Hands-on Activity 7.2 Webscraping using BeautifulSoup and Requests

Name: Dela Cruz, Eugene D.G.

Section: CPE22S3

Data Gathering

Sources of Data

A vast amount of historical data can be found in files such as:

- MS Word documents
- Emails
- Spreadsheets
- MS PowerPoints
- PDFs
- HTML
- and plaintext files

Public and Private Archives

CSV, JSON, and XML files use plaintext, a common format, and are compatible with a wide range of applications

The Web can be mined for data using a web scraping application

The IoT uses sensors create data

Sensors in smartphones, cars, airplanes, street lamps, and home appliances capture raw data

Open Data and Private Data

Example of gathering image data using webcam

```
import cv2
from google.colab.patches import cv2_imshow

webcam = cv2.VideoCapture(0)

while True:
    try:
        check, frame = webcam.read()
        if check:  # Check if frame is successfully captured
            print(check)  # prints true as long as the webcam is running
            print(frame)  # prints matrix values of each frame
            cv2_imshow(frame)  # Display the frame
            key = cv2.waitKey(1)

            if key == ord('s'):
                cv2.imwrite(filename='saved_img.jpg', img=frame)
                webcam.release()
                img_new = cv2.imread('saved_img.jpg', cv2.IMREAD_GRAYSCALE)
                img_new = cv2_imshow(img_new)
                cv2.waitKey(1650)
                cv2.destroyAllWindows()
                print("Processing image...")
                img_ = cv2.imread('saved_img.jpg', cv2.IMREAD_ANYCOLOR)
                print("Converting RGB image to grayscale...")
                gray = cv2.cvtColor(img_, cv2.COLOR_BGR2GRAY)
                print("Converted RGB image to grayscale...")
                print("Resizing image to 28x28 scale...")
                img_ = cv2.resize(gray, (28, 28))
                print("Resized...")
                img_resized = cv2.imwrite(filename='saved_img-final.jpg', img=img_)
                print("Image saved!")
                break

            elif key == ord('q'):
                print("Turning off camera.")
                webcam.release()
                print("Camera off.")
                print("Program ended.")
                cv2.destroyAllWindows()
                break
        else:
            print("Unable to capture frame. Check your webcam connection.")
            break

    except KeyboardInterrupt:
        print("Turning off camera.")
        webcam.release()
        print("Camera off.")
        print("Program ended.")
        cv2.destroyAllWindows()
        break
```

```
Unable to capture frame. Check your webcam connection.
```

