Data Scraping using Selenium

This guide will explain how to include "Selenium" related packages to use Selenium Chrome driver to simulate actions in chrome browser and do different types of actions in it. Also it explains how to extract information from a web page. This tutorial is only for learning purposes.

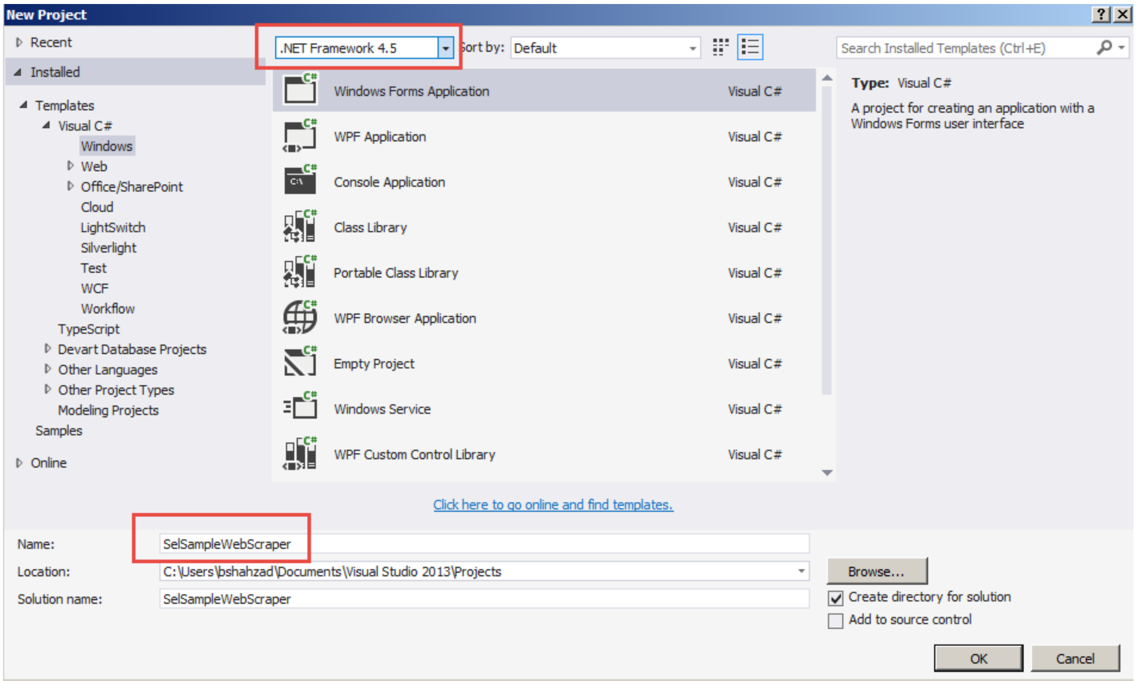
Selenium is an "Automation Testing" tool.

Version	**Last updated**	Comments	Modified By
V1.0	22-04-2016		Bilal Shahzad
V2.0	11-12-2018	If you face any issue at run time, it may be because of Chrome Driver version. Upgrade package by executing following command on Package Manager Console Install-Package Selenium.WebDriver.ChromeDriver -Version 2.44.0	

Here we are going to build a simple windows application which will use “Selenium” APIs and will extract basic job information from “rozee.pk” site.

Walkthrough

1- Create a new “Windows Forms” application with .NET Framework 4.5.


New Project
Recent
Installed
Templates
Visual C#
Windows
Web
Office/SharePoint
Cloud
LightSwitch
Silverlight
Test
WCF
Workflow
TypeScript
Devart Database Projects
Other Languages
Other Project Types
Modeling Projects
Samples
Online
Sort by: Default
Search Installed Templates (Ctrl+E)
Windows Forms Application Visual C#
WPF Application Visual C#
Console Application Visual C#
Class Library Visual C#
Portable Class Library Visual C#
WPF Browser Application Visual C#
Empty Project Visual C#
Windows Service Visual C#
WPF Custom Control Library Visual C#
Type: Visual C#
A project for creating an application with a Windows Forms user interface
Click here to go online and find templates.
Browse...
OK
Cancel

.NET Framework 4.5

Name: SelSampleWebScraper

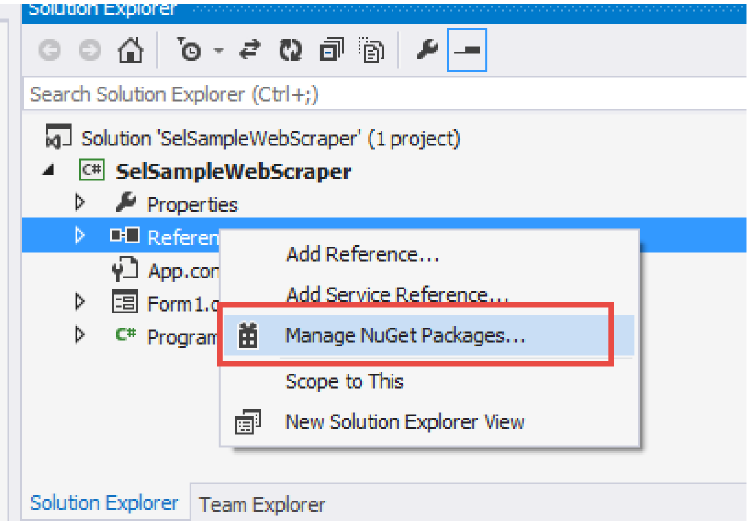
Location: C:\Users\bshahzad\Documents\Visual Studio 2013\Projects

