

关于生成的统计结果文件的说明

1 基本说明

1.1 爬取范围(爬取页面)

指从起始 url 开始爬取, 当 url 符合界面上设置的限定域范围时, 该 url 属于爬取范围(爬取页面)。只有当 url 属于爬取范围时, 该页面才会被继续跟踪爬取。

1.2 抓取范围(下载条目)

url 在属于爬取范围的基础上, 若同时符合界面上设置的条目抓取规则, 则该 url 属于抓取范围(下载条目)。

2 基于抓取条目的分析

2.1 运行结果

本文档记录了爬虫的运行结果, 包括: 运行时间、请求页面数、响应页面数、响应字节数、响应成功页面数(200)、抓取条目数、成功下载条目数。

2.2 基于抓取条目的链接列表

2.2.1 所有链接的列表

针对属于抓取范围内的条目, 若该页面上存在链接(链接可以属于抓取范围内外、爬取范围内外, 即页面上的全部链接), 则列举出来。

2.2.2 抓取范围内链接列表


针对属于抓取范围内的条目, 若该页面上存在链接(链接必须属于抓取范围内), 则列举出来。

2.2.3 抓取范围外链接列表

针对属于抓取范围内的条目, 若该页面上存在链接(链接必须属于抓取范围外, 可以属于爬取范围内外), 则列举出来。

2.3 基于起始 url 对应站点的链接统计

2.3.1

基于爬取页面的分析	
运行结果	
基于抓取条目的链接列表	▶
基于起始url对应站点的链接统计	▶
基于限定域的链接统计	▶
基于抓取页面的链接统计	▶
在这里输入	
添加分隔符	