Example of gathering voice data using microphone

```
!pip3 install sounddevice
```

```
Collecting sounddevice
  Downloading sounddevice-0.4.6-py3-none-any.whl (31 kB)
Requirement already satisfied: CFFI>=1.0 in /usr/local/lib/python3.10/dist-packages (from sounddevice) (1.16.0)
Requirement already satisfied: pycparser in /usr/local/lib/python3.10/dist-packages (from CFFI>=1.0->sounddevice) (2.21)
Installing collected packages: sounddevice
Successfully installed sounddevice-0.4.6
```

```
!pip3 install wavio
```

```
Collecting wavio
  Downloading wavio-0.0.8-py3-none-any.whl (9.4 kB)
Requirement already satisfied: numpy>=1.19.0 in /usr/local/lib/python3.10/dist-packages (from wavio) (1.25.2)
Installing collected packages: wavio
Successfully installed wavio-0.0.8
```

```
!pip3 install scipy
```

```
Requirement already satisfied: scipy in /usr/local/lib/python3.10/dist-packages (1.11.4)
Requirement already satisfied: numpy<1.28.0,>=1.21.6 in /usr/local/lib/python3.10/dist-packages (from scipy) (1.25.2)
```

```
!apt-get install libportaudio2
```

```
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
The following NEW packages will be installed:
  libportaudio2
0 upgraded, 1 newly installed, 0 to remove and 39 not upgraded.
Need to get 65.3 kB of archives.
After this operation, 223 kB of additional disk space will be used.
Get:1 http://archive.ubuntu.com/ubuntu jammy/universe amd64 libportaudio2 amd64 19.6.0-1.1 [65.3 kB]
Fetched 65.3 kB in 0s (146 kB/s)
Selecting previously unselected package libportaudio2:amd64.
(Reading database ... 121753 files and directories currently installed.)
Preparing to unpack .../libportaudio2_19.6.0-1.1_amd64.deb ...
Unpacking libportaudio2:amd64 (19.6.0-1.1) ...
Setting up libportaudio2:amd64 (19.6.0-1.1) ...
Processing triggers for libc-bin (2.35-0ubuntu3.4) ...
/sbin/ldconfig.real: /usr/local/lib/libtbbmalloc.so.2 is not a symbolic link

/sbin/ldconfig.real: /usr/local/lib/libtbbbind.so.3 is not a symbolic link

/sbin/ldconfig.real: /usr/local/lib/libtbbbind_2_5.so.3 is not a symbolic link

/sbin/ldconfig.real: /usr/local/lib/libtbb.so.12 is not a symbolic link

/sbin/ldconfig.real: /usr/local/lib/libtbbmalloc_proxy.so.2 is not a symbolic link

/sbin/ldconfig.real: /usr/local/lib/libtbbbind_2_0.so.3 is not a symbolic link
```

```
!pip install sounddevice --upgrade
```

```
Requirement already satisfied: sounddevice in /usr/local/lib/python3.10/dist-packages (0.4.6)
Requirement already satisfied: CFFI>=1.0 in /usr/local/lib/python3.10/dist-packages (from sounddevice) (1.16.0)
Requirement already satisfied: pycparser in /usr/local/lib/python3.10/dist-packages (from CFFI>=1.0->sounddevice) (2.21)
```

```
# import required libraries
import sounddevice as sd
from scipy.io.wavfile import write
import wavio as wv

# Sampling frequency
freq = 48000

# Recording duration
duration = 5

# Start recorder with the given values
# of duration and sample frequency
recording = sd.rec(int(duration * freq),
 samplerate=freq, channels=2)

# Record audio for the given number of seconds
sd.wait()

# This will convert the NumPy array to an audio
# file with the given sampling frequency
write("recording0.wav", freq, recording)
# Convert the NumPy array to audio file
wv.write("recording1.wav", recording, freq, sampwidth=2)
```

```
---------------------------------------------------------------------------
PortAudioError                            Traceback (most recent call last)
<ipython-input-7-3f46ebebbb4e> in <cell line: 14>()
     12 # Start recorder with the given values
     13 # of duration and sample frequency
---> 14 recording = sd.rec(int(duration * freq),
     15  samplerate=freq, channels=2)
     16

                              5 frames
/usr/local/lib/python3.10/dist-packages/sounddevice.py in query_devices(device, kind)
    567     info = _lib.Pa_GetDeviceInfo(device)
    568     if not info:
--> 569         raise PortAudioError(f'Error querying device {device}')
    570     assert info.structVersion == 2
    571     name_bytes = _ffi.string(info.name)

PortAudioError: Error querying device -1
```

Web Scraping

Web scraping, web harvesting, or web data extraction is data scraping used for extracting data from websites. The web scraping software may directly access the World Wide Web using the Hypertext Transfer Protocol or a web browser. While web scraping can be done manually by a software user, the term typically refers to automated processes implemented using a bot or web crawler. It is a form of copying in which specific data is gathered and copied from the web, typically into a central local database or spreadsheet, for later retrieval or analysis.

