



# Scraping with Selenium

by Dominique Theodore

# Overview

1. What is Selenium
2. Selenium vs. Scrapy
3. Download and installation
4. First steps
5. Finding elements
6. Using locators
7. Scraping dynamic pages
8. Infinite scrolling pages

# What is Selenium

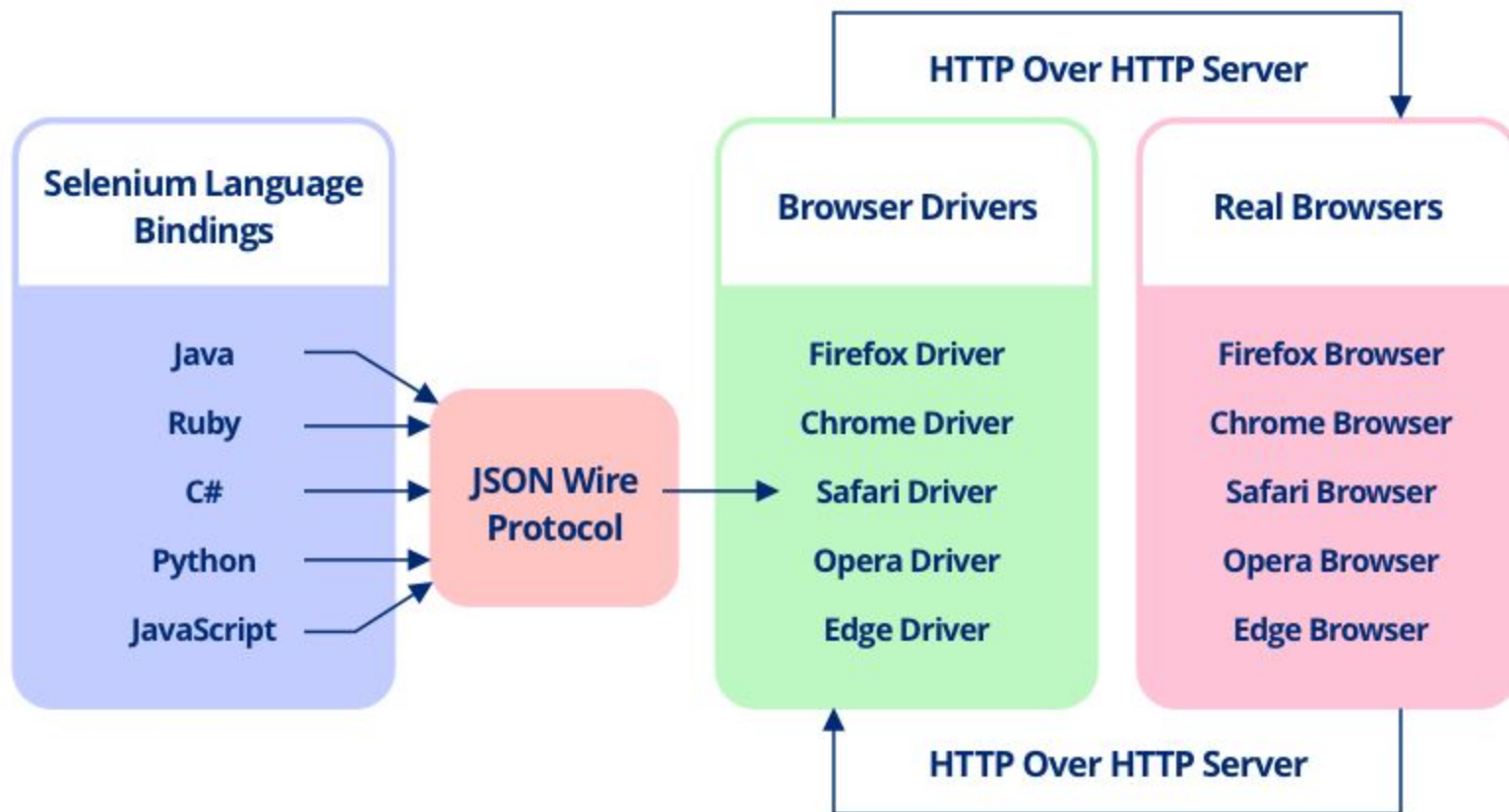
*"Selenium automates web browsers"*

<https://www.seleniumhq.org/>

- An open-source automated testing suite for web applications
- Selenium supports **all major browsers** (Chrome, Firefox, Internet Explorer), works on **many operating systems** (Linux, OS X, Windows) and can be used with **many programming languages** (Python, C#, Java, Haskell)
- 3 flavors: WebDriver, IDE, Grid

