Community Experience Distilled

Getting Started with Beautiful Soup

Build your own web scraper and learn all about web scraping with Beautiful Soup

Vineeth G. Nair

[PACKT] PUBLISHING open source* community experience distilled

Getting Started with Beautiful Soup

Build your own web scraper and learn all about web scraping with Beautiful Soup

Vineeth G. Nair

[PACKT] open source
community experience distilled
PUBLISHING

BIRMINGHAM - MUMBAI

Getting Started with Beautiful Soup

First published: January 2014

Production Reference: 1170114

Published by Packt Publishing Ltd.

Livery Place
35 Livery Street
Birmingham B3 2PB, UK.

ISBN 978-1-78328-955-4

www.packtpub.com

Cover Image by Mohamed Raoof (raoofpmajeed@gmail.com)

Credits

Author
Vineeth G. Nair

Reviewers
John J. Czaplewski
Christian S. Perone
Zhang Xiang

Acquisition Editor
Nikhil Karkal

Senior Commissioning Editor
Kunal Parikh

Commissioning Editor
Manasi Pandire

Technical Editors
Novina Kewalramani
Pooja Nair

Copy Editor
Janbal Dharmaraj

Project Coordinator
Jomin Varghese

Proofreader
Maria Gould

Indexer
Hemangini Bari

Graphics
Sheetal Aute
Abhinash Sahu

Production Coordinator
Adonia Jones

Cover Work
Adonia Jones

About the Author

Vineeth G. Nair completed his bachelors in Computer Science and Engineering from Model Engineering College, Cochin, Kerala. He is currently working with Oracle India Pvt. Ltd. as a Senior Applications Engineer.

He developed an interest in Python during his college days and began working as a freelance programmer. This led him to work on several web scraping projects using Beautiful Soup. It helped him gain a fair level of mastery on the technology and a good reputation in the freelance arena. He can be reached at `vineethgnair.mec@gmail.com`. You can visit his website at `www.kochi-coders.com`.

My sincere thanks to Leonard Richardson, the primary author of Beautiful Soup. I would like to thank my friends and family for their great support and encouragement for writing this book. My special thanks to Vijitha S. Menon, for always keeping my spirits up, providing valuable comments, and showing me the best ways to bring this book up. My sincere thanks to all the reviewers for their suggestions, corrections, and points of improvement.

I extend my gratitude to the team at Packt Publishing who helped me in making this book happen.

About the Reviewers

John J. Czaplewski is a Madison, Wisconsin-based mapper and web developer who specializes in web-based mapping, GIS, and data manipulation and visualization. He attended the University of Wisconsin – Madison, where he received his BA in Political Science and a graduate certificate in GIS. He is currently a Programmer Analyst for the UW-Madison Department of Geoscience working on data visualization, database, and web application development. When not sitting behind a computer, he enjoys rock climbing, cycling, hiking, traveling, cartography, languages, and nearly anything technology related.

Christian S. Perone is an experienced Pythonista, open source collaborator, and the project leader of Pyevolve, a very popular evolutionary computation framework chosen to be part of OpenMDAO, which is an effort by the NASA Glenn Research Center. He has been a programmer for 12 years, using a variety of languages including C, C++, Java, and Python. He has contributed to many open source projects and loves web scraping, open data, web development, machine learning, and evolutionary computation. Currently, he lives in Porto Alegre, Brazil.

Zhang Xiang is an engineer working for the Sina Corporation.

I'd like to thank my girlfriend, who supports me all the time.

---

www.PacktPub.com

Support files, eBooks, discount offers and more

You might want to visit `www.PacktPub.com` for support files and downloads related to your book.

Did you know that Packt offers eBook versions of every book published, with PDF and ePub files available? You can upgrade to the eBook version at `www.PacktPub.com` and as a print book customer, you are entitled to a discount on the eBook copy. Get in touch with us at `service@packtpub.com` for more details.

At `www.PacktPub.com`, you can also read a collection of free technical articles, sign up for a range of free newsletters and receive exclusive discounts and offers on Packt books and eBooks.

