

应用集成
2017年6月

应用集成 第二次作业报告

刘心蓓 141250078

曹姝玥 141250005

傅林华 141250036

刘璇琳 141250080

汤大业 141250124

1.项目简述

1.1. 项目主题

本次作业是应用集成课程的第二次作业，本小组在三个选题中选择了方案二：电影订购平台的信息集成。

1.2. 项目需求

项目要求实现基于互联网应用的数据集成，具体来讲，即分析已有的互联网应用数据异构的特征，实现多个同主题互联网应用的异构数据集成。

2.项目构建

2.1. 数据来源

本项目着眼于电影订购平台关于电影信息的不同进行多个网站的数据爬取和集成，主要选取了淘票票和格瓦拉作为数据爬取的来源，通过src/main/java/spider包中的爬虫类获取数据。

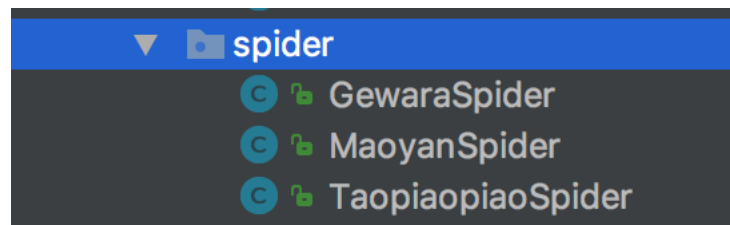
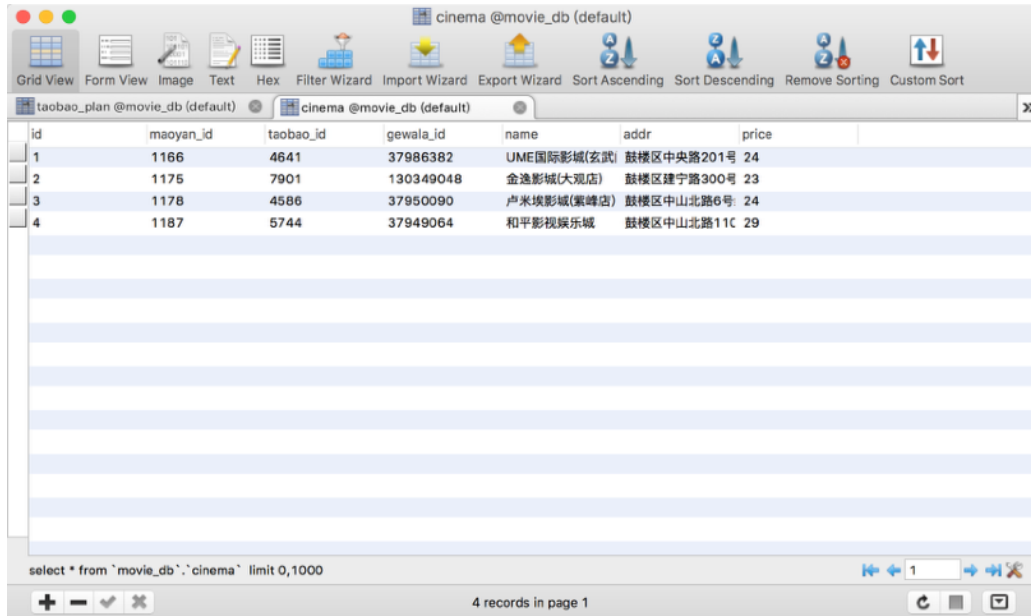


