有关python爬取热门电影的评论及数据分析的浅析——以《海王为例》

1. 前言

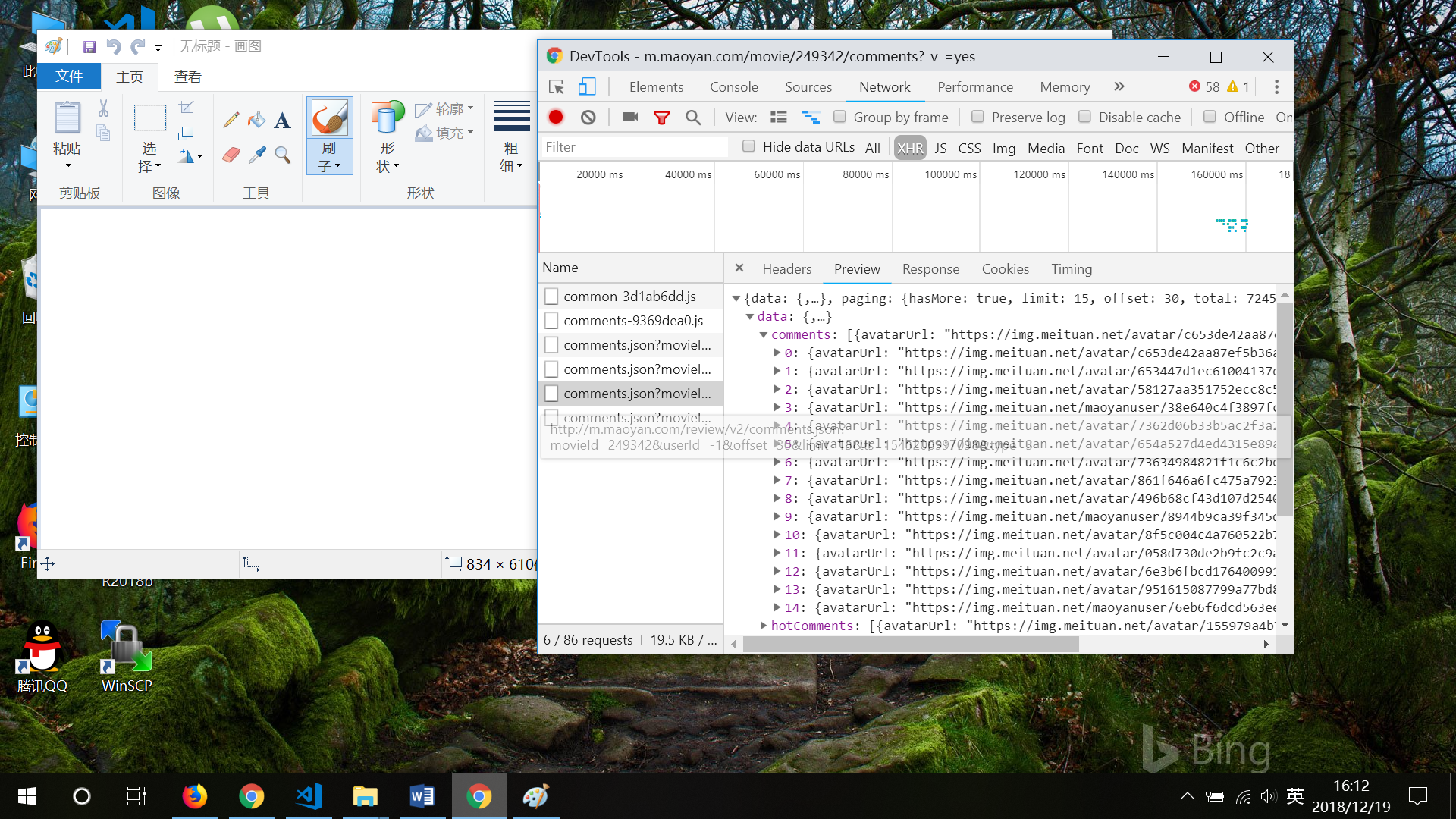
python是一种强有力的数据爬取和分析工具，本文试图通过电影的影评分析，以日前热门的电影《海王》为例综合评价电影。（数据来源：猫眼电影）

