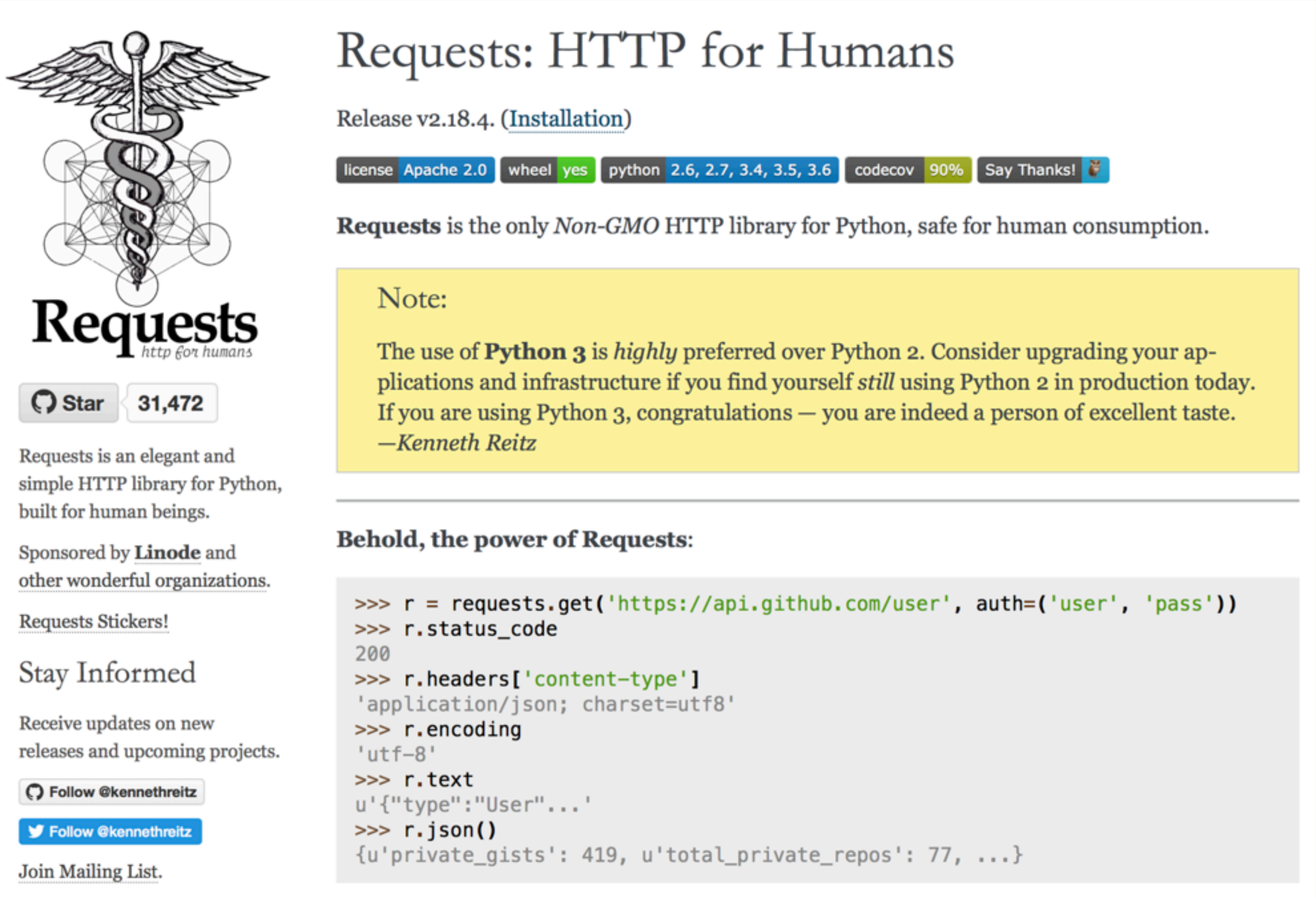
Scraping with Selenium

KnoxPy
April 5, 2018

Gavin Wiggins
gavinw.me

Requests

http://python-requests.org



Requests
http for humans

Star 31,472

Requests is an elegant and simple HTTP library for Python, built for human beings.

Sponsored by **Linode** and other wonderful organizations.

Requests Stickers!

Stay Informed

Receive updates on new releases and upcoming projects.

Follow @kennethreitz

Follow @kennethreitz

Join Mailing List.

Requests: HTTP for Humans

Release v2.18.4. (Installation)

license Apache 2.0 | wheel yes | python 2.6, 2.7, 3.4, 3.5, 3.6 | codecov 90% | Say Thanks!

Requests is the only *Non-GMO* HTTP library for Python, safe for human consumption.

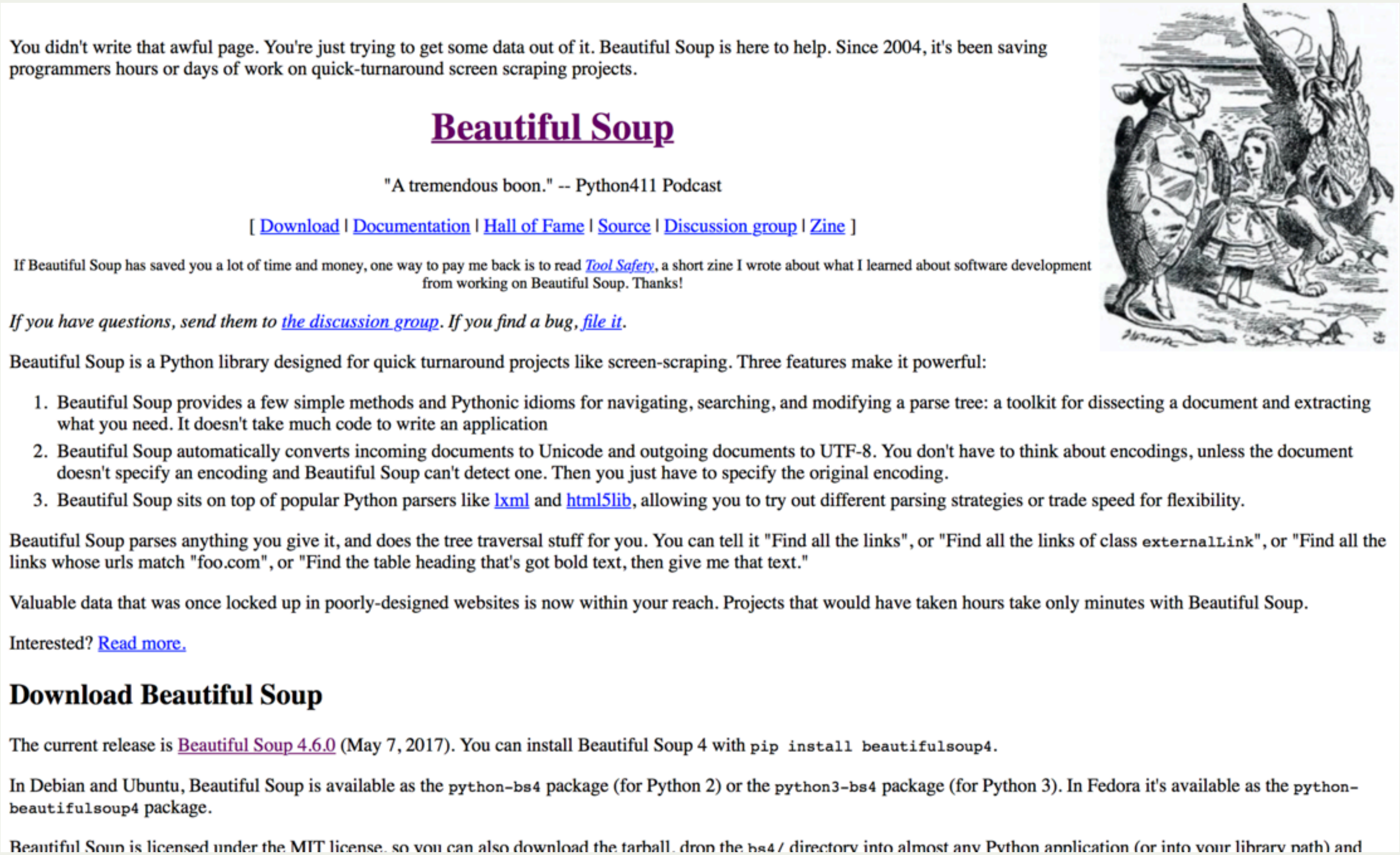
> Note:
>
> The use of **Python 3** is *highly* preferred over Python 2. Consider upgrading your applications and infrastructure if you find yourself *still* using Python 2 in production today. If you are using Python 3, congratulations — you are indeed a person of excellent taste. —*Kenneth Reitz*

Behold, the power of Requests:

```
>>> r = requests.get('https://api.github.com/user', auth=('user', 'pass'))
>>> r.status_code
200
>>> r.headers['content-type']
'application/json; charset=utf8'
>>> r.encoding
'utf-8'
>>> r.text
u'{"type":"User"...'
>>> r.json()
{u'private_gists': 419, u'total_private_repos': 77, ...}
```

Beautiful Soup

https://www.crummy.com/software/BeautifulSoup/



You didn't write that awful page. You're just trying to get some data out of it. Beautiful Soup is here to help. Since 2004, it's been saving programmers hours or days of work on quick-turnaround screen scraping projects.

Beautiful Soup

"A tremendous boon." -- Python411 Podcast

[Download | Documentation | Hall of Fame | Source | Discussion group | Zine]

If Beautiful Soup has saved you a lot of time and money, one way to pay me back is to read *Tool Safety*, a short zine I wrote about what I learned about software development from working on Beautiful Soup. Thanks!

If you have questions, send them to the discussion group. If you find a bug, file it.

Beautiful Soup is a Python library designed for quick turnaround projects like screen-scraping. Three features make it powerful:

1. Beautiful Soup provides a few simple methods and Pythonic idioms for navigating, searching, and modifying a parse tree: a toolkit for dissecting a document and extracting what you need. It doesn't take much code to write an application
2. Beautiful Soup automatically converts incoming documents to Unicode and outgoing documents to UTF-8. You don't have to think about encodings, unless the document doesn't specify an encoding and Beautiful Soup can't detect one. Then you just have to specify the original encoding.
3. Beautiful Soup sits on top of popular Python parsers like lxml and html5lib, allowing you to try out different parsing strategies or trade speed for flexibility.

Beautiful Soup parses anything you give it, and does the tree traversal stuff for you. You can tell it "Find all the links", or "Find all the links of class `externalLink`", or "Find all the links whose urls match "foo.com", or "Find the table heading that's got bold text, then give me that text."

Valuable data that was once locked up in poorly-designed websites is now within your reach. Projects that would have taken hours take only minutes with Beautiful Soup.

Interested? Read more.

Download Beautiful Soup

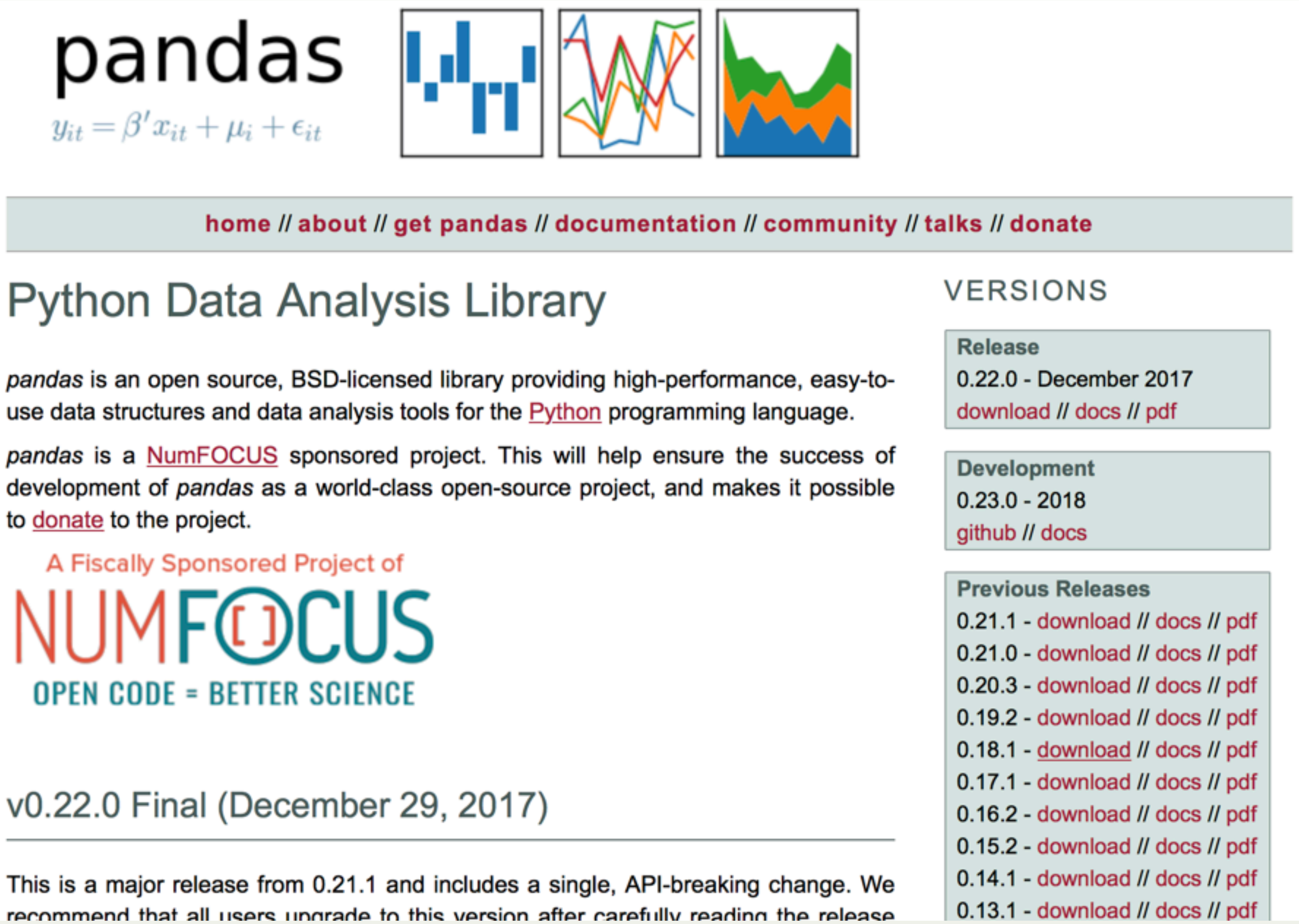
The current release is Beautiful Soup 4.6.0 (May 7, 2017). You can install Beautiful Soup 4 with `pip install beautifulsoup4`.

In Debian and Ubuntu, Beautiful Soup is available as the `python-bs4` package (for Python 2) or the `python3-bs4` package (for Python 3). In Fedora it's available as the `python-beautifulsoup4` package.

Beautiful Soup is licensed under the MIT license, so you can also download the tarball, drop the `bs4/` directory into almost any Python application (or into your library path) and

Pandas

https://pandas.pydata.org



pandas

$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$

home // about // get pandas // documentation // community // talks // donate

Python Data Analysis Library

pandas is an open source, BSD-licensed library providing high-performance, easy-to-use data structures and data analysis tools for the Python programming language.

pandas is a NumFOCUS sponsored project. This will help ensure the success of development of *pandas* as a world-class open-source project, and makes it possible to donate to the project.

A Fiscally Sponsored Project of NUMFOCUS OPEN CODE = BETTER SCIENCE

v0.22.0 Final (December 29, 2017)

This is a major release from 0.21.1 and includes a single, API-breaking change. We recommend that all users upgrade to this version after carefully reading the release

VERSIONS

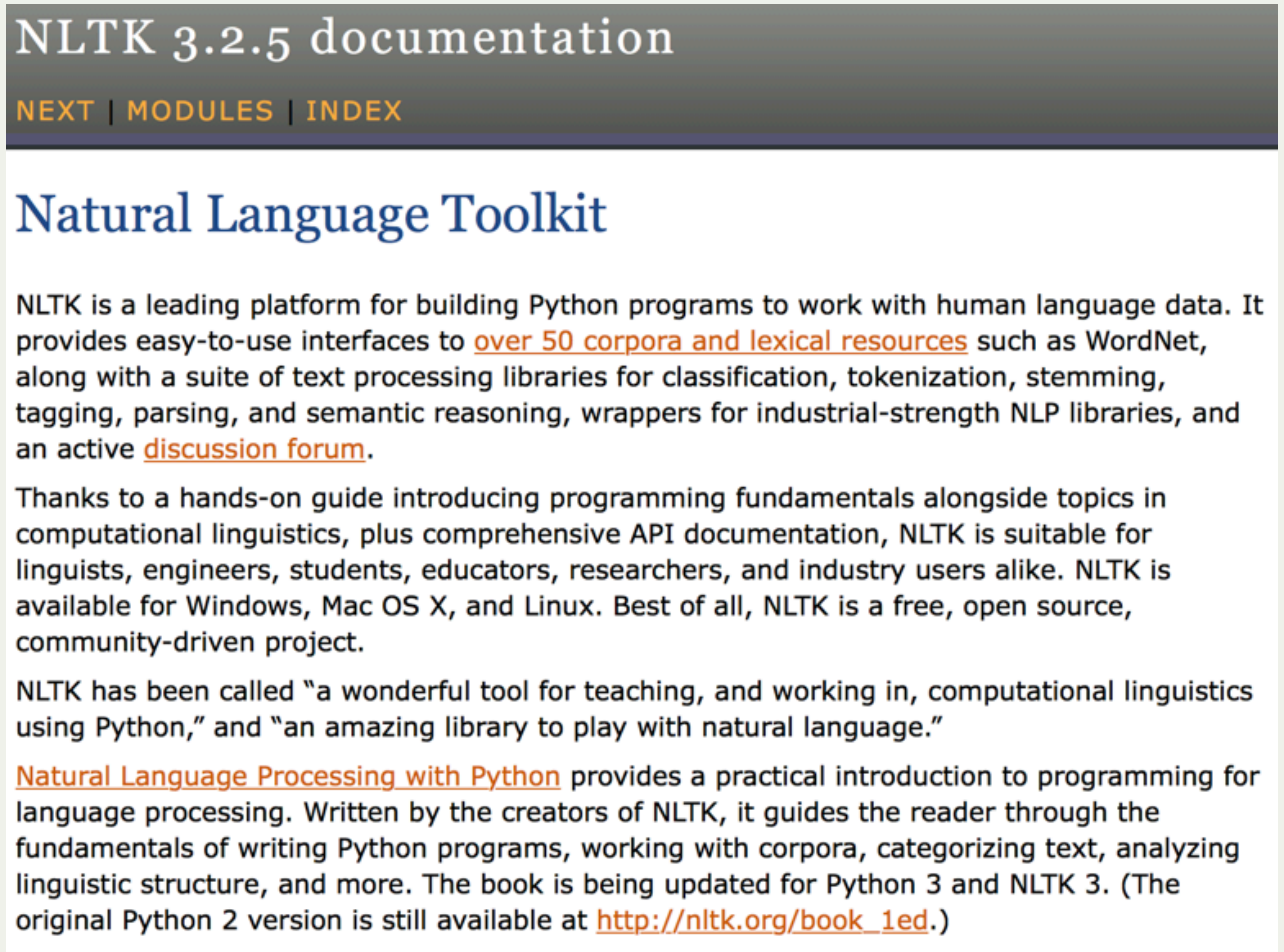
Release
0.22.0 - December 2017
download // docs // pdf

Development
0.23.0 - 2018
github // docs

Previous Releases
0.21.1 - download // docs // pdf
0.21.0 - download // docs // pdf
0.20.3 - download // docs // pdf
0.19.2 - download // docs // pdf
0.18.1 - download // docs // pdf
0.17.1 - download // docs // pdf
0.16.2 - download // docs // pdf
0.15.2 - download // docs // pdf
0.14.1 - download // docs // pdf
0.13.1 - download // docs // pdf

NLTK

http://www.nltk.org



NLTK 3.2.5 documentation

Natural Language Toolkit

NLTK is a leading platform for building Python programs to work with human language data. It provides easy-to-use interfaces to over 50 corpora and lexical resources such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning, wrappers for industrial-strength NLP libraries, and an active discussion forum.

Thanks to a hands-on guide introducing programming fundamentals alongside topics in computational linguistics, plus comprehensive API documentation, NLTK is suitable for linguists, engineers, students, educators, researchers, and industry users alike. NLTK is available for Windows, Mac OS X, and Linux. Best of all, NLTK is a free, open source, community-driven project.

NLTK has been called “a wonderful tool for teaching, and working in, computational linguistics using Python,” and “an amazing library to play with natural language.”

Natural Language Processing with Python provides a practical introduction to programming for language processing. Written by the creators of NLTK, it guides the reader through the fundamentals of writing Python programs, working with corpora, categorizing text, analyzing linguistic structure, and more. The book is being updated for Python 3 and NLTK 3. (The original Python 2 version is still available at http://nltk.org/book_1ed.)

Selenium

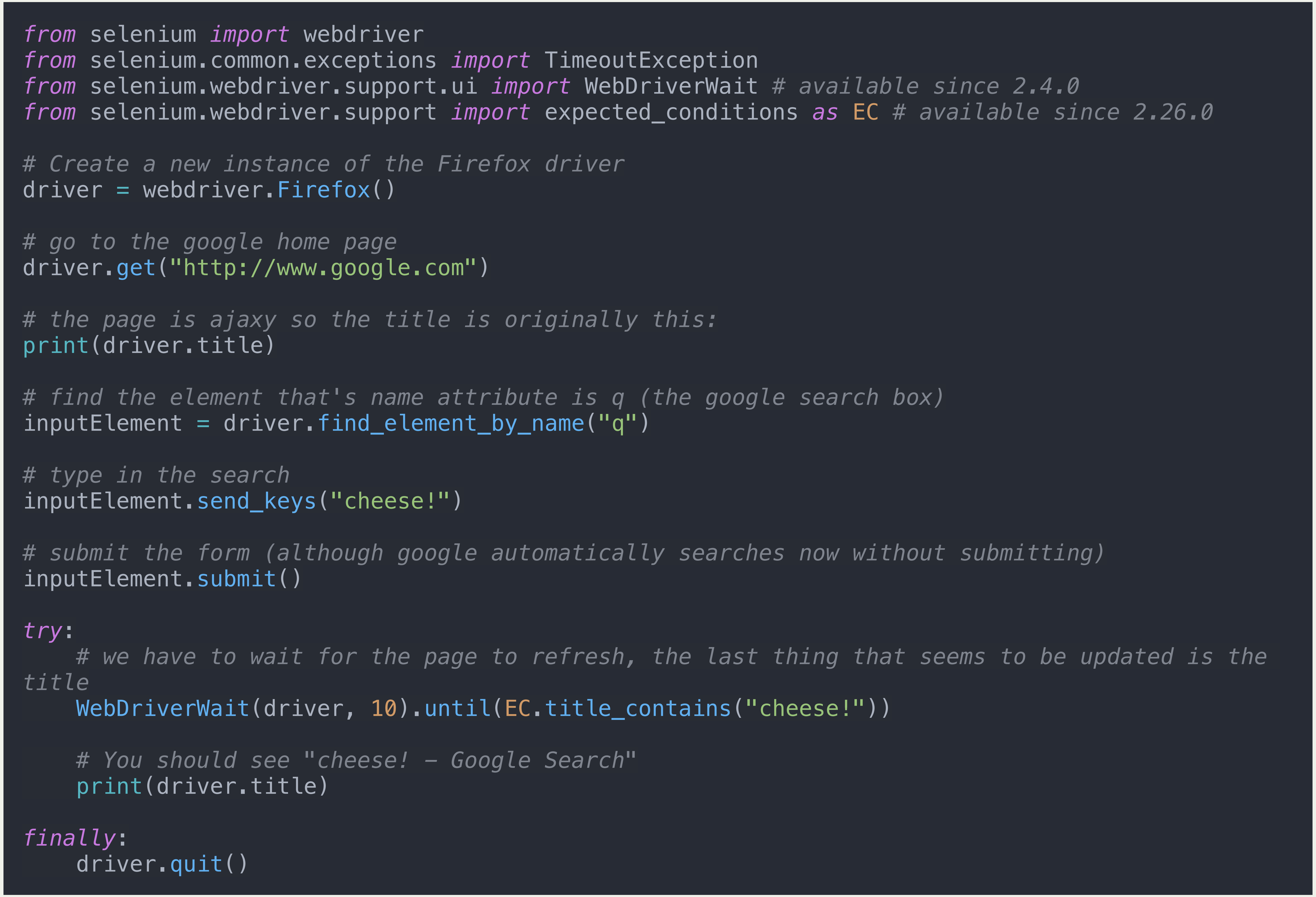
https://www.seleniumhq.org

Selenium WebDriver is a collection of bindings to drive a browser

- Operates a web browser natively just like a user would
- Language bindings available for Java, C#, Ruby, Python, Javascript

Selenium Grid runs tests on many servers at the same time

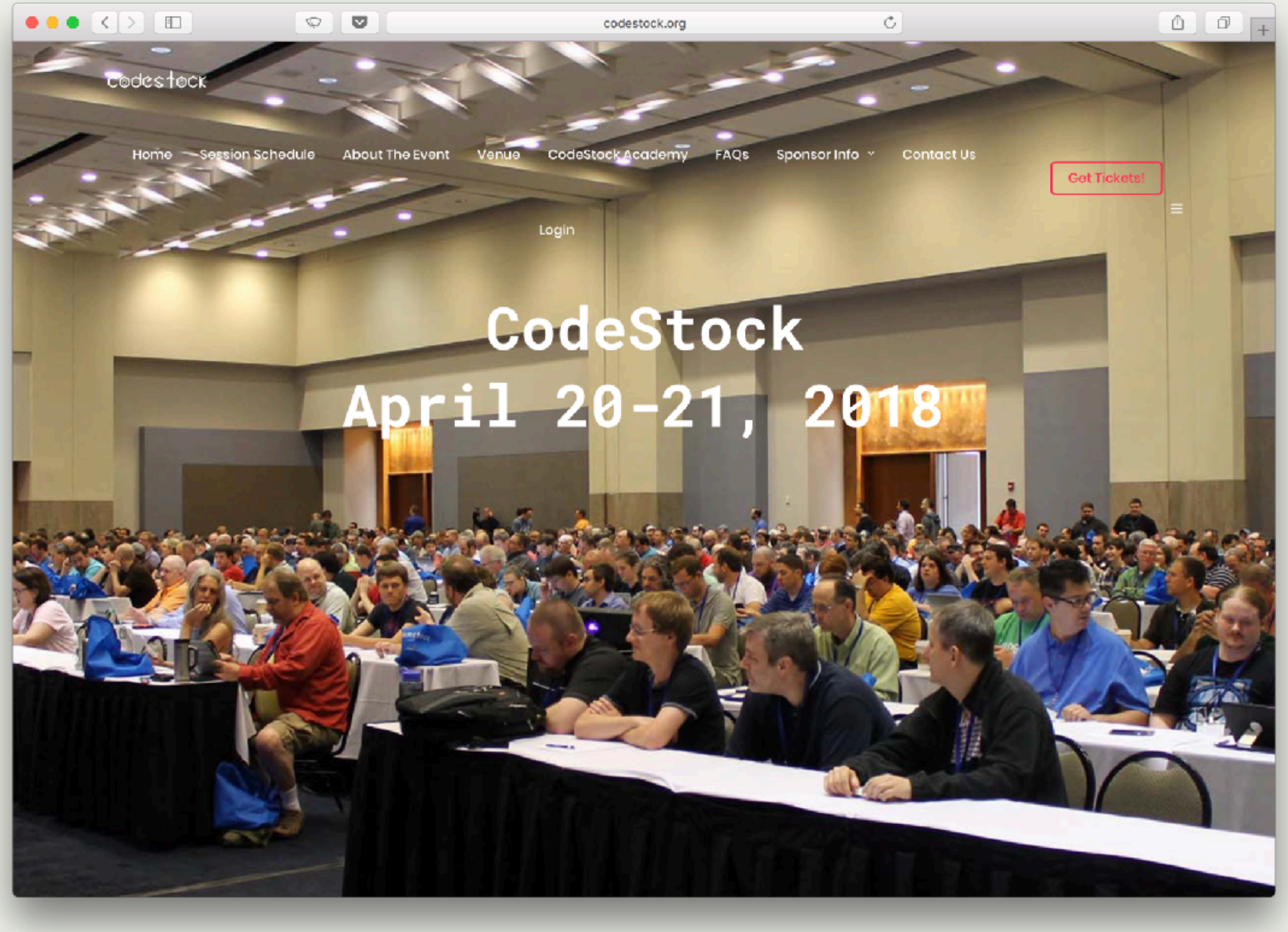
- Selenium IDE is a Firefox add-on to record and play back test
- **Selenium Remote Control** is a client/server system to control web browsers locally or remotely



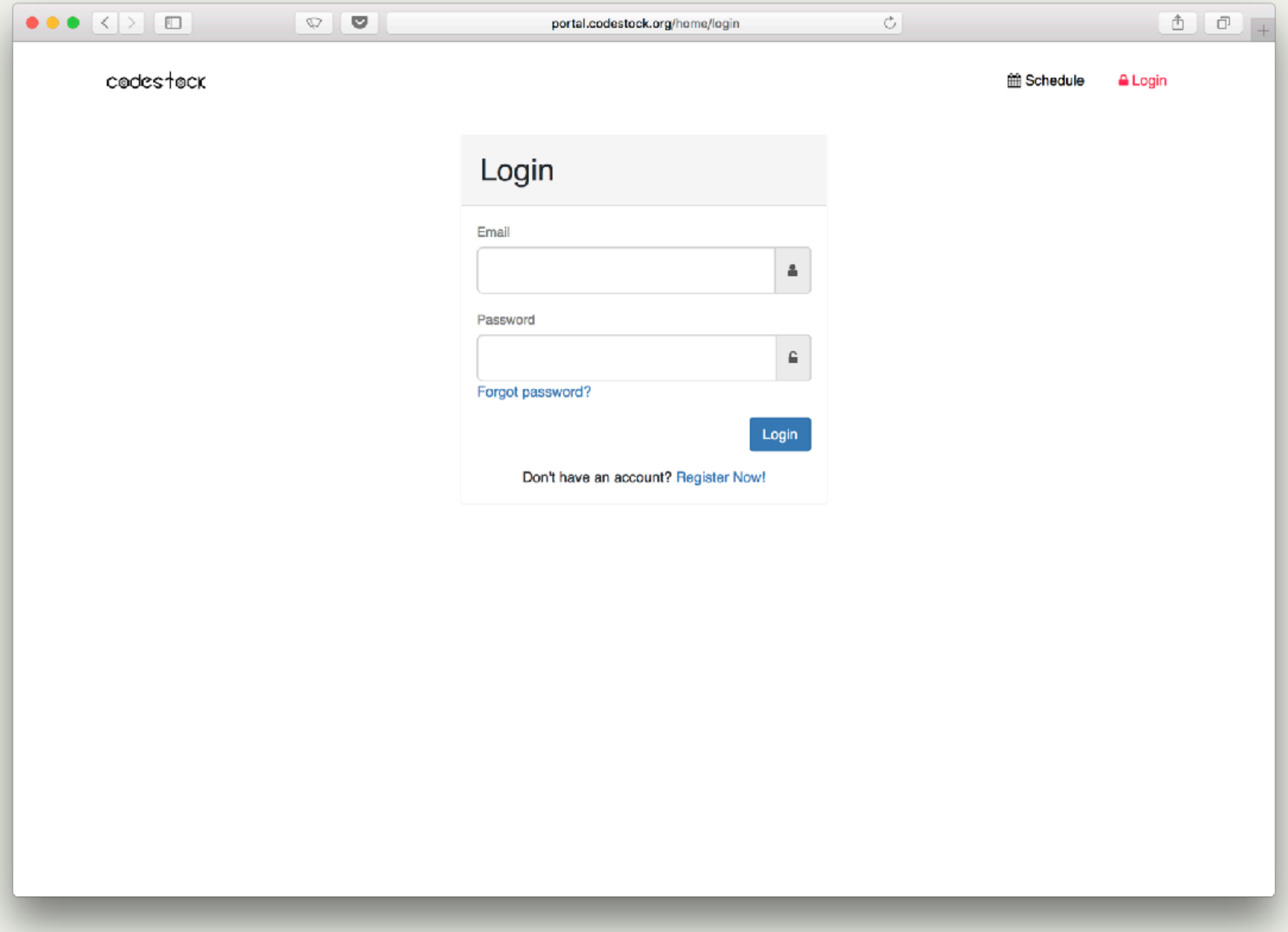
```python
from selenium import webdriver
from selenium.common.exceptions import TimeoutException
from selenium.webdriver.support.ui import WebDriverWait # available since 2.4.0
from selenium.webdriver.support import expected_conditions as EC # available since 2.26.0

# Create a new instance of the Firefox driver
driver = webdriver.Firefox()

# go to the google home page
driver.get("http://www.google.com")

# the page is ajaxy so the title is originally this:
print(driver.title)

# find the element that's name attribute is q (the google search box)
inputElement = driver.find_element_by_name("q")

# type in the search
inputElement.send_keys("cheese!")

# submit the form (although google automatically searches now without submitting)
inputElement.submit()

try:
    # we have to wait for the page to refresh, the last thing that seems to be updated is the
title
    WebDriverWait(driver, 10).until(EC.title_contains("cheese!"))

    # You should see "cheese! - Google Search"
    print(driver.title)

finally:
    driver.quit()
```

What else can we do with Selenium?

Scrape the ~~CodeStock~~ WebStock site



Must login to view submissions



portal.codestock.org/home/login

Login

Email

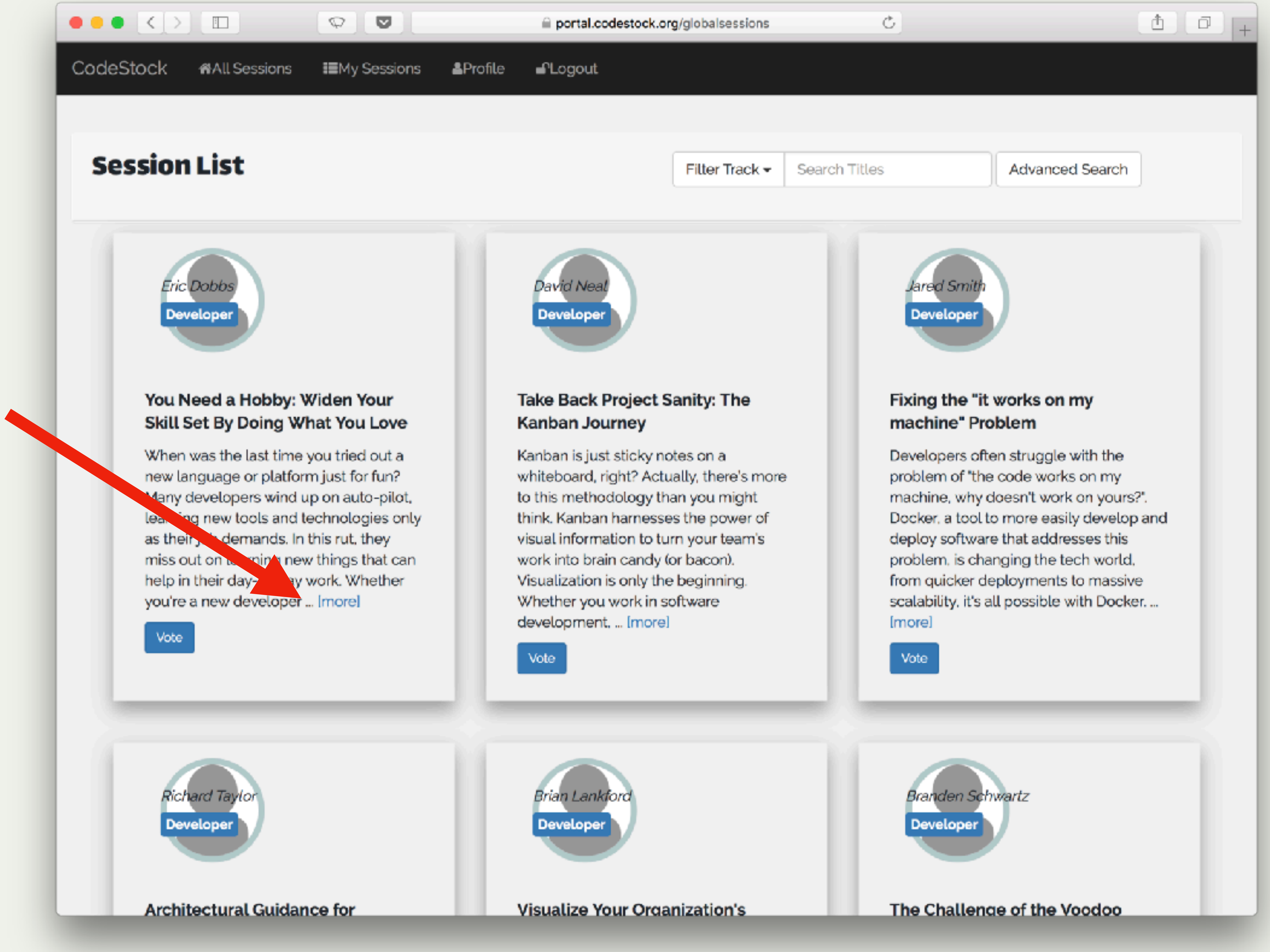
Password

Forgot password?

Don't have an account? Register Now!

Submissions page

Click the “more” button to view full abstract.



Session List

Eric Dobbs
Developer

You Need a Hobby: Widen Your Skill Set By Doing What You Love

When was the last time you tried out a new language or platform just for fun? Many developers wind up on auto-pilot, learning new tools and technologies only as their job demands. In this rut, they miss out on learning new things that can help in their day-to-day work. Whether you're a new developer ... [more]

Vote

David Neal
Developer

Take Back Project Sanity: The Kanban Journey

Kanban is just sticky notes on a whiteboard, right? Actually, there's more to this methodology than you might think. Kanban harnesses the power of visual information to turn your team's work into brain candy (or bacon). Visualization is only the beginning. Whether you work in software development, ... [more]

Vote

Jared Smith
Developer

Fixing the "it works on my machine" Problem

Developers often struggle with the problem of "the code works on my machine, why doesn't work on yours?". Docker, a tool to more easily develop and deploy software that addresses this problem, is changing the tech world, from quicker deployments to massive scalability, it's all possible with Docker. ... [more]

Vote

Richard Taylor
Developer

Architectural Guidance for

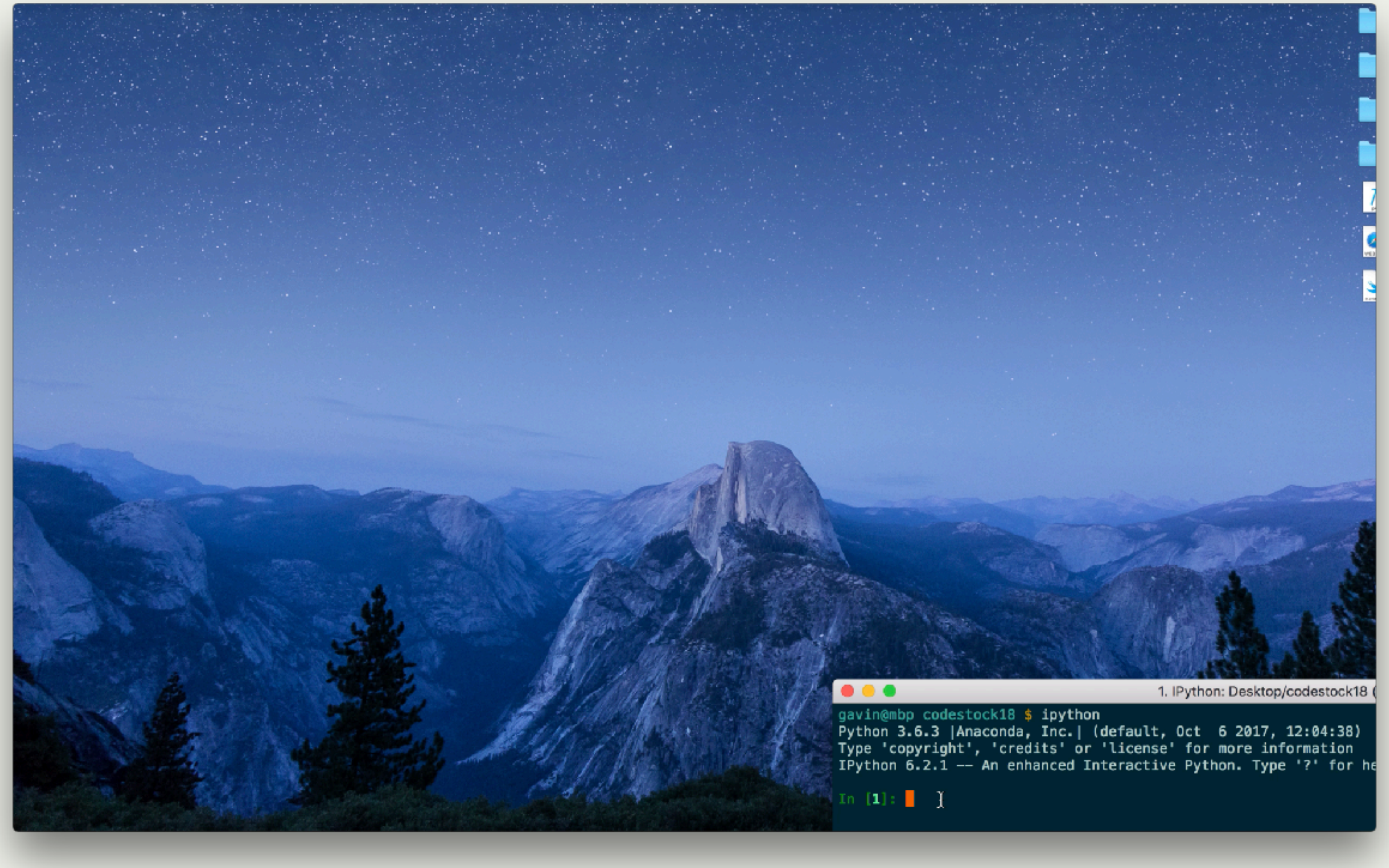
Brian Lankford
Developer

Visualize Your Organization's

Branden Schwartz
Developer

The Challenge of the Voodoo

Video of scraping the abstracts



```
gavin@mbp codestock18 $ ipython
Python 3.6.3 |Anaconda, Inc.| (default, Oct  6 2017, 12:04:38)
Type 'copyright', 'credits' or 'license' for more information
IPython 6.2.1 -- An enhanced Interactive Python. Type '?' for he

In [1]:
```

Demo …

Summary

Submissions

- Number of submissions = 370
- Max submissions per speaker = 15
- Most popular track = Developer
- Most common key words = Azure, .NET, ASP.NET, Angular, and SQL

Lineup

- Number of accepted talks = 89
- Max talks per speaker = 2
- Most popular track = ?
- Most common key words = .NET, C#, SQL, Elm, and ASP.NET

CodeStock is still WebStock :(