图1 工程文件src/main/java/spider包中的爬虫类

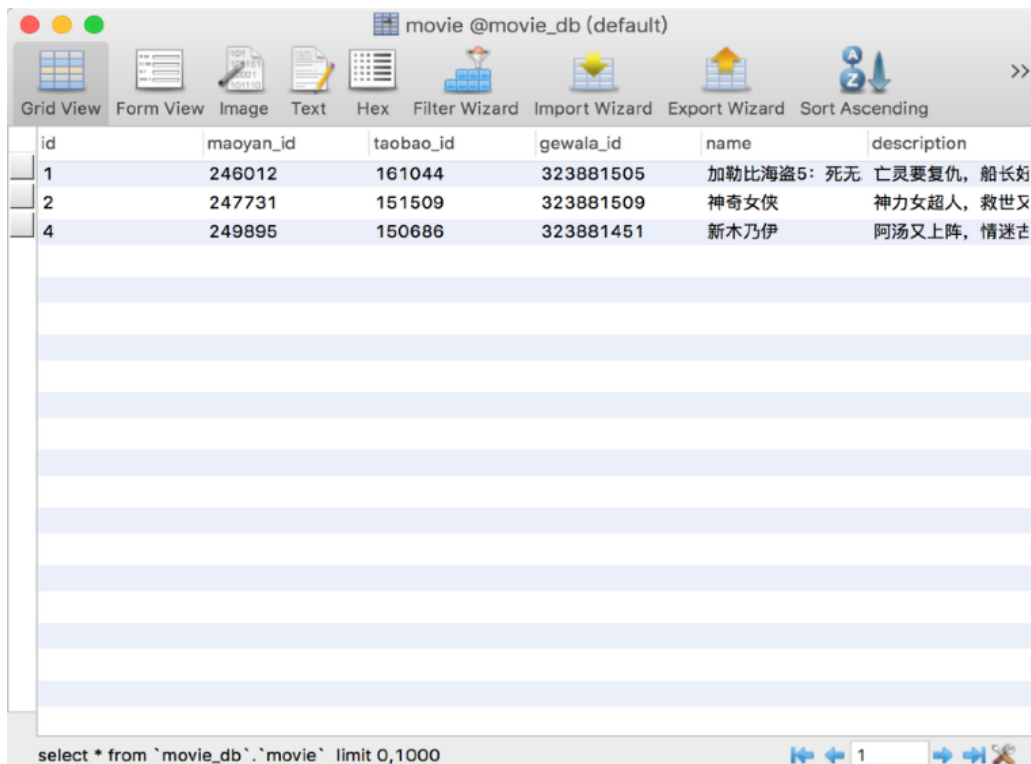
2.2. 数据处理

首先数据库中定义的cinema表和movie表，用于存放GewaraSpider和TaopiaopiaoSpider获取的数据。



id	maoyan_id	taobao_id	gewala_id	name	addr	price
1	1166	4641	37986382	UME国际影城(玄武)	鼓楼区中央路201号	24
2	1175	7901	130349048	金逸影城(大观店)	鼓楼区建宁路300号	23
3	1178	4586	37950090	卢米埃影城(紫峰店)	鼓楼区中山北路6号	24
4	1187	5744	37949064	和平影视娱乐城	鼓楼区中山北路11C	29

图2 数据库中的cinema表



id	maoyan_id	taobao_id	gewala_id	name	description
1	246012	161044	323881505	加勒比海盗5: 死无	亡灵要复仇, 船长好
2	247731	151509	323881509	神奇女侠	神力女超人, 救世又
4	249895	150686	323881451	新木乃伊	阿汤又上阵, 情迷古

图3 数据库中的movie表

获得不同的电影和影院在格瓦拉和淘票票网站对应的电影id和影院id，爬取最近几天的数据。

在工程文件src/main/java/entity中分别定义GewalaPlan和TaoppPlan，将数据注入，存放到数据库中的gewala_plan和taobaoPlan表中。

