到第一条搜索记录的网址，记录入gaokao_url数组中。

```
import requests
from bs4 import BeautifulSoup

gaokao_url = []
for (step,school) in enumerate(school_name,1):
  print(step,end=" ")
  url = "https://gaokao.chsi.com.cn/sch/search.do?searchType=1&yxmc=" + school + "
      ↪ &zymc=&sySsdm=&ssdm=&yxls=&yxlx=&xlcc="
  page = requests.get(url,timeout=30)
  soup = BeautifulSoup(page.content,'lxml')
  try:
    td = soup.find("td",attrs={"class":"js-yxk-yxmc"})
    tag = td.contents[1]
    link = tag["href"]
    gaokao_url.append("https://gaokao.chsi.com.cn" + link)
  except:
    gaokao_url.append(None)
```

四、 获取学校官网

通过下图可以看到，在教育部招考网上，每间学校都会有对应的官网放在页面上，因此通过分析源码结构，对其中的内容进行抓取即可。

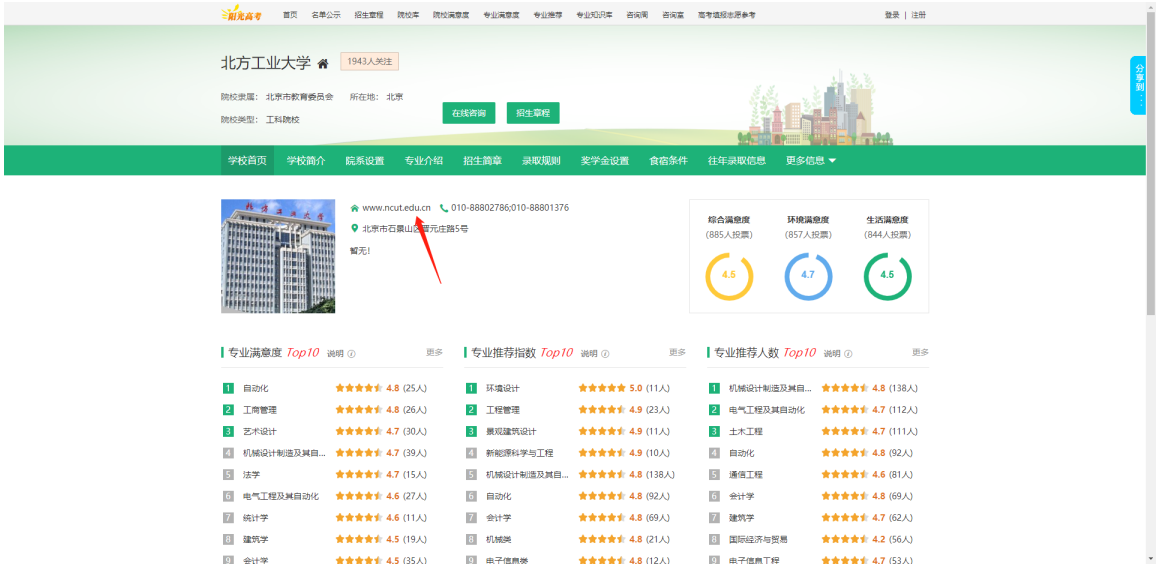


图 2: 官网网址所在位置

但是这里需要注意，有些学校连教育部都没有收录其官网，这些学校我们将忽略不计。在处理过程中则需要采用try...except...语句来处理这种情况。有记录官网的则将官网抓取出来加入official_url数组中，没有记录的则直接添加None进入数组。同时还要注意不同页面的布局格式不太一样，因此这里有两种不同的异常处理机制。

```
official_url = []
cnt_none = 0
for (step,url) in enumerate(gaokao_url,1):
  print(step,end=" ")
  if url == None:
    official_url.append(None)
    continue
  page = requests.get(url,timeout=30)
  soup = BeautifulSoup(page.content,'lxml')
  div = soup.find("div",attrs={"class":"msg"})
  try:
    tag = div.find("a")
    link = tag["href"]
    official_url.append(link)
  except:
    try:
      span = soup.find("span",attrs={"class":"judge-empty"})
      p = span.find("p")
      link = p.contents[0]
      official_url.append(link) # need human involution
      print(school_name[step-1],link)
    except:
```

```
        cnt_none += 1
        official_url.append(None)

print(official_url)
print(len(official_url))
outfile = open(official_url_file_name,"w")
outfile.write(str(official_url))
print(cnt_none)
```