1. 数据获取

python有诸如scrapy等很多实用的爬虫框架，本文选取较为入门的requests库作为爬取的工具。

猫眼电影的官网中只有区区10条评论，显然不足以作为分析样本。

通过f12审查工具对网页的数据来源予以审查，注意到pc端网页是通过直接返回html文件的形式，防止对网页评论的爬取。故使用谷歌浏览器模拟模拟移动端访问，

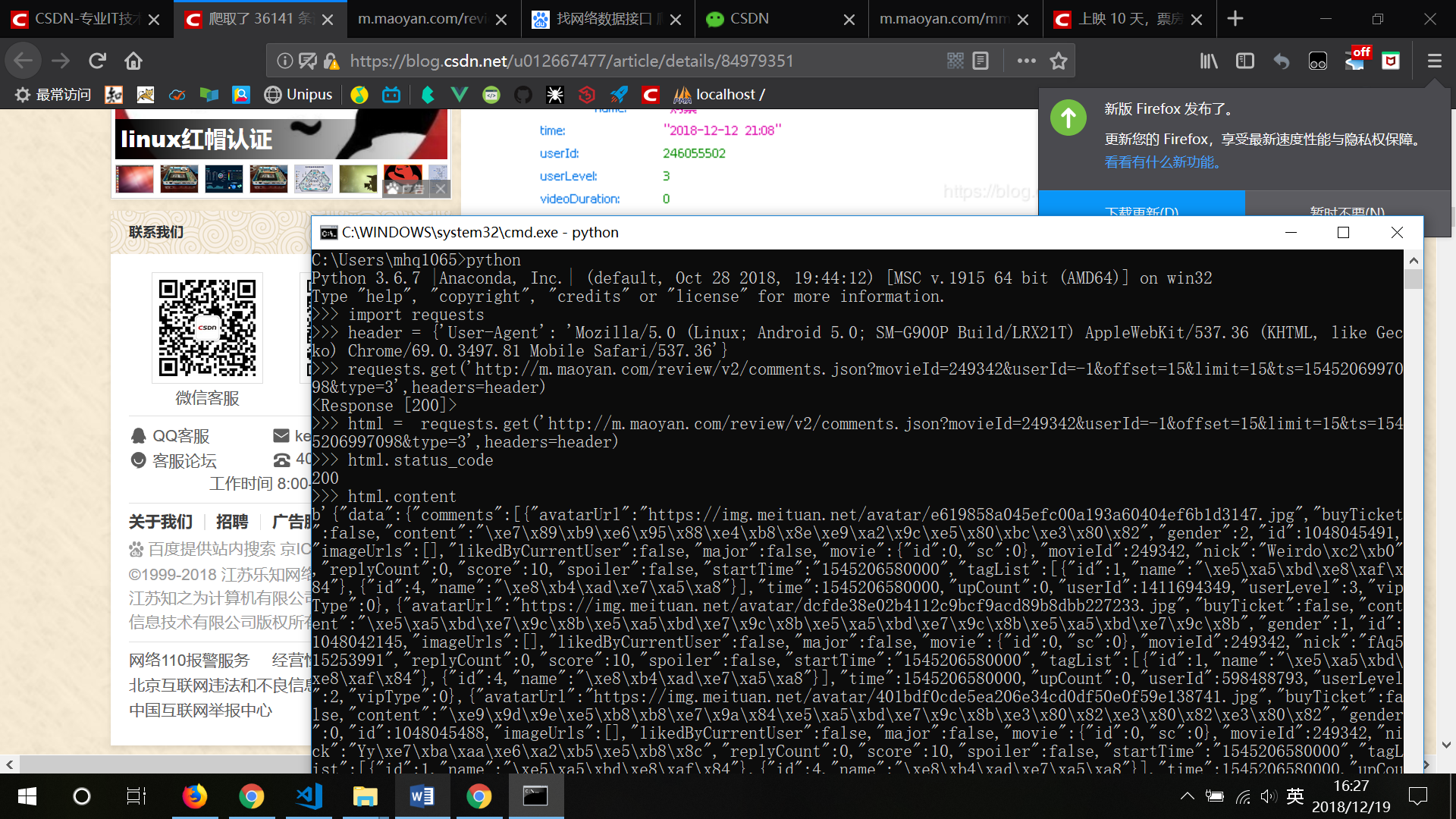


对数据网址予以分析



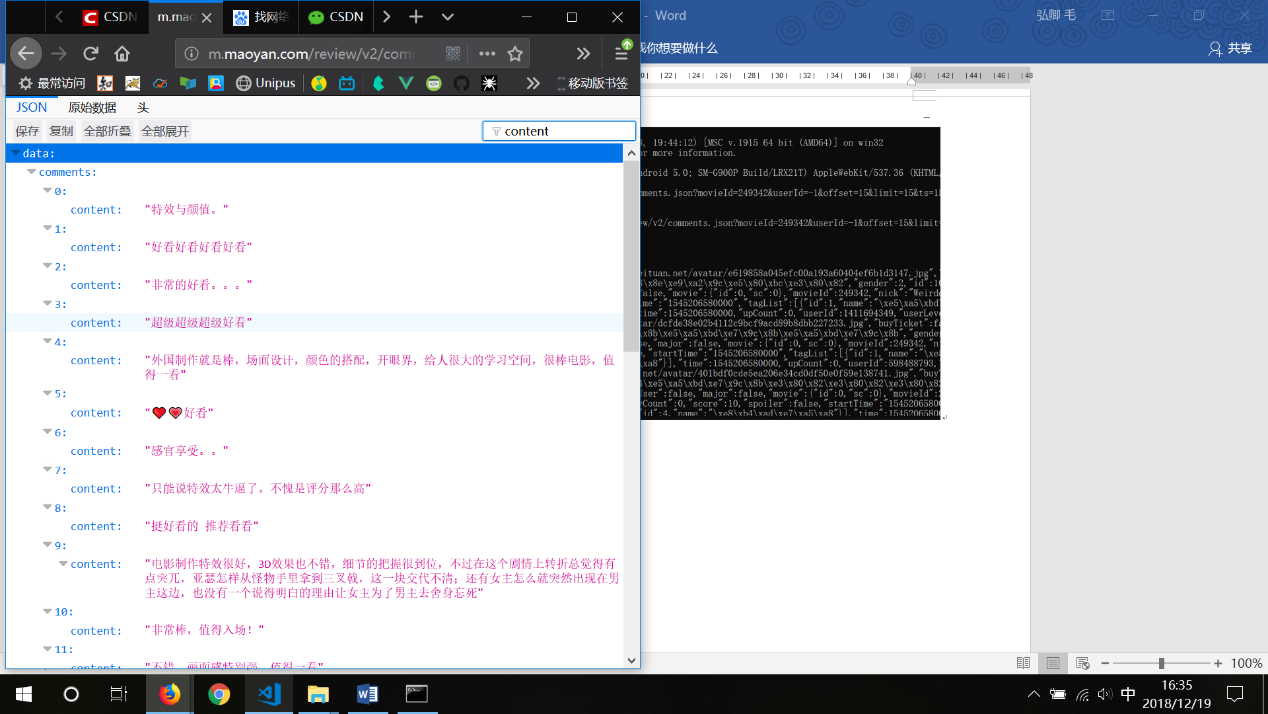
各个网址仅有offset的偏移量，利用这一规律编写python程序爬取数据。

以下为测试结果，确定方法可行。