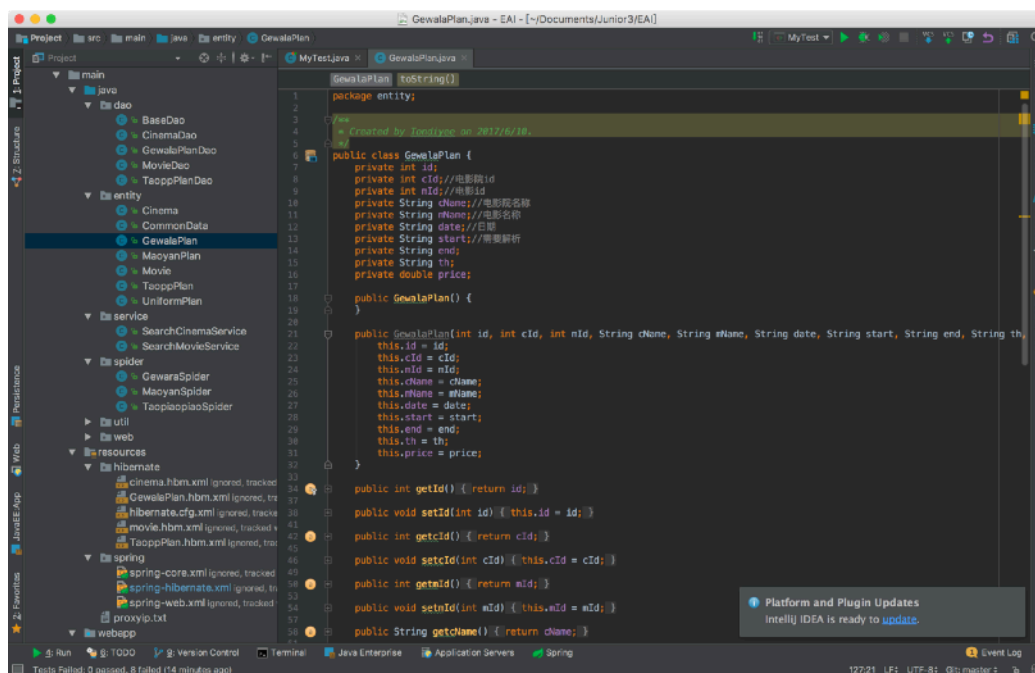


图4 工程文件src/main/java/entity中的GewalaPlan类

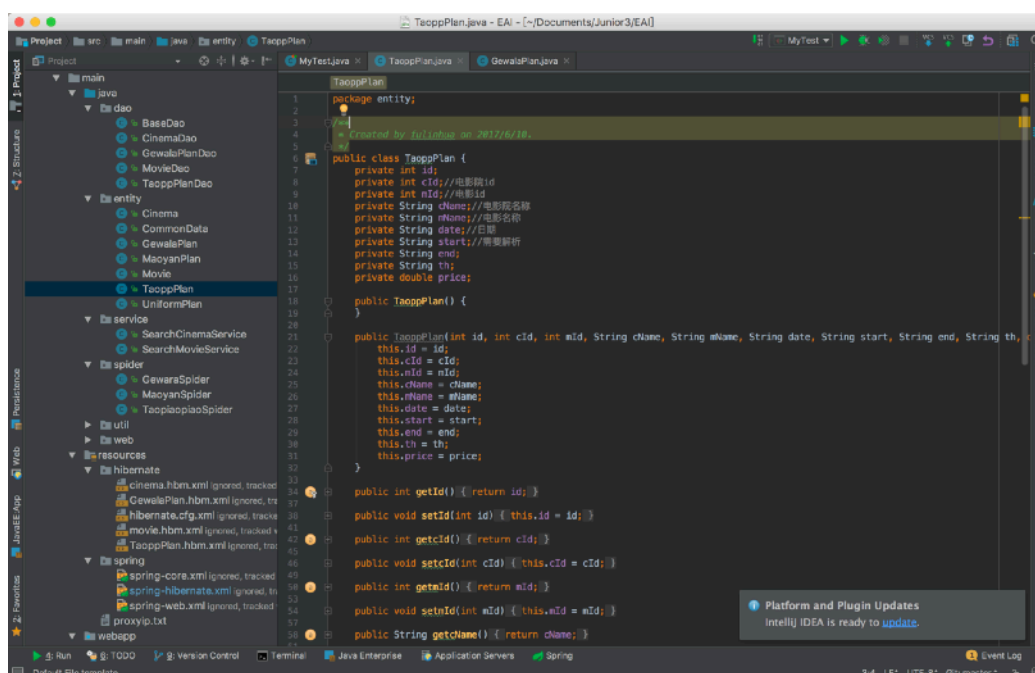


图5 工程文件src/main/java/entity中的TaoppPlan类

gewala_plan @movie_db (default)

Grid View Form View Image Text Hex Filter Wizard Import Wizard Export Wizard Sort Ascending Sort Descending Remove Sorting Custom Sort

id	m_id	c_id	m_name	c_name	start	end	th
1	323881505	37986382	加勒比海盗5：死无对证	UME国际影城(玄武门店)	10:30	预计12:39散场	7号4D厅
2	323881505	37986382	加勒比海盗5：死无对证	UME国际影城(玄武门店)	11:45	预计13:54散场	9号厅
3	323881505	37986382	加勒比海盗5：死无对证	UME国际影城(玄武门店)	13:20	预计15:29散场	11号厅
4	323881505	37986382	加勒比海盗5：死无对证	UME国际影城(玄武门店)	14:10	预计16:19散场	9号厅
5	323881505	37986382	加勒比海盗5：死无对证	UME国际影城(玄武门店)	15:45	预计17:54散场	11号厅
6	323881505	37986382	加勒比海盗5：死无对证	UME国际影城(玄武门店)	16:35	预计18:44散场	9号厅
7	323881505	37986382	加勒比海盗5：死无对证	UME国际影城(玄武门店)	19:00	预计21:09散场	9号厅
8	323881505	37986382	加勒比海盗5：死无对证	UME国际影城(玄武门店)	20:05	预计22:14散场	11号厅
9	323881505	37986382	加勒比海盗5：死无对证	UME国际影城(玄武门店)	21:25	预计23:34散场	9号厅
10	323881505	37986382	加勒比海盗5：死无对证	UME国际影城(玄武门店)	22:30	预计00:39散场	11号厅
11	323881509	37986382	神奇女侠	UME国际影城(玄武门店)	10:30	预计12:52散场	4号厅
12	323881509	37986382	神奇女侠	UME国际影城(玄武门店)	11:25	预计13:47散场	2号厅
13	323881509	37986382	神奇女侠	UME国际影城(玄武门店)	12:15	预计14:37散场	8号厅
14	323881509	37986382	神奇女侠	UME国际影城(玄武门店)	13:05	预计15:27散场	4号厅
15	323881509	37986382	神奇女侠	UME国际影城(玄武门店)	14:00	预计16:22散场	2号厅
16	323881509	37986382	神奇女侠	UME国际影城(玄武门店)	14:50	预计17:12散场	8号厅
17	323881509	37986382	神奇女侠	UME国际影城(玄武门店)	15:40	预计18:02散场	4号厅
18	323881509	37986382	神奇女侠	UME国际影城(玄武门店)	16:35	预计18:57散场	2号厅
19	323881509	37986382	神奇女侠	UME国际影城(玄武门店)	17:25	预计19:47散场	8号厅
20	323881509	37986382	神奇女侠	UME国际影城(玄武门店)	18:15	预计20:37散场	4号厅