Image Scraping using BeautifulSoup and Request

```
!pip install bs4
```

```
Collecting bs4
  Downloading bs4-0.0.2-py2.py3-none-any.whl (1.2 kB)
Requirement already satisfied: beautifulsoup4 in /usr/local/lib/python3.10/dist-packages (from bs4) (4.12.3)
Requirement already satisfied: soupsieve>1.2 in /usr/local/lib/python3.10/dist-packages (from beautifulsoup4->bs4) (2.5)
Installing collected packages: bs4
Successfully installed bs4-0.0.2
```

```
pip install requests
```

```
Requirement already satisfied: requests in /usr/local/lib/python3.10/dist-packages (2.31.0)
Requirement already satisfied: charset-normalizer<4,>=2 in /usr/local/lib/python3.10/dist-packages (from requests) (3.3.2)
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.10/dist-packages (from requests) (3.6)
Requirement already satisfied: urllib3<3,>=1.21.1 in /usr/local/lib/python3.10/dist-packages (from requests) (2.0.7)
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.10/dist-packages (from requests) (2024.2.2)
```

```
import requests
from bs4 import BeautifulSoup

def getdata(url):
 r = requests.get(url)
 return r.text

htmldata = getdata("https://www.google.com/")
soup = BeautifulSoup(htmldata, 'html.parser')
for item in soup.find_all('img'):
 print(item['src'])
```

```
/images/branding/googlelogo/1x/googlelogo_white_background_color_272x92dp.png
```

```
pip install selenium
```

```
Collecting selenium
  Downloading selenium-4.18.1-py3-none-any.whl (10.0 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 10.0/10.0 MB 24.8 MB/s eta 0:00:00
Requirement already satisfied: urllib3[socks]<3,>=1.26 in /usr/local/lib/python3.10/dist-packages (from selenium) (2.0.7)
Collecting trio~=0.17 (from selenium)
  Downloading trio-0.25.0-py3-none-any.whl (467 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 467.2/467.2 kB 36.7 MB/s eta 0:00:00
Collecting trio-websocket~=0.9 (from selenium)
  Downloading trio_websocket-0.11.1-py3-none-any.whl (17 kB)
Requirement already satisfied: certifi>=2021.10.8 in /usr/local/lib/python3.10/dist-packages (from selenium) (2024.2.2)
Requirement already satisfied: typing_extensions>=4.9.0 in /usr/local/lib/python3.10/dist-packages (from selenium) (4.10.0)
Requirement already satisfied: attrs>=23.2.0 in /usr/local/lib/python3.10/dist-packages (from trio~=0.17->selenium) (23.2.0)
Requirement already satisfied: sortedcontainers in /usr/local/lib/python3.10/dist-packages (from trio~=0.17->selenium) (2.4.0)
Requirement already satisfied: idna in /usr/local/lib/python3.10/dist-packages (from trio~=0.17->selenium) (3.6)
Collecting outcome (from trio~=0.17->selenium)
  Downloading outcome-1.3.0.post0-py2.py3-none-any.whl (10 kB)
Requirement already satisfied: sniffio>=1.3.0 in /usr/local/lib/python3.10/dist-packages (from trio~=0.17->selenium) (1.3.1)
Requirement already satisfied: exceptiongroup in /usr/local/lib/python3.10/dist-packages (from trio~=0.17->selenium) (1.2.0)
Collecting wsproto>=0.14 (from trio-websocket~=0.9->selenium)
  Downloading wsproto-1.2.0-py3-none-any.whl (24 kB)
Requirement already satisfied: pysocks!=1.5.7,<2.0,>=1.5.6 in /usr/local/lib/python3.10/dist-packages (from urllib3[socks]<3,>=1.26->selenium) (1.7.1)
Collecting h11<1,>=0.9.0 (from wsproto>=0.14->trio-websocket~=0.9->selenium)
  Downloading h11-0.14.0-py3-none-any.whl (58 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 58.3/58.3 kB 7.9 MB/s eta 0:00:00
Installing collected packages: outcome, h11, wsproto, trio, trio-websocket, selenium
Successfully installed h11-0.14.0 outcome-1.3.0.post0 selenium-4.18.1 trio-0.25.0 trio-websocket-0.11.1 wsproto-1.2.0
```

