选题目的（为什么选该课题）：

互联网是一个非常庞大的数据库，各种各样的数据与资源以各种形式存储在互联网上，在网页展示形式也非常多：文字、图片、Flash动画或者视频等等。当用户试图在网页上获取一些信息并点击链接的时候，往往会收到大量外带的数据，例如广告推送、网页浮动窗口等，不但容易干扰视线，使用户较难筛选出有效信息，而且某些广告弹窗可能会暗藏一些恶意链接，使用户造成不必要的损失。本课题的主要目的是设计一个可以主动定向收集天气信息的网络爬虫程序，使用户试图去在网络上获取信息的时候，避开这些可能存在的“坑”，高效、快速的完成信息的获取，此程序还可支持用户进行指定城市的天气信息的订阅，被订阅的城市近期的天气数据会主动推送到用户端进行显示，并可根据历史数据，得到每日的降水率，进行一定程度的天气预测。在满足功能的同时还需要考虑到对程序性能上的要求。要考虑到网络爬虫的各种需求，以及网络爬虫主体网站的特性。网络爬虫使用 scrapy 实现多线程，让爬虫具备更强大的抓取能力和灵活性。

前期基础（已学课程、掌握的工具，资料积累、软硬件条件等）：

已学课程: Python程序设计，数据库原理及应用，数据结构，计算机操作系统，Unix网络编程，Unix环境高级编程

掌握的工具:Python3.5，PyQt，VS2017，Mysql

资料积累: ① 《Python核心编程》 （美）丘恩（Chun，W.J.）[人民邮电出版社](http://baike.baidu.com/item/%E4%BA%BA%E6%B0%91%E9%82%AE%E7%94%B5%E5%87%BA%E7%89%88%E7%A4%BE/905716" \t "http://baike.baidu.com/_blank) I S B N ： 9787115178503

②《Unix网络编程》 （美）史蒂文斯 人民邮电出版社 I S B N ： 9787115215093

软件条件:windows操作系统(客户端程序)，Linux操作系统(服务器程序)，PyQt，Mysql

硬件条件:个人开发机一台(内存8G,CPU:i5-4200H)，阿里云服务器一台(内存1G,单核CPU,带宽100M)

要解决的问题（做什么）：

1.开发环境的搭建，包括：Python环境的搭建以及PyQt、VS2017、Mysql等软件的安装。

2.熟练掌握Python编程以及Linux操作系统的常用命令。

3.了解网络爬虫技术研究现状，学习网络爬虫的技术原理。

4.设计数据在服务器的存储格式，与数据的传输格式。

5.对所爬取网站的主体内容的分析。

6.若网站有防爬策略，比如IP限频、IP黑名单等，该如何解决。

6.对网络爬虫工具性能进行测试，对工具功能进行完善。

工作思路和方案（怎么做）：

思路： 1、在网上搜索环境搭建教程。

2、研读并深入学习关于客户端-服务器模型下的网络编程知识。

3、仔细阅读有关端到端实时性传输的有关论文。

4、熟悉网络爬虫基本编程，学习了解HTTP协议。

5、编写服务器程序，并做性能测试。根据客户端的请求，从网络上爬取全国主要城市天气数据，保存结果至服务器。

6、编写客户端程序，确定目标，客户端向服务器发起请求，可从服务器得到数据。

工作方案：2017年2月27日至2017年3月15日 学习Python语言的程序设计方法和网络爬虫相关技术。

2017年3月16日至2017年4月30日 设计和实现爬虫程序

2017年5月1日至2017年5月14日 对网络爬虫工具性能进行测试。

2017年5月15日至2017年5月30日 撰写论文，完成论文初稿。

2017年6月1日至2017年6月7日 完善并修改毕业论文。

2017年6月8日至2017年6月17日 准备答辩。