Problem：新闻爬取 实验报告

Stu:　18级自动化 181870078 黄廷基

Date：2021.4.9

---

1. 题目分析

1.1 原题要求

> 1、任选一个新闻网站，例如中国青年网，爬取【40条以上】【“新冠疫情”】相关内容
>
> 2、基于正则表达式等工具，提取以下内容：
>
> （1） 新闻标题 （2） 发表时间 （3） 作者或单位
>
> （4） 文本内容 （5） 新闻图片链接

1.2 需求分析

- 选取网站：中国青年网
- 新闻关键词：“新冠疫情”
- 爬取内容：新闻链接、标题、时间、作者、文本内容、图片链接
- 数量：40+

---

2. 概要设计

2.1 运行环境

> 软件：Pycharm2020（Python 3.9）
>
> 扩展包：requests、BeautifulSoup
>
> 模拟浏览器版本：Chrome/89.0.4389.114

2.2 参数设置

> news_list / data_list：所有的新闻的信息
>
> news_one：某一条新闻的信息
>
> title：某一条新闻标题
>
> date：某一条新闻时间
>
> author：某一条新闻作者
>
> content：某一条新闻文本内容
>
> pic：某一条新闻图片链接

2.3 程序模块

```
1）搜索结果读取模块：{
                        # 通过链接中的数字特点进行翻页；
                        # 使用requests.get()的方法获取整页搜索结果文本；
                        # 利用beautifulsoup的find()找到并储存每页的新闻链接；
                        # 逐条爬取信息
                    }
2）单条新闻读取模块：{
                        # 使用request.get()获取单条新闻文本；
                        # 使用beautifulsoup的find()找到各类信息；
                        # 返回结果
                    }
3）主模块：{
                # 调用搜索、读取模块；
                # 使用write进行文件写入；
                # 保存
            }
```

---

3. 拓展功能实现

3.1 任意条新闻读取

根据搜索结果的链接特点：

```
link = http://***&p=' + str(j) + '&s****'
```

可以通过更改 j 的数值范围，从而收集任意页的搜索结果。

4. 实验调试问题

1）不同网页的相同数据项，其命名规则不同

解决：一开始想提取text文本，然后用正则表达式去匹配，后来发现功底太浅搞不定，最后直接干脆用分类实现，爬不到的新闻就算了。

2）中文编码规则问题

解决：针对不同网站，提前设置编码规则为utf-8或gd18030。

3）有时遇到了网页访问失败，程序陷入未响应状态

解决：在requests.get()中加入timeout参数，防止网页未响应。