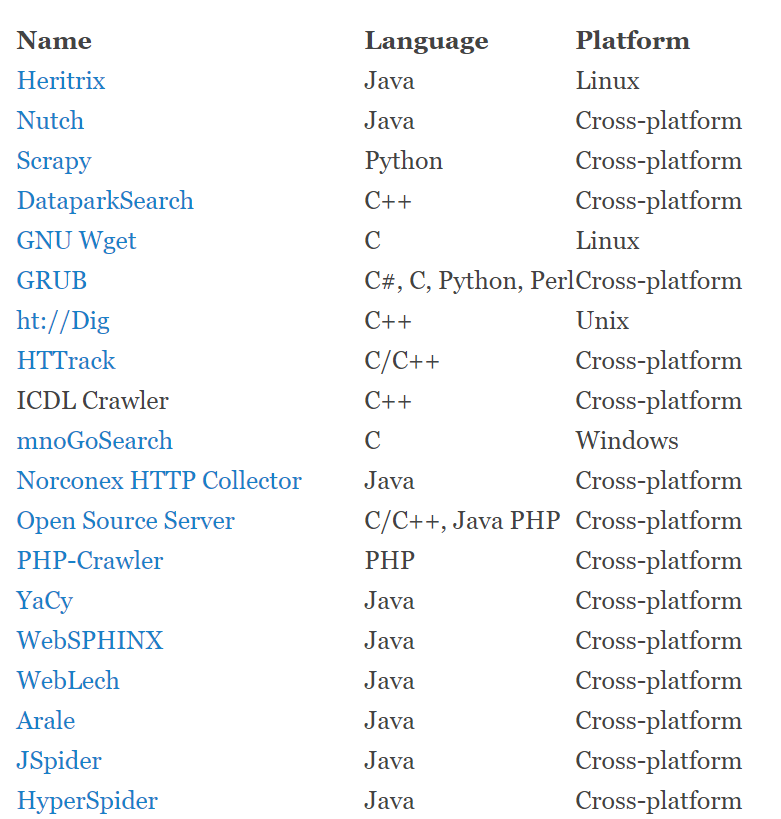
|  |
| --- |
|  |
| HW1.1 - 爬虫的调研 |
|  |
| PB15111604金泽文 |

[**Web Information Processing and Applications**](http://staff.ustc.edu.cn/~jpq/courses/webinfo.html)

HW1.1 - 爬虫的调研

PB15111604金泽文

1. 概述

当下主流的开源爬虫有如下图所示的几十个框架。

本次调研以[Heritrix](http://crawler.archive.org/)、[Nutch](http://nutch.apache.org/" \t "_blank)、[Scrapy](http://scrapy.org/" \t "_blank)为例。

1. 特点、局限性与区别
   1. Heritrix:
      1. 用java开发，开源、可扩展，严格按照robots.txt的指示和META robots标签，可用来获取完整的、精确的、站点内容的深度复制。
      2. 局限性与差别：照单全收，解析少，配置麻烦，功能没有Nutch多，不支持分布式爬取。针对获取网页完整的内容。
   2. Nutch:
      1. 用java开发，开源、功能多，类似于google搜索引擎，基于Hadoop任务的分布式处理模型，总体分为抓取和搜索两个部分。存储层剥离，支持存储HBase, Cassandra, MySql等数据库；基于插件式设计，扩展和定制比较方便；支持网页解析和索引，可以对接至Solr，搭建通用的搜索引擎。
      2. 局限性与差别：效率没有Heritrix高， 功能多。针对搜索引擎功能。
   3. Scrapy:
      1. 用python开发，开源、可扩展，爬虫规则定制简单，支持抓取和抽取，数据抽取结构化抽取支持xpath和css提取网页数据。
      2. 局限性与差别：单机多线程实现，不支持分布式，默认中间过程的网页不保存，只保存抽取结果。针对网站的结构性数据，用于数据挖掘与信息处理。