Solution name: SelSampleWebScraper

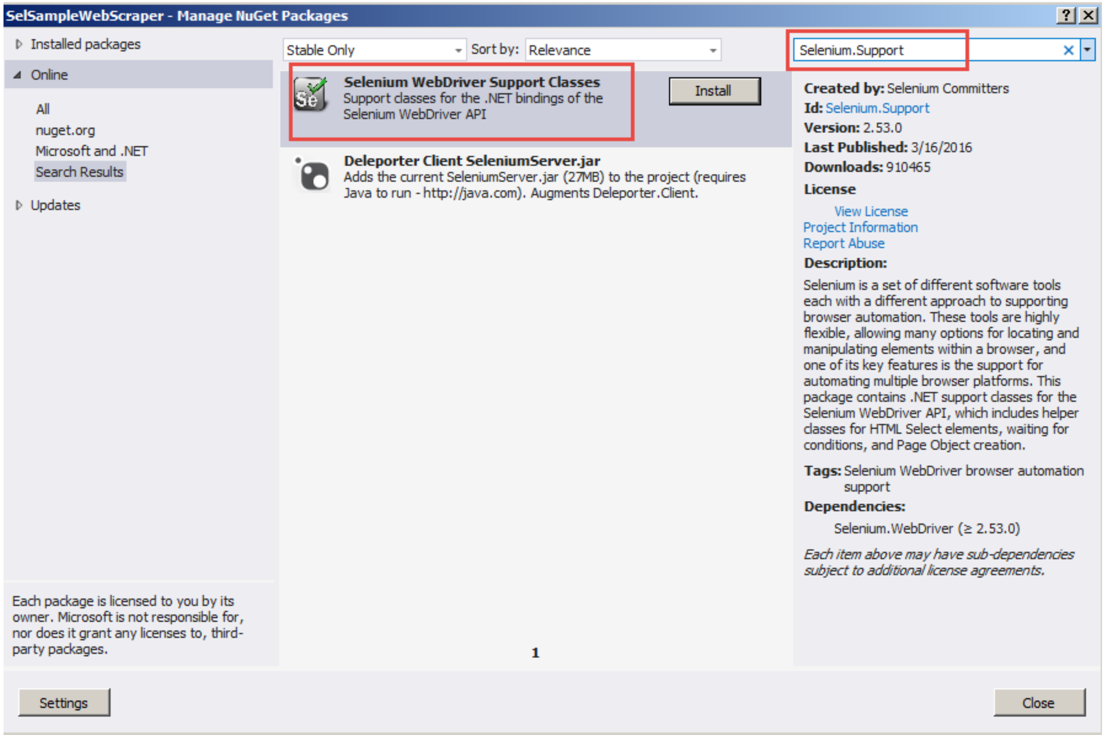
☑ Create directory for solution

□ Add to source control

2- Now we need to add some references to access “Selenium” API. Right click on “References” and select “Manage NuGet Packages…”



3- In search box type “Selenium.Support” and Install the package as highlighted in following screenshot.



Selenium WebDriver Support Classes
Support classes for the .NET bindings of the Selenium WebDriver API

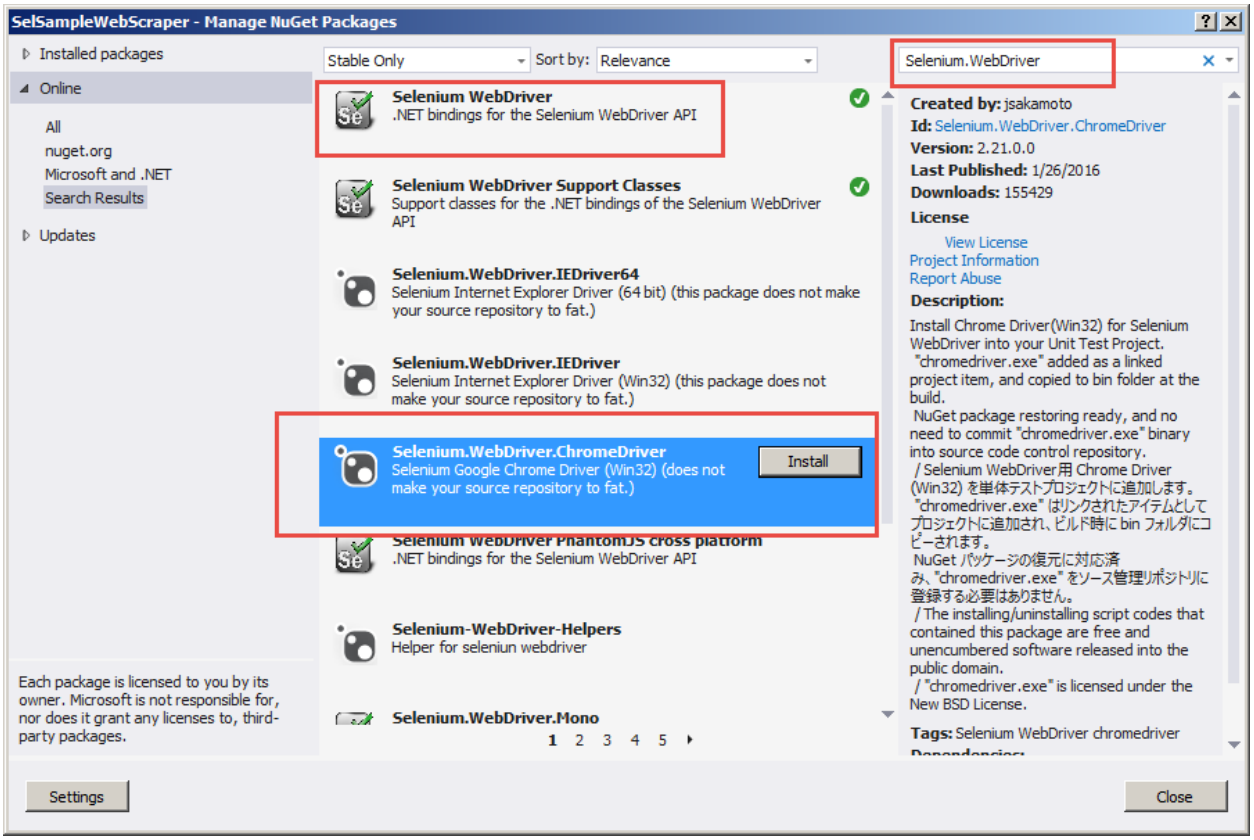
Deleporter Client SeleniumServer.jar
Adds the current SeleniumServer.jar (27MB) to the project (requires Java to run - http://java.com). Augments Deleporter.Client.

