票房爬虫：

使用python爬取猫眼票房的当天票房信息，每天定时任务

数据项：

datetime 票房日期

total\_incom 当日总票房

moviename 电影名

income\_now 电影当日票房

income\_ratio 票房占比

row\_ratio 排片占比

occupancy 上座率

豆瓣电影爬虫：

因为数据量大，链接多，所以使用python的异步爬虫框架scrapy，对豆瓣电影的数据进行爬取，定时任务。

目录结构：

spider：

movie： 电影爬虫

misc:

\_\_init\_\_.py

bloomfilter.py 网页规则过滤，防重复爬取

helper.py 帮助

middlewares.py 中间件

spiders:

\_\_init\_\_.py

movie.py 电影爬虫规则

\_\_init\_\_.py

items.py 爬取数据项定义

pipelines.py 数据库写入，利用twisted异步写入数据库，防止io阻塞，提高效率

scrapy.cfg 爬虫配置

settings.py 爬虫设置

piaofang： 票房爬虫

piaofang.py

数据项定义：

class MovieItem(Item):

douban\_id = Field() //豆瓣id

name = Field() //电影名

name\_other = Field() //其他电影名

directors = Field() //导演

actors = Field() //主演

countries = Field() //国家

genres = Field() //电影类型

languages = Field() //电影语言

runtime = Field() //播放时长

description = Field() //电影描述

release\_date = Field() //发布时间

tags = Field() //电影tag

image = Field() //电影首页图链接

douban\_rating = Field() //豆瓣评分