项目什么时候开始的忘了，就说说我记得的。

第一周：

工作内容:初步了解了xpath的使用语法，对其和正则的难易程度有一个粗略的感知。使用利用谷歌浏览器理清了瑞安论坛帖子的网页代码结构，并且经过参考写了一个请求头和伪装头以及帖子标题的爬取，并且成功爬取标题。

问题：保存文件应该如何使用文件操作（这个问题不大，但是没想到这个小问题最后影响到收尾工作）；另外论坛将最后的评论时间和发布时间整合到了一个结构中，如何分离。

第二周：

工作内容：通过fp实现了文件操作，初步文件格式为csv，后来改为表格文件存储，另外利用数据切片完成了第一周留下的数据分离问题。并且初步实现了对帖子信息的完全爬取以及数据分离。

问题：某些数据爬取产生了乱码问题。

第三周：

工作内容：将帖子对应评论的爬取代码也写了，产生了一个啼笑皆非的问题（我直接复制了爬取帖子的代码框架，然后就出现函数入口不对且数据变量被乱改）。并且将评论代码和帖子代码作了整合合并到了main.py文件中。

问题：帖子和评论跳转（已解决），爬取的评论只有最后一个帖子的评论，感觉是代码的循环结构出了问题，但是终端调试之前的也有爬取内容！！！发了许多信息，咋老师没回。睡前突发奇想可能是文件写入的覆盖问题。

第四周：

工作内容：睡前的突发奇想对了，更改了文件操作（帖子的为覆盖写入，修改评论的写入模式为追加），所以一定要注意细节。解决完评论的写入后该爬虫基本完善，陆续解决写入的文字排版已经爬取的元组和列表的去括号问题。完善readme和开发过程记录基本可上交。以及完善了抛出异常功能，避免长时间工作遇到异常数据前期化为乌有，暂定图形化评论内容为异常数据。

后续有时间会完善的内容：某些写入的字体可采用红色，将所有板块都可进行爬取（现在完整可爬取的是瑞安说事板块）。有时间将会将其他板块通过列表的形式存储板块号，修改请求头的格式即可完成整个论坛的爬取。

已经完善了所有板块号的存储和修改，现在已经算是可以爬取整个瑞安论坛的数据了

总结：

其实说是四周，但是我感觉真实的时间其实更少，主要还是编程时间离散的问题。编程一定要先理清项目的调用结构，然后再切分问题，由小块推导到大块（爬虫类），另外python一定要注意各个函数之间的接口和变量修改问题，很有可能就是变量未被修改或者接口未被调用导致问题。另外对于文件操作一定要细，不要因为细节导致收尾难搞。