Created by: Selenium Committers
Id: Selenium.Support
Version: 2.53.0
Last Published: 3/16/2016
Downloads: 910465
License
View License
Project Information
Report Abuse
Description:
Selenium is a set of different software tools each with a different approach to supporting browser automation. These tools are highly flexible, allowing many options for locating and manipulating elements within a browser, and one of its key features is the support for automating multiple browser platforms. This package contains .NET support classes for the Selenium WebDriver API, which includes helper classes for HTML Select elements, waiting for conditions, and Page Object creation.
Tags: Selenium WebDriver browser automation support
Dependencies:
Selenium.WebDriver (≥ 2.53.0)

Each item above may have sub-dependencies subject to additional license agreements.

Each package is licensed to you by its owner. Microsoft is not responsible for, nor does it grant any licenses to, third-party packages.

4- In search box type “Selenium.WebDriver” and Install the packages as highlighted in following screenshot.



Selenium WebDriver
.NET bindings for the Selenium WebDriver API

Selenium WebDriver Support Classes
Support classes for the .NET bindings of the Selenium WebDriver API

Selenium.WebDriver.IEDriver64
Selenium Internet Explorer Driver (64 bit) (this package does not make your source repository to fat.)

Selenium.WebDriver.IEDriver
Selenium Internet Explorer Driver (Win32) (this package does not make your source repository to fat.)

Selenium.WebDriver.ChromeDriver
Selenium Google Chrome Driver (Win32) (does not make your source repository to fat.)

Selenium WebDriver PhantomJS cross platform
.NET bindings for the Selenium WebDriver API

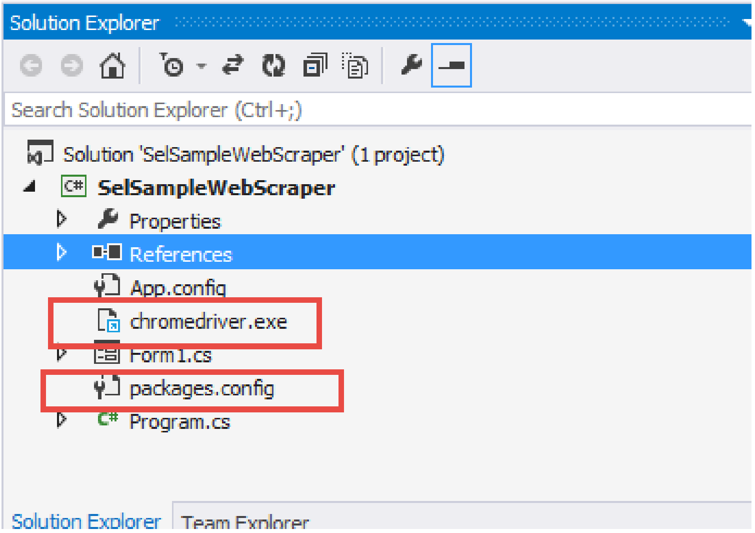
Selenium-WebDriver-Helpers
Helper for selenium webdriver

Selenium.WebDriver.Mono

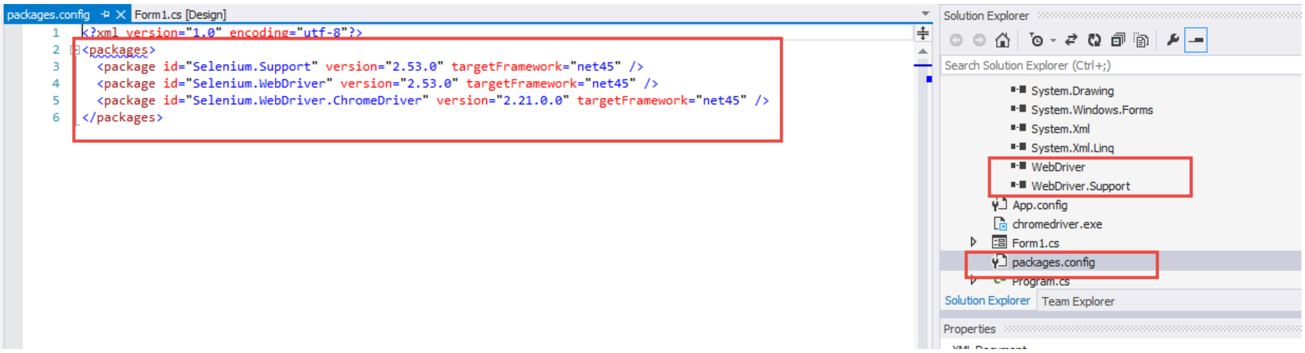
Created by: jsakamoto
Id: Selenium.WebDriver.ChromeDriver
Version: 2.21.0.0
Last Published: 1/26/2016
Downloads: 155429
License
View License
Project Information
Report Abuse
Description:
Install Chrome Driver(Win32) for Selenium WebDriver into your Unit Test Project. "chromedriver.exe" added as a linked project item, and copied to bin folder at the build. NuGet package restoring ready, and no need to commit "chromedriver.exe" binary into source code control repository. / Selenium WebDriver用 Chrome Driver (Win32) を単体テストプロジェクトに追加します。"chromedriver.exe" はリンクされたアイテムとしてプロジェクトに追加され、ビルド時に bin フォルダにコピーされます。NuGet パッケージの復元に対応済み、"chromedriver.exe" をソース管理リポジトリに登録する必要はありません。/ The installing/uninstalling script codes that contained this package are free and unencumbered software released into the public domain. / "chromedriver.exe" is licensed under the New BSD License.
Tags: Selenium WebDriver chromedriver

Each package is licensed to you by its owner. Microsoft is not responsible for, nor does it grant any licenses to, third-party packages.

5- You will see “chromdriver.exe” and “Packages.config” file.

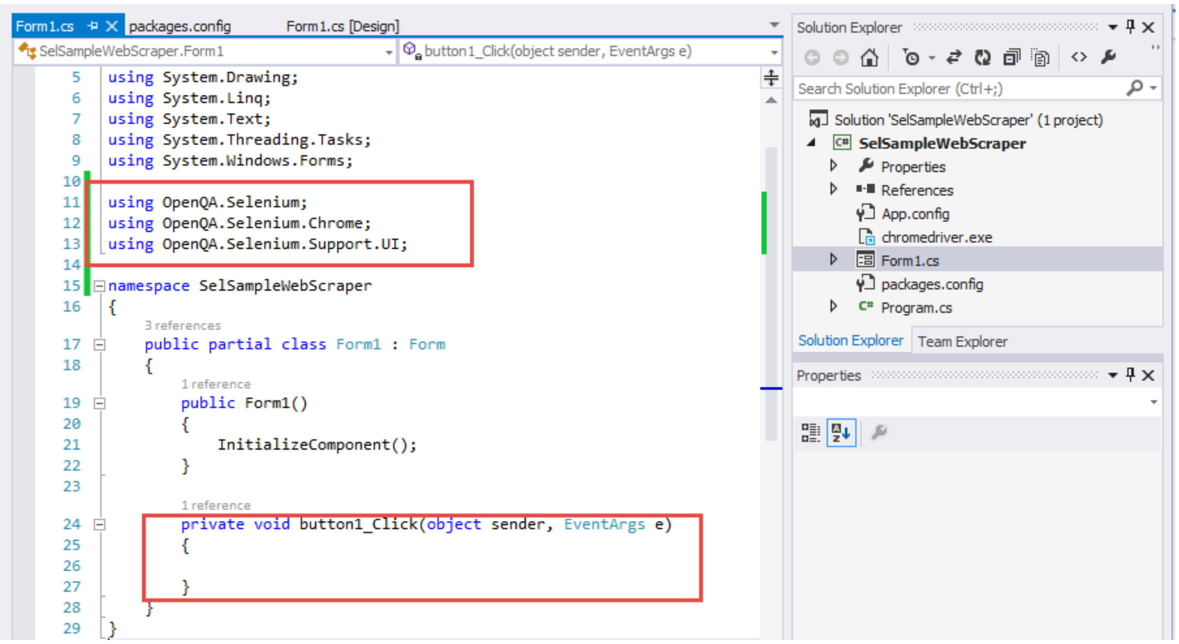


6- In “Packages.config”, you will see following packages are added. Also check the reference of DLLs shown in right side panel.



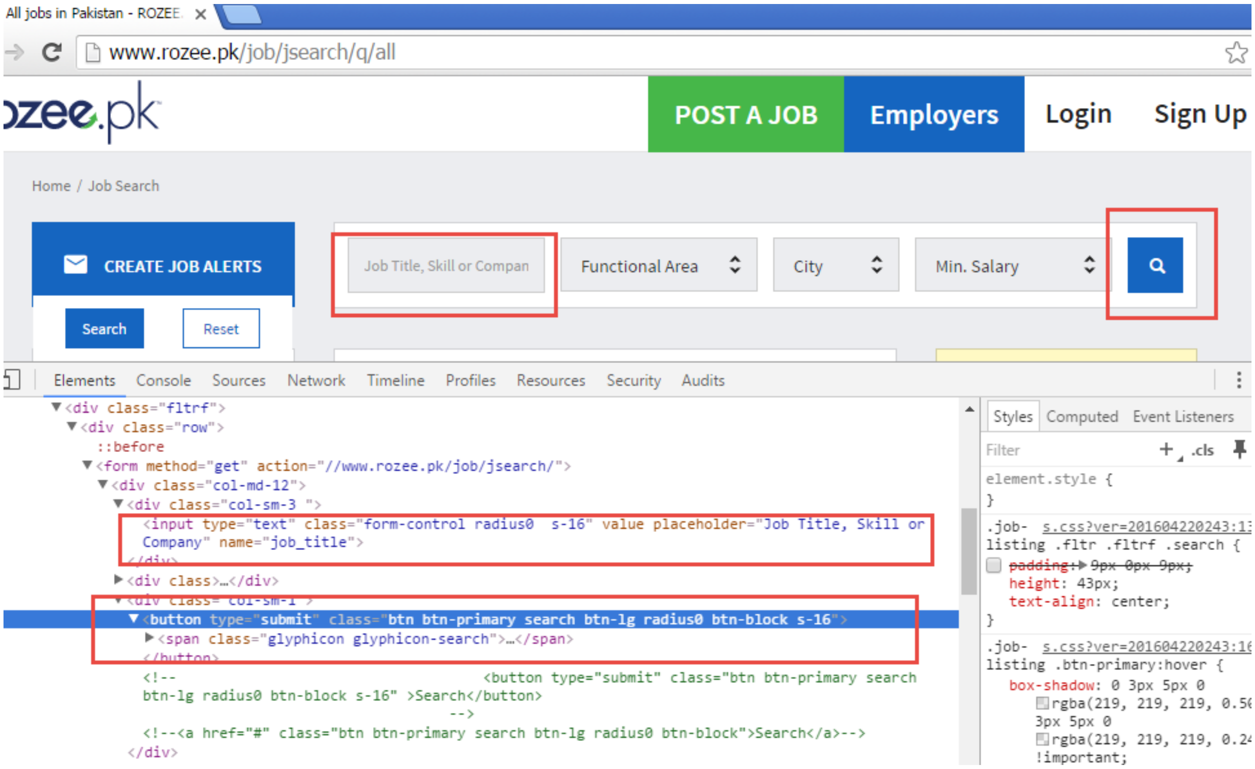
```
<?xml version="1.0" encoding="utf-8"?>
<packages>
  <package id="Selenium.Support" version="2.53.0" targetFramework="net45" />
  <package id="Selenium.WebDriver" version="2.53.0" targetFramework="net45" />
  <package id="Selenium.WebDriver.ChromeDriver" version="2.21.0.0" targetFramework="net45" />
</packages>
```

7- Now add a button on a form and generate click event of it. In Code, add the reference of namespaces as shown below.



```
using System.Drawing;
using System.Linq;
using System.Text;
using System.Threading.Tasks;
using System.Windows.Forms;

using OpenQA.Selenium;
using OpenQA.Selenium.Chrome;
using OpenQA.Selenium.Support.UI;

namespace SelSampleWebScraper
{
    3 references
    public partial class Form1 : Form
    {
        1 reference
        public Form1()
        {
            InitializeComponent();
        }

        1 reference
        private void button1_Click(object sender, EventArgs e)
        {

        }
    }
}
```

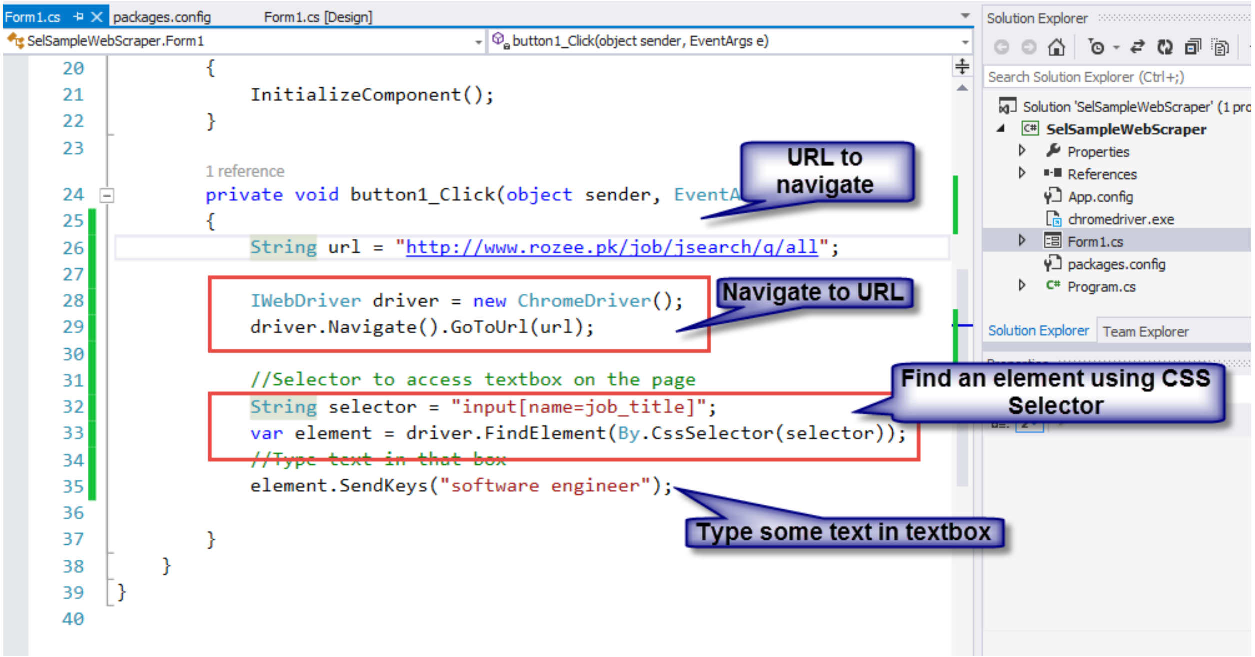
8- We are going to scrap data from following http://www.rozee.pk/job/jsearch/q/all. On this link, we'll provide some text to type in highlighted textbox. Also HTML of textbox and button is highlighted to access them via code using CSS selector approach.



```
▼<div class="fltrf">
  ▼<div class="row">
      ::before
    ▼<form method="get" action="//www.rozee.pk/job/jsearch/">
      ▼<div class="col-md-12">
        ▼<div class="col-sm-3 ">
            <input type="text" class="form-control radius0  s-16" value placeholder="Job Title, Skill or
            Company" name="job_title">
          </div>
        ▶<div class>…</div>
        ▼<div class="col-sm-1 ">
          ▼<button type="submit" class="btn btn-primary search btn-lg radius0 btn-block s-16">
            ▶<span class="glyphicon glyphicon-search">…</span>
            </button>
            <!--                                    <button type="submit" class="btn btn-primary search
            btn-lg radius0 btn-block s-16" >Search</button>
                                            -->
            <!--<a href="#" class="btn btn-primary search btn-lg radius0 btn-block">Search</a>-->
          </div>
```

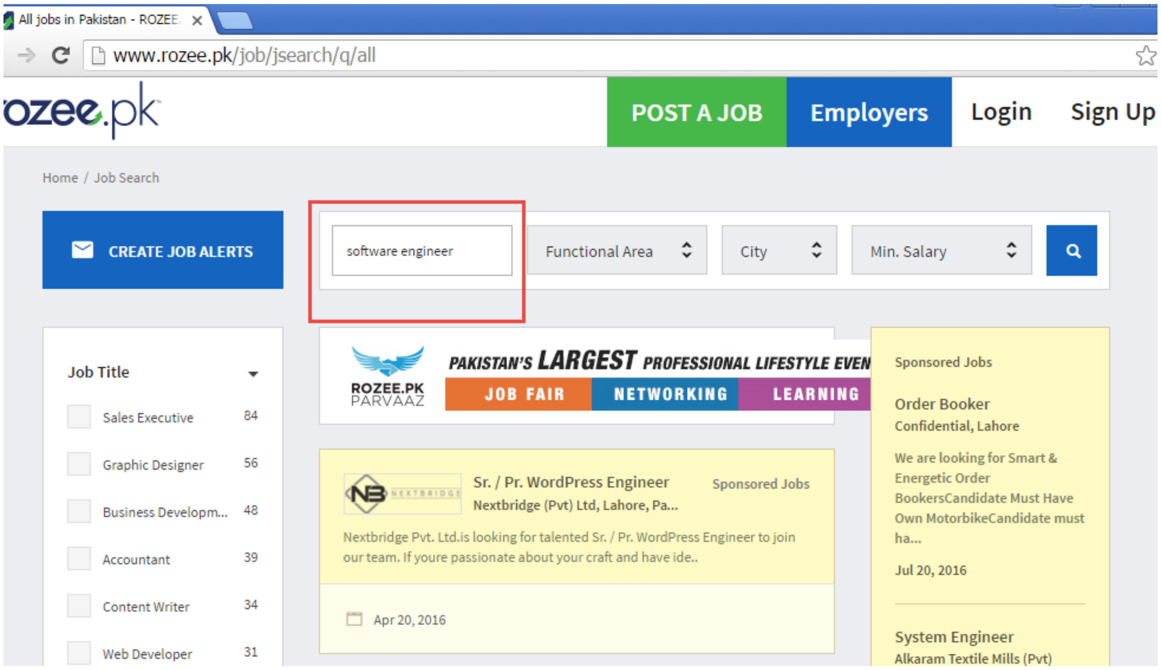
```
element.style {
}
.job-  s.css?ver=201604220243:13
listing .fltr .fltrf .search {
  ~~padding: ▶ 9px 0px 9px;~~
  height: 43px;
  text-align: center;
}
.job-  s.css?ver=201604220243:16
listing .btn-primary:hover {
  box-shadow: 0 3px 5px 0
    rgba(219, 219, 219, 0.5
    3px 5px 0
    rgba(219, 219, 219, 0.24
    !important;
```

9- Add following code in button click event handler. This code will open a chrome browser and will navigate it to the given URL. After page is loaded, it will send keys (i.e. software engineer) to specific text box.



```
        {
            InitializeComponent();
        }

        1 reference
        private void button1_Click(object sender, EventA
        {
            String url = "http://www.rozee.pk/job/jsearch/q/all";

            IWebDriver driver = new ChromeDriver();
            driver.Navigate().GoToUrl(url);

            //Selector to access textbox on the page
            String selector = "input[name=job_title]";
            var element = driver.FindElement(By.CssSelector(selector));
            //Type text in that box
            element.SendKeys("software engineer");

        }
    }
}
```

10- If you run above code and click button, you will see that a new browser window will open and will start loading the page. Once page is loaded, you will see that “software engineer” will get typed in highlighted box.



software engineer

Sr. / Pr. WordPress Engineer
Nextbridge (Pvt) Ltd, Lahore, Pa...

Nextbridge Pvt. Ltd.is looking for talented Sr. / Pr. WordPress Engineer to join our team. If youre passionate about your craft and have ide..

Apr 20, 2016

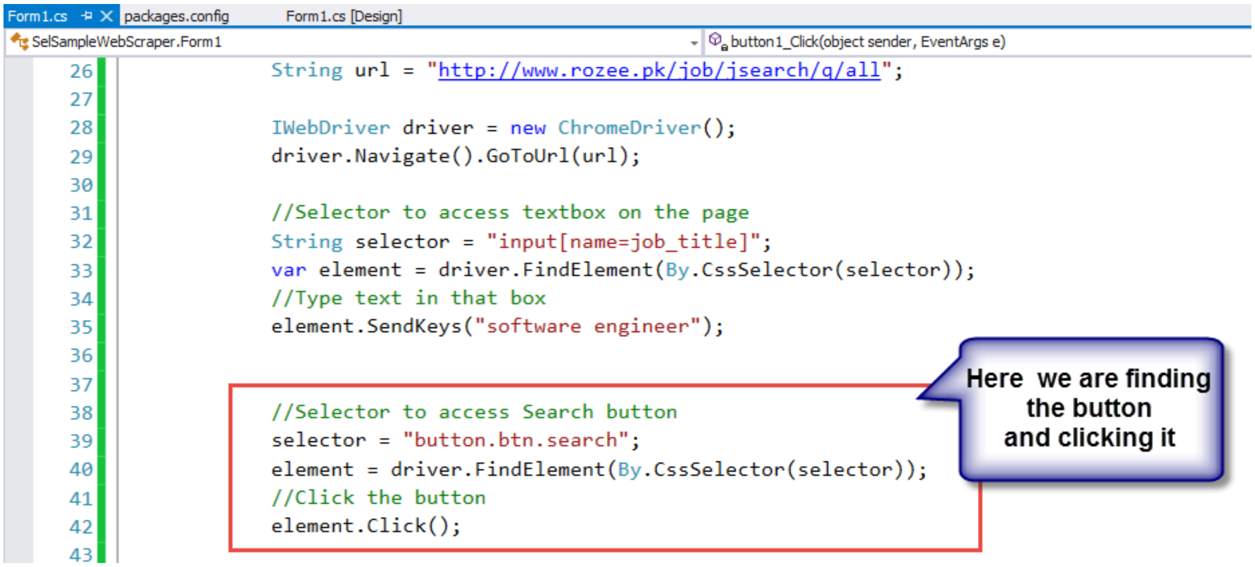
Order Booker
Confidential, Lahore

We are looking for Smart & Energetic Order BookersCandidate Must Have Own MotorbikeCandidate must ha...

Jul 20, 2016

System Engineer
Alkaram Textile Mills (Pvt)

11- Following code finds the button and clicks it.

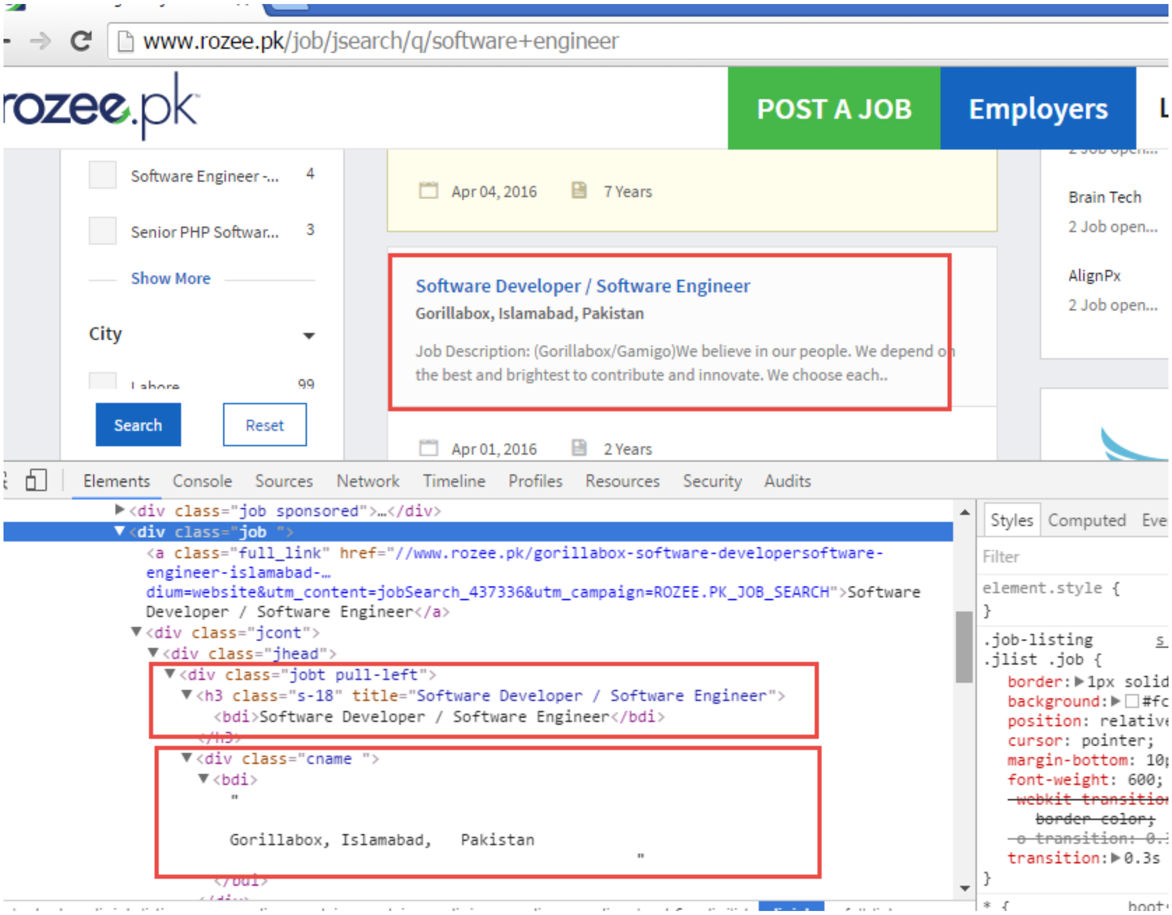


Form1.cs
packages.config
Form1.cs [Design]
SelSampleWebScraper.Form1
button1_Click(object sender, EventArgs e)

```
String url = "http://www.rozee.pk/job/jsearch/q/all";

IWebDriver driver = new ChromeDriver();
driver.Navigate().GoToUrl(url);

//Selector to access textbox on the page
String selector = "input[name=job_title]";
var element = driver.FindElement(By.CssSelector(selector));
//Type text in that box
element.SendKeys("software engineer");


//Selector to access Search button
selector = "button.btn.search";
element = driver.FindElement(By.CssSelector(selector));
//Click the button
element.Click();

```

Here we are finding the button and clicking it

12- Now when button is clicked, you will see that URL is updated in browser and results are also filtered now. Here by inspecting our results, we can investigate the HTML of our results. Each result is in “div.job”. Inside div, we can see the selectors of Title and company and so on.



www.rozee.pk/job/jsearch/q/software+engineer

Apr 04, 2016 7 Years

Software Developer / Software Engineer
Gorillabox, Islamabad, Pakistan

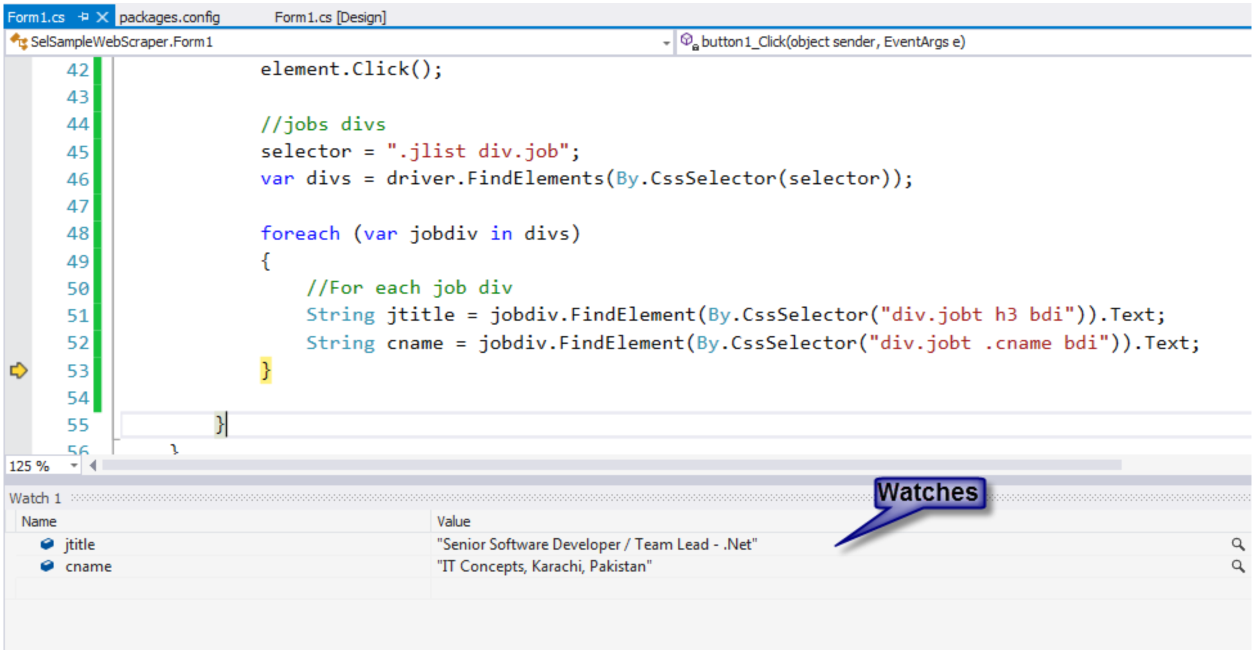
Job Description: (Gorillabox/Gamigo)We believe in our people. We depend on the best and brightest to contribute and innovate. We choose each..

Apr 01, 2016 2 Years

```
▶<div class="job sponsored">…</div>
▼<div class="job ">
   <a class="full_link" href="//www.rozee.pk/gorillabox-software-developersoftware-
   engineer-islamabad-…
   dium=website&utm_content=jobSearch_437336&utm_campaign=ROZEE.PK_JOB_SEARCH">Software
   Developer / Software Engineer</a>
  ▼<div class="jcont">
    ▼<div class="jhead">
      ▼<div class="jobt pull-left">
        ▼<h3 class="s-18" title="Software Developer / Software Engineer">
            <bdi>Software Developer / Software Engineer</bdi>
          </h3>
        ▼<div class="cname ">
          ▼<bdi>
              "

              Gorillabox, Islamabad,   Pakistan
                                                  "
            </bdi>
          </div>
```

```
element.style {
}
.job-listing
.jlist .job {
   border:▶1px solid
   background:▶□#fc
   position: relative
   cursor: pointer;
   margin-bottom: 10
   font-weight: 600;
   ~~-webkit-transition~~
       ~~border-color;~~
   ~~-o-transition: 0.~~
   transition:▶0.3s
}
```

13- Her is the code to find all DIVs and then iterating those divs to extract information from it.



Form1.cs
packages.config
Form1.cs [Design]
SelSampleWebScraper.Form1
button1_Click(object sender, EventArgs e)
125 %
Watch 1
Watches

```
                element.Click();

                //jobs divs
                selector = ".jlist div.job";
                var divs = driver.FindElements(By.CssSelector(selector));

                foreach (var jobdiv in divs)
                {
                    //For each job div
                    String jtitle = jobdiv.FindElement(By.CssSelector("div.jobt h3 bdi")).Text;
                    String cname = jobdiv.FindElement(By.CssSelector("div.jobt .cname bdi")).Text;
                }

            }
        }
```

Name	Value
jtitle	"Senior Software Developer / Team Lead - .Net"
cname	"IT Concepts, Karachi, Pakistan"

References

https://saucelabs.com/resources/selenium/css-selectors

Testing Using Selenium

http://learnseleniumtesting.com/basic-webdriver-and-c-sharp-script-for-beginners/

https://www.udemy.com/selenium-with-c/