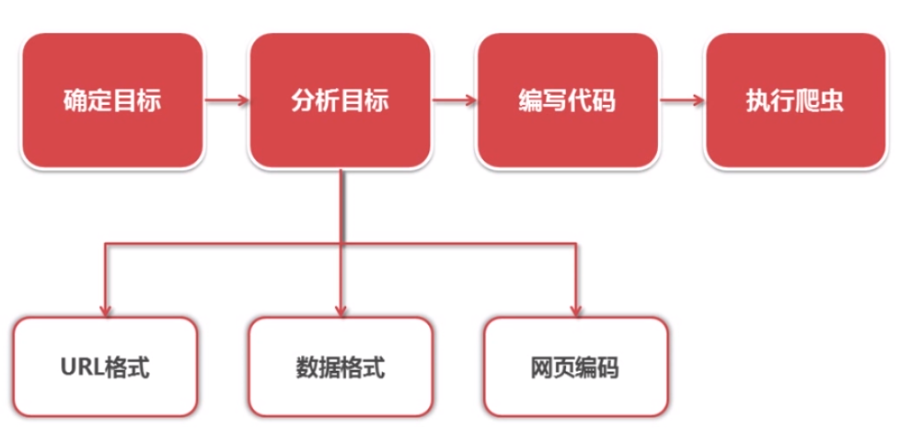
本次课程设计主要带领同学们设计实现一个简单的网页爬虫。首先，简单介绍爬虫的概念，然后，讲解一个网页爬虫的基础架构，接着，讲解爬虫的三大组成，URL管理器，网页下载器和网页解析器，最后，介绍一个完整的爬虫实例。

**实例**

演示如何真正开发实现一个爬虫。开发一个爬虫包含如下图所示的步骤，先确定目标，就是要爬取哪个网站的哪些网页的哪部分数据，本实例确定要爬取郑州轻工业大学新闻站点中新闻的标题，时间以及点击次数。然后，分析目标，就是需要制定爬取这个网站数据的策略，有三个部分需要分析，第一，要分析要抓取的目标页面的URL格式，用来限定要抓取的页面的范围，如果范围不限定的话，爬虫就会爬取很多跟我们目标不相干的网页，第二，要分析抓取数据的格式，本实例中就是要分析每个新闻页面中标题，日期和时间这些数据所在的标签格式，第三，要分析页面编码，在代码的网页解析器部分，需要指定网页的编码，然后才能进行正确的解析。然后，就是编写代码，执行爬虫。



打开爬虫入口页，也就是我校新闻网站首页，<http://news.zzuli.edu.cn/>。通过浏览器的审查元素工具，网页上点击右键，就出现审查元素菜单，可以确定URL格式，数据格式和网页编码等爬取策略信息。

目标：爬取我校新闻的标题，日期和点击次数

入口页：<http://news.zzuli.edu.cn/>

URL格式：

新闻页面URL：/2019/0613/c284a199931/page.htm

数据格式：

标题：<h1 class="arti-title">\*\*\*\*\*</h1>

日期：<span class="arti-update">时间：2019-06-11</span>

点击次数：<span class="WP\_VisitCount" ...>356</span>

页面编码：UTF-8

代码结构：

spider\_main.py 调度器

urls\_manager.py URL管理器

html\_downloader.py 网页下载器

html\_parser.py 网页解析器

html\_outputer.py 网页输出器，把爬取的有价值信息以HTML格式输出