Google Play App信息爬蟲

1. 程式語言:

Python 2.7

1. 使用module:
2. Pandas

(<http://pandas.pydata.org/> )

1. Requests

(<http://docs.python-requests.org/en/master/> )

1. BeautifulSoup

(<https://www.crummy.com/software/BeautifulSoup/bs4/doc/> )

1. Threading

(<https://docs.python.org/2/library/threading.html> )

1. 抓取的Key值:

['ClkUrl', 'Thumbnails', 'mail', 'Title', 'similarApps', 'comp', 'price', 'Gp

Category', 'datePublished', 'Recommenders', 'genre2', 'fileSize', 'LogoUrl', 'ge

nre', 'GpTag', 'inAppMsg', 'numDownloads', 'Rating', 'Lan', 'PostId', 'PkgName',

'ads', 'AppType', 'Reviewers', 'contentRating', 'name', 'softwareVersion', 'sco

re', 'ratingHistogram', 'ratingValue', 'Country', 'whatsNew', 'GpCategory2', 'De

sc', 'operatingSystems']

1. 程式method說明：
2. 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。

1. 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。

1. 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。

1. 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。

1. 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。