Image Scraping using Selenium

```
!pip install selenium
!apt-get update # to update ubuntu to correctly run apt install
!apt install chromium-chromedriver
!cp /usr/lib/chromium-browser/chromedriver /usr/bin
```

```
import sys
sys.path.insert(0,'/usr/lib/chromium-browser/chromedriver')

from selenium import webdriver
import time
import requests
import shutil
import os
import getpass
import urllib.request
import io
import time
from PIL import Image
user = getpass.getuser()
chrome_options = webdriver.ChromeOptions()
chrome_options.add_argument('--headless')
chrome_options.add_argument('--no-sandbox')
chrome_options.add_argument('--disable-dev-shm-usage')
driver = webdriver.Chrome('chromedriver',chrome_options=chrome_options)

search_url = "https://www.google.com/search?q={q}&tbm=isch&tbs=sur%3Afc&hl=en&ved=0CAIQpwVqFwoTCKCa1c6s4-oCFQAAAAAdAAAAABAC&biw=1251&bih=568"
driver.get(search_url.format(q='Car'))

def scroll_to_end(driver):
  driver.execute_script("window.scrollTo(0, document.body.scrollHeight);")
  time.sleep(5)# dding a pause for 5 seconds to let the page load.

def getImageUrls(name,totalImgs,driver): # Function to get image URLs for a given search term.
  search_url = "https://www.google.com/search?q={q}&tbm=isch&tbs=sur%3Afc&hl=en&ved=0CAIQpwVqFwoTCKCa1c6s4-oCFQAAAAAdAAAAABAC&biw=1251&bih=568"
  driver.get(search_url.format(q=name))
  img_urls = set()
  img_count = 0
  results_start = 0

  while(img_count<totalImgs): #loop to extract actual images until the desired number is reached.

    scroll_to_end(driver)

    thumbnail_results = driver.find_elements_by_xpath("//img[contains(@class,'Q4LuWd')]")
    totalResults=len(thumbnail_results)
    print(f"Found: {totalResults} search results. Extracting links from{results_start}:{totalResults}")

    for img in thumbnail_results[results_start:totalResults]:

      img.click()# clicking on the thumbnail image.
      time.sleep(2) #pause for 2 sec
      actual_images = driver.find_elements_by_css_selector('img.n3VNCb')
      for actual_image in actual_images:
        if actual_image.get_attribute('src') and 'https' in actual_image.get_attribute('src'):
          img_urls.add(actual_image.get_attribute('src'))

      img_count=len(img_urls)

      if img_count >= totalImgs:
        print(f"Found: {img_count} image links")
        break
      else:
        print("Found:", img_count, "looking for more image links ...")
        load_more_button = driver.find_element_by_css_selector(".mye4qd")
        driver.execute_script("document.querySelector('.mye4qd').click();")
        results_start = len(thumbnail_results)
  return img_urls

def downloadImages(folder_path,file_name,url):
    try:
      image_content = requests.get(url).content
    except Exception as e:
      print(f"ERROR - COULD NOT DOWNLOAD {url} - {e}")

    try:
      image_file = io.BytesIO(image_content)
      image = Image.open(image_file).convert('RGB')

      file_path = os.path.join(folder_path, file_name)

      with open(file_path, 'wb') as f:
        image.save(f, "JPEG", quality=85)
      print(f"SAVED - {url} - AT: {file_path}")
    except Exception as e:
      print(f"ERROR - COULD NOT SAVE {url} - {e}")

def saveInDestFolder(searchNames,destDir,totalImgs,driver):
    for name in list(searchNames):
      path=os.path.join(destDir,name)
      if not os.path.isdir(path):
        os.mkdir(path)
      print('Current Path',path)
      totalLinks=getImageUrls(name,totalImgs,driver)
      print('totalLinks',totalLinks)
    if totalLinks is None:
        print('images not found for :',name)

    else:
      for i, link in enumerate(totalLinks):
        file_name = f"{i:150}.jpg"
        downloadImages(path,file_name,link)

searchNames=['cat']
destDir=f'/content/drive/My Drive/Colab Notebooks/Dataset/'
totalImgs=5

saveInDestFolder(searchNames,destDir,totalImgs,driver)
```

```
Requirement already satisfied: selenium in /usr/local/lib/python3.10/dist-packages (4.18.1)
Requirement already satisfied: urllib3[socks]<3,>=1.26 in /usr/local/lib/python3.10/dist-packages (from selenium) (2.0.7
Requirement already satisfied: trio~=0.17 in /usr/local/lib/python3.10/dist-packages (from selenium) (0.25.0)
Requirement already satisfied: trio-websocket~=0.9 in /usr/local/lib/python3.10/dist-packages (from selenium) (0.11.1)
Requirement already satisfied: certifi>=2021.10.8 in /usr/local/lib/python3.10/dist-packages (from selenium) (2024.2.2)
Requirement already satisfied: typing_extensions>=4.9.0 in /usr/local/lib/python3.10/dist-packages (from selenium) (4.10
Requirement already satisfied: attrs>=23.2.0 in /usr/local/lib/python3.10/dist-packages (from trio~=0.17->selenium) (23.
Requirement already satisfied: sortedcontainers in /usr/local/lib/python3.10/dist-packages (from trio~=0.17->selenium) (
