这次能够参与大创是在机缘巧合下促成的。当时正在台湾作交换生，如果参加大创的话，没办法和队友及时交流，因此也没想到会得到负责人的邀请，对此真的非常感谢。同时因为加入的时间比较晚，初次审核的内容基本上团队成员都已经准备好，匆匆忙忙的看了材料以及在一些很小的方面，如论文修改等提出自己的建议。初次答辩，队友也准备的十分充分，顺利通过。在暑假之前，我们团队负责人初步安排了每位成员的大创学习任务。我们在完成学校规定的学习课程之外，也在为大创做着准备。并且把各自的学习进度同步到了git仓库。

此次大创，我主要负责的是编写程序对CSDN、博客园中关于人工智能的三个领域：机器学习、自然语言处理、机器视觉进行爬取以及图谱可视化的PHP操作和少部分MYSQL。暑假时负责人将我们聚集在了一起，互相鼓励学习以及开始各自的任务。

在选择编写爬虫的语言方面，我选择了python。python不单用于抓取网页文档的接口简洁，同时其访问网页文档的API也相当完整(如urllib2库)；除此之外，很多网站会封杀爬虫。这时我们需要构造合适的请求来模拟浏览器的行为，而python提供了非常优秀的第三方包如Requests，mechanize。在抓取了网页之后，还需进一步的处理（比如过滤html标签，提取文本等）。而python的beautifulsoap库等能让我们编写非常简洁的代码即可完成大部分文档的处理。虽然一些其它语言或工具来实现以上功能，但是python是最快速简洁的语言。

在编写爬虫的时候也遇到各种各样的问题，记录如下:

1.robots.txt

robots.txt通常存放于网站根目录下，负责告知我们哪些内容是不能被获取的，哪些是能被获取的。其实际上并不是一个必须遵守的规范。如果爬虫不遵守，那么网站的隐私是不能通过这种方式来防止泄露的。

2.头文件headers

浏览器和爬虫程序，通常会向服务器发一个头文件：headers 以发送网络请求。这里面的大多数的字段都是浏览器对服务器的标识，因此大多数时候我们编写的爬虫，都需要伪造headers以掩盖爬虫的身份，向服务器发送请求。而符合该网站要求的headers，可打开浏览器开发者工具->选择Network->刷新页面->选择最上方的文件->找到request headers 即可

3.验证码

部分网站会要求输入验证码，应使用python的selenium库，模拟浏览器，网站会要求扫二维码进行验证，扫二维码即可解决，但这个策略并不总是有效，因而验证码一直是反爬虫一个非常有效的手段，

4.IP限制

封ip和封账号也是非常有效的反爬虫策略封掉了IP，网页返回的状态码为302，解决方法:

(1)限制程序抓取频率，每隔几秒登录一次（这种方法工作量巨大）

(2)用多个账号、或者多台机器抓取，既处理了爬虫问题，也等同于做了分流处理，可以降低单台机器带宽压力（此方法在此次项目实现有点困难，不现实，仅供参考）

(3)使用代理IP，先写一个爬虫程序对免费代理IP的相关网站进行爬取。此爬虫程序较为复杂，这次选择的是使用在github上开源的程序(github上有详细的使用说明，以及需要的环境，非常方便，有效的解决了此次项目的问题)，虽然免费网站的代理IP，每一天能用的数量少，质量不高，爬取速度非常慢。但是也足以解决这次项目中IP被限制的问题

5.cookie

通过cookie限制抓取信息。 此问题在抓取博客园，csdn时均遇到，我们选择的是在伪造请求头时，因为headers里包含cookie，同时伪造cookie即可，通过浏览器的F12查看器，然后逐个尝试在整个过程都请求的URL(主要包括HTML、JS、XHR)，直到成功为止，一般情况下是最上方的HTML包含了我们所需要的cookie。

6.JS渲染

采用JS渲染页面。 有些时候返回的页面是由JS操作DOM得到，我们实际上拿不到正确数据。这个也是非常有效的反爬虫策略。我们选择使用selenium库模拟浏览器请求返回渲染完JS的页面。

在PHP方面，我是负责用MYSQL语言编写存储过程与触发器，并让PHP与MYSQL交互，从而使得形成前端->PHP->MYSQL动态的数据交互，完成可视化操作。

在这次完成大创项目的过程中，我熟悉了Python，PHP，MYSQL等三种语言。同时因为在对数据进行爬取后会大概的浏览一下内容,我对人工智能相关知识有了进一步的认识，同时也看到了一些比较优秀的文章。最难忘的便是暑假的时候，负责人把我们聚在一起。这是第一次体会到团队的感觉，大家在一起学习，互相交流思想，互相提醒监督，一起成长。非常感谢团队里的其它成员，尤其是负责人。规划了我们这个项目的基本方向，以及及时的与成员交流，推动项目的进行。在暑假里，大家都借了许多书籍，我个人也在负责人的帮助下，对scala以及spark有了一定的了解。而在暑假过后的那一学期，我们某项课程的实验便要求用Python来完成，得益于在大创的时候已经对Python使用的比较熟练了，实验完成的十分顺利。与此同时，我们凭借着这个项目参加了安徽省ican创新创业大赛，并最终获得了省二等奖。

这次大创不仅仅收获了素质方面的提高，能力的培养，更重要的是明白了团队协作的重要性。以及在一个团队里，每一个人都要付出，每个人都需要尽力为团队做出贡献。我们的团队只有三个人，很多另外的项目团队人数超过我们许多。但是我们还是最终完成了这个大创项目。非常感谢，以及非常荣幸能和两位同专业的同学成为队友。我从他们身上学习到了很多优秀的品质。无论是面对难题时的冷静思考，还是答辩场上的镇定自若。他们都在闪闪发光。