哈爾濱工業大學

毕业设计（论文）中期报告

题 目：面向电影领域的微信聊天机器人

专　　业　计算机科学与技术

学　　生　马　晶　义

学　　号　**1130310723**

指导教师　杨　沐　昀

日　　期　**2017.4.22**

哈尔滨工业大学教务处制

1．论文工作是否按开题报告预定的内容及进度安排进行

1.1 课题的主要研究内容

本课题的主要研究内容是面向电影领域的微信聊天机器人研究与实现。本课题将从近期热映电影的影评入手，构建电影的知识库，力图实现较为准确的电影领域的问答系统，并与微信进行结合，在微信公众号和微信群中实现与用户之间的交互问答。本课题的研究的整体系统框架如图 Figure 1。主要研究内容有：

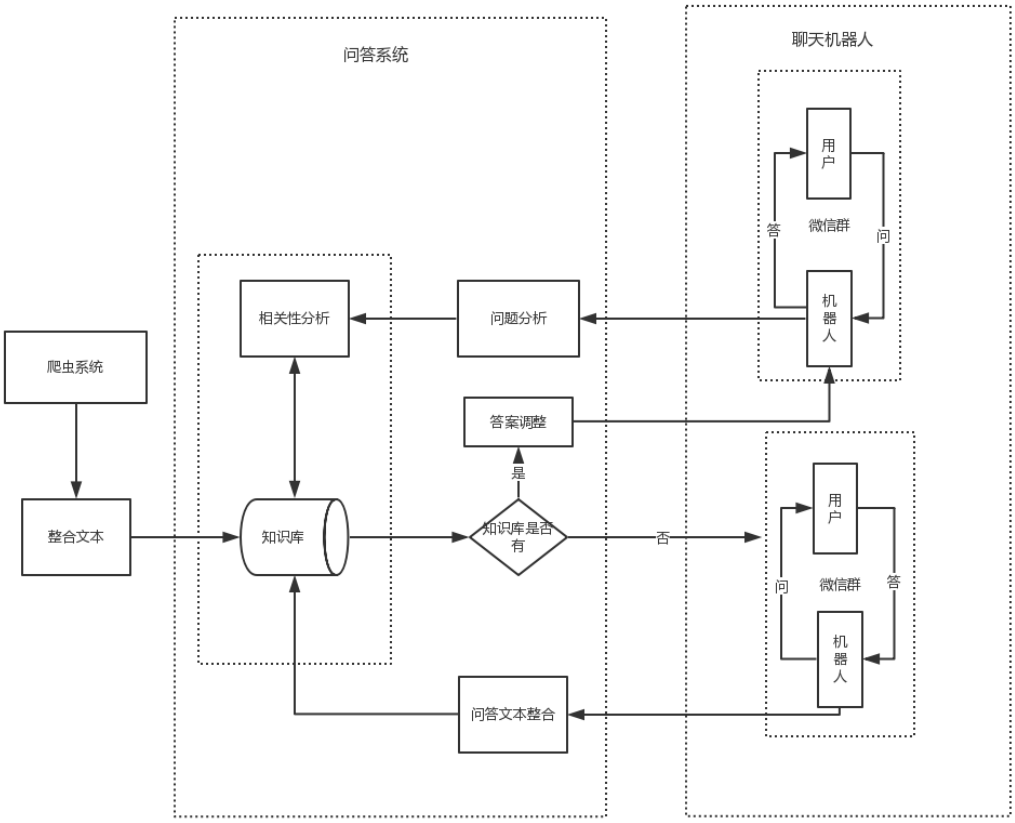

Figure 1 系统架构图

1）影评数据获取与数据训练、测试。由于影评的获取主要是通过爬虫在豆瓣影评、微博电影、猫眼影评多个数据源上爬取，该阶段主要编写爬虫的代码进行爬取数据即可；而在数据训练与测试中，该阶段则需要选取合适的训练测试方案进行调整系统的参数。

