Google Play App 信息爬蟲

I. 程式語言:

Python 2.7

II. 使用 module:

1. Pandas
   (http://pandas.pydata.org/)
2. Requests
   (http://docs.python-requests.org/en/master/)
3. BeautifulSoup
   (https://www.crummy.com/software/BeautifulSoup/bs4/doc/)
4. Threading
   (https://docs.python.org/2/library/threading.html)

III. 抓取的 Key 值:

['ClkUrl', 'Thumbnails', 'mail', 'Title', 'similarApps', 'comp', 'price', 'Gp Category', 'datePublished', 'Recommenders', 'genre2', 'fileSize', 'LogoUrl', 'genre', 'GpTag', 'inAppMsg', 'numDownloads', 'Rating', 'Lan', 'PostId', 'PkgName', 'ads', 'AppType', 'Reviewers', 'contentRating', 'name', 'softwareVersion', 'score', 'ratingHistogram', 'ratingValue', 'Country', 'whatsNew', 'GpCategory2', 'Desc', 'operatingSystems']

IV. 程式 method 說明：

1. GetLangList(id)
   Id: 該 App 的包名
   從該 App 的 html 中抓出所有支援語言的 URL 代碼。
2. app_detail(id,lang)
   Id: 該 App 的包名
   Lang: 語言的 URL 代碼
   透過 App id 與支援語言的 URL 代碼抓取該 App 的所有資訊，資訊為上述的 Key 值。
3. CrawlPackage(fileindex1=0,fileindex2=99,save_file=False)
   fileindex1: 儲存包名文檔的起始檔案代碼
   fileindex1: 儲存包名文檔的最終檔案代碼
   save_file: 是否將抓下來的資訊存成 Json 格式
   由於全球 App 量太大，只根據 file 內的 App ID 抓取特定 App。

4. CrawlByPkgName(pkgname='com.facebook.katana',save_file=False)
   Pkgname: 抓取 App 的包名
   save_file: 是否將抓下來的資訊存成 Json 格式
   抓取單一 App 的所有語言版本資訊。
5. MultiThreadCrawlPackage(fileindex1=0,fileindex2=99,n_job=3)
   fileindex1: 儲存包名文檔的起始檔案代碼
   fileindex1: 儲存包名文檔的最終檔案代碼
   n_job: 啟用幾個執行緒
   多執行緒的執行 CrawlPackage。