关于爬虫脚本的使用说明

第一次使用时:

1、确认已安装Python 3。

2、确认已安装谷歌浏览器Chrome。（目前支持版本为v75）

范例:

1、在 https://www.python.org/downloads/下载系统对应的最新版的Python 3安装包（本例使用的是 Windows 64-bit 下的 Python 3.7.3。）

第一步中，需选择将Python路径加入OS的环境变量path。其他选项都可留在默认设定，点击加一步及安装。安装完毕后，可打开cmd，输入python，确定是否已安装。如果确认返回了最新Python 3版本号（3.7.3）等信息，则输入quit() 退出Python环境。

2、安装Google Chrome。（直到今日2019/06/13，最新可下载版本为v75。文件夹内已包括了对应v75版本的chromedriver驱动，使用未来版本的用户将需要上指定网站下载此驱动。（链接在底下的注意事项中）

3、在cmd中进入webspider目录。（cd C:\……\webspider）

4、cmd中输入python setup.py install，等待进程结束。

5、cmd中输入go （作用是进入脚本主函数以执行），不换行直接打空格之后输入所需参数，分别为：保存路径 chromedriver路径 搜索页数 搜索关键词(可加不同限定) 例如：

go C:\Desktop\result\ .\chromedriver 3 PCS9000系统 site:wenku.baidu.com

6、运行结束后，在保存文件夹里打开刚刚生成的csv文件。需要注意，使用excel开启csv时需创建新空白工作簿，点击最上方菜单栏当中的数据按钮。点击“自文本” 引入csv文件，并把编码格式改为UTF-8。点击下一步，以逗号分隔文件。

注意:

1、一台机器上，范例中的step1-step3只要做一遍。以后每次只需要进到py脚本所在路径下，执行step4-6即可。

2、目前所提供的C:\webspider\chromedriver.exe驱动可能只支持75版本的Chrome浏览器。如果浏览器版本号继续升级，可能需要到https://selenium-python.readthedocs.io/installation.html#drivers下载对应链接。由于谷歌浏览器官方驱动在google.com域名下，因此下载时可能需要打开VPN。 非官方免翻墙驱动下载地址: http://chromedriver.storage.googleapis.com/index.html

3、由于百度搜索引擎限制，在不同指定网站下搜索关键词时，必须重新启动此程序。不支持一次搜素多个指定网站下的结果。

4. cmd传入的一些参数如果有一些特殊字符（如：括号），可能会导致程序报语法错误（syntax error）。在这种情况下只需要在该关键字前后加双引号即可。比如：PCS9000系统 “-(操作面板)” “-(测试百科网)”

5. 本程序的特性是对于当前文件的保存路径进行遍历，如果发现有相同关键词相同限定的搜索记录，则在本次搜索结果中不写入相同项的搜索结果。例如，在当次搜索中，关键词为pcs9000，限定为site:wenku.baidu.com，保存路径为C:\Desktop\result\，则对桌面上的result文件夹进行遍历，若发现其中有同样为pcs9000 site:wenku.baidu.com搜索后

保存的文件，则当次搜索后保存的结果不显示和前次相同的项目。若想在此次搜索中保留重复项，则可变更保存路径，并保证该路径下没有和本次搜索关键词相同的结果文件即可。

6. 按照本说明中安装python setup.py install时，最后执行到这两步，要是干等结果是等不到正确结束，会报错。错误原因是访问python的第三方库官网时间过长导致下载失败。必须网页浏览器手动下载这两个文件到安装包的\webspiderv2.egg-info下才可以。

Downloading https://files.pythonhosted.org/packages/51/bd/23c926cd341ea6b7dd0b2a00aba99ae0f828be89d72b2190f27c11d4b7fb/requests-2.22.0-py2.py3-none-any.whl#sha256=9cf5292fcd0f598c671cfc1e0d7d1a7f13bb8085e9a590f48c010551dc6c4b31

Downloading https://files.pythonhosted.org/packages/c6/22/a43126b87020c325fac159bb3b7f4e7ea99e7b2594ce5b8fa23cfa6ee90d/lxml-4.3.4-cp37-cp37m-win_amd64.whl#sha256=dd9f0e531a049d8b35ec5e6c68a37f1ba6ec3a591415e6804cbdf652793d15d7

执行时cmd终端报错信息： error: Download error for https://files.pythonhosted.org/packages/51/bd/23c926cd341ea6b7dd0b2a00aba99ae0f828be89d72b2190f27c11d4b7fb/requests-2.22.0-py2.py3-none-any.whl#sha256=9cf5292fcd0f598c671cfc1e0d7d1a7f13bb8085e9a590f48c010551dc6c

4b31: [WinError 10060] 由于连接方在一段时间后没有正确答复或连接的主机没有反应，连接尝试失败。