**辅导员聊天机器人**

**想法来源：**

之前有一次有个较为紧急的事情要咨询辅导员，由于当时不方便只能在网上问他，但是辅导员隔了很久才回我，导致事情有点耽搁。但实际上后面发现我的问题在学校相关网站都有说明，只是不知道如何获取。正好那段时间自己在学习写爬虫，并看到了关于聊天机器人的博客，所以我就想能不能做个爬虫专门爬取学校各方面信息，再进行一定的数据清理和整合，放入数据库，当学生有什么问题不了解的时候，可以直接通过聊天机器人进行提问，聊天机器人通过之前的处理好的相关数据能及时回答学生的各种问题。

**项目概要**：

有了这个想法后，我就去查阅了相关资料，加上自己之前的经验，大致捋了一下思路，大致分三步：

第一：写一个爬虫程序，从学校各个网站（教务，学校官网，各学院网站等）爬取有用的没用的信息，初步处理后放入数据库

第二：建立一个网站，写好聊天机器人框架，用户可以直接提问来获取初步的较为简陋的答案（后期不知能否以小程序的方式接入微信）

第三步 利用自然语言处理相关技术，处理爬取的信息，这样聊天机器人能更好的从爬取的信息中组织语言，更加准确的回答

这些还只是我的初步想法，由于能力有限，很多问题还需要考虑，也还有很多方面可以更好，希望老师能不惜指导。谢谢！

**一月5号和小五第一次讨论：**

初步将项目实现定为三个阶段：

第一阶段：只是实现信息的匹配与排序功能，其中信息分为两类，一类是学生个人信息（包括正方教务系统、信息门户、图书馆等），一类是不属于学生个人信息的学校各种信息。在检索时通过分词、相关度匹配对搜索结果进行排序，稍微区别的是，第一类一般返回较为精准的匹配信息，如成绩、借阅情况等，第二类返回的是依据关键词匹配的各种通知，每项包含通知的所有内容。

要实现的技术：问题的语义识别，分词，匹配排序

第二阶段： 实现信息的整合，依据学生提出的问题，对不同的信息进行文本挖掘（较为初级，块级的语言理解），然后依据问题，从不同的信息来源获取信息整合成更准确的答案

实现技术：暂时不太了解

第三步：第二阶段的升级版，程序能实现能自动了解信息源中的所有信息，自动学习强化，然后根据不同的问题（以及提问者信息、提问时间等特征），自己组织语言进行回答

实现技术：暂时不太了解

**1月17号 爬虫及数据说明**

本爬虫主要爬取的对象是学校主页上的各个分栏（如上理要闻，校园快讯等）的所有通知以及学校教务处的所有通知公告，将其各个通知的信息存储为 id,title,body,url,crawltime,releasetime 六个字段，其中爬虫需要爬取的是title，代表的通知标题，body代表的通知正文（学校主页上的通知正文为该通知网页中正文的源码，教务处的通知正文为网页中通知源码的text部分），releasetime代表的通知发布时间。其他的id字段为自增，crawltime字段为数据存储时数据库自动生成，爬虫具体程序见代码。爬虫的存储格式为：

|  |  |  |  |
| --- | --- | --- | --- |
| 字段名 | 数据类型 | 约束 | 说明 |
| Id | Int(11) | 自增，主键 |  |
| Titile | Varchar(200) |  | 标题 |
| Body | Varchar(20000) |  | 通知正文 |
| Crawltime | Timestamp | 默认为插入时间 | 爬取时间 |
| Releasetime | Varchar(50) |  | 通知发布时间按 |
| url | Varchar(200) |  | 通知网址 |

这个爬虫目前只能实现学校公共信息类别的部分数据爬取，对于学生个人信息类别的数据由于现在不能登入相应网站，暂时未实现。