2）问答系统的研究与设计。问答系统的核心是答案检索和答案抽取。则这部分的主要研究内容为：

问题答案的检索。问题的答案需要在知识库中进行检索得到，因此优秀的检索方案是问答系统中必不可少的。问答检索的方法有 LDA 主题模型方法和 Word2Vec 中词向量模

型方法，本系统采用的是佐治亚州立大学的 Zhibo Wang 和 Yanqing Zhang 在 2016 年提出的整合 LDA 模型和 Word2Vec 模型方法，其性能的 F 值能达到 85.6%，优于 LDA 模型（68.0%）和 Word2Vec 模型（80.75%）。

问题答案的抽取。在知识库中检索得到的答案句子有多条，需要进行相关性排序，并提取出合理的答案。本课题采用 BM25 算法对检索的结果进行排序，提取排名靠前的答案作为对用户的回答。

3）微信机器人系统的实现。利用热映电影的影评构成的知识库完成基于电影的微信机器人系统设计，将问题分类、语义解析、答案检索、答案提取扩展融合到问答系统中去，实现能够针对用户的自然语言问题，给出较合理的回答，并且通过微信这一聊天工具与用户进行对话交流。

1.2 论文工作是否按开题报告预定的内容及进度安排进行

论文工作在按照预定内容上有部分的改动：一是在数据来源上由单一的数据来源（豆瓣影评），换成了多个数据来源；二是在问答系统设计中，去除对问题的语义分析、关键词和限定词提取及用关键词检索答案这些部分，改成直接用问句在知识库中检索的方式，在检索的结果中进行提取问题的答案。由于数据获取部分内容增加，进度相对开题安排的较慢，但问答系统的内容有所筛减，因此后期工作量相对减少，后期的进度会相对加快。

2．已完成的研究工作及成果

由于开题较晚，现已完成的研究工作主要是数据获取部分：豆瓣影评、微博电影、猫眼电影的数据获取。

Table1 已完成的研究成果

已完成	类别	影评数据量	备注
数据获取	豆瓣影评	21012 条	部分刚上映电影的影评量少
	微博电影	24058 条	微博数据有的不能获取完全
	猫眼电影	11854 条	网站限制单部电影 1675 条

2.1 豆瓣影评

2.1.1 模拟登录

豆瓣的影评数据量多，用户未登录的情况下只能查看少量的数据，因此程序需要进行模拟用户登录，来查看全部的影评数据。

模拟登录的方法主要采用 python 的 requests 包，requests 的 post 方法能够将登录的数据参数通过网络发送给登录请求页面，当登录成功会返回正常的状态码。豆瓣模拟登录主

要的难题是验证码的问题，如图 figure2。验证码的获取方式可以通过解析登录页面的 html 文件得到验证码的下载链接，下载验证码进行识别。采用代码进行图像识别，难度较大，最终选择程序下载并打开验证码图片，人工进行识别输入。最终如图中所标注的，将数据进行封装成 Form data 通过 post 方法进行模拟登录。

