第二章：爬虫的实现原理和技术

- 爬虫的实现原理
  - 通用（全网）爬虫

    START-->初始URL-->获取网页-->分析网页、提取新的URL、存入队列-->判断是否满足条件循环-->结束
  - 聚焦（主题）爬虫

    START-->初始URL-->获取网页-->分析网页、提取新的URL、存入队列-->限定策略（分析算法、评价网页及URL）-->判断是否满足条件循环（循环同时增加限定策略）-->结束
  - 搜索引擎

    爬取网页-->数据存储-->预处理-->检索和排名（PageRank算法排序）
- 爬虫爬取页面的详细流程
  - 种子URL-->待爬取URL队列
  - DNS解析-->读取URL
  - 互联网-->下载网页
  - 网页内容-->抽取URL-->待爬取URL队列
  - 网页内容-->已爬取URL队列
  - 网页内容-->下载网页库
- 通用爬虫中的网页分类
  - 已下载网页
    - 未过期网页
    - 已过期网页
  - 待下载网页
  - 可知网页
  - 不可知网页
- 通用爬虫的相关网站文件
  - robots.txt文件
  - Sitemap.xml文件（小型数据库、存储文件）

    <lastmod>最近访问时间</lastmod>
- 防爬虫应对策略
  - 用户代理（User-Agent）
  - 代理IP
  - 降低访问频率
  - 验证码限制
- 选择Python做爬虫的原因
  - 爬取网页本身的接口（urllib2包提供）

- 网页爬取后的处理
- 开发效率高
- 上手快

以上内容整理于 幕布文档