为避免重复抓取，当所有学校官网都被抓取后，会被保存在一个文件中。如果程序发现该文件存在，则下次运行时将直接读取，而不会再到网页上抓取，如下代码所示。

```
if official_url_file_name not in os.listdir():
  # execute the program above
else:
  url_file = open(official_url_file_name,"r")
  official_url = eval(url_file.read())
```

五、 抓取学校网页并保存

下面的代码则是本次实验的核心内容。

一开始我想着通过浏览器模拟来直接保存页面，即使用selenium调用Chrome的内核，并执行页面保存的操作，如下代码所示。但是后来发现这种方式实在太过低效，因为每次调用浏览器内核都会产生巨大的运行时开销，且这些操作都是在前台运行的，对用户的其他工作非常不友好。

```
from selenium import webdriver
from selenium.webdriver.common.action_chains import ActionChains
from selenium.webdriver.common.keys import Keys

br = webdriver.Chrome()
br.get(url)
save_me = ActionChains(br).key_down(Keys.CONTROL).key_down('s').key_up('s').key_up
    ↪ (Keys.CONTROL)
save_me.perform()
print("done save")
browser.implicitly_wait(5)
enter = ActionChains(br).key_down(Keys.ENTER).key_up(Keys.ENTER)
enter.perform()
```

因此后来我还是采用了最为原始的方式，即分析网页源码，从中获取内网所有页面资源并逐一进行下载，具体流程如下：

- 首先抓取学校官网首页，并保存为`index.html`

- 然后对当前网页进行html解析，从中获取所有href标签，添加入下一级需要遍历的页面中。这里需要注意几点
  - href中的网页往往是相对地址，需要使用urljoin与官网url进行整合获得绝对地址
  - 这里判断是否学校内部网站一个简单但不一定准确的途径是看网页地址中的学校缩写是否相同，若不相同则判定为外网，在下一轮迭代中将不会对其进行分析
- 不断循环遍历新页面中出现的网页（宽度优先搜索），将当前页面保存并解析链接地址。注意这里页面文件保存的地址与网页url的相对地址相同。如http:www.sysu.edu.cn\folder1\file，则file.html将会被在www.sysu.edu.cn\foler1的目录下，这样在本地打开离线网页时依然可以进行相对寻址。
- 直到不再产生新页面链接或达到最大深度时停止。（尽管进行了环检测[2]，但是有些学校的网站实在太过庞大，不设置最大深度根本无法爬完）

```
import os, sys
from urllib.parse import urljoin

def crawl(pages, school_abbr, school_chinese_name, pagefolder="pages"):
  pagefolder += sep + school_chinese_name
  if not os.path.exists(pagefolder):
    os.makedirs(pagefolder)
  try:
    folder = pagefolder + sep + "index.html"
    html = requests.get(pages[0],timeout=30)
    with open(folder,"wb") as file:
      file.write(html.content)
    logger.info("Saved {}".format(folder))
  except:
    logger.info("Cannot create {}".format(folder))
    print("Cannot create {}".format(school_chinese_name))
    return

  indexed_url = []
  curr_depth = 0
  while len(pages) != 0:
    curr_depth += 1
    if curr_depth > MAX_DEPTH:
      break
    new_pages = []
    for page in pages:
      if page not in indexed_url:
```

