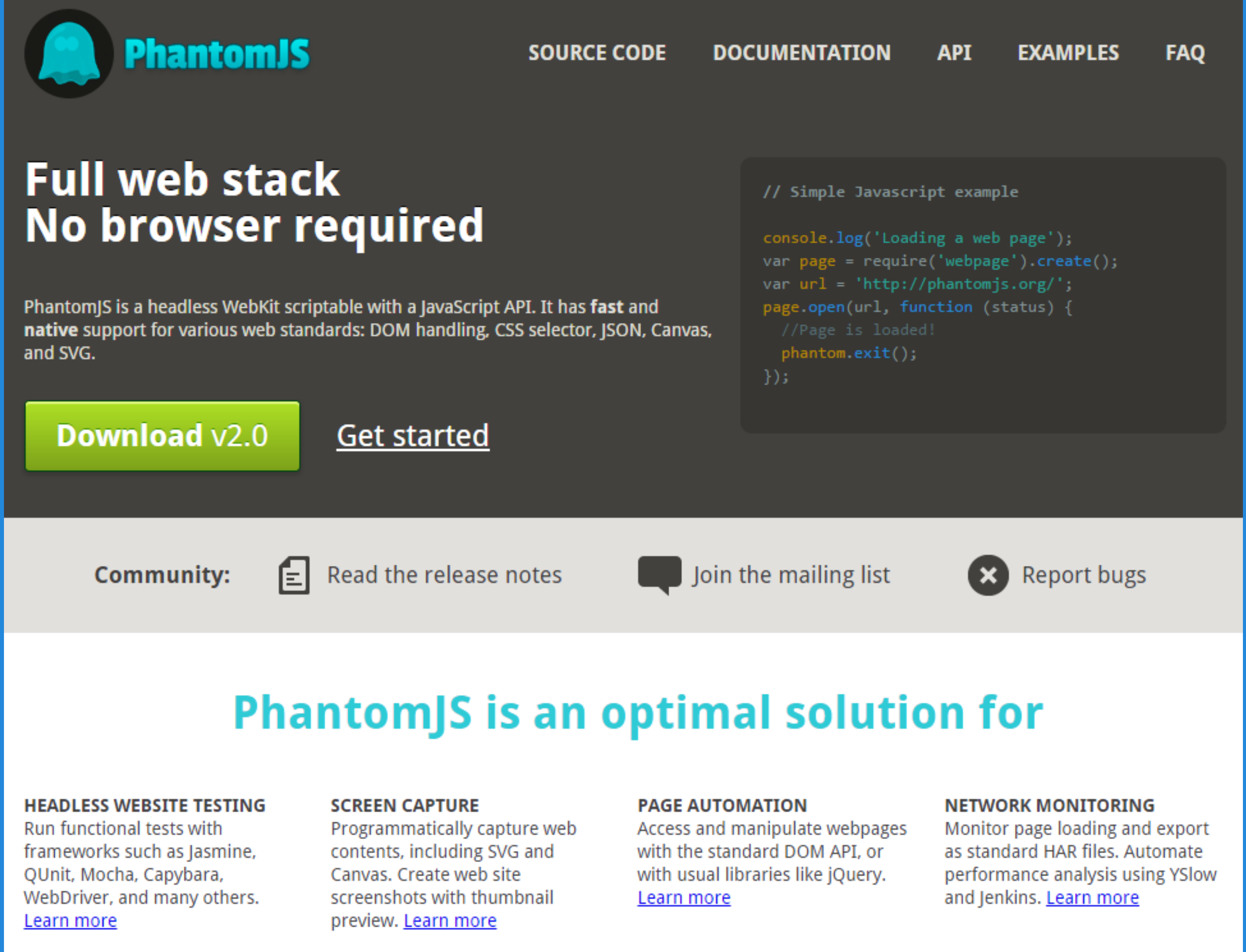
# Web Scraping

Python, PhantomJS, & Selenium



# Full web stack
# No browser required

PhantomJS is a headless WebKit scriptable with a JavaScript API. It has **fast** and **native** support for various web standards: DOM handling, CSS selector, JSON, Canvas, and SVG.

```
// Simple Javascript example

console.log('Loading a web page');
var page = require('webpage').create();
var url = 'http://phantomjs.org/';
page.open(url, function (status) {
  //Page is loaded!
  phantom.exit();
});
```

**Community:** Read the release notes · Join the mailing list · Report bugs

## PhantomJS is an optimal solution for

**HEADLESS WEBSITE TESTING**
Run functional tests with frameworks such as Jasmine, QUnit, Mocha, Capybara, WebDriver, and many others. Learn more

**SCREEN CAPTURE**
Programmatically capture web contents, including SVG and Canvas. Create web site screenshots with thumbnail preview. Learn more

**PAGE AUTOMATION**
Access and manipulate webpages with the standard DOM API, or with usual libraries like jQuery. Learn more

**NETWORK MONITORING**
Monitor page loading and export as standard HAR files. Automate performance analysis using YSlow and Jenkins. Learn more

# Why PhantomJS?

Headless WebKit Browser

Runs JavaScript

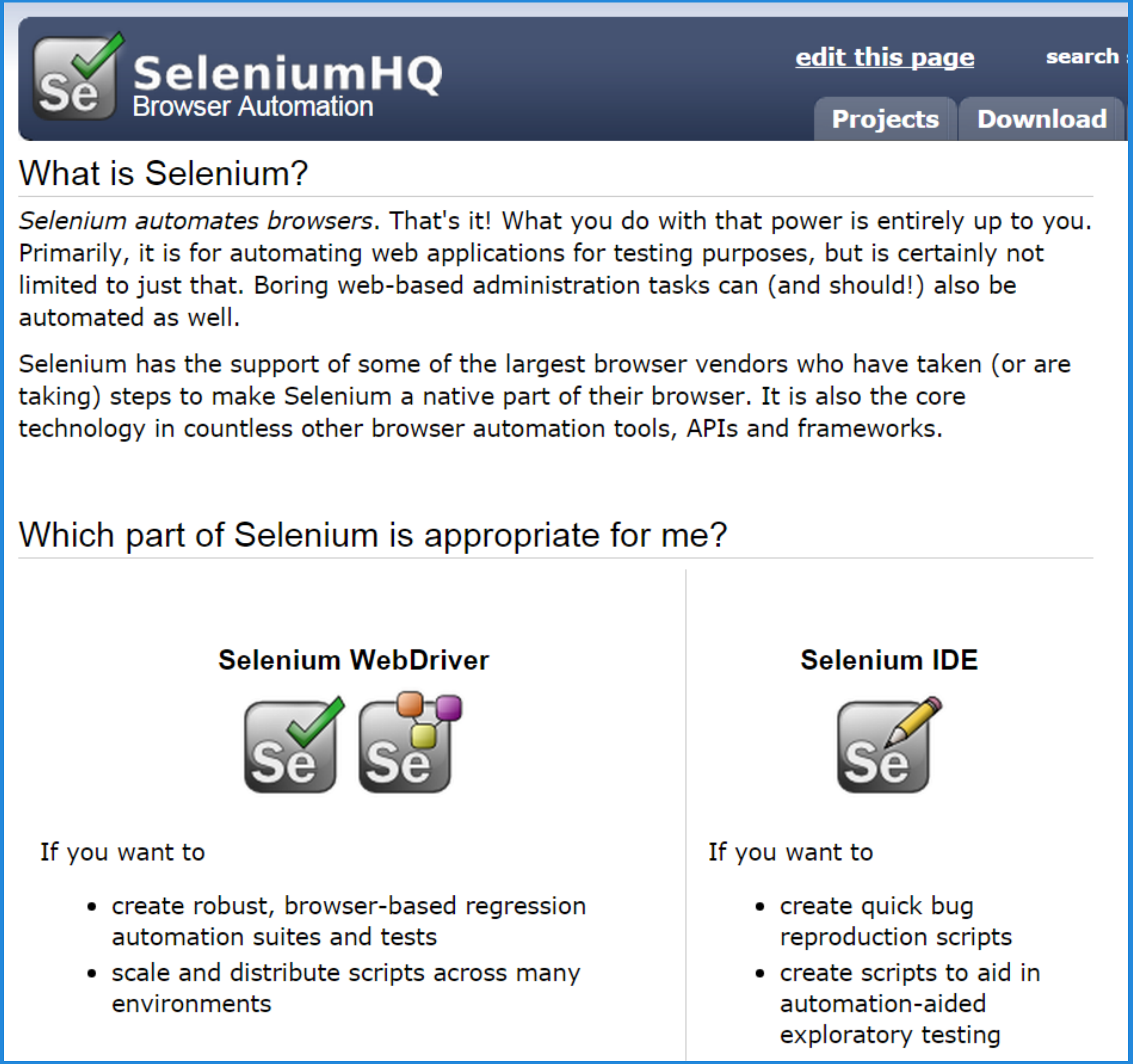
Inject JavaScript

Interact with the page (forms, etc)

Take screenshots

# Selenium automates browsers

http://docs.seleniumhq.org/



## What is Selenium?

*Selenium automates browsers*. That's it! What you do with that power is entirely up to you. Primarily, it is for automating web applications for testing purposes, but is certainly not limited to just that. Boring web-based administration tasks can (and should!) also be automated as well.

Selenium has the support of some of the largest browser vendors who have taken (or are taking) steps to make Selenium a native part of their browser. It is also the core technology in countless other browser automation tools, APIs and frameworks.

## Which part of Selenium is appropriate for me?

**Selenium WebDriver**

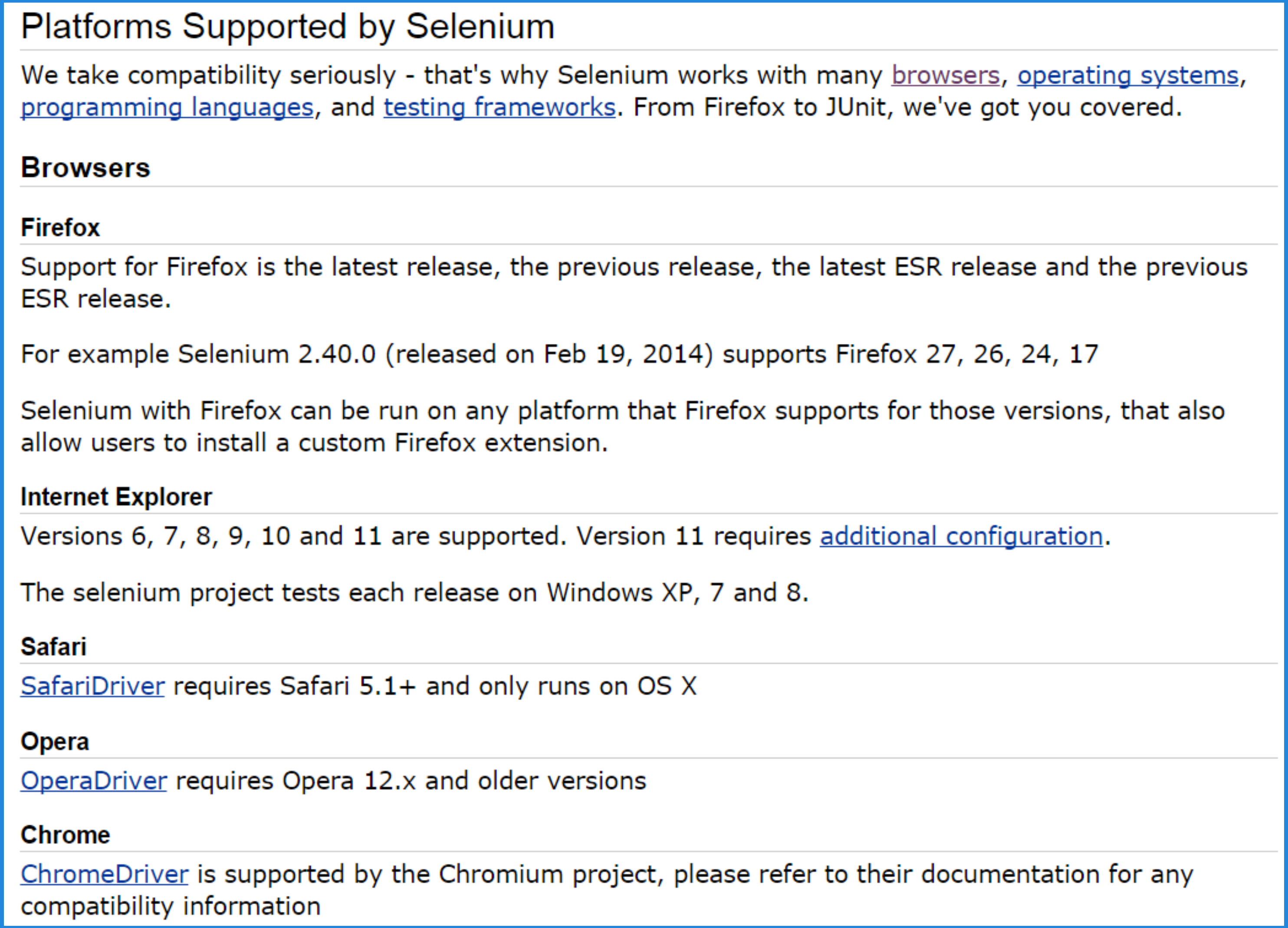
If you want to

- create robust, browser-based regression automation suites and tests
- scale and distribute scripts across many environments

**Selenium IDE**

If you want to

- create quick bug reproduction scripts
- create scripts to aid in automation-aided exploratory testing



# Platforms Supported by Selenium

We take compatibility seriously - that's why Selenium works with many browsers, operating systems, programming languages, and testing frameworks. From Firefox to JUnit, we've got you covered.

## Browsers

### Firefox

Support for Firefox is the latest release, the previous release, the latest ESR release and the previous ESR release.

For example Selenium 2.40.0 (released on Feb 19, 2014) supports Firefox 27, 26, 24, 17

Selenium with Firefox can be run on any platform that Firefox supports for those versions, that also allow users to install a custom Firefox extension.

### Internet Explorer

Versions 6, 7, 8, 9, 10 and 11 are supported. Version 11 requires additional configuration.

The selenium project tests each release on Windows XP, 7 and 8.

### Safari

SafariDriver requires Safari 5.1+ and only runs on OS X

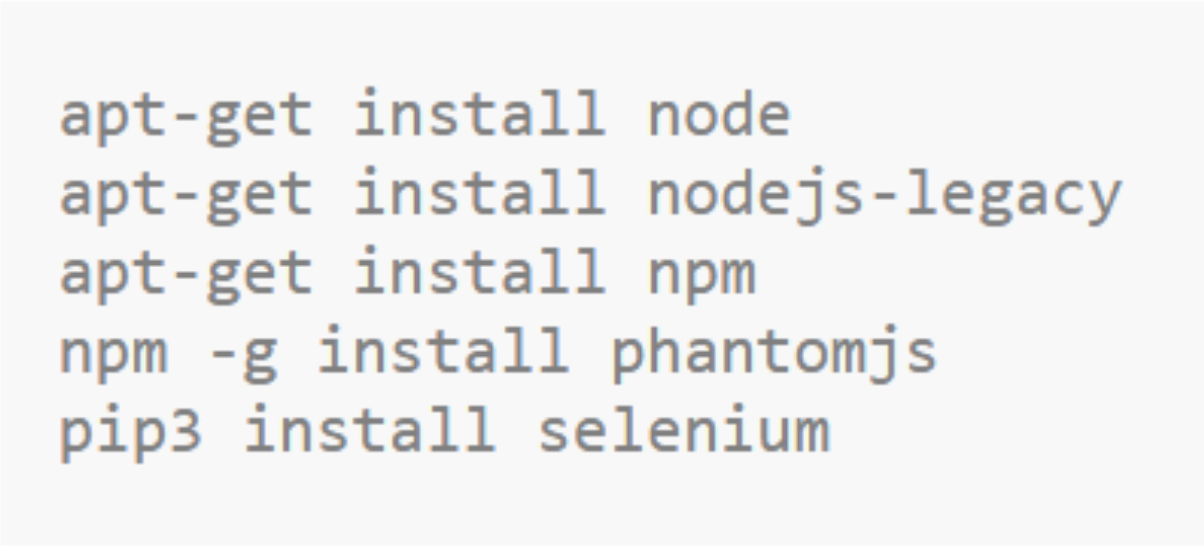
### Opera

OperaDriver requires Opera 12.x and older versions

### Chrome

ChromeDriver is supported by the Chromium project, please refer to their documentation for any compatibility information

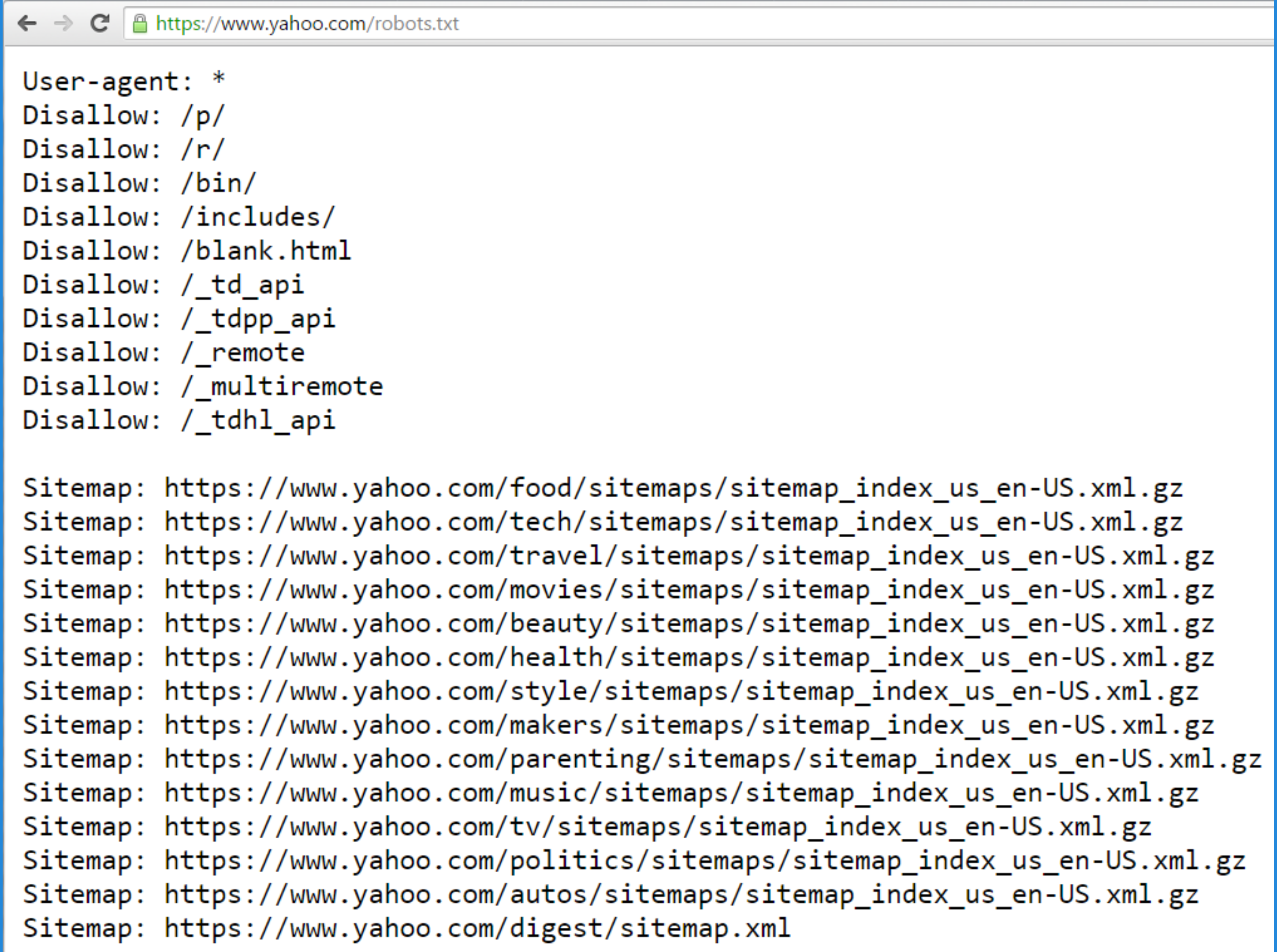
# Install...



```
apt-get install node
apt-get install nodejs-legacy
apt-get install npm
npm -g install phantomjs
pip3 install selenium
```

https://simpletutorials.com/c/2191/Installing+Selenium+and+PhantomJS+for+Python+3+on+Ubuntu+14.04

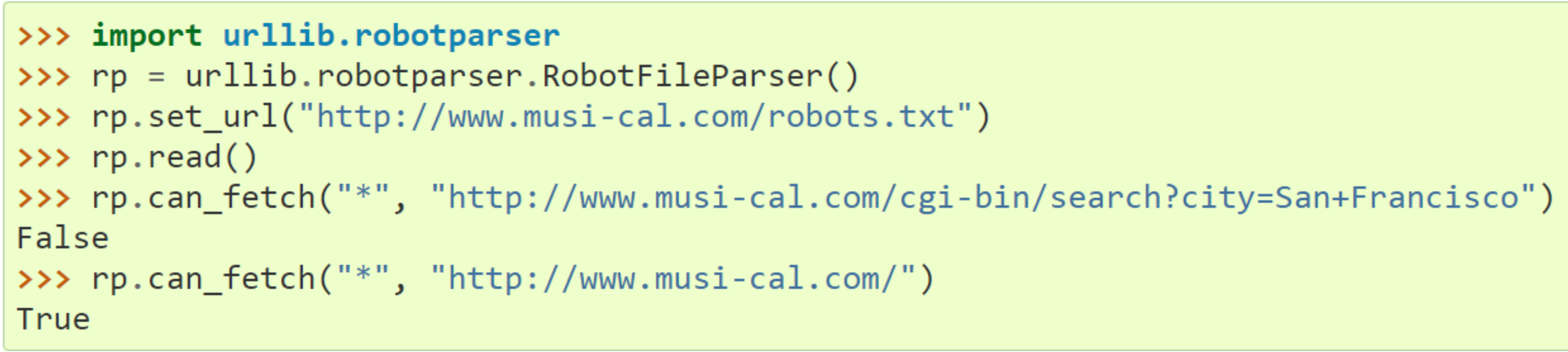
# robots.txt



```
User-agent: *
Disallow: /p/
Disallow: /r/
Disallow: /bin/
Disallow: /includes/
Disallow: /blank.html
Disallow: /_td_api
Disallow: /_tdpp_api
Disallow: /_remote
Disallow: /_multiremote
Disallow: /_tdhl_api

Sitemap: https://www.yahoo.com/food/sitemaps/sitemap_index_us_en-US.xml.gz
Sitemap: https://www.yahoo.com/tech/sitemaps/sitemap_index_us_en-US.xml.gz
Sitemap: https://www.yahoo.com/travel/sitemaps/sitemap_index_us_en-US.xml.gz
Sitemap: https://www.yahoo.com/movies/sitemaps/sitemap_index_us_en-US.xml.gz
Sitemap: https://www.yahoo.com/beauty/sitemaps/sitemap_index_us_en-US.xml.gz
Sitemap: https://www.yahoo.com/health/sitemaps/sitemap_index_us_en-US.xml.gz
Sitemap: https://www.yahoo.com/style/sitemaps/sitemap_index_us_en-US.xml.gz
Sitemap: https://www.yahoo.com/makers/sitemaps/sitemap_index_us_en-US.xml.gz
Sitemap: https://www.yahoo.com/parenting/sitemaps/sitemap_index_us_en-US.xml.gz
Sitemap: https://www.yahoo.com/music/sitemaps/sitemap_index_us_en-US.xml.gz
Sitemap: https://www.yahoo.com/tv/sitemaps/sitemap_index_us_en-US.xml.gz
Sitemap: https://www.yahoo.com/politics/sitemaps/sitemap_index_us_en-US.xml.gz
Sitemap: https://www.yahoo.com/autos/sitemaps/sitemap_index_us_en-US.xml.gz
Sitemap: https://www.yahoo.com/digest/sitemap.xml
```

# Robots.txt Parser



```
>>> import urllib.robotparser
>>> rp = urllib.robotparser.RobotFileParser()
>>> rp.set_url("http://www.musi-cal.com/robots.txt")
>>> rp.read()
>>> rp.can_fetch("*", "http://www.musi-cal.com/cgi-bin/search?city=San+Francisco")
False
>>> rp.can_fetch("*", "http://www.musi-cal.com/")
True
```

https://docs.python.org/3/library/urllib.robotparser.html

# Docs

https://selenium-python.readthedocs.org/

http://selenium-python.readthedocs.org/en/latest/api.html

# User Agent

```
...
{
  "headers": {
    "Connection": "close",
    "Host": "httpbin.org",
    "Accept-Encoding": "gzip",
    "Accept-Language": "ru-RU",
    "User-Agent": "Mozilla/5.0 (Unknown; Linux i686) AppleWebKit/534.34 (KHTML, like
Gecko) PhantomJS/1.10.0 (development) Safari/534.34",
    "Accept": "text/html,application/xhtml+xml,application/xml;q=0.9,*/*;q=0.8"
  }
 ...
```

http://stackoverflow.com/questions/17858663/custom-headers-in-phantomjs-selenium-webdriver

# User Agent

```
DesiredCapabilities.PHANTOMJS['phantomjs.page.settings.userAgent'] = \
        'Mozilla/5.0 (Windows NT 6.1; Win64; x64; rv:16.0) Gecko/20121026 Firefox/16.0'
```

http://stackoverflow.com/questions/28532347/selenium-with-phantomjs-yahoo-login-form-not-submitting-python-bindings?rq=1