

# Web scraping

Sugarkhuu Radnaa

Py4Econ in Ulaanbaatar

*py4econ@gmail.com*

9 January, 2021

## Week 6: Learning objectives

- 1 Setup selenium and chromedriver on your machine
- 2 Extract data from websites!

# Why Selenium?

There are many, many automation testing tools. We will use Selenium because it does pretty much all that I needed. There is no reason that the others are better at different tasks. Do let me know if you know or find such cases.

- Selenium
- BeautifulSoup
- Scrapy
- Urllib
- MechanicalSoup
- LXML
- Python Requests

# Chromedriver and Selenium

- To control Chrome through Python
- Chromedriver should be in your system path
  - Know your chrome version: `chrome://settings/help`
  - Download a compatible chromedriver from:  
`https://chromedriver.chromium.org/downloads`
  - Unzip and place it somewhere easy to access like:  
`C:\Users\sugarkhuu\chromedriver_win32`
  - Mac - brew install cask selenium
- Selenium (`pip install selenium`)



# HTML elements

- Tags (html, head, body, a, nav, title, div, h1, h3, p, button, ...)
- Attributes (id, class, text, link, name, type, href, ...)

```
1 <!DOCTYPE html>
2 <html>
3   <head>
4     <title>Example</title>
5     <link rel="stylesheet" href="st
6   </head>
7   <body>
8     <h1>
9       <a href="/">Header</a>
10    </h1>
11    <nav>
12      <a href="one/">One</a>
13      <a href="two/">Two</a>
14      <a href="three/">Three</a>
15    </nav>
```

## HTML + CSS + Javascript = Web

See basics of HTML -

[https://www.w3schools.com/html/html\\_paragraphs.asp](https://www.w3schools.com/html/html_paragraphs.asp)

# Locating HTML elements

- id
- name
- xpath (full xpath)
- class name
- text (with contains)
- tag
- css selector
- link\_text

# Navigating

- send\_keys
- clear
- click
- get
- forward
- back
- text
- switch\_to\_window



# No Homework, but Project!

Онлайн худалдааны ямар нэг платформ сонгон нэг төрлийн бүтээгдэхүүний хувьд нэмэлт бүтээгдэхүүн орж ирэхэд тодорхой имейл хаяг бүхий хэрэглэгчид рүү имейл илгээдэг байх код бичнэ үү.

- 1 Шалгах давтамжийг сонгох – Task scheduler (cron job), next week
- 2 Хуучин дата хадгалах – csv, pickle, json
- 3 Имейл нь шинэ бүтээгдэхүүний мэдээллийг агуулах хэрэгтэй – Email, next week

Дараах үйлдлүүд хийгдэх болов уу

- 1 Ямар нэг бүтээгдэхүүн сонгоно. Хөргөгч, угаалгын машин г.м.
- 2 Тухайн төрлийн бүх бүтээгдэхүүний мэдээллийг түүнэ.
- 3 Түүсэн мэдээллээ файл эсвэл өгөгдлийн санд хадгална.
- 4 Тогтоосон хугацааны дараа вэбсайтаа дахин шалган, мэдээллийг түүнэ.
- 5 Шинэ мэдээллийг өмнөх хадгалсан мэдээлэлтэй харьцуулна.
- 6 Хоёр мэдээлэл ялгаагүй бол зогсоно.
- 7 Хоёр мэдээлэл ялгаатай байвал шинээр орж ирсэн бүтээгдэхүүнүүдийг ялгаж авна.
- 8 Шинэ бүтээгдэхүүний мэдээлэл бүхий имейл загвар үүсгэнэ.
- 9 Үүсгэсэн имейл загвараа имейлийн жагсаалт руу илгээнэ.
- 10 Шинэ болон хуучин датаг нэгтгэн, давхардлыг арилгаад хадгална.
- 11 Тогтсон хугацааны дараа дахин вэбсайтаа шалгана ...

Thank you!