Scrapy是python的一個抓取網頁和萃取資料的應用程式框架，當使用該框架進行開發時，會將專案分成Scrapy所設定好的資料夾和檔案進行建立，主要由spider.py的檔案進行網頁的抓取及資料的收集，再透過Scrapy的Engine進行一連串的處理，將整個Scrapy運行起來，由於專案已經結構化，所以在專案上的合作開發或維護上顯得更加容易。

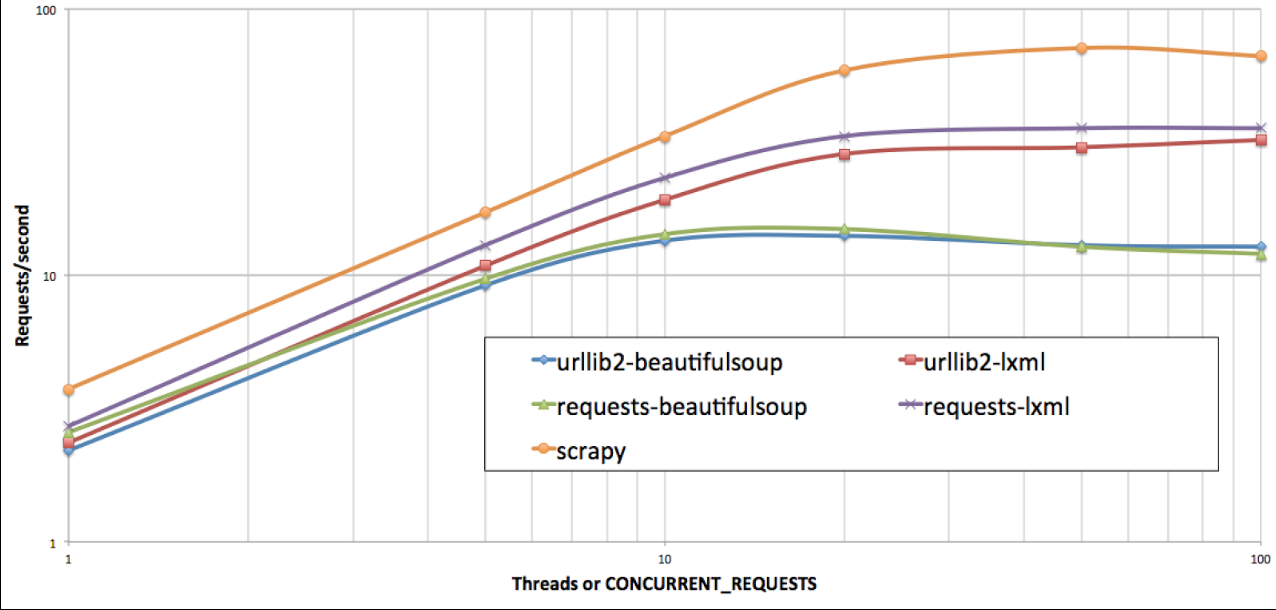
Scrapy本身擁有selectors去讀取DOM樹的結構化資料，不須再搭配其他的函式庫來幫助Scrpay抓取資料，也由於該框架的設計是採用非同步式設計，不須一個一個等待抓取的網頁回傳成功後才接續抓取，此設計是Scrapy最大的優點，在網頁抓取所花費時間遠低於其他python所提供的網頁抓取函式庫，另外，Scrapy所抓取到的資料可以用很簡易的指令去產生檔案，如：JSON、XML和CSV檔，也不須再搭配其他函式庫就可以完成該項工作。

urllib2是python的一個抓取網頁和萃取資料的函式庫，提供了許多簡易的函式供開發者使用，但缺點是少了selectors去抓取DOM樹的資料，必須搭配其他的函式庫來抓取DOM樹的資料，例如：BeautifulSoup4或lxml，urllib2是基於同步式設計，所以每抓取一個網頁時，必須等待上一網頁抓取完才能抓取下一個網頁，在抓取很多的網頁下時間的花費必須等待很長的時間才能夠完成。

比較：

工具

|  |  |  |
| --- | --- | --- |
| 項目 | Scrapy | urllib2 |
| 主要應用類型 | 應用程式框架 | 函式庫 |
| selectors支援 | 有 | 無 |
| 檔案輸出 | 有 | 無 |
| 非同步式設計 | 是 | 否 |
| 撰寫風格 | 像在寫設定檔 | 一般程式撰寫 |
| 團隊開發 | 較適合 | 較不適合 |
| 抓取網頁速度 | 快 | 慢 |



參考資料：

1. <https://doc.scrapy.org/en/1.2/>

2. <https://docs.python.org/2/library/urllib2.html>

3. <http://www.scrapinginsider.com/2016/01/scrapy-urllib2-requests-beautifulsoup-lxml.html>