`http://PacktLib.PacktPub.com`

Do you need instant solutions to your IT questions? PacktLib is Packt's online digital book library. Here, you can access, read and search across Packt's entire library of books.

Why Subscribe?

- Fully searchable across every book published by Packt
- Copy and paste, print and bookmark content
- On demand and accessible via web browser

Free Access for Packt account holders

If you have an account with Packt at `www.PacktPub.com`, you can use this to access PacktLib today and view nine entirely free books. Simply use your login credentials for immediate access.

Table of Contents

Preface

Web scraping is now widely used to get data from websites. Whether it be e-mails, contact information, or selling prices of items, we rely on web scraping techniques as they allow us to collect large data with minimal effort, and also, we don't require database or other backend access to get this data as they are represented as web pages.

Beautiful Soup allows us to get data from HTML and XML pages. This book helps us by explaining the installation and creation of a sample website scraper using Beautiful Soup. Searching and navigation methods are explained with the help of simple examples, screenshots, and code samples in this book. The different parser support offered by Beautiful Soup, supports for scraping pages with encodings, formatting the output, and other tasks related to scraping a page are all explained in detail. Apart from these, practical approaches to understanding patterns on a page, using the developer tools in browsers will enable you to write similar scrapers for any other website.

Also, the practical approach followed in this book will help you to design a simple web scraper to scrape and compare the selling prices of various books from three websites, namely, Amazon, Barnes and Noble, and PacktPub.

What this book covers

Chapter 1, *Installing Beautiful Soup*, covers installing Beautiful Soup 4 on Windows, Linux, and Mac OS, and verifying the installation.

Chapter 2, *Creating a BeautifulSoup Object*, describes creating a `BeautifulSoup` object from a string, file, and web page; discusses different objects such as `Tag`, `NavigableString`, and parser support; and specifies parsers that scrape XML too.

Chapter 3, *Search Using Beautiful Soup*, discusses in detail the different search methods in Beautiful Soup, namely, `find()`, `find_all()`, `find_next()`, and `find_parents()`; code examples for a scraper using search methods to get information from a website; and understanding the application of search methods in combination.

Chapter 4, *Navigation Using Beautiful Soup*, discusses in detail the different navigation methods provided by Beautiful Soup, methods specific to navigating downwards and upwards, and sideways, to the previous and next elements of the HTML tree.

Chapter 5, *Modifying Content Using Beautiful Soup*, discusses modifying the HTML tree using Beautiful Soup, and the creation and deletion of HTML tags. Altering the HTML tag attributes is also covered with the help of simple examples.

Chapter 6, *Encoding Support in Beautiful Soup*, discusses the encoding support in Beautiful Soup, creating a `BeautifulSoup` object for a page with specific encoding, and the encoding supports for output.

Chapter 7, *Output in Beautiful Soup*, discusses formatted and unformatted printing support in Beautiful Soup, specifications of different formatters to format the output, and getting just text from an HTML page.

Chapter 8, *Creating a Web Scraper*, discusses creating a web scraper for three websites, namely, Amazon, Barnes and Noble, and PacktPub, to get the book selling price based on ISBN. Searching and navigation methods used to create the parser, use of developer tools so as to identify the patterns required to create the parser, and the full code sample for scraping the mentioned websites are also explained in this chapter.

What you need for this book

You will need Python Version 2.7.5 or higher and Beautiful Soup Version 4 for this book.

For *Chapter 3*, *Search Using Beautiful Soup* and *Chapter 8*, *Creating a Web Scraper*, you must have an Internet connection to scrape different websites using the code examples provided.

Who this book is for

This book is for beginners in web scraping using Beautiful Soup. Knowing the basics of Python programming (such as functions, variables, and values), and the basics of HTML, and CSS, is important to follow all of the steps in this book. Even though it is not mandatory, knowledge of using developer tools in browsers such as Google Chrome and Firefox will be an advantage when learning the scraper examples in chapters 3 and 8.