jsoup

作者：Kagura

原文链接：https://ld246.com/article/1492683499413

来源网站：链滴

jsoup: Java HTML Parser

jsoup is a Java library for working with real-world HTML. It provides a very convenient API for extracting and manipulating data, using the best of DOM, CSS, and jquery-like methods.

jsoup implements the WHATWG HTML5 specification, and parses HTML to the same DOM as modern browsers do.

- scrape and parse HTML from a URL, file, or string
- find and extract data, using DOM traversal or CSS selectors
- manipulate the HTML elements, attributes, and text
- clean user-submitted content against a safe white-list, to prevent XSS attacks
- output tidy HTML

jsoup is designed to deal with all varieties of HTML found in the wild; from pristine and validat ng, to invalid tag-soup; jsoup will create a sensible parse tree.

See **jsoup.org** for downloads and the full API documentation.

Example

Fetch the Wikipedia homepage, parse it to a DOM, and select the headlines from the In the N ws section into a list of Elements (online sample):

```
Document doc = Jsoup.connect("http://en.wikipedia.org/").get();
Elements newsHeadlines = doc.select("#mp-itn b a");
```

Open source

jsoup is an open source project distributed under the liberal MIT license. The source code is a ailable at GitHub.

Getting started

1. Download the latest jsoup jar (or it add to your Maven/Gradle build)
2. Read the cookbook
3. Enjoy!

Development and support

If you have any questions on how to use jsoup, or have ideas for future development, please et in touch via the mailing list.

If you find any issues, please file a bug after checking for duplicates.

The colophon talks about the history of and tools used to build jsoup.

Status

原文链接：jsoup

jsoup is in general, stable release.

原文链接：jsoup