select * from `movie_db`.`gewala_plan` limit 0,1000

163 records in page 1

图6 数据库表中的gewala_plan表

taobao_plan @movie_db (default)

Grid View Form View Image Text Hex Filter Wizard Import Wizard Export Wizard Sort Ascending Sort Descending Remove Sorting Custom Sort

id	m_id	c_id	m_name	c_name	start	end	th
1	161044	4641	加勒比海盗5：死无	UME国际影城(玄武门店)	10:30	10:30 预计12:39散场	4D厅
2	161044	4641	加勒比海盗5：死无	UME国际影城(玄武门店)	11:45	11:45 预计13:54散场	9号厅
3	161044	4641	加勒比海盗5：死无	UME国际影城(玄武门店)	13:20	13:20 预计15:29散场	11号厅
4	161044	4641	加勒比海盗5：死无	UME国际影城(玄武门店)	14:10	14:10 预计16:19散场	9号厅
5	161044	4641	加勒比海盗5：死无	UME国际影城(玄武门店)	15:45	15:45 预计17:54散场	11号厅
6	161044	4641	加勒比海盗5：死无	UME国际影城(玄武门店)	16:35	16:35 预计18:44散场	9号厅
7	161044	4641	加勒比海盗5：死无	UME国际影城(玄武门店)	19:00	19:00 预计21:09散场	9号厅
8	161044	4641	加勒比海盗5：死无	UME国际影城(玄武门店)	20:05	20:05 预计22:14散场	11号厅
9	161044	4641	加勒比海盗5：死无	UME国际影城(玄武门店)	21:25	21:25 预计23:34散场	9号厅
10	161044	4641	加勒比海盗5：死无	UME国际影城(玄武门店)	22:30	22:30 预计次日00:39散场	11号厅
11	151509	4641	神奇女侠	UME国际影城(玄武门店)	10:30	10:30 预计12:52散场	4号厅
12	151509	4641	神奇女侠	UME国际影城(玄武门店)	11:25	11:25 预计13:47散场	2号厅
13	151509	4641	神奇女侠	UME国际影城(玄武门店)	12:15	12:15 预计14:37散场	8号厅
14	151509	4641	神奇女侠	UME国际影城(玄武门店)	13:05	13:05 预计15:27散场	4号厅
15	151509	4641	神奇女侠	UME国际影城(玄武门店)	14:00	14:00 预计16:22散场	2号厅
16	151509	4641	神奇女侠	UME国际影城(玄武门店)	14:50	14:50 预计17:12散场	8号厅
17	151509	4641	神奇女侠	UME国际影城(玄武门店)	15:40	15:40 预计18:02散场	4号厅
18	151509	4641	神奇女侠	UME国际影城(玄武门店)	16:35	16:35 预计18:57散场	2号厅
19	151509	4641	神奇女侠	UME国际影城(玄武门店)	17:25	17:25 预计19:47散场	8号厅
20	151509	4641	神奇女侠	UME国际影城(玄武门店)	18:15	18:15 预计20:37散场	4号厅

select * from `movie_db`.`taobao_plan` limit 0,1000

166 records in page 1

图7 数据库表中的taobao_plan表

2.3. 数据集成与展示

本项目通过网页展示具体电影信息，包括片名、影院、场次、来源和价格。可通过搜索电影获得不同影院场次；也可通过搜索一家影院获得该影院所有电影的信息。为用户提供多个平台的数据，进行价格或位置等更具相对优势的选择。

