Wrangling Court Data on a National Level

Michael Lissner, Founder & Lead Developer, Juriscraper
LVI 2012, 2012-03-15

Access to case law has recently become easier than ever: By simply visiting a court's website it is now possible to find and read thousands of cases without ever leaving your home. At the same time, there are nearly a hundred court websites, many of these websites suffer from poor funding or prioritization, and gaining a higher-level view of the law can be challenging. "Juriscraper" is a new project designed to ease these problems for all those that wish to collect these court opinions daily.[1] The project is under active development, and we are looking for others to get involved.

Juriscraper is a liberally-licensed[2] open source library that can be picked up and used by any organization to scrape the case data from court websites. In addition to a simply scraping the websites and extracting metadata from them, Juriscraper has a number of other design goals:

- Extensibility to support video, oral argument audio, and other media types
- Support for all metadata provided by court websites
- Extensibility to support varied geographies and jurisdictions
- Generalized object-oriented architecture with little or no code repetition
- Standardized coding techniques using the latest libraries and standards (Python, xpath, lxml, requests, chardet)
- Simple installation, configuration, and API
- Friendly and transparent to court websites

As well as a number of features:

- Harmonization of metadata (US, USA, United States of America, etc → United States; et al, et. al., etc. get eliminated; vs., v, vs → v.; all dates are Python objects; etc.)
- Smart title-casing of case names (several courts provide case names in uppercase only)
- Sanity checking and sorting of metadata values returned by court websites

Once implemented, Juriscraper is part of a two-part system. The second part is the caller, which uses the API, and which itself solves some interesting questions:

- How are duplicates detected and avoided?
- How can the impact on court websites be minimized?
- How can mime type detection be completed successfully so that textual contents can be extracted?
    - What should we do if it is an image-based PDF?
    - How should HTML be tidied?
- How often should we check a court website for new content?
- What should we do in case of failure?

Juriscraper is currently deployed by CourtListener.com to scrape all of the Federal Appeals courts, and we are slowly adding additional state courts over the coming weeks.

We have been scraping these sites in various ways for several years, and Juriscraper is the culmination of what we've learned. We hope that by presenting our work at LVI 2012, we will be able to share what we have learned and gain additional collaborators in our work.

---

1 Juriscraper is available online at: https://bitbucket.org/mlissner/juriscraper/

2 Juriscraper is licensed under the BSD license: http://www.opensource.org/licenses/bsd-license.php

Presenter Biography

Michael Lissner is the lead developer and founder of CourtListener.com. He has a diverse background that includes literature, technology, policy, and law. When he's not behind a computer working on a project, he's usually out doing long-distance hiking or biking, sleeping in a snow cave or summiting a mountain.