# 题目**要求**

**网页 “****[http://movie.shtimes.com/FilmId/ ”中包含观众对电影的评分信息](http://movie.xtime.com/FilmId/%20)，请编写程序抓取网页（网页样本保存于task0202目录中）上电影的评分信息并计算其统计信息（统计方法指对某部电影的评分求极值或求平均值），本题的赛前抽取参数是统计方法以及网页样本，请参赛学生将本题的答案保存于****ans0202.txt文件中，注意ans0202.txt文件中只能包含一个浮点型数字，保留4位小数**

## 爬虫思路：

1. **读取本地task0202目录中网页内容**
2. **通过正则表达式获取电影id（代码见handel.py）**
3. **根据电影id构造评分网页url（评分页面url规律为：<https://movie.douban.com/subject/> + 影片id + /reviews）**
4. **编写scrapy代码（spiders文件夹中gradeSpider.py文件），将构造的评分网页url放入start\_urls**
5. **由于豆瓣网具有反扒虫（尝试过设置代理，但代理质量差速度太慢，所以使用cookies,但cookies具有时效性），在settins.py文件中开启延时（DEOWNLOAD\_DELAY=3），**

**爬虫类中重写start\_requests函数（设置cookies和headers）**

1. **编写parse函数获取电影名和评分（评分由文字转为分数，格式为{'movie\_name': '麦兜·当当伴我心的影评 (584)', 'grade': 5},循环获取下一页）**
2. **编写中间件pipelines的函数process\_item，将所有电影名和评分放入变量result**
3. **编写中间件pipelines的\_\_del\_\_函数，用于在爬虫结束时将result中结果转为{"金刚：骷髅岛的影评 (1175)": {"avg\_grade": "3.6000", "max\_grade": 5, "min\_grade": 2}**

## 运行结果(文件为**ans0202.txt**)