# How it works

## Selenium WebDriver Architecture



# Selenium vs Scrapy

- Selenium built for browser automation, not for scraping, therefore slower performance to be expected
- Less steep learning curve
- Easier to scrape dynamic pages with JavaScript elements

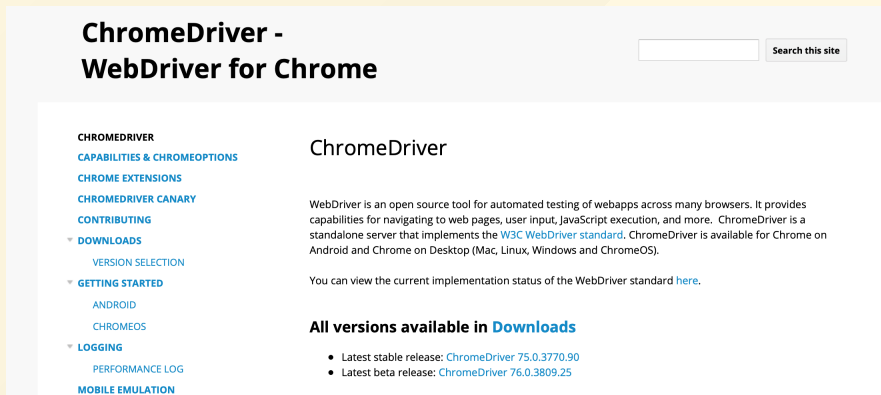
# Download and installation

1. Add Selenium to your project:

```
pip install selenium
```

2. Download your WebDriver and ensure it is in your \$PATH

<http://chromedriver.chromium.org>

A screenshot of the ChromeDriver website. The header shows "ChromeDriver - WebDriver for Chrome" with a search bar. A left sidebar contains a list of links: CHROMEDRIVER, CAPABILITIES & CHROME OPTIONS, CHROME EXTENSIONS, CHROMEDRIVER CANARY, CONTRIBUTING, DOWNLOADS (highlighted), VERSION SELECTION, GETTING STARTED, ANDROID, CHROME OS, LOGGING, PERFORMANCE LOG, and MOBILE EMULATION. The main content area is titled "ChromeDriver" and contains a paragraph describing it as an open-source tool for automated testing, followed by a link to the WebDriver standard. Below this, it states "All versions available in Downloads" and lists two bullet points: "Latest stable release: ChromeDriver 75.0.3770.90" and "Latest beta release: ChromeDriver 76.0.3809.25".

**ChromeDriver -  
WebDriver for Chrome**

Search this site

**CHROMEDRIVER**  
CAPABILITIES & CHROME OPTIONS  
CHROME EXTENSIONS  
CHROMEDRIVER CANARY  
CONTRIBUTING  
\* **DOWNLOADS**  
VERSION SELECTION  
\* **GETTING STARTED**  
ANDROID  
CHROME OS  
\* **LOGGING**  
PERFORMANCE LOG  
MOBILE EMULATION

## ChromeDriver

WebDriver is an open source tool for automated testing of webapps across many browsers. It provides capabilities for navigating to web pages, user input, JavaScript execution, and more. ChromeDriver is a standalone server that implements the [W3C WebDriver standard](#). ChromeDriver is available for Chrome on Android and Chrome on Desktop (Mac, Linux, Windows and ChromeOS).

You can view the current implementation status of the WebDriver standard [here](#).

**All versions available in [Downloads](#)**

- Latest stable release: ChromeDriver 75.0.3770.90
- Latest beta release: ChromeDriver 76.0.3809.25

# First steps

```
from selenium import webdriver
from selenium.webdriver.common.keys import Keys

# path to ChromeDriver
chromedriver = "/usr/local/bin/chromedriver"

# create an instance of Chrome Webdriver
driver = webdriver.Chrome()

# navigate to a url
driver.get("http://www.google.com/search?q=pymug");

# wait 5 seconds and close the browser
time.sleep(5)
driver.close()
```

# Finding elements on the page

- Selenium provides several locators to interact with UI elements:
  - ID
  - Name
  - Link Text
  - CSS Selector
  - DOM (Document Object Model)
  - XPath



# Using locators

```
# find an element using XPath  
elem = driver.find_element_by_xpath(  
    "//div[@class='r'][1]/a"  
)  
  
# interacting with an element  
elem.click()  
elem.send_keys("pymug")  
elem.send_keys(Keys.RETURN)
```

# Scraping dynamic webpages

Our target:



Ministry of Labour, Industrial Relations, Employment and Training

Home About Us ▾ Jobseeker Employer Jobs Search Contact Us

**SEARCH FOR A JOB** *More Than 390 Jobs Available*

Keyword

Economic Sector

 Show local jobs only ☐

 Show International Jobs Only ☐

[ADVANCED SEARCH +](#)

... promoting and sustaining decent work

# Scraping dynamic webpages

```
<td width="225">ICT/BPO</td>
<td width="175">GLOBAL EDGE SOFTWARE LTD</td>
<td width="125">Mauritius</td>
<td width="100">28/06/2019</td>
<td align="center"></td>
</tr>

<tr><td colspan="7" class="hidden-td">
<div class="show_details" id="17063">
<b>Web Developer</b>
<table class="job_details" width="100%">
<tr>
<td width="300">Employer</td>
<td>GLOBAL EDGE SOFTWARE LTD</td>
</tr>
<tr>
```

```
<div class="main-content">
<h1>Jobs Search</h1>

<script type="text/javascript">function jobdetails(val){$("#"+val).slideToggle();}
function applyjob(val)
{$(this).scrollTop(0);$('#popup').show();$('#popup_content').show();$.ajax({type:"POST",url:"ht
on(msg){document.getElementById("popup_content").innerHTML=msg;}});}
```

# Scraping dynamic webpages

```
from selenium import webdriver
from selenium.webdriver.common.keys import Keys

chromedriver = "/usr/local/bin/chromedriver"

driver = webdriver.Chrome()
driver.get("https://www.mauritiusjobs.mu/jobsearch");

input = driver.find_element_by_name("search_by_keyword")
input.send_keys("developer")
input.send_keys(Keys.RETURN)

result = driver.find_element_by_xpath(
    "//img[contains(@id, 'view_job')]")
result.click
```

# Infinite scrolling pages

The screenshot displays a LinkedIn profile for Dominique Theodore, Team Lead at AGILEUM. The page layout includes a left-hand navigation pane with profile statistics (46 profile views, 234 connections), a PYMUG (Python) group listing (6 followers), and a 'Recent' section. The main content area features a 'Start a post' section with options to write an article or upload media, followed by a 'Jobs recommended for you' section. This section lists two job opportunities: 'Project Manager - Upcoming' at Dubai Airports and 'CAT A Aircraft Engineers' at Aer Lingus, both noting that Imperial College London alumni work there. Below these is a promoted post from Dropbox, highlighting new integrations with Zoom. The right-hand sidebar contains an 'Add to your feed' section with hashtags like #startups and recommendations for Gary Vaynerchuk and Medine Limited. At the bottom of the sidebar, there is a section for Dominique to explore relevant opportunities with MCB Group. The page footer includes links to 'About', 'Help Center', and 'Privacy & Terms'.

https://www.linkedin.com/feed/

Search

Home My Network Jobs Messaging Notifications Me Work

Online Master in 1 Year - Online Master degree from University of Salford. Get more information! Ad ...

**Dominique Theodore**  
Team Lead at AGILEUM

Who's viewed your profile 46  
Connections 234  
Grow your network

Access exclusive tools & insights  
Reactivate Premium

**PYMUG (Python ...)**  
6 followers

Page activity 0  
Recent visitors 3

Share an update

Recent  
Smartly EMBA May 2019 ...

Groups  
Smartly EMBA May 2019 ...

Start a post

Write an article on LinkedIn

Sort by: Top

Jobs recommended for you

**Project Manager - Upcoming**  
Dubai Airports • Dubai, AE  
2 Imperial College London alumni work here

**CAT A Aircraft Engineers**  
Aer Lingus • Dublin, IE  
4 Imperial College London alumni work here

See more

**Dropbox**  
232,121 followers  
Promoted

New integrations with Zoom make video conferencing seamless. Add and join meetings directly from Dropbox and easily share Dropbox files during a Zoom meeting.

Add to your feed

#startups

**Gary Vaynerchuk**  
Chairman of VaynerX, CEO of VaynerMedia, 5-...

**Medine Limited**  
Company • Executive Office

View all recommendations

Get the latest jobs and industry n

Dominique, explore relevant opportunities with MCB Gro

Follow

About Help Center Privacy & Terms

# Infinite scrolling pages

- Use expected conditions to wait for the page to load and prevent Element Not Visible Exception

## Selenium Waits

```
from selenium.webdriver.support.ui import WebDriverWait  
from selenium.webdriver.support import expected_conditions
```

- Selenium can execute Javascript and scroll the page for us

```
driver.execute_script("window.scrollTo(0,  
    document.body.scrollHeight);")
```

**Thank you!**

