专栏：013：我要你知道实时票房.

E - 爬虫技术 爬虫知识

> **用理工科思维看待这个世界**

系列爬虫专栏

初学者，尽力实现最小化学习系统

主题：selenium + PhantomJS + sqlalchemy

selenium + PhantomJS 教程
SQLALchemy 教程

0：说明

实现编程学习的最小化系统。

使用`selenium + PhantonJS`获取网页源代码，此工具在异步加载处网页中很好用。之前使用的不多，觉得尝试使用此工具操作爬虫，目的是抓取中国票房首页的数据， 采用Xpath对数据进行解析。使用ORM技术实现自动创建数据表，并将数据存储入MySQL数据库中。

任务：抓取图示内容：

CBO实时票房榜

2016-5-10 周二 今日大盘 ： 5408.4万

	影片名称	实时票房（万）	票房占比	累计票房（万）	排片占比	上映天数	
1	美国队长3：英...	3840.7	71.01%	72254.1	53.35%	5	--
2	北京遇上西雅图...	908.0	16.79%	65960.6	19.99%	12	--
3	奇幻森林	246.4	4.56%	96568.8	8.97%	26	--
4	魔宫魅影	87.6	1.62%	8472.1	4.00%	12	↑1
5	梦想合伙人	69.7	1.29%	7912.8	2.90%	12	↓1
6	百鸟朝凤	69.2	1.28%	264.8	1.65%	5	--
7	妄想症	39.8	0.74%	144.6	1.21%	5	↑2
8	谁的青春不迷茫	33.2	0.61%	17847.5	1.57%	19	--
9	大唐玄奘	30.3	0.56%	3160.9	1.37%	12	↓2
10	判我有罪	25.2	0.47%	111.6	1.86%	5	--

1：任务分解

- 抓取网页源代码
- 对网页源代码进行解析，抓取需要的数据
- 数据结构化
- 创建数据表
- 将结构化数据存储入数据库中

技能需求：

- selenium 的基本使用
- unittest 的基本使用
- sqlalchemy的基本使用
- xpath语法的掌握
- MySQL数据基本知识

---

2. 实战

- selenium 使用：
  参考：点我试试
- xpath 的使用
  全部数据：`//div[@id="top_list"]/table/tbody/tr/td`

图示：

- sqlalchemy 的使用
    - 创建连接
    - 声明映射文件
    - 创建模式
    - 初始化映射类实例
    - 创建回话
    - 持久化实例对象

```
# 核心代码
engine = create_engine("mysql://root:123456@localhost:3306/test?charset=utf
8", echo = True) # 创建连接
Base = declarative_base()
metadata = MetaData(engine)
sql_table = Table("Realtime_film", metadata,
                  Column("id", Integer, primary_key=True),
                  Column("Rank", String(32)),
                  Column("Moviename", String(32)),
                  Column("Realtime", String(12)),
                  Column("Ratio_of_movie", String(16)),
                  Column("sum_movie", String(128)),
                  Column("Ration_of_open", String(128)),
                  Column("Screen_time", String(128)),
                    mysql_engine='InnoDB',
                    mysql_charset='utf8')  # 表声明，定义字段及类型
sql_table.create()  # 创建数据库表
sql_table_2 = Table("Realtime_film", metadata, autoload=True)
i = sql_table_2.insert()
# for one in Movie_datas:
#     i.execute(one)
con = engine.connect()
con.execute(i, Movie_datas)  # 插入全部数据
```

效果展示：
自动在本地数据库创建数据表，并把数据插入数据库中.（省去了编写了sql语句）

id	Rank	Moviename	Realtime	Ratio_of_movie	sum_movie	Ration_of_open	Screen_time
1	1	美国队长3：英...	4421.6	71.31%	72835.0	53.35%	5
2	2	北京遇上西雅图...	1062.0	17.13%	66114.5	19.99%	12
3	3	奇幻森林	264.1	4.26%	96586.5	8.97%	26
4	4	魔宫魅影	103.3	1.67%	8487.8	4.00%	12
5	5	梦想合伙人	75.9	1.22%	7919.0	2.90%	12
6	6	百鸟朝凤	72.1	1.16%	267.6	1.65%	5
7	7	妄想症	43.1	0.69%	147.9	1.21%	5
8	8	谁的青春不迷茫	36.7	0.59%	17851.0	1.57%	19
9	9	大唐玄奘	33.4	0.54%	3163.9	1.37%	12
10	10	判我有罪	26.7	0.43%	113.1	1.86%	5
(Auto)	(NULL)	(NULL)	(NULL)	(NULL)	(NULL)	(NULL)	(NULL)

完整版代码

3：总结

崇尚的思维是：
了解这是什么。
知道应该怎么做。
学会亲自动手。

最怕陷入学而不思则罔，思而不学则殆的地步

> 关于本人：
> 只有一个职业：学生
> 只有一个任务：学习
> 在这条路上，充满无尽的困境，我希望成为一个精神世界丰满的人。