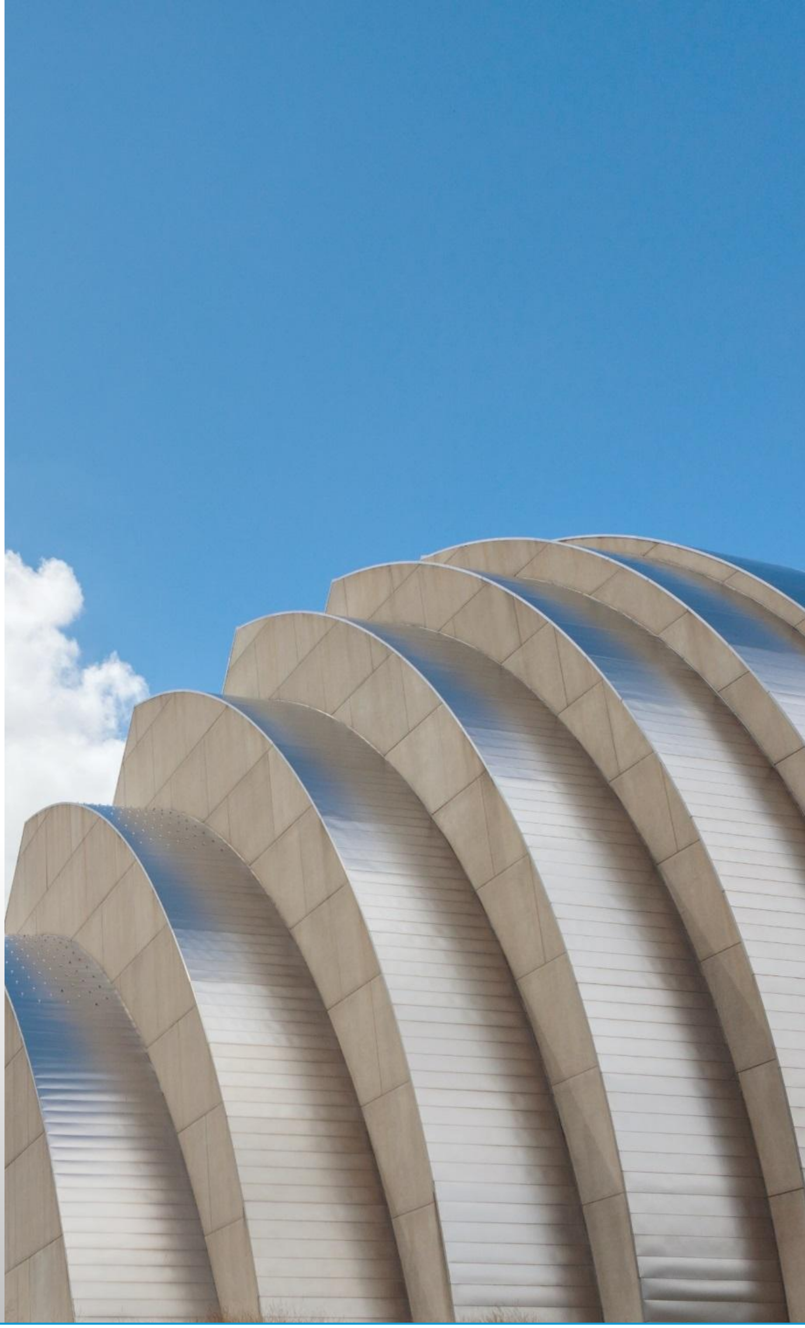
Scrape and Verify

Jsoup

Nick Gover
Sr. Software Quality Engineer | NIPR

09/20/2018



What is Jsoup?

- Jsoup is an open source Java library
  - used mainly for extracting data from HTML
- Convenient API for extracting and manipulating data
- It has a steady development line
- It has great documentation
- A fluent and flexible API
- Jsoup can also be used to parse and build XML.

Why Jsoup?

- Jsoup is an open source Java library used mainly for extracting data from HTML. It provides a very convenient API for extracting and manipulating data, using the best of DOM, CSS, and jquery-like methods.

- It has a steady development line, great documentation, and a fluent and flexible API. Jsoup can also be used to parse and build XML.

Getting started with Jsoup

- Compatible with JVM projects
  - Java, Groovy, Kotlin, Scala
- Build with Ant, Maven or Gradle
- Single import statement

How Does jsoup work?

- Jsoup loads the page HTML and builds the corresponding DOM tree. This tree works the same way as the DOM in a browser, offering methods similar to jQuery and vanilla JavaScript to select, traverse, manipulate text/HTML/attributes and add/remove elements.

What is the DOM?

- The Document Object Model (DOM) is a programming interface for HTML and XML documents. It represents the page so that programs can change the document structure, style, and content. The DOM represents the document as nodes and objects.
- A Web page is a document. This document can be either displayed in the browser window or as the HTML source.

Traversing the DOM

- Documents consist of Elements and Nodes
- The inheritance chain is:
  - Document extends Element extends Node.
  - TextNode extends Node.
- An Element contains a list of children Nodes, and has one parent Element. They also have provide a filtered list of child Elements only

What if my HTML is invalid?

- The parser will make every attempt to create a clean parse from the HTML you provide, regardless of whether the HTML is well-formed or not. It handles…
  - unclosed tags
  - implicit tags
  - reliably creating the document structure

Unclosed Tags

- <p>Lorem <p>Ipsum
  parses to
- <p>Lorem</p> <p>Ipsum</p>

Implicit Tags

- A naked <td>Table data</td>

is wrapped into

- a <table><tr><td>

Reliably Creating the Document

- Will create the html containing a <head> and <body>, and only create appropriate elements within the <head>

JSoup Demo

Parse HTML From String

- Jsoup makes it easy to parse HTML in a String that is acquired from various sources.
- Use the static Jsoup.parse(String html) method
- or…
- Jsoup.parse(String html, String baseUri) if the page came from the web
- The parse(String html, String baseUri) method parses the input HTML into a new Document.

Parse HTML From a URL

- Jsoup makes it easy to parse HTML in a String that is acquired from various sources.
- Use the static Jsoup.parse(String html) method
- or…
- Jsoup.parse(String html, String baseUri) if the page came from the web
- The parse(String html, String baseUri) method parses the input HTML into a new Document.

Modify Elements

Problem

You have a parsed document that you would like to update attribute values on, before saving it out to disk, or sending it on as a HTTP response.

Solution

Use the attribute setter methods Element.attr(String key, String value), and Elements.attr(String key, String value).

Sanitize Unsafe HTML

- Perhaps you have a website that allows users to leave comments with HTML formatting.
- Malicious users can use this feature to send cross-site scripting attacks (XSS).

How Does the Sanitizer Work?

- The jsoup whitelist sanitizer works by parsing the input HTML (in a safe, sand-boxed environment), and then iterating through the parse tree and only allowing known-safe tags and attributes (and values) through into the cleaned output.

What will you do with JSoup?

Learn More...

- Homepage
  - https://jsoup.org/
- Downloads
  - https://jsoup.org/download
- API Docs
  - https://jsoup.org/apidocs/overview-summary.html
- JSoup Cookbook
  - https://jsoup.org/cookbook/
- MoTKC Github Repo
  - https://github.com/seleniumkc/jsoup-java-gradle