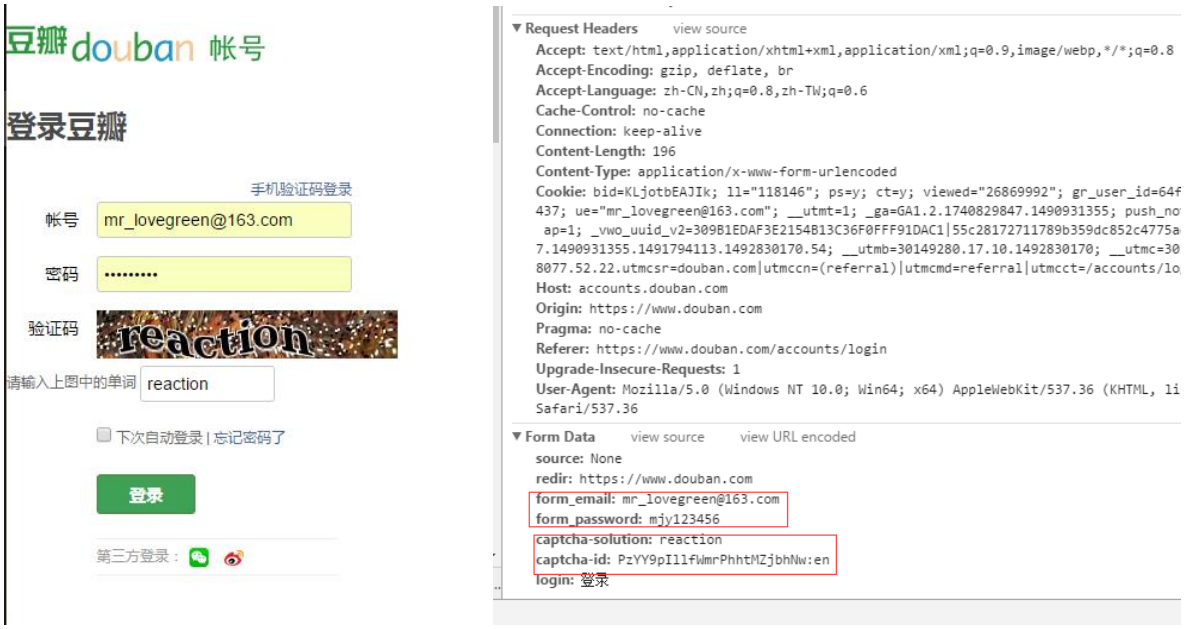

Figure 2 豆瓣模拟登录

2.1.2 电影信息获取

电影信息页面，本来是只需要解析出 html 中的电影信息进行保存即可，但遇到一个问题，如图 Figure3 中，电影信息的每一块在 html 中没有明显的标签区分，而且各个电影信息量也不是都很全，导致用固定的列表的第几项不能准确代表第几块的内容，因此采集了大量电影信息的块长，采用判断块长与选用正则表达式的方法进行选择电影信息解析并存储。

