# 专利爬虫项目说明

2020年02月06日，作者：Kelvin He

## 1. 项目背景

本项目，旨在通过软件来爬取中国知识产权局的专利信息，并将爬取后的信息写入到Excel中。

其意义在于可以将大量网页内容整齐地写入到一个文档中。提高效率，减轻手工量。

## 2. 软件相关

### 2.1 编程语言

使用MacBook Pro + OSX 15开发。

采用Python开发；

IDE为Pycharm。

### 2.2 浏览器

主要使用了selenium + firefox浏览器。

### 2.3 其他工具

Anaconda：库管理器。

Selenium：requests库无法解决登录问题，cookies一直不对，返回412.

Pandas：方便处理字典类型数据，方便写入Excel中。

## 文件介绍

### 3.1 firefoxSelenium.py

项目入口函数。

主要配置都在这个文件中。如用户名和密码，登录网址，以及要爬取多少页的数据。

可以指定爬取1~x页的数据，当x=0时，爬取所有网页数据。

在终端中运行该文件后，会打开Firefox浏览器，然后人工识别验证码，输入验证码后，按下回车键或点击“登录”按钮。

登录后，搜索结果。可以使用普通搜索，也可以可以高级搜索。当搜索出结果后，回到终端，按下任意键继续。

之后，就开始自动爬取数据了。

### 3.2 parseHtml.py

该文件，主要是对网页数据进行解析。包括申请日、申请号、申请人等。这些信息都是根据网页的HTML内容进行高度定制的，如果网页排版发生变化，这些内容可能也需要做调整。

### 3.3 excelDeal.py

内容较少，只是对爬取到的Excel文档进行整理，调整格式等。

## 版本变更说明

|  |  |  |  |
| --- | --- | --- | --- |
| 序号 | 版本号 | 日期 | 变更内容 |
| 1 | V0.1 | 2020-02-06 | 首次发布，实现基本功能。 |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |