Web2.0爬虫说明:

高效多线程基于web2.0标准url爬取,基于webkit,并且抓取post请求包。

功能:

爬取方式:

带referer、cookie Html A标签多线程抽取

三种模式:

深度爬取

Json/cgi类站点爬取

伪静态站点爬取

Domian url去重保存

网络请求抓取

不抓取url黑名单

需要环境:

Python2.7 java mysql browsermobproxy PhantomJS

用到技术：

Python mysql

数据库建表:

\*\_List\_url:

id url url\_re state

**CREATE TABLE data\_url  
 (id integer primary key autoincrement ,  
 url TEXT NOT NULL,**

**url\_re TEXT NOT NULL,  
 state INT NOT NULL)**

\*\_net\_url

id url url\_re state method.

**CREATE TABLE** net\_url **(id integer primary key autoincrement ,  
 url TEXT NOT NULL,**

**url\_re TEXT NOT NULL,**

method **TEXT NOT NULL,  
 state INT NOT NULL)**

高效多线程基于web2.0标准url爬取,基于webkit,并且抓取network请求包。

爬取方式

带referer、cookie Html A标签多线程抽取

三种爬取方式:

深度爬取(while)

伪静态取重爬取(re\_while)

Jsonp去重爬取(json\_while)

三个结果：

爬虫爬取页面a标记

爬虫爬取domian

Webkit network