任务一 技术综述

1. 爬虫步骤：
   下载网页内容，分析网页内容，整合下所需的数据输出
2. 代码组成：
   总调度函数
   url 管理器
   网页下载器
   内容解析器
   结果输出器
3. 大致实现：
   使用了 python，url 管理器是通过两个 set 实现，一个存储待爬的 url，一个存储已经爬过的 url，这样可以避免重复页面的爬取；
   网页下载器时通过 python 中一个库 urllib2 的 urlopen 函数来获取内容
   网页分析器，是用了一个比较低效的网页解析框架 beautifulsoup，原理是按照网页生成的 DOM 树来查找特定的元素