First attempt to write a spider

This program is going to get the data of the toppest 250 movies from douban

Ask a request

In this section, we are going to ask a request to the url

In [3]:

```
import requests
```

In [4]:

```
url = 'https://movie.douban.com/top250'
response = requests.get(url)
print(response)
```

```
<Response [418]>
```

> Response means request is denied, because there are anti-spider method to solve this, adding user-agent

Asking a request by add a user-agent

> the user-agent is copied from the chrome

In [5]:

```
header = {
    'User-Agent':
    'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) C
hrome/80.0.3987.132 Safari/537.369'
}
response = requests.get(url, headers=header)
print(response.status_code)
```

```
200
```

Save the html file

In [7]:

```
file = open('I:\Programming\Python\python-workspace\Workspace for Python\studying file
\Spider\douban_top_250_movies.html', 'w', encoding='utf-8')
file.write(response.text)
file.close
html = response.text
```

Analyze the html file

> through beautifulSoup

In [10]:

```
from bs4 import BeautifulSoup
```

In [11]:

```
soup = BeautifulSoup(html, 'html.parser')
```

'html.parse' is html analyzer. there are some method could be used:

1. find(condition) > find the first satified the condition
2. find_all(condition) > find all
3. select(css_selector) > find css_selector

In [38]:

```
movie_list = soup.find('ol', class_='grid_view') # attrs = ['class': 'grid_view']
```

Find information for each movie

In [13]:

```
movies = movie_list.find_all('li')
print(type(movie_list))
```

```
<class 'bs4.element.Tag'>
```

In [15]:

```
names = []

for movie in movies:
    name = movie.find('span', class_='title').get_text()
    names.append(name)

print(names)
```

```
['肖申克的救赎', '霸王别姬', '阿甘正传', '这个杀手不太冷', '美丽人生', '泰坦尼克
号', '千与千寻', '辛德勒的名单', '盗梦空间', '忠犬八公的故事', '海上钢琴师',
'机器人总动员', '三傻大闹宝莱坞', '楚门的世界', '放牛班的春天', '星际穿越', '大
话西游之大圣娶亲', '熔炉', '疯狂动物城', '无间道', '龙猫', '教父', '当幸福来敲
门', '怦然心动', '触不可及']
```

In [27]:

```
e_names = []
c_names = []

for movie in movies:
    name = movie.find_all('span', class_='title')

    print(name[0].get_text())
    c_names.append(name[0].get_text())
    if len(name) > 1:
        e_names.append(name[1].get_text())
    else:
        e_names.append('')

print(e_names)
print(c_names)
```

```
肖申克的救赎
霸王别姬
阿甘正传
这个杀手不太冷
美丽人生
泰坦尼克号
千与千寻
辛德勒的名单
盗梦空间
忠犬八公的故事
海上钢琴师
机器人总动员
三傻大闹宝莱坞
楚门的世界
放牛班的春天
星际穿越
大话西游之大圣娶亲
熔炉
疯狂动物城
无间道
龙猫
教父
当幸福来敲门
怦然心动
触不可及
['\xa0/\xa0The Shawshank Redemption', '', '\xa0/\xa0Forrest Gump', '\xa0/
\xa0Léon', '\xa0/\xa0La vita è bella', '\xa0/\xa0Titanic', '\xa0/\xa0千と千
尋の神隠し', "\xa0/\xa0Schindler's List", '\xa0/\xa0Inception', "\xa0/\xa0H
achi: A Dog's Tale", "\xa0/\xa0La leggenda del pianista sull'oceano", '\xa
0/\xa0WALL·E', '\xa0/\xa03 Idiots', '\xa0/\xa0The Truman Show', '\xa0/\xa0
Les choristes', '\xa0/\xa0Interstellar', '\xa0/\xa0西遊記大結局之仙履奇緣',
'\xa0/\xa0도가니', '\xa0/\xa0Zootopia', '\xa0/\xa0無間道', '\xa0/\xa0となり
のトトロ', '\xa0/\xa0The Godfather', '\xa0/\xa0The Pursuit of Happyness',
'\xa0/\xa0Flipped', '\xa0/\xa0Intouchables']
['肖申克的救赎', '霸王别姬', '阿甘正传', '这个杀手不太冷', '美丽人生', '泰坦尼克
号', '千与千寻', '辛德勒的名单', '盗梦空间', '忠犬八公的故事', '海上钢琴师',
'机器人总动员', '三傻大闹宝莱坞', '楚门的世界', '放牛班的春天', '星际穿越', '大
话西游之大圣娶亲', '熔炉', '疯狂动物城', '无间道', '龙猫', '教父', '当幸福来敲
门', '怦然心动', '触不可及']
```

In [20]:

```
urls = []
for movie in movies:
    url = movie.find('a')['href']
    urls.append(url)

print(urls)
```

```
['https://movie.douban.com/subject/1292052/', 'https://movie.douban.com/su
bject/1291546/', 'https://movie.douban.com/subject/1292720/', 'https://mov
ie.douban.com/subject/1295644/', 'https://movie.douban.com/subject/129206
3/', 'https://movie.douban.com/subject/1292722/', 'https://movie.douban.co
m/subject/1291561/', 'https://movie.douban.com/subject/1295124/', 'http
s://movie.douban.com/subject/3541415/', 'https://movie.douban.com/subject/
3011091/', 'https://movie.douban.com/subject/1292001/', 'https://movie.dou
ban.com/subject/2131459/', 'https://movie.douban.com/subject/3793023/', 'h
ttps://movie.douban.com/subject/1292064/', 'https://movie.douban.com/subje
ct/1291549/', 'https://movie.douban.com/subject/1889243/', 'https://movie.
douban.com/subject/1292213/', 'https://movie.douban.com/subject/5912992/',
'https://movie.douban.com/subject/25662329/', 'https://movie.douban.com/su
bject/1307914/', 'https://movie.douban.com/subject/1291560/', 'https://mov
ie.douban.com/subject/1291841/', 'https://movie.douban.com/subject/184903
1/', 'https://movie.douban.com/subject/3319755/', 'https://movie.douban.co
m/subject/6786002/']
```

Save the data into file or database

In [36]:

```
with open('movies.csv', 'w', encoding = 'utf-8')as f:
    for i in range(len(urls)):
        f.write(c_names[i]+','+e_names[i]+','+urls[i]+'\n')
```