大作业技术报告

姓名：胡文博，学号：1120111828，班级：07111102

一、 选取题目

论坛信息获取

二、选取网站

虎嗅网 http://www.huxiu.com/

二、 主要代码和运行环境

1， 针对虎嗅网<h4>标签，获取相应的 URl 和标题，进入获取其<meta>标签中 description 对象

2， 建立数据库 huxiu.db

3， 将用正则表达式匹配出的标题和内容存入数据库

4， 运行环境 Python2.7.6

5， 运用 Python 的 sched 库 schedule 来定时获取内容

三、 主要代码文件

1、Grab.py 主要抓取代码

2、TimingGrab.py 定时抓取代码

3、Read_DB.py 读取数据库内容代码

4、huxiu.db 数据库文件

四、遇到的问题及解决方法

1、数据写入数据库是出现 operationerror，最后将 content 里的双引号去掉就没有出现问题。用 replace()函数

五、运行前置条件

将三个文件放在一个文件下即可运行