[2]indexed_url数组用于查重

```
        indexed_url.append(page)
        try:
          html = requests.get(page,timeout=30).content
        except:
          logger.info("Could not open {}".format(page))
          continue
        filename = page[7:].split("/")
        if filename[-1] == "" or filename[-1][-3:] == ".cn" or filename[-1][-4:] ==
            ↪  ".com":
          folder = pagefolder + sep + filename[0]
          try:
            if not os.path.exists(folder):
              os.makedirs(folder)
            folder += sep + "index.html"
            if not os.path.isfile(folder):
              with open(folder,"wb") as file:
                file.write(html)
              logger.info("Saved {}".format(folder))
          except:
            logger.info("Cannot create {}".format(folder))
        else:
          for i,path in enumerate(filename):
            folder = pagefolder + sep + sep.join(filename[:i+1])
            if i == len(filename) - 1:
              if not os.path.isfile(folder):
                try:
                  with open(folder,"wb") as file:
                    file.write(html)
                  logger.info("Saved {}".format(folder))
                except:
                  logger.info("Cannot save {}".format(folder))
            else:
              try:
                if not os.path.exists(folder):
                  os.makedirs(folder)
              except:
                logger.info("Cannot create {}".format(folder))
        # get next urls
        soup = BeautifulSoup(html,'lxml')
        links = soup.find_all("a") # find all sub links
        for link in links:
          if "href" in dict(link.attrs):
            rel_path = link['href']
            tmp_url = urljoin(page,link['href'])
            if tmp_url.find("'") != -1 or tmp_url.find(school_abbr) == -1:
```

```
            continue
          tmp_url = tmp_url.split("#")[0]
          if tmp_url[0:4] == "http":
            new_pages.append(tmp_url)
    pages = new_pages
  print("Finish {}".format(school_abbr))
```

六、 其他设施

1. 平台适应

考虑到我的程序可能在不同的操作系统上运行，因此这里采用了platform包对操作系统进行判定。如果发现程序运行所在的操作系统是Linux，则采用Linux的文件路径命名方式，即以/作为分隔符；若发现操作系统是Windows，则以\\[3]作为分隔符。

```
import platform

sep = "\\" if 'Windows' in platform.system() else "/"
```

2. 并发加速

由于单核单进程进行爬虫任务的执行实在太过缓慢，因此这里采用了Python的多进程设施[4]来并行执行爬虫任务。

具体实施则是通过创建一个进程池Pool，然后将所有爬虫任务都添加到进程池中，交由操作系统及运行时系统进行调度。

```
import multiprocessing

if __name__ == "__main__":
  pool = multiprocessing.Pool()
  for i in range(len(official_url)):
    url = official_url[i]
    if url == None:
      continue
    # if url == None or i < 10 or i > 20:
    #   continue
    print("Downloading {}...".format(url))
    pool.apply_async(crawl,args=([url],url.split(".")[1],str(task[i])+"-"+
        ↪ school_name[i]))

  pool.close()
  pool.join()
```

[3]这里进行了转义

[4]之所以不使用多线程是因为Python采用了全局解释器锁(Global Interpreter Lock, GIL)机制，导致其只能使用单核CPU

3. 日志记录

即使开启了并发处理，爬取100所学校的网页依然是件十分耗时的工作，如果中间出现什么问题而没法跟踪的话是非常麻烦的。因此在我的程序中采用了标准库中的`logging`设施进行日志记录。通过将重要的网页抓取信息输出到日志文件中，我就可以知道爬虫到底有没有出错，并且哪些学校的页面没有被正常抓取。

```
import logging

logger = logging.getLogger(__name__)
logger.setLevel(level = logging.INFO)
handler = logging.FileHandler("log.txt")
handler.setLevel(logging.INFO)
formatter = logging.Formatter('%(asctime)s - %(name)s - %(levelname)s - %(message)
    ↪ s')
handler.setFormatter(formatter)
logger.addHandler(handler)

# example
logger.info("Saved {}".format(folder))
```