Requirement already satisfied: idna in /usr/local/lib/python3.10/dist-packages (from trio~=0.17->selenium) (3.6)
Requirement already satisfied: outcome in /usr/local/lib/python3.10/dist-packages (from trio~=0.17->selenium) (1.3.0.pos
Requirement already satisfied: sniffio>=1.3.0 in /usr/local/lib/python3.10/dist-packages (from trio~=0.17->selenium) (1.
Requirement already satisfied: exceptiongroup in /usr/local/lib/python3.10/dist-packages (from trio~=0.17->selenium) (1.
Requirement already satisfied: wsproto>=0.14 in /usr/local/lib/python3.10/dist-packages (from trio-websocket~=0.9->selen
Requirement already satisfied: pysocks!=1.5.7,<2.0,>=1.5.6 in /usr/local/lib/python3.10/dist-packages (from urllib3[sock
Requirement already satisfied: h11<1,>=0.9.0 in /usr/local/lib/python3.10/dist-packages (from wsproto>=0.14->trio-websoc
Get:1 http://security.ubuntu.com/ubuntu jammy-security InRelease [110 kB]
Get:2 https://cloud.r-project.org/bin/linux/ubuntu jammy-cran40/ InRelease [3,626 B]
Get:3 https://developer.download.nvidia.com/compute/cuda/repos/ubuntu2204/x86_64  InRelease [1,581 B]
Hit:4 http://archive.ubuntu.com/ubuntu jammy InRelease
Get:5 http://archive.ubuntu.com/ubuntu jammy-updates InRelease [119 kB]
Hit:6 http://archive.ubuntu.com/ubuntu jammy-backports InRelease
Hit:7 https://ppa.launchpadcontent.net/c2d4u.team/c2d4u4.0+/ubuntu jammy InRelease
Hit:8 https://ppa.launchpadcontent.net/deadsnakes/ppa/ubuntu jammy InRelease
Get:9 https://developer.download.nvidia.com/compute/cuda/repos/ubuntu2204/x86_64  Packages [773 kB]
Hit:10 https://ppa.launchpadcontent.net/graphics-drivers/ppa/ubuntu jammy InRelease
Hit:11 https://ppa.launchpadcontent.net/ubuntugis/ppa/ubuntu jammy InRelease
Get:12 http://archive.ubuntu.com/ubuntu jammy-updates/main amd64 Packages [1,898 kB]
Get:13 http://archive.ubuntu.com/ubuntu jammy-updates/universe amd64 Packages [1,356 kB]
Fetched 4,261 kB in 4s (1,106 kB/s)
Reading package lists... Done
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
The following additional packages will be installed:
  apparmor chromium-browser libfuse3-3 liblzo2-2 libudev1 snapd squashfs-tools systemd-hwe-hwdb
  udev
Suggested packages:
  apparmor-profiles-extra apparmor-utils fuse3 zenity | kdialog
The following NEW packages will be installed:
  apparmor chromium-browser chromium-chromedriver libfuse3-3 liblzo2-2 snapd squashfs-tools
  systemd-hwe-hwdb udev
The following packages will be upgraded:
  libudev1
1 upgraded, 9 newly installed, 0 to remove and 38 not upgraded.
Need to get 26.4 MB of archives.
After this operation, 116 MB of additional disk space will be used.
Get:1 http://archive.ubuntu.com/ubuntu jammy-updates/main amd64 apparmor amd64 3.0.4-2ubuntu2.3 [595 kB]
Get:2 http://archive.ubuntu.com/ubuntu jammy/main amd64 liblzo2-2 amd64 2.10-2build3 [53.7 kB]
Get:3 http://archive.ubuntu.com/ubuntu jammy/main amd64 squashfs-tools amd64 1:4.5-3build1 [159 kB]
Get:4 http://archive.ubuntu.com/ubuntu jammy-updates/main amd64 libudev1 amd64 249.11-0ubuntu3.12 [78.2 kB]
Get:5 http://archive.ubuntu.com/ubuntu jammy-updates/main amd64 udev amd64 249.11-0ubuntu3.12 [1,557 kB]
Get:6 http://archive.ubuntu.com/ubuntu jammy/main amd64 libfuse3-3 amd64 3.10.5-1build1 [81.2 kB]
Get:7 http://archive.ubuntu.com/ubuntu jammy-updates/main amd64 snapd amd64 2.58+22.04.1 [23.8 MB]
Get:8 http://archive.ubuntu.com/ubuntu jammy-updates/universe amd64 chromium-browser amd64 1:85.0.4183.83-0ubuntu2.22.04
Get:9 http://archive.ubuntu.com/ubuntu jammy-updates/universe amd64 chromium-chromedriver amd64 1:85.0.4183.83-0ubuntu2.
Get:10 http://archive.ubuntu.com/ubuntu jammy-updates/main amd64 systemd-hwe-hwdb all 249.11.5 [3,228 B]
Fetched 26.4 MB in 1s (20.7 MB/s)
Preconfiguring packages ...
Selecting previously unselected package apparmor.
(Reading database ... 121759 files and directories currently installed.)
Preparing to unpack .../apparmor_3.0.4-2ubuntu2.3_amd64.deb ...
Unpacking apparmor (3.0.4-2ubuntu2.3) ...
Selecting previously unselected package liblzo2-2:amd64.
Preparing to unpack .../liblzo2-2_2.10-2build3_amd64.deb ...
Unpacking liblzo2-2:amd64 (2.10-2build3) ...
Selecting previously unselected package squashfs-tools.
Preparing to unpack .../squashfs-tools_1%3a4.5-3build1_amd64.deb ...
Unpacking squashfs-tools (1:4.5-3build1) ...
Preparing to unpack .../libudev1_249.11-0ubuntu3.12_amd64.deb ...
Unpacking libudev1:amd64 (249.11-0ubuntu3.12) over (249.11-0ubuntu3.10) ...
Setting up libudev1:amd64 (249.11-0ubuntu3.12) ...
Selecting previously unselected package udev.
(Reading database ... 121967 files and directories currently installed.)
Preparing to unpack .../udev_249.11-0ubuntu3.12_amd64.deb ...
Unpacking udev (249.11-0ubuntu3.12) ...
Selecting previously unselected package libfuse3-3:amd64.
Preparing to unpack .../libfuse3-3_3.10.5-1build1_amd64.deb ...