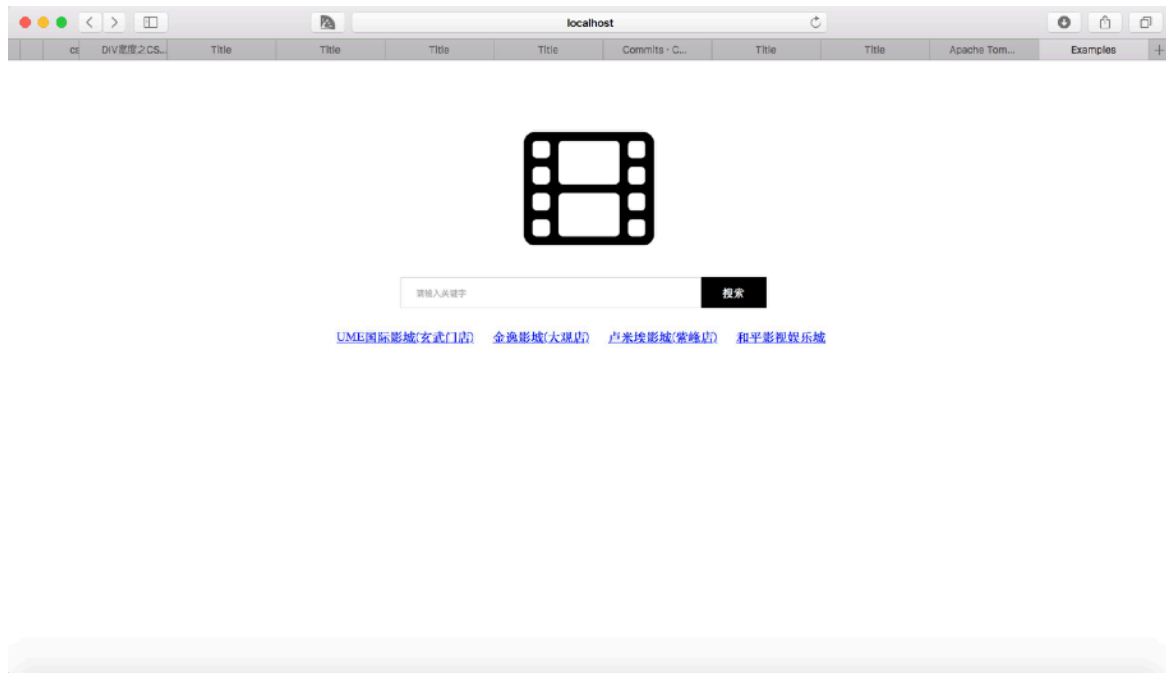


图8 搜索影院获得电影信息

UME国际影城(玄武门店)							
电影名称	放映日期	开始时间	结束时间	格瓦拉场次	格瓦拉价格	淘宝场次	淘宝价格
加勒比海盗5：死无对证	2017-06-12	10:30	预计12:39散场	7号4D厅	39.0	4D厅	39.5
加勒比海盗5：死无对证	2017-06-12	11:45	预计13:54散场	9号厅	29.0	9号厅	29.5
加勒比海盗5：死无对证	2017-06-12	13:20	预计15:29散场	11号厅	34.0	11号厅	34.5
加勒比海盗5：死无对证	2017-06-12	14:10	预计16:19散场	9号厅	34.0	9号厅	34.5
加勒比海盗5：死无对证	2017-06-12	15:45	预计17:54散场	11号厅	34.0	11号厅	34.5
加勒比海盗5：死无对证	2017-06-12	16:35	预计18:44散场	9号厅	34.0	9号厅	34.5
加勒比海盗5：死无对证	2017-06-12	19:00	预计21:09散场	9号厅	39.0	9号厅	39.5
加勒比海盗5：死无对证	2017-06-12	20:05	预计22:14散场	11号厅	39.0	11号厅	39.5
加勒比海盗5：死无对证	2017-06-12	21:25	预计23:34散场	9号厅	39.0	9号厅	39.5
加勒比海盗5：死无对证	2017-06-12	22:30	预计00:39散场	11号厅	39.0	11号厅	39.5
神奇女侠	2017-06-12	10:30	预计12:52散场	4号厅	29.0	4号厅	29.5
神奇女侠	2017-06-12	11:25	预计13:47散场	2号厅	29.0	2号厅	29.5
神奇女侠	2017-06-12	12:15	预计14:37散场	8号厅	29.0	8号厅	29.5
神奇女侠	2017-06-12	13:05	预计15:27散场	4号厅	34.0	4号厅	34.5
神奇女侠	2017-06-12	14:00	预计16:22散场	2号厅	34.0	2号厅	34.5
神奇女侠	2017-06-12	14:50	预计17:12散场	8号厅	34.0	8号厅	34.5
神奇女侠	2017-06-12	15:40	预计18:02散场	4号厅	34.0	4号厅	34.5