七、 实验结果

通过分析，下面2所学院在阳光招考网上找不到对应学校，猜测学校可能更名或者被撤销了。

2021 广西师范学院师园学院

2080 海南科技职业学院

下面3所学校则是在阳光招考网上找不到对应的学校官网，教育部网站上显示“暂无”字样。

1563 河南艺术职业学院

2546 甘肃卫生职业学院

2550 甘肃能源化工职业学院

下面33所学校则无法正常访问官网[5]，可能是网站太久没维护，域名过期等原因导致出错。

9 北方工业大学

222 河北能源职业技术学院

230 唐山工业职业技术学院

[5]即首页都访问不了，这里抓取时采用的网络是我校的有线网。

306 山西艺术职业学院
337 山西运城农业职业技术学院
433 中国刑事警察学院
485 沈阳职业技术学院
546 长春理工大学光电信息学院
611 哈尔滨石油学院
623 黑龙江职业学院
693 上海建桥学院
724 上海体育职业学院
809 江苏工程职业技术学院
829 南通科技职业学院
870 苏州百年职业学院
965 浙江艺术职业学院
1035 安徽信息工程学院
1077 安徽工商职业学院
1134 三明学院
1235 南昌工学院
1319 青岛农业大学
1403 山东外贸职业学院
1410 山东化工职业学院
1618 荆楚理工学院
1684 武汉科技职业学院
1703 湖北青年职业学院
1744 湖南女子学院
1829 湖南国防工业职业技术学院
1977 广东舞蹈戏剧职业学院
2188 四川工商学院
2266 安顺学院
2322 贵州装备制造职业学院
2482 陕西邮电职业技术学院

即一共有38所高校无法获取官网网页，其余62所高校均可以正常抓取。

最终结果如下图所示，我采用学校序号+学校中文名的方式对文件夹命名，下图展示了所有95所**有官网**学校的文件夹。

```
chz@HAS-T640:~/MyTest/Database/pages_new$ ls
1035-安徽信息工程学院          1618-荆楚理工学院              2266-安顺学院                  485-沈阳职业技术学院
1077-安徽工商职业学院          1647-湖北工程学院新技术学院    2291-安顺职业技术学院          504-大连航运职业技术学院
1134-二明学院                  1684-武汉科技职业学院          230-唐山工业职业技术学院       546-长春理工大学光电信息学院
1202-泉州轻工职业学院          1703-湖北青年职业学院          2322-贵州装备制造职业学院      547-长春财经学院
1235-南昌工学院                1741-湖南财政经济学院          2419-西安科技大学              561-长春医学高等专科学校
1238-华东交通大学理工学院      1744-湖南女子学院              2420-西安石油大学              571-松原职业技术学院
1240-南昌航空大学科技学院      1792-湘潭医卫职业技术学院      2439-商洛学院                  595-齐齐哈尔大学
1245-江西师范大学科学技术学院  181-防灾科技学院               2472-西安航空职业技术学院      611-哈尔滨白油学院
1319-青岛农业大学              1829-湖南国防工业职业技术学院  2479-宝鸡职业技术学院          623-黑龙江职业学院
1330-鲁东大学                  182-河北经贸大学               2481-陕西电子信息职业技术学院  688-上海电机学院
1356-烟台大学文经学院          1849-广东药科大学              2482-陕西邮电职业技术学院      690-上海政法学院
1403-山东外贸职业学院          1879-华南理工大学广州学院      2520-甘肃医学院                693-上海建桥学院
1410-山东化工职业学院          1954-肇庆医学高等专科学校      293-山西农业大学信息学院       724-上海体育职业学院
1416-潍坊工商职业学院          1957-广州华南商贸职业学院      300-太原工业学院               771-东南大学成贤学院
1435-山东传媒职业学院          1960-广东工程职业技术学院      306-山西艺术职业学院           785-南京中医药大学翰林学院
1439-山东文化产业职业学院      1972-广州华夏职业学院          333-忻州职业技术学院           809-江苏工程职业技术学院
1470-许昌学院                  1977-广东舞蹈戏剧职业学院      337-山西运城农业职业技术学院  822-常州信息职业技术学院
1480-孟乡学院                  197-河北工业大学城市学院       372-呼和浩特职业学院           829-南通科技职业学院
1516-黄河水利职业技术学院      2089-重庆师范大学              392-乌兰察布医学高等专科学校  855-南京铁道职业技术学院
1537-高山少林武术职业学院      2108-重庆第二师范学院          398-赤峰工业职业技术学院       870-苏州自年职业学院
1552-郑州城市职业学院          214-张家口职业技术学院         421-中国医科大学               900-浙江中医药大学
156-河北建筑工程学院           2159-四川农业大学              433-中国刑事警察学院           965-浙江艺术职业学院
1575-郑州黄河护理职业学院      2188-四川工商学院              454-锦州医科大学医疗学院       9-北万工业大学
15-北京建筑大学                222-河北能源职业技术学院       476-辽阳职业技术学院
```