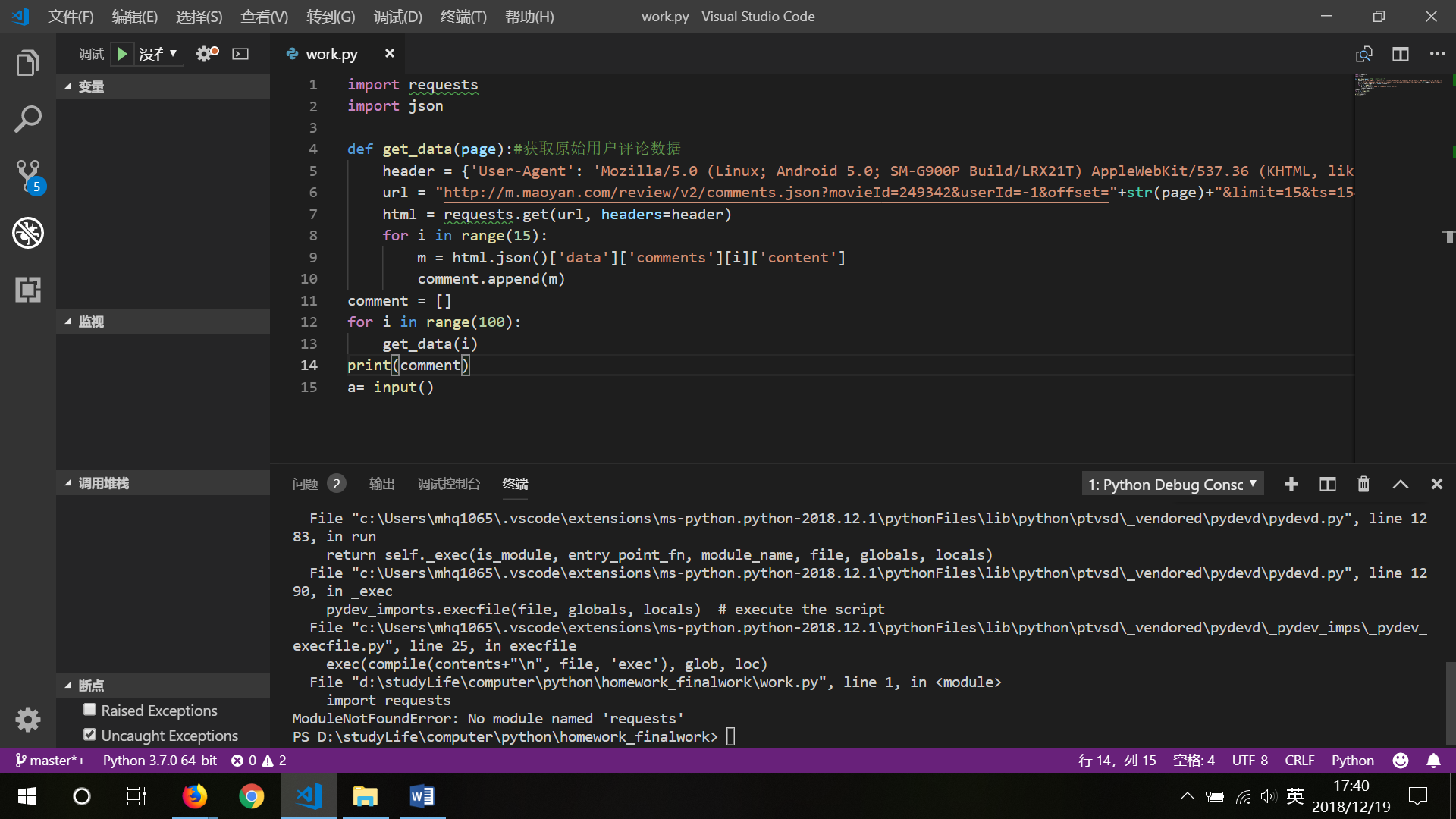


1. 数据来源分析和数据储存

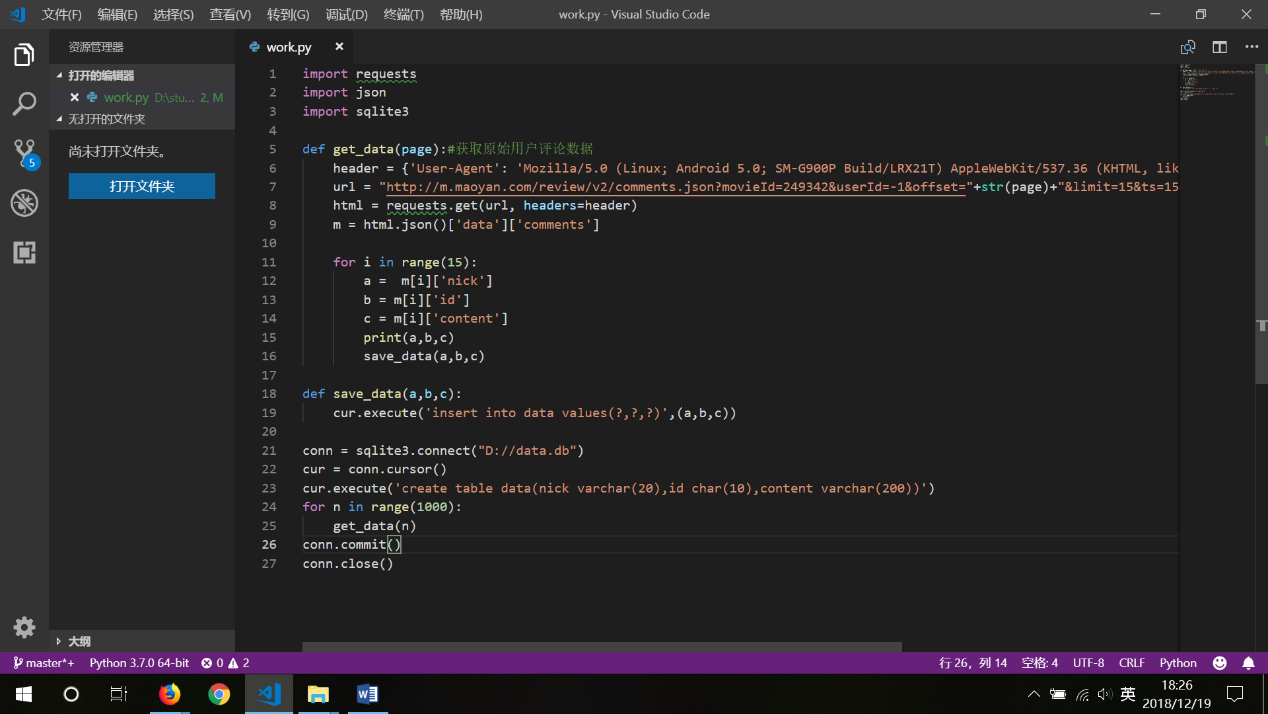
用火狐浏览器打开链接，注意到数据以json格式接收，引入json库。



观察数据结构，只需要提取content的内容即可。初步程序如下：