Unpacking libfuse3-3:amd64 (3.10.5-1build1) ...
Selecting previously unselected package snapd.
Preparing to unpack .../snapd_2.58+22.04.1_amd64.deb ...
Unpacking snapd (2.58+22.04.1) ...
Setting up apparmor (3.0.4-2ubuntu2.3) ...
Created symlink /etc/systemd/system/sysinit.target.wants/apparmor.service → /lib/systemd/system/apparmor.service.
Setting up liblzo2-2:amd64 (2.10-2build3) ...
Setting up squashfs-tools (1:4.5-3build1) ...
Setting up udev (249.11-0ubuntu3.12) ...
invoke-rc.d: could not determine current runlevel
invoke-rc.d: policy-rc.d denied execution of start.
Setting up libfuse3-3:amd64 (3.10.5-1build1) ...
Setting up snapd (2.58+22.04.1) ...
Created symlink /etc/systemd/system/multi-user.target.wants/snapd.aa-prompt-listener.service → /lib/systemd/system/snapd
Created symlink /etc/systemd/system/multi-user.target.wants/snapd.apparmor.service → /lib/systemd/system/snapd.apparmor.
Created symlink /etc/systemd/system/multi-user.target.wants/snapd.autoimport.service → /lib/systemd/system/snapd.autoimp
Created symlink /etc/systemd/system/multi-user.target.wants/snapd.core-fixup.service → /lib/systemd/system/snapd.core-fi
Created symlink /etc/systemd/system/multi-user.target.wants/snapd.recovery-chooser-trigger.service → /lib/systemd/system
Created symlink /etc/systemd/system/multi-user.target.wants/snapd.seeded.service → /lib/systemd/system/snapd.seeded.serv
Created symlink /etc/systemd/system/cloud-final.service.wants/snapd.seeded.service → /lib/systemd/system/snapd.seeded.se
Unit /lib/systemd/system/snapd.seeded.service is added as a dependency to a non-existent unit cloud-final.service.
Created symlink /etc/systemd/system/multi-user.target.wants/snapd.service → /lib/systemd/system/snapd.service.
Created symlink /etc/systemd/system/timers.target.wants/snapd.snap-repair.timer → /lib/systemd/system/snapd.snap-repair.
Created symlink /etc/systemd/system/sockets.target.wants/snapd.socket → /lib/systemd/system/snapd.socket.
Created symlink /etc/systemd/system/final.target.wants/snapd.system-shutdown.service → /lib/systemd/system/snapd.system-
Selecting previously unselected package chromium-browser.
(Reading database ... 122200 files and directories currently installed.)
Preparing to unpack .../chromium-browser_1%3a85.0.4183.83-0ubuntu2.22.04.1_amd64.deb ...
=> Installing the chromium snap
==> Checking connectivity with the snap store
===> System doesn't have a working snapd, skipping
Unpacking chromium-browser (1:85.0.4183.83-0ubuntu2.22.04.1) ...
Selecting previously unselected package chromium-chromedriver.
Preparing to unpack .../chromium-chromedriver_1%3a85.0.4183.83-0ubuntu2.22.04.1_amd64.deb ...
Unpacking chromium-chromedriver (1:85.0.4183.83-0ubuntu2.22.04.1) ...
Selecting previously unselected package systemd-hwe-hwdb
```

```
Selecting previously unselected package systemd-hwe-hwdb.
Preparing to unpack .../systemd-hwe-hwdb_249.11.5_all.deb ...
Unpacking systemd-hwe-hwdb (249.11.5) ...
Setting up systemd-hwe-hwdb (249.11.5) ...
Setting up chromium-browser (1:85.0.4183.83-0ubuntu2.22.04.1) ...
update-alternatives: using /usr/bin/chromium-browser to provide /usr/bin/x-www-browser (x-www-browser) in auto mode
update-alternatives: using /usr/bin/chromium-browser to provide /usr/bin/gnome-www-browser (gnome-www-browser) in auto m
Setting up chromium-chromedriver (1:85.0.4183.83-0ubuntu2.22.04.1) ...
Processing triggers for udev (249.11-0ubuntu3.12) ...
Processing triggers for hicolor-icon-theme (0.17-2) ...
Processing triggers for libc-bin (2.35-0ubuntu3.4) ...
/sbin/ldconfig.real: /usr/local/lib/libtbbmalloc.so.2 is not a symbolic link

/sbin/ldconfig.real: /usr/local/lib/libtbbbind.so.3 is not a symbolic link

/sbin/ldconfig.real: /usr/local/lib/libtbbbind_2_5.so.3 is not a symbolic link

/sbin/ldconfig.real: /usr/local/lib/libtbb.so.12 is not a symbolic link

/sbin/ldconfig.real: /usr/local/lib/libtbbmalloc_proxy.so.2 is not a symbolic link

/sbin/ldconfig.real: /usr/local/lib/libtbbbind_2_0.so.3 is not a symbolic link

Processing triggers for man-db (2.10.2-1) ...
Processing triggers for dbus (1.12.20-2ubuntu4.1) ...
cp: '/usr/lib/chromium-browser/chromedriver' and '/usr/bin/chromedriver' are the same file
---------------------------------------------------------------------------
TypeError                                 Traceback (most recent call last)
<ipython-input-12-49711713c5eb> in <cell line: 23>()
     21 chrome_options.add_argument('--no-sandbox')
     22 chrome_options.add_argument('--disable-dev-shm-usage')
---> 23 driver = webdriver.Chrome('chromedriver',chrome_options=chrome_options)
     24
     25 search_url = "https://www.google.com/search?q={q}&tbm=isch&tbs=sur%3Afc&hl=en&ved=0CAIQpwVqFwoTCKCa1c6s4-
oCFQAAAAAdAAAAABAC&biw=1251&bih=568"

TypeError: WebDriver.__init__() got an unexpected keyword argument 'chrome_options'
```