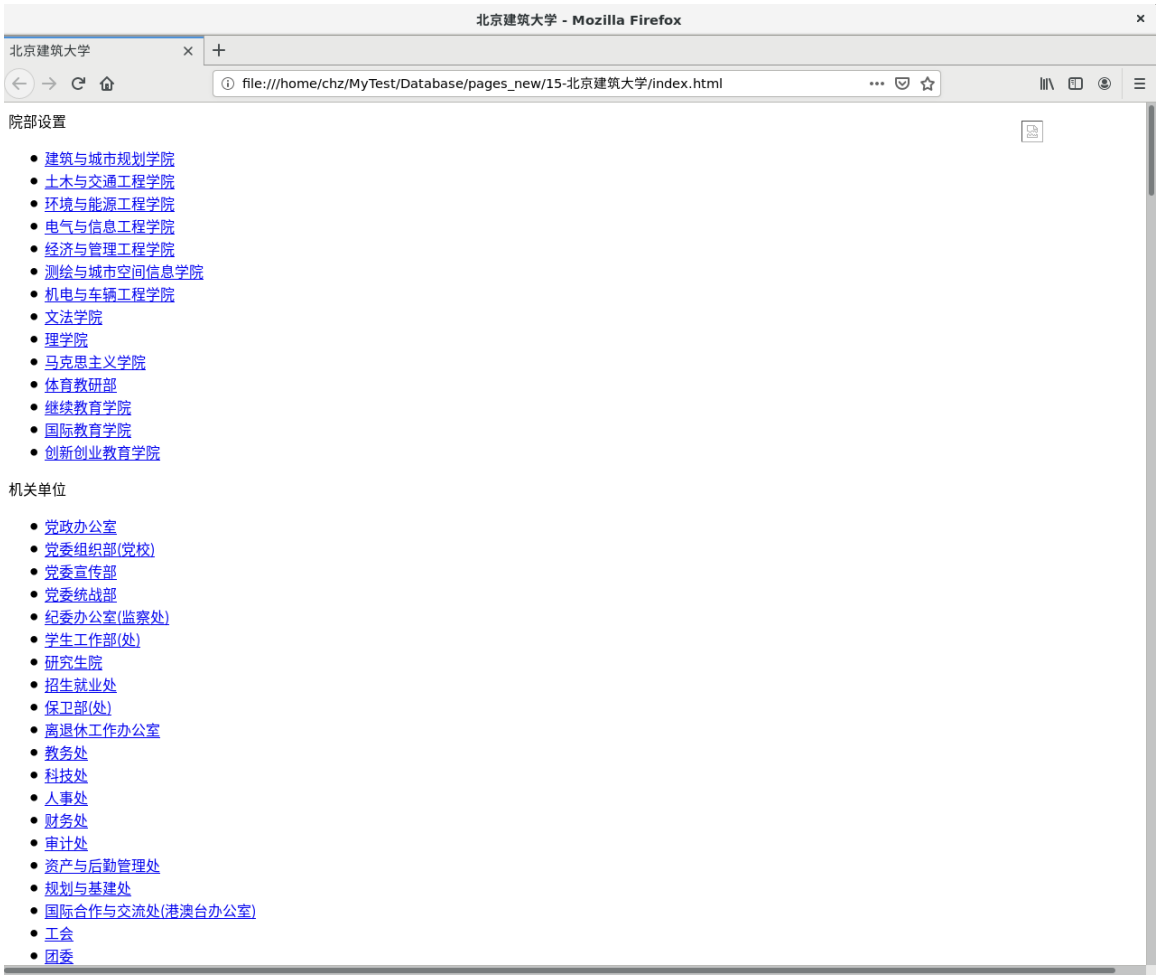
对于每一所学校，所有含有学校英文缩写的域名都会被抓取下来，下图是北京建筑大学的例子。可以看到最外层会有一个名为index.html的网页，这个即为学校官网的主页面。其他文件夹均为与主页面同级的页面，可能是不同的学院官网，也可能是信息网页，文件夹命名方式与域名相同，方便索引。

```
chz@HAS-T640:~/MyTest/Database/pages_new/15-北京建筑大学$ ls
aah.bucea.edu.cn           dzb.bucea.edu.cn          jingguan.bucea.edu.cn  qzlx.bucea.edu.cn       www.bucea.edu.cn
ada.bucea.edu.cn           english.bucea.edu.cn      jwbgs.bucea.edu.cn     rsc.bucea.edu.cn        wzb.bucea.edu.cn
brbfac.bucea.edu.cn        gccxzx.bucea.edu.cn       jwc.bucea.edu.cn       sci.bucea.edu.cn        xb.bucea.edu.cn
bucea.fanya.chaoxing.com   gczx.bucea.edu.cn         jxjy.bucea.edu.cn      sie.bucea.edu.cn        xcgl.bucea.edu.cn
bucea.jincin.com           gjy.bucea.edu.cn          jxzlgc.bucea.edu.cn    sie.bucea.edu.cn:8080   xsgz.bucea.edu.cn
bwc.bucea.edu.cn           gzhy.bucea.edu.cn         jyjjh.bucea.edu.cn     sjc.bucea.edu.cn        xww.bucea.edu.cn
bzdemo.elenut.com          hnxy.bucea.edu.cn         jzxy.bucea.edu.cn      tiet.bucea.edu.cn       xxq.bucea.edu.cn
cgyx.bucea.edu.cn          hq.bucea.edu.cn           kjc.bucea.edu.cn       tuanwei.bucea.edu.cn    xyh.bucea.edu.cn
chxy.bucea.edu.cn          hvac.bucea.edu.cn         ltb.bucea.edu.cn       tu.bucea.edu.cn         yjsc.bucea.edu.cn
cwc.bucea.edu.cn           i.bucea.edu.cn            mail.bucea.edu.cn      tumu.bucea.edu.cn       zhc.bucea.edu.cn
cxcy.bucea.edu.cn          index.html                mail.stu.bucea.edu.cn  tyb.bucea.edu.cn        zjc.bucea.edu.cn
ddh.bucea.edu.cn           indoorinfo.cn             material.bucea.edu.cn  udc.bucea.edu.cn        zuzhibu.bucea.edu.cn
door.bucea.edu.cn          int.bucea.edu.cn          mayuan.bucea.edu.cn    updi.bucea.edu.cn
dwtzb.bucea.edu.cn         jdxy.bucea.edu.cn         nic.bucea.edu.cn       usswe.bucea.edu.cn
dxxy.bucea.edu.cn          jg.bucea.edu.cn           pinggu.bucea.edu.cn    wenfa.bucea.edu.cn
```

