Web Crawling on DouBan Movie

Yang, J.

1 Introduction

THE web crawling is a automated process which browses web in a methodical, automated manner. It analyzes the web, then extracts the target information which people need. In this task, I will write a script to extract all posters related with "Cheng Long" in the website "DouBan Moive" and analyze the web crawling procedure.

2 Analyze the website

First, go to DouBan movie, https://movie.douban.com/;

Then search "Cheng Long", return the web https://search.douban.com/movie/subject_search?search_text=成龙&cat=1002;

Next inspect the source and find the location of the posters. Specifically, find the posters by from locating the outer div classes to inner div classes, until find the img tag. The class name of the img tag is "cover".

Last, the number of pages may be not only 1.

3 Algorithm implemented

3.1 Multi-thread

I create a main thread to produce movie which contains the url and the name of the movie, besides, 5 threads to download posters

3.2 Mock login

By using selenium and google driver, I can mock login on DouBan Movie.

3.3 Download the poster

Check whether the poster has already been saved, if not, download this poster by using urlretrieve.

3.4 Next information page

In main function, I create a loop to iterate over up to 50, if there are related information in the current source, I download thoses posters, if not, break the loop and terminate main thread and other threads.

4 Problem and Resolving Method

During the implementation, I encounterred much problem

1. By using requests.get(), I cannot find the movies' information on the return page.
Resolve: selenium and google driver

2. Download is fast, IP may be blocked
Resolve: add sleep()

3. KeyError
Resolve: I add wrong elements into the queue

4. The downloaded posters are not jpg, but file.
Resolve: add '.jpg' as the suffix

5. Sometimes raise error 'FileNotFound' when downloading.
Resolve: Becausee some posters' name contain '/', I simply use string.replace() to remove '/'

6. Download speed is so low.
Resolve: multi-thread

7. Thread never stop

4 Procedure of Web Crawling

1. Start from one intial URL (first http request): https://search.douban.com/movie/subject_search?search_text=%E6%88%90%E9%BE%99&cat=1002&start=0

2. By using Requests.get() to get the resource of the web. There are two usually used methods for Requests, one is requests.get() which sent request to a web, and return an object along with its status code. There are two parameters in get() method, one is URL, the other one is headers, sometimes different websites have different requirements on headers, hence need to change headers when needed. Requests.post() is used for writing data, submitting data to be processed, specifically the functionalities of get() and post() are similar, but post() is safer because data sent is part of the URL when using get(). After I get the response object, I can use response.text to get the plain text of the object.

3. Using json.loads() to transfer response.text to dict, and process the python objects(dict) later.

4. Then there are several ways to store the data: 1. directly write data into a file; 2. json.dumps; 3. store them into csv files

5. The next URL is simply each time increasing the parameter after 'start=' by 15. Continue to crawl next URL until the related data of return page is None.