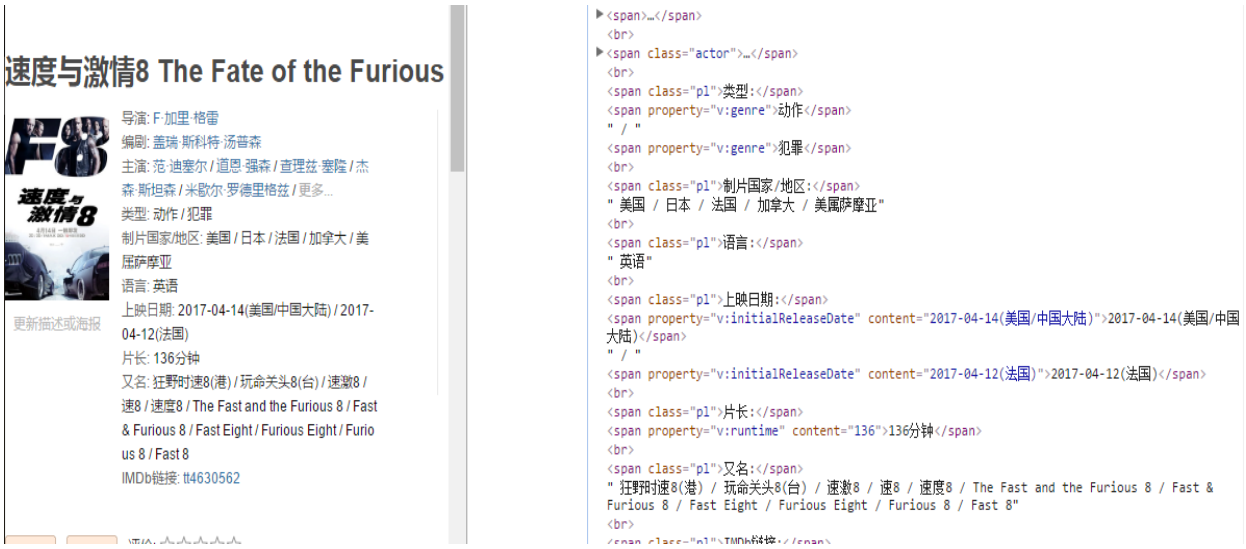

Figure 3 豆瓣电影信息

2.1.3 影评数据获取

豆瓣影评的数据分为长、短影评。由于影评的数据量较多，豆瓣都作了分页处理，除了解析每一页，获取下一页的链接是关键。由于第一页是首页，页面内没有前一页和首页的链接，和后续的 html 页面有所不同，因此对第一页进行单独解析。解析得到下一页链接经过封装，使用 python 的 requests 库的 get 方法获取下一页的 html 内容，从而一直能够访问到最后一页。

2.1.4 数据库存储展示

构建了四个表：better_than、comments、 movies、 score_box 分别存储电影好于同类电影的信息、影评信息、电影信息及电影评分信息。下图是展示一条评论信息的存储状态，除了存储评论的文本，还额外存储了其他信息保证数据的无损性。

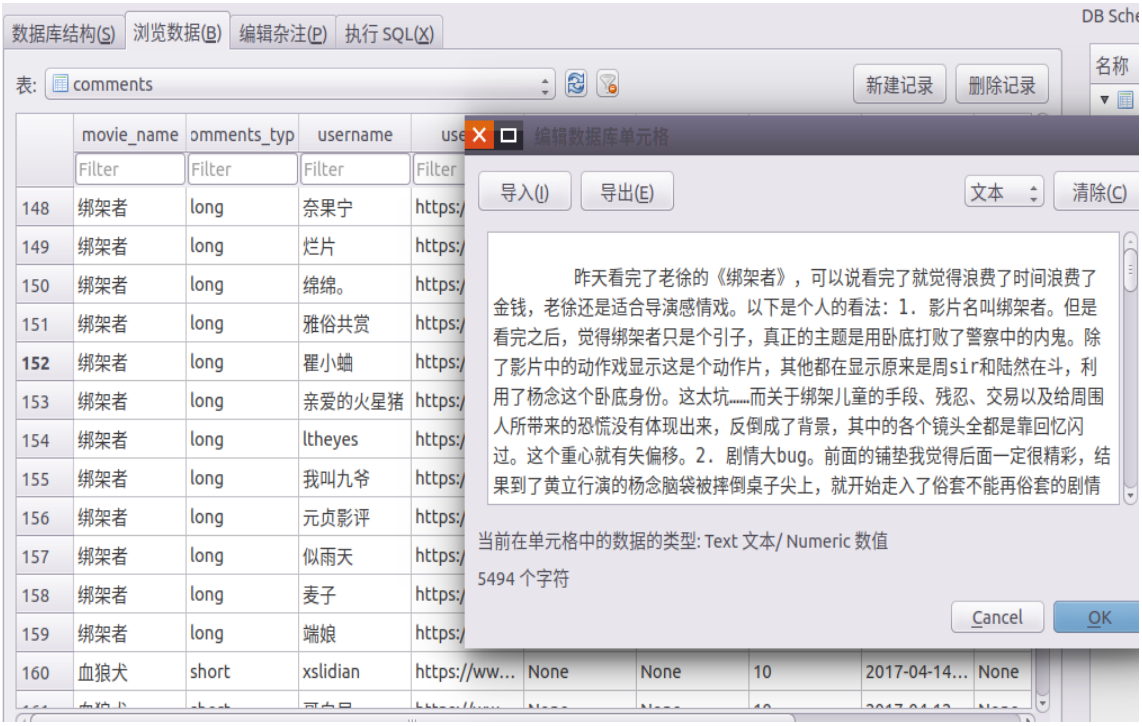

Figure 4 豆瓣数据库存储展示

2.2 微博电影

2.2.1 模拟登录

微博的数据获取相对豆瓣来说较复杂，网页端的数据安全机制及反爬虫手段较高级，因此选择爬取手机端的数据。微博手机端的模拟登录，相对于豆瓣的模拟登录，除了增加一些参数，最重要的是在获取验证码的时候，需要通过对用户名及当前时间进行加密，得到一个参数，需要通过这个参数去获取验证码。加密的过程是通过 base64 进行加密，因此采用 python 的包 base64 及 urllllib 的 quote_plus 进加密即可。

2.2.2 电影信息获取

微博手机端的数据都是通过 js 获取 json 数据的，因此找到 json 的请求链接，对获取

的 json 数据进行解析，从而得到电影信息的数据。如图 figure5，获取参数进行拼接成请求链接，构造请求头的 Host 及 User-Agent 字段，然后请求数据。由于 cookies 可以在 session 中保存，采用 session 的 get 方法请求就自动带有用户的 cookie。

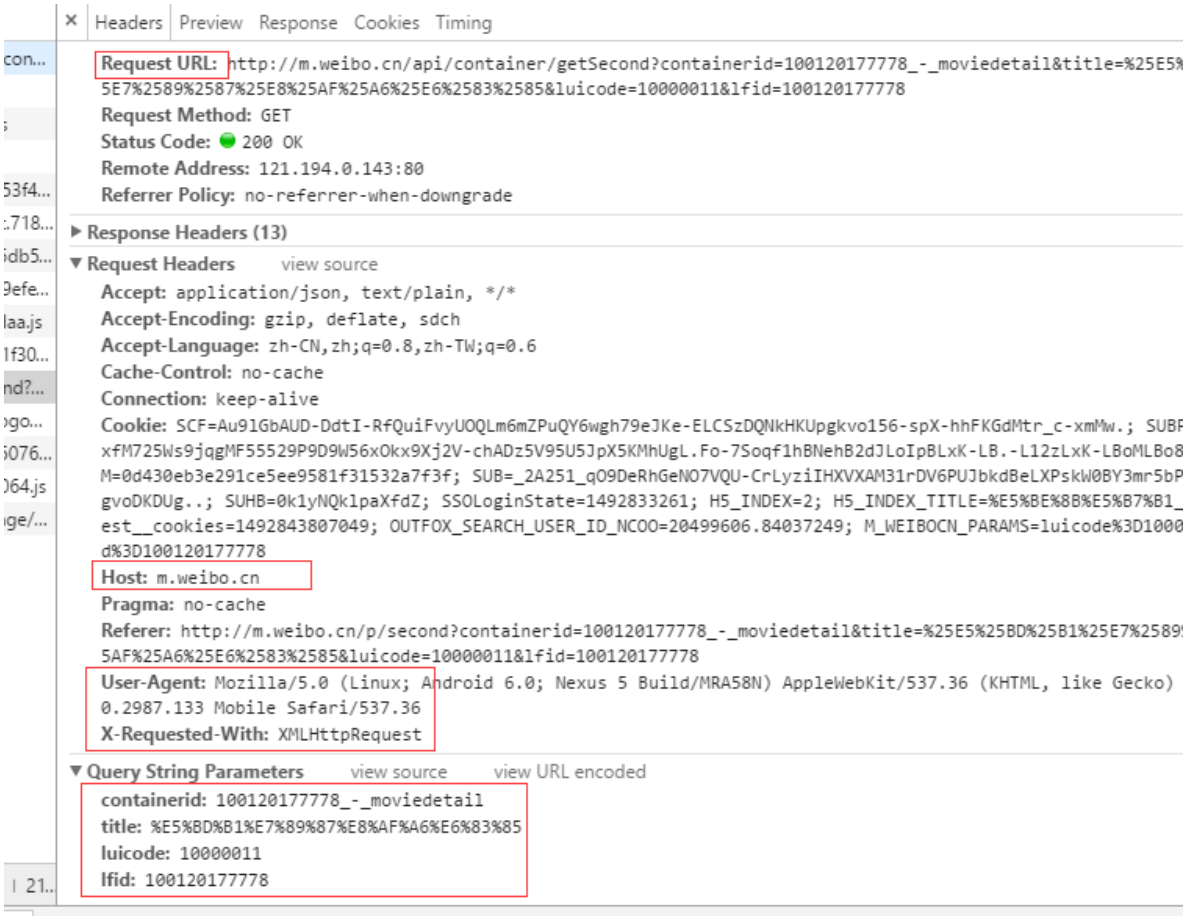

Figure 5 电影详情页面请求

2.2.3 电影微博的分类及微博内容获取

通过对电影页面的分析，关于电影微博分为以下几类：点评团影评微博、网友影评微博、相关话题微博、好评微博及差评微博。因此在电影页面获取的 json 数据，解析出这五种微博的链接，再分别在这五种微博的链接上加入参数进行构造成 json 的请求链接，通过上述的 session 的 get 方法分别获取对应的 json 数据，解析出对应的内容。在返回的 json 数据中有微博的内容以及每条的微博的链接，微博的链接就可以作为获取以下微博评论的数据参数。一次请求的 json 中包含 15 条微博，因此需要一直进行获取，当获取的 json 数据中”ok”参数为 0 时，代表能访问的页面结束。

2.2.4 对微博的评论数据获取

由上述微博内容获取中获得微博的链接，进行访问可以获取关于此微博的评论数目及评论的初始请求链接，因为有评论的总数目，每页的评论数目是固定的 25 条，因此可以直接计算出评论的页数，然后请求每一页的 json 数据，解析 json 提取评论的信息存入数据库。

2.2.5 “钓鱼”功能

微博的“钓鱼”功能是指通过程序模拟用户去发一些微博，并艾特一些用户，然后获取用户的转发或评论内容。功能层面如图 Figure6 已经可用，这里的核心是：需要有吸引人的微博内容来吸引用户来进行转发和回答；需要艾特那些在电影领域乐于分享影评的用户。

Figure 6 微博钓鱼功能

2.3 猫眼电影

2.3.1 电影信息获取

猫眼影评的数据相对于豆瓣和微博要容易获取，其一是不需要模拟登录都可以访问影评的页面，其二是手机端的页面是静态 html 文件，相对于微博的 js 请求要简单许多。电影信息可以通过直接解析电影页面，获取对应的数据。

2.3.2 影评数据获取

影评数据也是相对来说比较单一，只有一类影评数据。因此在电影页面获取评论的链接及评论的数目。访问评论的链接，当下拉评论的时候会刷新影评，从而获取下一页的评论内容，解析内容即可得到数据。目前发现当影评数量较多的时候，用户最多能够访问前 67 页，每一页是 15 条影评，即只有 1675 条影评，因此限制了总的数据量。

2.3.3 数据库存储展示

构建的数据库设计了四张表：movies、box、celebrities、comments 分别存储了电影信息、电影评分信息、演职员表及电影的影评数据。如下图 Figure7 展示的是猫眼影评的一条数据的评论信息在数据库中的存储情况。

Figure 7 猫眼影评数据库展示

3. 后期拟完成的研究工作及进度安排

3.1 后期拟完成的研究工作

1）知识库构建及数据去重

2）答案检索模型搭建与性能测试

3）微信聊天机器人设计

4）系统整体功能测试与评估

5）撰写毕业论文准备结题答辩

3.2 进度安排

Table 2 预期进度安排

起始时间	完成时间	计划完成内容
2017.4.26	2017.5.3	知识库构建及数据去重
2017.5.4	2017.5.18	答案检索模型搭建与性能测试
2017.5.19	2017.5.31	微信聊天机器人设计
2017.6.1	2017.6.5	系统整体功能测试与评估
2017.6.6	2017.6.13	撰写毕业论文准备结题答辩

