## Web scraping

Sugarkhuu Radnaa

Py4Econ in Ulaanbaatar

py4econ@gmail.com

9 January, 2021



### Week 6: Learning objectives

- 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
- Beautiful Soup
- Scrapy
- Urllib
- MechanicalSoup
- LXML
- Python Requests

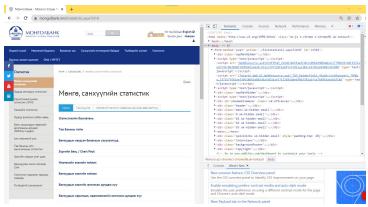


#### Chromedriver and Selenium

- 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)

# Chrome Inspect (Ctrl + Shift + I)

#### Elements



#### HTML elements

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

```
<!DOCTYPE html>
   <html>
       <head>
           <title>Example</title>
           k rel="stylesheet" href="st
       </head>
       <body>
            <h1>
                <a href="/">Header</a>
           </h1>
           <nav>
                <a href="one/">One</a>
13
               <a href="two/">Two</a>
               <a href="three/">Three</a>
           </nav>
```

#### HTML + CSS + Javascript = Web

See basics of HTML -

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 vv.

- 💶 Шалгах давтамжийг сонгох Task scheduler (cron job), next week
- Хуучин дата хадгалах csv, pickle, ison
- Имейл нь шинэ бүтээгдэхүүний мэдээллийг агуулах хэрэгтэй — Email, next week

Дараах үйлдлүүд хийгдэх болов уу

- 🚺 Ямар нэг бүтээгдэхүүн сонгоно. Хөргөгч, угаалгын машин г.м. Тухайн төрлийн бүх бүтээгдэхүүний мэдээллийг түүнэ.
  - Туусэн мэдээллээ файл эсвэл өгөгдлийн санд хадгална.
- Тогтоосон хугацааны дараа вэбсайтаа дахин шалган, мэдээллийг туунэ,
- Шинэ мэдээллийг өмнөх хадгалсан мэдээлэлтэй харьцуулна.
- Хоёр мэдээлэл ялгаагүй бол зогсоно.
- Хоёр мэдээлэл ялгаатай байвал шинээр орж ирсэн бүтээгдэхүүнүүдийг ялгаж авна.
- Шинэ бүтээгдэхүүний мэдээлэл бүхий имейл загвар үүсгэнэ. Уусгэсэн имейл загвараа имейлийн жагсаалт руу илгээнэ.
  - Шинэ болон хуучин датаг нэгтгэн, давхардлыг арилгаад хадгална.
    - Тогтсон хугацааны дараа дахин вэбсайтаа шалгана ...

# Thank you!