图9 搜索影院获得电影信息_数据展示

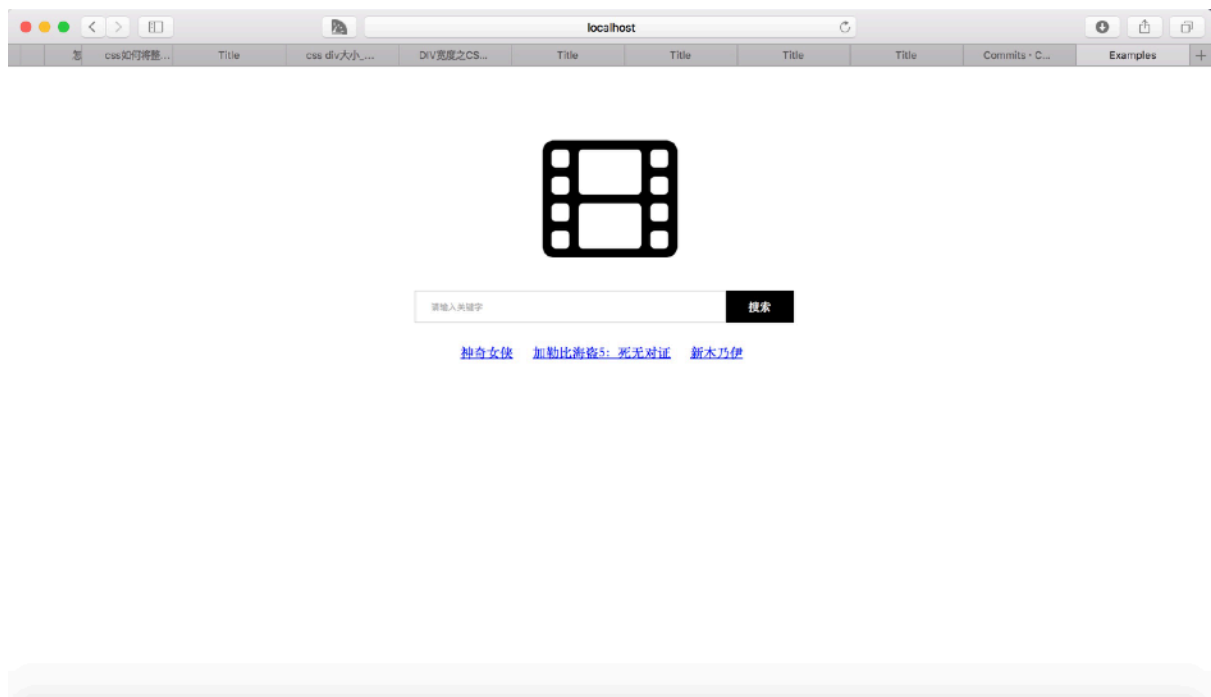


图10 搜索电影获得不同影院场次

加勒比海盗5: 死无对证							
影院名称	放映日期	开始时间	结束时间	格瓦拉场次	格瓦拉价格	淘宝场次	淘宝价格
UME国际影城(玄武门店)	2017-06-12	10:30	预计12:39散场	7号4D厅	39.0	4D厅	39.5
UME国际影城(玄武门店)	2017-06-12	11:45	预计13:54散场	9号厅	29.0	9号厅	29.5
UME国际影城(玄武门店)	2017-06-12	13:20	预计15:29散场	11号厅	34.0	11号厅	34.5
UME国际影城(玄武门店)	2017-06-12	14:10	预计16:19散场	9号厅	34.0	9号厅	34.5
UME国际影城(玄武门店)	2017-06-12	15:45	预计17:54散场	11号厅	34.0	11号厅	34.5
UME国际影城(玄武门店)	2017-06-12	16:35	预计18:44散场	9号厅	34.0	9号厅	34.5
UME国际影城(玄武门店)	2017-06-12	19:00	预计21:09散场	9号厅	39.0	9号厅	39.5
UME国际影城(玄武门店)	2017-06-12	20:05	预计22:14散场	11号厅	39.0	11号厅	39.5
UME国际影城(玄武门店)	2017-06-12	21:25	预计23:34散场	9号厅	39.0	9号厅	39.5
UME国际影城(玄武门店)	2017-06-12	22:30	预计00:39散场	11号厅	39.0	11号厅	39.5
金逸影城(大观店)	2017-06-12	12:00	预计14:09散场	一号厅	28.0	无	无
金逸影城(大观店)	2017-06-12	12:50	预计14:59散场	五号厅	28.0	无	无
金逸影城(大观店)	2017-06-12	16:10	预计18:19散场	六号厅	28.0	无	无
金逸影城(大观店)	2017-06-12	17:50	预计19:59散场	五号厅	28.0	无	无
金逸影城(大观店)	2017-06-12	20:40	预计22:49散场	六号厅	28.0	7号厅	29.0
卢米埃影城(蜜蜂店)	2017-06-12	13:20	预计15:29散场	三号厅	34.0	三号厅	29.0
卢米埃影城(蜜蜂店)	2017-06-12	15:50	预计17:59散场	三号厅	34.0	三号厅	29.0

