# <BaiduSearchSpider>开发文档

版本修订日志

|  |  |  |
| --- | --- | --- |
| 修订时间 | 修订人 | 修订内容 |
| 2016年10月09日 | 杨恬 | 创建 |
|  |  |  |
|  |  |  |
|  |  |  |

## 1．1 BaiduSearchSpider 百度搜索词条数爬取程序

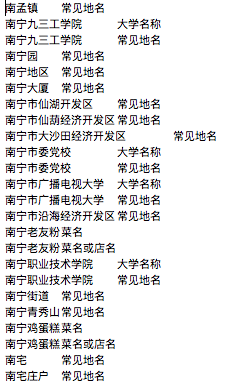
1. 程序编号： BaiduSearchSpider
2. 程序名称：百度搜素词条数爬取程序
3. 程序说明：本程序为百度搜索爬取程序，主要涉及三个搜索的方向：百度搜索，百度视频搜索以及百度新闻搜索。目前是为爬取某词的百度词条数量以判断该词权重功能研发，主要研发思路：先获取三种搜索方式结果页的网址模式，然后根据该模式根据拼出百度搜索某词条后的网址，获取网页源代码，然后提取其中的搜索词条数。包含的程序模块如下：

模块A：从文件中提取关键词并加以处理

模块B：将关键词作为搜索关键字获取百度结果页代码并提取搜索词条数

1. 程序功能可视化说明：

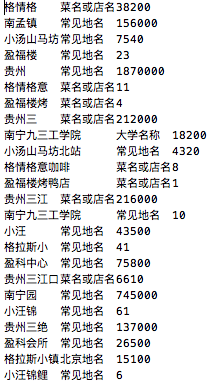
(1) 提取百度元搜索词条数。对于以下的输入文件：



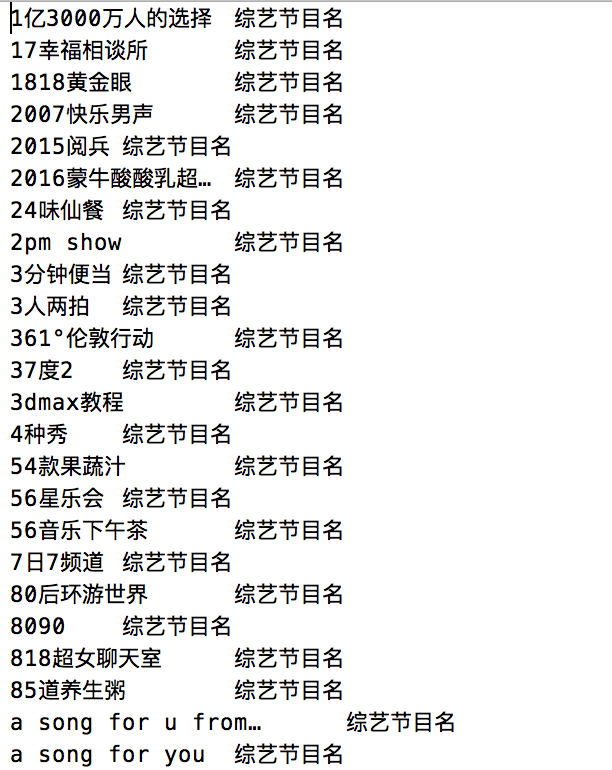
程序访问以下页面：



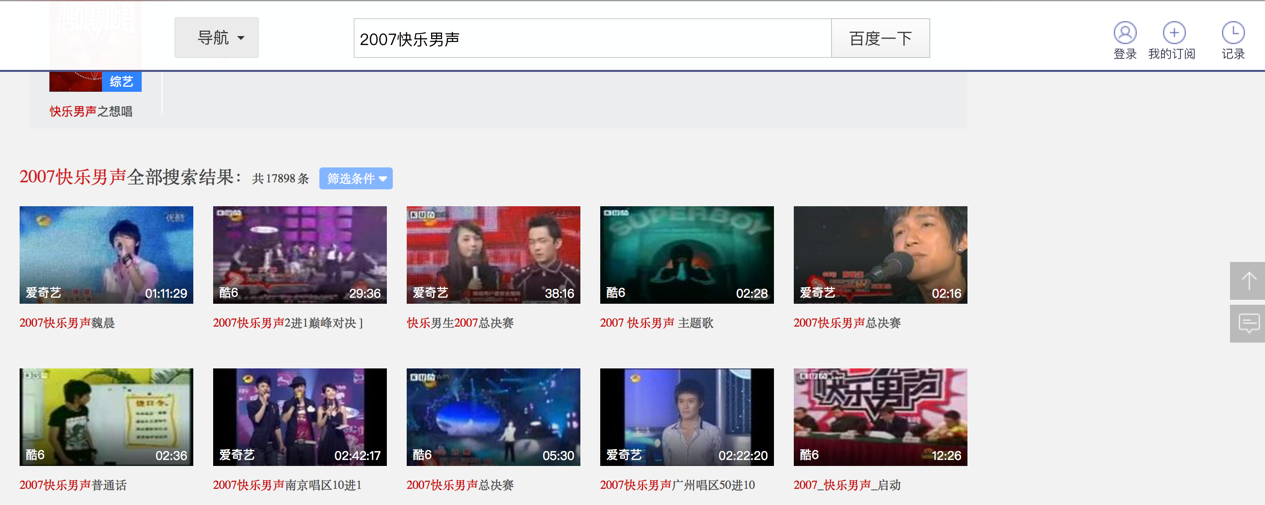
提取出其中相关结果的数量，得到如下的输出文件：



(2) 提取百度视频词条数。对于如下格式的输入文件，程序可以读取其中的节目名称，将该名称作为百度视频搜索的关键词，提取搜索到的词条数量。

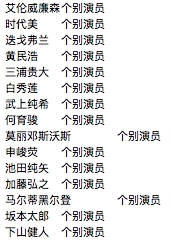


目标数据如下图：



本程序提取网页中全部搜索结果的数量，并且按以下形式

(3) 提取百度新闻词条数。对于以下格式的输出文件：



程序访问如下页面：



并得到输出文件格式如下：



1. 程序输入与输出说明：

参考如上，程序读入格式如下图的输入文件：

keyword nature

输出如下格式的文件：

keyword nature number

1. 处理过程流程图：

程序操作流程图：

../../../未命名.pdf

程序逻辑流程图：

../../未命名.pdf

1. 开放接口（若被调用，需要提供开放的接口说明）：

暂无

1. 引入的包及接口：

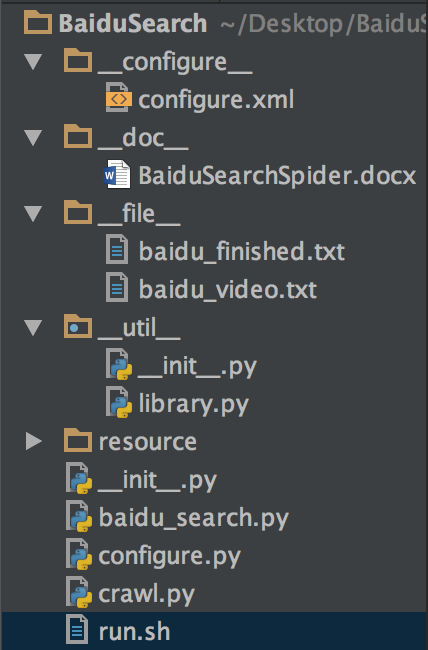
requests : 简单易用的网络访问模块

selenium : 用于模拟浏览器行为以获取合法cookie

scrapy : 爬虫封装模块，借用其中的Selector类定位元素

