# Mining Web Pages

Social web APIs that return data conforming to a well-defined schema are essential, but the most basic currency of human communication is natural language data such as the words that you are reading on this page, Facebook posts, web pages linked into tweets, and so forth. Human language is by far the most ubiquitous kind of data available to us, and the future of data-driven innovation depends largely upon our ability to effectively harness machines to understand digital forms of human communication.

## Scraping, Parsing, and Crawling the Web

Although it's trivial to use a programming language or terminal utility such as curl or wget to fetch an arbitrary web page, extracting the isolated text that you want from the page isn't quite as trivial. Although the text is certainly in the page, so is lots of other boilerplate content such as navigation bars, headers, footers, advertisements, and other sources of noise that you probably don't care about. Hence, the bad news is that the problem isn't quite as simple as just stripping out the HTML tags and processing the text that is left behind, because the removal of HTML tags would have done nothing to remove the boilerplate itself. In some cases, there may actually be more boilerplate in the page that contributes noise than the signal you were looking for in the first place.

The good news is that the tools for helping to identify the content you're interested in have continued to mature over the years, and there are some excellent options for isolating the material that you'd want for text-mining purposes. Additionally, the relative ubiquity of feeds such as RSS and Atom can often aid the process of retrieving clean text without all of the cruft that's typically in web pages, if you have the foresight to fetch the feeds while they are available.

One excellent tool for web scraping (the process of extracting text from a web page) is the Java-based boilerpipe library, which is designed to identify and remove the boilerplate from web pages. The boilerpipe library is based on a published paper entitled "Boilerplate Detection Using Shallow Text Features," which explains the efficacy of using supervised machine learning techniques to bifurcate the boilerplate and the content of the page. Supervised machine learning techniques involve a process that creates a predictive model from training samples that are representative of its domain, and thus, boilerpipe is customizable should you desire to tweak it for increased accuracy.

Even though the library is Java-based, it's useful and popular enough that a Python package wrapper called python-boilerpipe is available.. Installation of this package is predictable: use pip install boilerpipe. Assuming you have a relatively recent version of Java on your system, that's all that should be required to use boilerpipe.

## Crawling

Crawling websites is a logical extension of the same concepts already presented in this section: it typically consists of fetching a page, extracting the hyperlinks in the page, and then systematically fetching all of those pages that are hyperlinked. This process is repeated to an arbitrary depth, depending on your objective. This very process is the way that the earliest search engines used to work, and the way most search engines that index the Web still continue to work today. Although a crawl of the Web is far outside our scope, it is helpful to have a working knowledge of the problem, so let’s briefly think about the computational complexity of harvesting all of those pages.