Web Scraping of Movies Information using BeautifulSoup

```
from requests import get
url = 'https://hurawatch.cc/search/logan'
response = get(url) # sends a get request to the specified URL
print(response.text[:500]) # prints the first 500 characters of the HTML content of the response
```

```
<!DOCTYPE html>
<html xmlns="http://www.w3.org/1999/xhtml" xml:lang="en" lang="en">
<head>
    <title>Search results for &#39;logan&#39; Movies &amp; Tv Series Hurawatch</title>
<meta http-equiv="Content-Type" content="text/html; charset=UTF-8"/>

    <meta name="robots" content="index, follow">
    <meta name="revisit-after" content="1 days">

<meta http-equiv="content-language" content="en"/>
<link rel="dns-prefetch" href="//www.google-analytics.com">
<link rel="dns-prefetch" href="//www.gstat
```

```
from bs4 import BeautifulSoup
html_soup = BeautifulSoup(response.text, 'html.parser')
headers = {'Accept-Language': 'en-US,en;q=0.8'} # defines the headers for the HTTP request.
type(html_soup) # returns the type of the parsed HTML content
```

```
bs4.BeautifulSoup
def __call__(*args, **kwargs)

/usr/local/lib/python3.10/dist-packages/bs4/__init__.py
A data structure representing a parsed HTML or XML document.

Most of the methods you'll call on a BeautifulSoup object are inherited from
PageElement or Tag.
```

```
movie_containers = html_soup.find_all('div', class_ = 'flw-item')
print(type(movie_containers))
print(len(movie_containers))
```

```
<class 'bs4.element.ResultSet'>
6
```

```
first_movie = movie_containers[0]
first_movie
```

```
<div class="flw-item">
<div class="film-poster">
<div class="pick film-poster-quality">HD</div>
<img alt="Logan" class="film-poster-img lazyload" data-src="https://img.hurawatch.cc/xxrz/250x400/348/16/f7/16f7b48a3df281f25cb746394488ea9d/16f7b48a3df281f25cb746394488ea9d.jpg"
title="Logan"/>
<a class="film-poster-ahref flw-item-tip" href="/movie/watch-logan-online-19754" title="Logan"><i class="fa fa-play"></i></a>
</div>
<div class="film-detail">
<h2 class="film-name"><a href="/movie/watch-logan-online-19754" title="Logan">Logan</a>
</h2>
<div class="fd-infor">
<span class="fdi-item">2017</span>
<span class="dot"></span>
<span class="fdi-item fdi-duration">137m</span>
<span class="float-right fdi-type">Movie</span>
</div>
<div class="clearfix"></div>
</div>
<div class="clearfix"></div>
</div>
```

```
first_movie.div # accessing the first movie with "div" tag with in first_movie object
```

```
<div class="film-poster">
<div class="pick film-poster-quality">HD</div>
<img alt="Logan" class="film-poster-img lazyload" data-src="https://img.hurawatch.cc/xxrz/250x400/348/16/f7/16f7b48a3df281f25cb746394488ea9d/16f7b48a3df281f25cb746394488ea9d.jpg"
title="Logan"/>
<a class="film-poster-ahref flw-item-tip" href="/movie/watch-logan-online-19754" title="Logan"><i class="fa fa-play"></i></a>
</div>
```

```
first_movie.a # accessing the first movie with "a" tag with in first_movie object
```

```
<a class="film-poster-ahref flw-item-tip" href="/movie/watch-logan-online-19754" title="Logan"><i class="fa fa-play"></i></a>
```

```
first_movie.h2 #accessing the first movie with "h2" tag with in first_movie object
```

```
<h2 class="film-name"><a href="/movie/watch-logan-online-19754" title="Logan">Logan</a>
</h2>
```

```
first_movie.h2.a #accessing the first movie with "a" tag with in first_movie with "h2" with in the first_move object
```

```
<a href="/movie/watch-logan-online-19754" title="Logan">Logan</a>
```

```
first_name = first_movie.h2.a.text
first_name
```

```
'Logan'
```

```
first_year = first_movie.find('span', class_='fdi-item')
if first_year:
    print(first_year.text)
else:
    print("Year information not found")
```

```
2017
```

```
first_year = first_year.text
first_year
```

```
'2017'
```