如下图所示，可以看到每个子域名下面同样是按照这种层次结构排布网页内容。由于限制了网页抓取的深度，因此每个文件夹下的子文件夹深度也不会太大，这里仅仅是作为示意。事实上，仅仅抓取最大深度为5的学校页面，所有学校的网页都已经达到了27G的大小。

```
chz@HAS-T640:~/MyTest/Database/pages_new/15-北京建筑大学$ cd tuanwei.bucea.edu.cn/
chz@HAS-T640:~/MyTest/Database/pages_new/15-北京建筑大学/tuanwei.bucea.edu.cn$ ls
cms  docs  ggl  index.htm  index.html  jcxx  twjj  twzs  wjhb  zhxw  zhxx  zthd  zzjg
chz@HAS-T640:~/MyTest/Database/pages_new/15-北京建筑大学/tuanwei.bucea.edu.cn$ cd jcxx
chz@HAS-T640:~/MyTest/Database/pages_new/15-北京建筑大学/tuanwei.bucea.edu.cn/jcxx$ ls
index1.htm  index5.htm  index.htm
chz@HAS-T640:~/MyTest/Database/pages_new/15-北京建筑大学/tuanwei.bucea.edu.cn/jcxx$ cd ../zhxw
chz@HAS-T640:~/MyTest/Database/pages_new/15-北京建筑大学/tuanwei.bucea.edu.cn/zhxw$ ls
102111.htm  102513.htm  102690.htm  102919.htm  103666.htm  104249.htm  104378.htm  142473.htm  143196.htm  index33.htm
102512.htm  102515.htm  102767.htm  102973.htm  104247.htm  104363.htm  112647.htm  142866.htm  index1.htm  index.htm
```

