Crawling Framework

Open Source

Apache Nutch

Apache Nutch

Programming Language – Java

Pros

Highly extensible and Flexible system for web crawling Implements search when combined with open source search platforms like Apache Lucene or Apache Solr Dynamically scalable with Hadoop

Cons

Difficult to setup Poor documentation Some operations take longer, as the size of crawler grows

Heritrix

- Heritrix

Programming Language – Java

Pros

Excellent user documentation and easy setup Extensible, good performance and decent support for distributed crawls Respects robot.txt

Cons

Not dynamically scalable

출처

- https://www.scrapehero.com/best-web-crawling-tools-and-frameworks/

관련 문서

Plugin Backlinks: 아무 것도 없습니다.