項目功能:利用代理IP網站的IP爬取鏈家二手房信息，並存到MongoDB數據庫中

使用工具:

文檔編寫:word

代碼管理工具:git

編譯工具: pycharm sublime or Anaconda, Spyder, Fiddler

Chrome 插件:Proxy Switch Omega(代理切換插件),Xpath Helper:網頁數據解析插件，JSON view

資料庫:MongoDB

適用作業系統:window 10, Ubuntu Linux

爬取數據步驟：

1.確定需要爬取URL地址

2.通過HTTP/HTTPS協議來獲取相應HTML頁面

3.提取HTML頁面有用數據

4.寫正則表達式

5.定義類，寫程序框架

6.補全代碼

Crawl操作說明:

1. 代理ip位置:proxies = {"http": "http://309435365:szayclhp@123.206.119.108:16817"}
2. 爬取北京小區名稱及對應價格
3. Url: "https://bj.lianjia.com/ershoufang/pg"
4. 正則表達式:

'<div class="houseInfo".\*?data-el="region">(.\*?)</a>.\*?

<div class="totalPrice">.\*?<span>(.\*?)</span>(.\*?)</div>’

1. 用類結構寫此次代碼框架
2. 圖片說明如下:

|  |  |
| --- | --- |
|  | 先確定需要爬取URL地址及內容 |
|  | 確定url的頁數的起始號碼 |
|  | 確認要爬取的內容，並用正則表達式篩選出來，  左圖以”houseInfo”為基準，取出region內容 |
|  | 左圖以”unitPrice”為基準取出房區名的價格資訊 |
|  | 啟動程式後按”y”,開始爬取，按”q”結束程式。 |
|  | 進入mongodb指令列，並使用指定儲存的資料庫名。 |
|  | 表列出爬取的數據無誤對照鏈家網的二手房第一頁。 |
|  |  |
|  |  |
|  |  |