下图是抓取下来的网页示意，可以看到页面打开后可以正常显示，仅仅包括文字内容。如果网页中的超链接(href)采用**相对链接**的方式，则可以直接离线访问文件夹中的其他子网页，否则是无法正常跳转的，这也与各学校网页的开发人员有关。



我在每一文件目录下都创建了README.md文档，提供说明帮助，一个例子如下。由于采用了python元语言方式生成，因此每间学校的说明文档也是不一样的。

```
## 15-北京建筑大学

* 学校官网： <http://www.bucea.edu.cn/>
* 使用说明：
    1. `index.html` 为学校官网首页
    2. 学校官网也可以从 `www.bucea.edu.cn/index.html` 进行访问
    3. 若网页中的超链接(`href`)
    采用相对链接的方式，则可以直接离线访问文件夹中的其他子网页；否则只能进入每一个文件夹逐一进行查看
```

所有中间结果文件也都附在附件中，可以查阅。

附录 A. 学校名称及官网对应

序号	学校名称	学校官网
9	北方工业大学	http://www.ncut.edu.cn/
15	北京建筑大学	http://www.bucea.edu.cn/

156	河北建筑工程学院	http://www.hebiace.edu.cn/
181	防灾科技学院	http://www.cidp.edu.cn
182	河北经贸大学	http://www.heuet.edu.cn/
197	河北工业大学城市学院	http://cc.hebut.edu.cn/
214	张家口职业技术学院	http://www.zhz.cn/
222	河北能源职业技术学院	http://www.hbnyxy.cn/
230	唐山工业职业技术学院	http://www.tsgzy.edu.cn/
293	山西农业大学信息学院	http://www.cisau.com.cn/
300	太原工业学院	http://www.tit.edu.cn/
306	山西艺术职业学院	http://www.sxyz.com/
333	忻州职业技术学院	http://www.xzvtc.com/
337	山西运城农业职业技术学院	http://www.ycnxy.com/
372	呼和浩特职业学院	http://www.hhvc.edu.cn/
392	乌兰察布医学高等专科学校	http://www.wlcbyz.org.cn
398	赤峰工业职业技术学院	http://www.nmgfzxx.cn/
421	中国医科大学	http://www.cmu.edu.cn/
433	中国刑事警察学院	http://www.cipuc.edu.cn
454	锦州医科大学医疗学院	http://www.jymu.edu.cn
476	辽阳职业技术学院	http://www.419.com.cn/
485	沈阳职业技术学院	http://www.vtcsy.com/
504	大连航运职业技术学院	http://www.dlsc.net.cn/
546	长春理工大学光电信息学院	http://www.csoei.com/
547	长春财经学院	http://www.ccufe.com/
561	长春医学高等专科学校	http://www.ccmc.edu.cn/
571	松原职业技术学院	http://www.sypt.cn/
595	齐齐哈尔大学	http://www.qqhru.edu.cn/
611	哈尔滨石油学院	http://www.hip.edu.cn/
623	黑龙江职业学院	http://www.hljp.edu.cn
688	上海电机学院	http://www.sdju.edu.cn/
690	上海政法学院	http://www.shupl.edu.cn
693	上海建桥学院	http://61.172.146.40/