图11 搜索电影获得不同影院场次_数据展示

由于图6、图7所示的两张数据表中的数据是异构的，而我們所需要的数据需要同时有同一场次电影的格瓦拉价格和位置及淘票票价格和位置。根据爬取到的数据信息可以发现，两个表的电影名称字段、影院字段、开始时间字段是复合主键，即可以唯一确定一个场次和价格。

因此建立了一个包含公共信息的类 src/main/java/entity/CommonData类，通过读取gewala_plan表中的每一个格瓦拉场次信息中抽取以上字段注入CommonData类，用该类的三个字段搜索taobao_plan中的淘票票场次，将两个表中不同的场次和价格信息注入src/main/java/entity/UniformPlan类，该类用于展示不同平台的位置、场次、价格等信息。

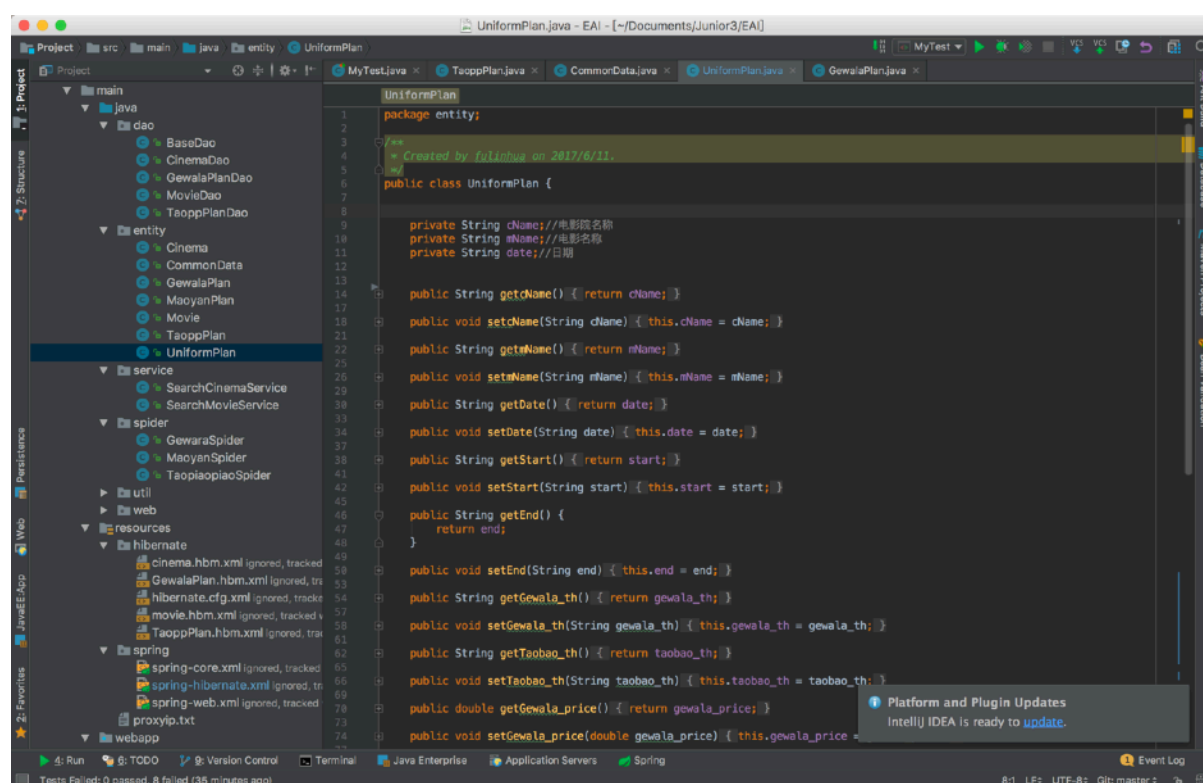


图12 工程文件src/main/java/entity/UniformPlan类

3. 影响和意义

本项目爬取了格瓦拉和淘票票网站的电影数据进行集成，并发现了实际互联网应用中存在的异构数据。



The screenshot shows a web browser window with the address bar set to 'localhost'. The page title is '金逸影城(大观店)'. Below the title is a table with 8 columns: 电影名称 (Movie Name), 放映日期 (Show Date), 开始时间 (Start Time), 结束时间 (End Time), 格瓦拉场次 (Gwara Showtimes), 格瓦拉价格 (Gwara Price), 淘宝场次 (Taobao Showtimes), and 淘宝价格 (Taobao Price). The table lists showtimes for movies like '加勒比海盗5: 死无对证' and '神奇女侠'.

