技术路线：requests->re->BeautifulSoup->Sqlite

Python版本：2.7

爬虫简介：这是一个豆瓣读书爬虫，获取图书信息并写入如数据库。

实现思路：

先从某一门图书开始，在该图书页面爬取其他图书的URL地址，并放入队列，再从当前页面爬取该图书信息并插入数据库。为了防止对某一页面重复爬取，使用哈希表对需要爬取的页面进行检查，当某页面加入队列前，先在哈希表中对该页面进行查找，若在哈希表中查找到了该地址，则跳过该地址，若没有找到该地址，则该地址加入队列以及哈希表。

分析过程：

首先打开某一本图书页面的URL地址，<https://book.douban.com/subject/20432061/>，经过分析每一本图书的URL地址均为<https://book.douban.com/subject/XXXXX/>格式，其中XXX为一组长度不等的数字，由此可以得出该数字为该本图书的ID，可以用来当做数据库主键使用，在分析具体html页面，发现每本图书页面可能有其他相关图书的连接，由此可以使用正则表达式获取其他图书的ID，并依次爬取相关内容。

使用方法：

在main函数中实例化DoubanBookCrawler，调用startCrawl函数，参数中填写需要爬取页面的豆瓣ID，也就是爬虫爬取的起点，例如我要从<https://book.douban.com/subject/20432061/>页面开始爬取，则输入以下代码 crawler = DoubanBookCrawler()

crawler.startCrawl("20432061")