然后，利用sqlite3写入数据库，修改程序



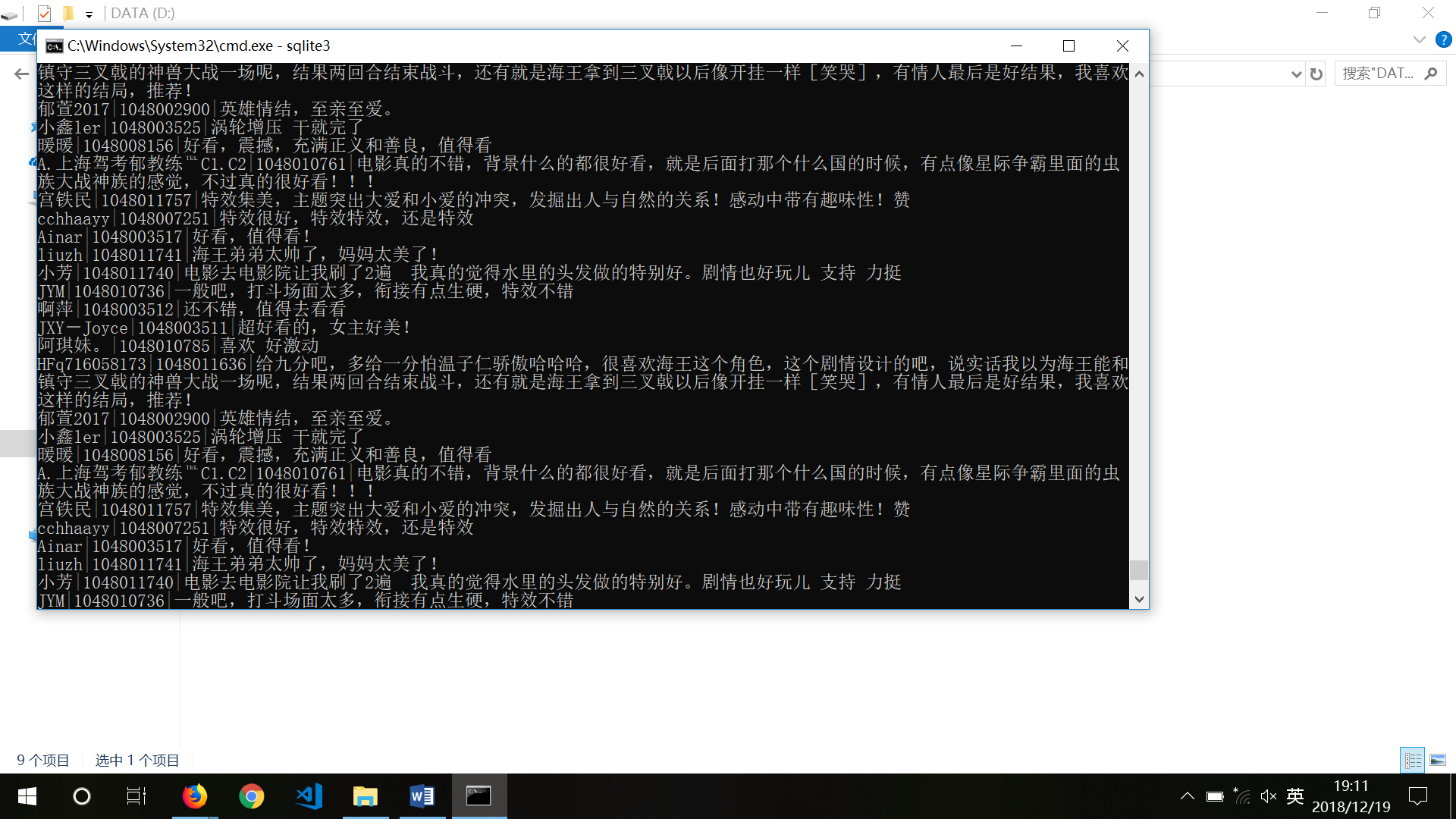
得到最终程序。（文件保存于D：//）

1. 源代码



1. 结果展示





1. 结语

利用python成功爬取电影影评数据，并同时存入sqlite数据库。但是所有爬虫必须遵循robot协议，合理地爬取网络数据。