电影名称	放映日期	开始时间	结束时间	格瓦拉场次	格瓦拉价格	淘宝场次	淘宝价格
加勒比海盗5: 死无对证	2017-06-12	12:00	预计14:09散场	一号厅	28.0	无	无
加勒比海盗5: 死无对证	2017-06-12	12:50	预计14:59散场	五号厅	28.0	无	无
加勒比海盗5: 死无对证	2017-06-12	16:10	预计18:19散场	六号厅	28.0	无	无
加勒比海盗5: 死无对证	2017-06-12	17:50	预计19:59散场	五号厅	28.0	无	无
加勒比海盗5: 死无对证	2017-06-12	20:40	预计22:49散场	六号厅	28.0	7号厅	29.0
神奇女侠	2017-06-12	11:20	预计13:42散场	四号厅	28.0	无	无
神奇女侠	2017-06-12	13:20	预计15:42散场	六号厅	28.0	4号激光厅	29.0
神奇女侠	2017-06-12	14:00	预计16:22散场	四号厅	28.0	无	无
神奇女侠	2017-06-12	16:40	预计19:02散场	四号厅	28.0	无	无
神奇女侠	2017-06-12	19:20	预计21:42散场	四号厅	28.0	无	无
神奇女侠	2017-06-12	22:00	预计00:22散场	四号厅	28.0	无	无
神奇女侠	2017-06-12	22:45	预计01:07散场	五号厅	28.0	无	无
新木乃伊	2017-06-12	09:50	预计11:36散场	二号厅	28.0	无	无
新木乃伊	2017-06-12	11:00	预计12:46散场	三号厅	28.0	1号双机激光厅	29.0
新木乃伊	2017-06-12	11:45	预计13:31散场	二号厅	28.0	无	无
新木乃伊	2017-06-12	13:00	预计14:46散场	三号厅	28.0	1号双机激光厅	29.0
新木乃伊	2017-06-12	13:40	预计15:26散场	二号厅	28.0	无	无

图13 异构数据展现_1

神奇女侠	2017-06-12	20:15	预计22:37散场	无	无	6号VIP厅	44.0
神奇女侠	2017-06-12	21:25	预计23:47散场	无	无	3号厅	29.0
新木乃伊	2017-06-12	10:00	预计11:46散场	无	无	2号双机激光厅	29.0
新木乃伊	2017-06-12	12:00	预计13:46散场	无	无	2号双机激光厅	29.0
新木乃伊	2017-06-12	14:00	预计15:46散场	无	无	2号双机激光厅	29.0
新木乃伊	2017-06-12	16:00	预计17:46散场	无	无	2号双机激光厅	29.0
新木乃伊	2017-06-12	17:30	预计19:16散场	无	无	5号激光厅	29.0
新木乃伊	2017-06-12	18:00	预计19:46散场	无	无	2号双机激光厅	29.0
新木乃伊	2017-06-12	18:30	预计20:16散场	无	无	4号激光厅	29.0
新木乃伊	2017-06-12	19:30	预计21:16散场	无	无	5号激光厅	29.0
新木乃伊	2017-06-12	20:00	预计21:46散场	无	无	2号双机激光厅	29.0

图14 异构数据展现_2

如上图11所示，由于金逸影院在不同平台上的数据不同，对于同样的电影《加勒比海盗5: 死无对证》、《神奇女侠》，有些场次在格瓦拉有数据，在淘票票却无数据；而如图12所示，电影如《神奇女侠》、《新木乃伊》在格瓦拉无数据，但在淘票票有数据。

在实际应用中，相同的位置信息在不同的网站中可能有不同的表述方式，由此可能引起歧义，数据集成可以在一定程度上避免该种歧义。

本项目的主要目的是分析已有的互联网应用数据异构的特征，实现多个同主题互联网应用的异构数据集成。同时也为应用集成课程的第三次作业“基于机器学习的方法实现对产品的评分或打标签”提供了选题并打下基础，将在后期对本项目进一步填充信息和功能。