year 用户手册

1. 使用步骤

1.1 获得 cookies 值

打开浏览器，登录百度账号。百度账号网址：

https://github.com/lpg-it/BaiduYunVIP/blob/master/baiduYunVIP.md

登陆成功，点击浏览器的【...】键，【更多工具】->【开发人员工具】，点击 network/网络，键盘输入【ctrl+R】，出现【www.baidu.com】，点击，复制 cookies 里的内容，到 config.py 中的 cookie 粘贴。这里要放入多个 cookies，3-5 个吧，能多不能少。

COOKIES = ['cookies1','cookies2',……]

注：cookies1 指的是刚刚复制的值

```
COOKIES = [
    BAIDUID=2C8B1515D6119C1FB6A42CA6A96CA968:FG=1; BIDUPSID=2C8B1515D6119C1FB6A42CA6A96CA968; PSTM=1569847469; BD_UPN=12314753; BDORZ=B490B5EBF6F3CD402E515D22BCDA1598; bdindexid=7sslvctfk4
]
```

1.2 代理 ip

打开 get_index.py，找到【_http_get】模块，更改里面的 proxy_list，格式为【http://<ip 地址>:<端口号>】(<>里是要被替代的值，<>不要出现)。其他的不用更改。

!!! 最好加一个自己的 ip 地址，端口号写 1080，8080，8888 都可以!!!

代理 ip 网址：https://www.xicidaili.com/ ，最好选新的

```
def _http_get(self, url, cookies=COOKIES):
    """
        发送get请求，程序中所有的get都是调这个方法
        如果想使用多cookies抓取，和请求重试功能
        在这自己添加
    """
    i = random.randint(0,len(cookies)-1)
    #headers['user-agent']= "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_7_0) AppleWebKit/535.11 (KHTML, like Gecko) Chrome/17.0.963.56 Safari/535.11"#可有可无
    headers['Cookie'] = cookies[i]
    proxy_list = [
        {"https": "http://127.0.0.1:1080"},
        {"https":"http://223.198.128.105:1080"},
        {"https": "http://124.93.201.59:8888"}
    ]
    proxy = random.choice(proxy_list)  # 随机选择一个ip地址
    response = requests.get(url, headers=headers,proxies=proxy, timeout=5)
    #response = requests.get(url, headers=headers,timeout=5)
    if response.status_code != 200:
        raise requests.Timeout
    return response.text
```

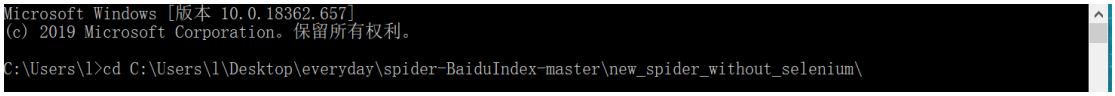
1.3 更改关键词

打开 demo.py，将孔子更改为关键词。支持多关键词，但是!! 最好不要使用，被反爬了就得不偿失了

```
from get_index import BaiduIndex
import pandas as pd
import csv
import time
import random
if __name__ == "__main__":
    keywords = ['孔子']

    PROVINCE_CODE = {'山东': '901', '贵州': '902', '江西': '903', '重庆': '904', '内蒙古': '905', '湖北': '906', '辽宁': '907', '湖南': '908', '福建': '909', '上海': '910', '北京': '911', '广西': '912'

    years = ['2019']

    result = []
    yearIndex = 0;
    for yearIndex in range(len(years)):
        for keys,value in PROVINCE_CODE.items():
            startDay = years[yearIndex]+'-01-01'
            endDay = years[yearIndex]+'-12-31'
            index = -1
            for j in range(10):#处理0值
                if j%2 == 0:
                    #虚买一枪
                    tempDay = random.randint(int(years[yearIndex]),2019)
                    tempDay = str(tempDay)+'-12-31'
                    baidu_index = BaiduIndex(keywords, startDay, tempDay, value)
                    baidu_index.get_index(startDay)
                else:
                    #正式代码
                    baidu_index = BaiduIndex(keywords,startDay, endDay,value)
                    index = baidu_index.get_index(startDay)
                    #正式代码
                    if (index[0] != 0):
                        break


            temp = [];
            temp.append(index[0]);
            index = temp;
            index.append(keys)
            index.append(years[yearIndex])
            result.append(index)
            print(result)
        pd.DataFrame(result).to_csv('孔子.csv', encoding='gbk',mode='a')
        print("{0} in  is successfully saved" .format(years[yearIndex]))
        time.sleep(60)#避免反爬，暂停1min
```

1.4 运行脚本

打开命令行窗口，输入【cd <demo.py 所在的路径>】，实例：



（注意，<>不要输入，不要到 demo.py，只需到达 demo.py 的上一级）

【回车】，输入【python demo.py】，坐等运行结束

2 注意事项

2.1 运行过程会打印“2011 is successful saved”，这是为了告诉你程序 keep working，不用管.

2.2 每次更改关键词，建议更改 ip，更改 ip 只需要断开网络，重连。可以百度搜索【ip 地址】看是否改变。一定要更改 cookies，如何更改见上。这是为了反反爬

2.3 禁止在一台电脑上运行多个进程，会遭受猛烈的反爬攻击。但是可以在多台电脑上运行，多台电脑运行一定要选择不同的 cookies。

2.4 更改代码建议使用正规的代码编辑器，由于 python 是缩进敏感的语言，使用 notepad 修改可能会破坏缩进结构。（代码里所有的缩进都是 TAB 键）

2.5 有问题随时增加