from lxml import etree

import requests

import re

import pymysql

import time

headers = {

'User-Agent':'Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/49.0.2623.112 Safari/537.36'

}

conn = pymysql.connect(host='140.131.115.87',user='root',passwd='109504109504',db='testdb',port=3306,charset='utf8')

cursor = conn.cursor()

def get\_movie\_url(url):

html = requests.get(url,headers=headers)

selector = etree.HTML(html.text)

movie\_hrefs = selector.xpath('//div[@class="hd"]/a/@href')

for movie\_href in movie\_hrefs:

get\_movie\_info(movie\_href)

def get\_movie\_info(url):

html = requests.get(url,headers=headers)

selector = etree.HTML(html.text)

try:

video\_name = selector.xpath('//\*[@id="content"]/h1/span[1]/text()')[0]

type = re.findall('<span property="v:genre">(.\*?)</span>',html.text,re.S)[0]

time = re.findall('上映日期:</span>.\*?>(.\*?)</span>',html.text,re.S)[0]

cursor.execute(

"insert into doubanmovie (video\_name,type,type) values (%s,%s,%s)",

(str(video\_name),str(type),str(time))

)

except IndexError:

pass

if \_\_name\_\_ =='\_\_main\_\_':

urls = ['https://movie.douban.com/top250?start={}&filter='.format(str(i)) for i in range(0,250,25)]

for url in urls:

get\_movie\_url(url)

time.sleep(2)

conn.commit()

conn.close()