phantomjs: 伪浏览器，可用sudo apt-get install phantomjs安装

1. 程序源码目录结构说明：



\_\_configure\_\_: 包含配置文件configure.xml

\_\_doc\_\_: 包含程序文档

\_\_file\_\_: 存放程序输入输出文件

\_\_util\_\_: 存放一些经常被调用的常用函数

\_\_resource\_\_: 存放模拟浏览器所需的phantomjs执行文件

baidu\_search.py: 程序执行的主文件，入口

configure.py: 读取配置文件参数

crawl.py : 主要爬虫代码

run.sh: 执行入口

1. 程序启动方式说明

配置好输入输出文件，进入程序目录，执行: chmod +x ./run.sh;./run.sh即可

1. 程序配置文件及说明

配置文件：configure.xml

主要配置参数：

crawl\_type: 分为：s, v, n。分别对应百度搜索，百度视频搜索，百度新闻搜索

thread\_control: 抓取线程数量控制

input\_file: 输入文件目录

output\_file : 输出文件目录

base\_url : 抓取时所需的入口网址（若无变化无需修改）

pattern: 抓取数据的正则表达式（若无变化无需修改）

1. 部署程序打包方式：

将文件夹整个放置于服务器即可

1. 程序部署说明：

执行之前需要先安装selenium库以及scrappy库。需安装phantomjs环境。

1. 程序潜在的问题，推荐解决方案及其他特殊说明

(1) 线程数量控制问题

推荐最大不超过5个线程，对服务器压力最小，速度也不慢，抓取10000以内的数据3个线程即可。

(2) Phantomjs使用问题

因为平台的复杂性，在执行之前需要确保安装了phantomjs程序。在Mac系统上可以用brew install phantomjs命令安装。Ubuntu上可以用sudo apt-get install phantomjs命令安装。

(3) 输入输出路径问题

为了防止产生文件找不到的错误，建议将输入输出文件都放置于\_\_file\_\_目录下，配置时以./\_\_file\_\_/开头。

1. 关键技术说明

基于百度会封禁爬取过快的爬虫程序，本程序提出了一个解决方案。在尝试过程中发现百度是基于cookie封禁爬虫的，若是在一个cookie被封禁后重新尝试一个合法的cookie就可以重新进行爬取。因此程序中添加了一个通过模拟浏览器phantomjs获取合法cookie的功能（get\_cookie函数）。当判断爬虫被封禁之后重新获取cookie，以达到继续爬取的目的，提高爬取速度。

1. 版本号和CHANGES

BaiduSearchSpider 1.0