4. 存在的问题与困难

目前主要遇到的问题，采用了以下给出的解决方案

4.1 对网页数据的提取

1) 静态网页数据提取

静态网页是 html 数据，其内容是多而复杂的。提取方式有两种，一是直接用正则表达式，通过正则匹配提取对应的内容；二是采用 python 的 bs4 的包，可以将 html 进行 xml 解析，然后通过提取对应的 html 标签，获取对应的内容。第一种方式较复杂但准确，针对每一条数据都要写一个正则表达式太过繁琐，第二种方式虽然可以通过 html 标签快速找到数据位置，但是如果遇到标签相同而数据不同的情况，就需要结合第一种方式来提取数据。

2) 动态网页数据提取

动态网页数据是不能通过下载 html 文件获得，因为所需的数据都是通过 js 在浏览时从数据库动态获取并嵌入 html 文件中，下载的 html 是没有这些数据的。通过抓包工具，查找出 js 请求数据的请求链接及参数，通过程序请求并获取对应的 json 数据，针对 json 数据的格式，解析出对应的数据信息，从而达到提取动态网页数据的目的。

4.2 网页数据的编码问题

1) 中文网页编码格式转换问题

在网页编码中经常会遇到各种编码，教常见的是 utf-8、gbk、gb2312 等。在解析获得的数据中，由于编码混乱中文经常是乱码的状态。想要统一编码就需要知道网页的编码格式，一查看网页开始部分的编码设置声明，但不是所有网页都有，二通过 python 的 chardet 包进行检测数据的编码，然后通过解码成 unicode 再编码为 utf-8 格式。

2) 中文数据存储编码格式问题

当数据的编码格式确定之后，存储时对数据库的编码也有要求。在 windows 的环境下，使用 mysql 数据库，设置的 utf-8 编码，但存储及查看的时候中文都为乱码，而 windows 的命令行编码也是 gbk 编码，查看 utf-8 编码的中文也是乱码。最终选择了 sqlite3 数据库，并在 ubantu 的环境下存储查看数据。

4.3 关于网站的模拟登录问题

模拟登录是通过 requests 的 post 方法将登录的数据参数传给登录请求，虽然模拟登录成功，但想要访问其他页面的时候需要带有 cookie 访问是一个问题。大多数的页面的 cookie 的内容不一，如果采用每次访问的时候，构造 cookie 请求头进行请求，太过繁琐。采用 requests 的 Request 模块及 session 来解决，session 能够保留每次访问的 cookie 信息，在访问链接时可以自带这些信息。

4.4 用户代理问题

为了程序能够更像浏览器访问页面，因此采用用户代理是必要的，而且这也是最简单的反反爬虫的方法。通过浏览器的抓包工具，可以得到浏览器的代理信息，在每次访问链

接时，在请求头加入对应的代理信息即可。

4.5 反爬虫问题

1) 验证码问题

验证码问题目前遇到主要是加密请求及验证码识别。微博验证码的加密方式通过搜索得出，因此采用相同的加密方式即可得到加密的请求参数；验证码识别，可以采用图像识别的方法进行处理，但增加了额外的工作量，效果也不一定理想，选择了最简答的人工识别并输入验证码的方法。

2) ip 限制问题

在爬取豆瓣的数据时，在短时间内访问了大量页面时，豆瓣的反爬虫就会启动，限制请求的 ip。不过限制 ip 的时长是几个小时，初期可以选择隔几个小时爬一次，不过这样限制比较大，数据易出错。因而采用 ip 代理的方式，ip 代理是在免费 ip 代理网站上获取有效的 ip，将获得的 ip 封装成请求的 proxy，然后请求链接时就是代理的 ip 进行的请求。

5．论文按时完成的可能性

论文在老师的指导下，现已设计对应的论文大纲初稿，对毕业论文有一个简单的整体把握。在毕业设计完成的过程中，逐渐完成论文的各个部分初稿，最终整合修改成毕业论文。目前的进度稍慢，后期会尽量加快进度，相信在后期的努力及老师的指导下能够按时完成毕业论文。