724	上海体育职业学院	http://www.ssi.edu.cn/
771	东南大学成贤学院	http://cxxy.seu.edu.cn/
785	南京中医药大学翰林学院	http://www.hlxy.edu.cn/
809	江苏工程职业技术学院	http://www.jcet.edu.cn/
822	常州信息职业技术学院	http://zjczs.ccit.js.cn/
829	南通科技职业学院	http://www.ntac.edu.cn/
855	南京铁道职业技术学院	http://www.njrts.edu.cn/
870	苏州百年职业学院	http://www.hkuspace.edu.cn/
900	浙江中医药大学	http://www.zcmu.edu.cn
965	浙江艺术职业学院	http://zhaosheng.zj-art.com/
1035	安徽信息工程学院	http://www.ahpumec.edu.cn/
1077	安徽工商职业学院	http://ahbvc.cn/
1134	三明学院	mailto:smxyzsb@163.com
1202	泉州轻工职业学院	http://www.qzqgxy.com/
1235	南昌工学院	http://www.ncpu.edu.cn
1238	华东交通大学理工学院	http://www.ecjtuit.com.cn/
1240	南昌航空大学科技学院	http://www.nckjxy.cn/
1245	江西师范大学科学技术学院	http://kjxy.jxnu.edu.cn/
1319	青岛农业大学	http://www.qau.edu.cn/
1330	鲁东大学	http://www.ldu.edu.cn/
1356	烟台大学文经学院	http://wenjing.ytu.edu.cn/
1403	山东外贸职业学院	http://www.sdwm.cn
1410	山东化工职业学院	http://www.qledu.net/
1416	潍坊工商职业学院	http://www.wfgsxy.com/
1435	山东传媒职业学院	http://www.sdcmc.net/
1439	山东文化产业职业学院	http://www.sdcivc.com/
1470	许昌学院	http://www.xcu.edu.cn/
1480	新乡学院	http://www.xxu.edu.cn
1516	黄河水利职业技术学院	http://www.yrcti.edu.cn
1537	嵩山少林武术职业学院	http://www.shaolinkungfu.edu.cn/
1552	郑州城市职业学院	http://www.zcu.edu.cn/

1563	河南艺术职业学院	None
1575	郑州黄河护理职业学院	http://www.zyrnvc.com/
1618	荆楚理工学院	http://www.jcut.edu.cn/
1647	湖北工程学院新技术学院	http://www.hbeutc.cn
1684	武汉科技职业学院	http://www.whuvt.com
1703	湖北青年职业学院	http://www.hbqnxy.com/
1741	湖南财政经济学院	http://www.hufe.edu.cn/
1744	湖南女子学院	http://www.hnwu.edu.cn/
1792	湘潭医卫职业技术学院	http://www.xtzy.com
1829	湖南国防工业职业技术学院	http://www.hnkgzy.com/
1849	广东药科大学	http://www.gdpu.edu.cn/
1879	华南理工大学广州学院	http://www.gcu.edu.cn/
1954	肇庆医学高等专科学校	http://www.zqyz.gd.cn/
1957	广州华南商贸职业学院	http://www.hnsmxy.com/
1960	广东工程职业技术学院	http://www.gpc.net.cn/
1972	广州华夏职业学院	http://www.gzhxtc.cn/
1977	广东舞蹈戏剧职业学院	http://www.gdddc.cn
2021	广西师范学院师园学院	None
2080	海南科技职业学院	None
2089	重庆师范大学	http://www.cqnu.edu.cn/
2108	重庆第二师范学院	http://www.cque.edu.cn
2159	四川农业大学	http://www.sicau.edu.cn/
2188	四川工商学院	http://www.cdxy.edu.cn/
2266	安顺学院	http://www.asu.edu.cn/
2291	安顺职业技术学院	http://www.asotc.cn/
2322	贵州装备制造职业学院	http://www.gzzbzy.cn
2419	西安科技大学	http://www.xust.edu.cn
2420	西安石油大学	http://www.xsyu.edu.cn/
2439	商洛学院	http://www.slxy.cn/
2472	西安航空职业技术学院	http://www.xihang.com.cn/
2479	宝鸡职业技术学院	http://www.bjvtc.com/

2481	陕西电子信息职业技术学院	http://www.sxitu.com/
2482	陕西邮电职业技术学院	http://www.sptc.sn.cn/
2520	甘肃医学院	http://www.plmc.edu.cn/
2546	甘肃卫生职业学院	None
2550	甘肃能源化工职业学院	None

附录 B. 附件说明

- `spider`：爬虫核心程序
- `analysis`：网页爬取后分析及`README.md`写入
- `all_school_url`：附录A
- `gaokao_url`：阳光招考网上对应网站
- `official_url_new`：学校官网数组
- `school_chinese`：学校中文名数组
- `log.txt`：程序运行日志