**TOP 7 PYTHON WEB SCRAPING TOOLS FOR DATA SCIENTISTS**

12/11/2019

In this article, we list down the top seven web scraping frameworks in Python.

(The list is in alphabetical order)

**1) Beautiful Soup**

Beautiful Soup is a Python library for pulling data out of HTML and XML files. It is mainly designed for projects like screen-scraping. This library provides simple methods and Pythonic idioms for navigating, searching, and modifying a parse tree. This tool automatically converts incoming documents to Unicode and outgoing documents to UTF-8.

**2) LXML**

The lxml is a Python tool for C libraries libxml2 and libxslt. It is recognised as one of the feature-rich and easy-to-use libraries for processing XML and HTML in Python language. It is unique in the case that it combines the speed and XML feature of these libraries with the simplicity of a native Python API and is mostly compatible but superior to the well-known ElementTree\_API.

**3) MechanicalSoup**

MechanicalSoup is a Python library for automating interaction with websites. This library automatically stores and sends cookies, follows redirects and can follow links and submit forms. MechanicalSoup provides a similar API, built on Python giants Requests (for HTTP sessions) and BeautifulSoup (for document navigation). However, this tool became unmaintained for several years as it didn’t support Python 3.

**4) Python Requests**

Python Requests is the only Non-GMO HTTP library for Python language. It allows the user to send HTTP/1.1 requests and there is no need to manually add query strings to your URLs, or to form-encode your POST data. There are a number of feature support such as browser-style SSL verification, automatic decompression, automatic content decoding, HTTP(S) proxy support and much more. Requests officially support Python 2.7 & 3.4–3.7 and runs on PyPy.

**5) Scrapy**

Scrapy is an open-source and collaborative framework for extracting the data a user needs from websites. Written in Python language, Scrapy is a fast high-level web crawling & scraping framework for Python. It can be used for a wide range of purposes, from data mining to monitoring and automated testing. It is basically an application framework for writing web spiders that crawl web sites and extract data from them. Spiders are the classes that a user defines and Scrapy uses the Spiders to scrape information from a website (or a group of websites).

**6) Selenium**

Selenium Python is an open-source web-based automation tool which provides a simple API to write functional or acceptance tests using Selenium WebDriver. Selenium is basically a set of different software tools each with a different approach to supporting test automation. The entire suite of tools results in a rich set of testing functions specifically geared to the needs of testing of web applications of all types. With the help of Selenium Python API, a user can access all functionalities of Selenium WebDriver in an intuitive way. The currently supported Python versions are 2.7, 3.5 and above.

**7) Urllib**

The urllib is a Python package which can be used for opening URLs. It collects several modules for working with URLs such as urllib.request for opening and reading URLs which are mostly HTTP, urllib.error module defines the exception classes for exceptions raised by urllib.request, urllib.parse module defines a standard interface to break Uniform Resource Locator (URL) strings up in components and urllib.robotparser provides a single class, RobotFileParser, which answers questions about whether or not a particular user agent can fetch a URL on the Web site that published the robots.txt file.