The IMDB rating

```
first_movie.strong
```

```
first_imdb = float(first_movie.strong.text)
first_imdb
```

```
---------------------------------------------------------------------------
AttributeError                            Traceback (most recent call last)
<ipython-input-25-92faeb51c9f2> in <cell line: 1>()
----> 1 first_imdb = float(first_movie.strong.text)
      2 first_imdb

AttributeError: 'NoneType' object has no attribute 'text'
```

the metascore

```
first_mscore = first_movie.find('span', class_ = 'metascore favorable')
first_mscore = int(first_mscore.text)
print(first_mscore)
```

```
---------------------------------------------------------------------------
AttributeError                            Traceback (most recent call last)
<ipython-input-26-889bc009bd72> in <cell line: 2>()
      1 first_mscore = first_movie.find('span', class_ = 'metascore favorable')
----> 2 first_mscore = int(first_mscore.text)
      3 print(first_mscore)

AttributeError: 'NoneType' object has no attribute 'text'
```

the number of votes

```
first_votes = first_movie.find('span', attrs = {'name':'nv'})
first_votes
```

```
first_votes['data-value']
```

```
---------------------------------------------------------------------------
TypeError                                 Traceback (most recent call last)
<ipython-input-28-2d836d02a09a> in <cell line: 1>()
----> 1 first_votes['data-value']

TypeError: 'NoneType' object is not subscriptable
```

```
first_votes = int(first_votes['data-value'])
```

```
---------------------------------------------------------------------------
TypeError                                 Traceback (most recent call last)
<ipython-input-29-e337b21fe258> in <cell line: 1>()
----> 1 first_votes = int(first_votes['data-value'])

TypeError: 'NoneType' object is not subscriptable
```

```
first_duration = first_movie.find('span', class_='fdi-item fdi-duration')
if first_duration:
    print(first_duration.text)
else:
    print("Duration information not found")
```

```
137m
```

the script

```
# Lists to store the scraped data in
names = []
years = []
durations = []


for container in movie_containers:
    if container.find('div', class_='fd-infor') is not None:
        # Name
        name = container.h2.a.text
        names.append(name)

        # Year
        year = container.find('span', class_='fdi-item').text
        years.append(year)

        # Duration
        duration_element = container.find('span', class_='fdi-item fdi-duration')
        if duration_element is not None:
            duration = duration_element.text
        else:
            duration = 'Not available'
        durations.append(duration)

print(names)
print(years)
print(durations)
```

```
['Logan', 'Logan Lucky', 'The Taking of Deborah Logan', 'The Night Logan Woke Up', 'The Two Worlds of Jennie Logan', "Logan's Run"]
['2017', '2017', '2014', 'SS 1', '1979', '1976']
['137m', '119m', '90m', 'Not available', '94m', '119m']
```

```
import pandas as pd
test_df = pd.DataFrame({
'movie': names,
'year': years,
'duration': duration,
})
print(test_df.info())
test_df
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6 entries, 0 to 5
Data columns (total 3 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   movie     6 non-null      object
 1   year      6 non-null      object
 2   duration  6 non-null      object
dtypes: object(3)
memory usage: 272.0+ bytes
None
```

	movie	year	duration
0	Logan	2017	119m
1	Logan Lucky	2017	119m
2	The Taking of Deborah Logan	2014	119m
3	The Night Logan Woke Up	SS 1	119m
4	The Two Worlds of Jennie Logan	1979	119m
5	Logan's Run	1976	119m

Next steps: View recommended plots

The script for multiple pages

```
[]
[]
[]
```

```
movie_ratings = pd.DataFrame({
```

```
'movie': names,
'year': years,
'duration': duration,
})
print(movie_ratings.info)
movie_ratings.head(10)
```

```
<bound method DataFrame.info of                                  movie  year duration
0                           Logan  2017     119m
1                      Logan Lucky  2017     119m
2     The Taking of Deborah Logan  2014     119m
3         The Night Logan Woke Up  SS 1     119m
4  The Two Worlds of Jennie Logan  1979     119m
5                      Logan's Run  1976     119m>
```

	movie	year	duration
0	Logan	2017	119m
1	Logan Lucky	2017	119m
2	The Taking of Deborah Logan	2014	119m
3	The Night Logan Woke Up	SS 1	119m
4	The Two Worlds of Jennie Logan	1979	119m
5	Logan's Run	1976	119m

Next steps: View recommended plots

```
import pandas as pd
movie_df = pd.DataFrame({
'movie': names,
'year': years,
'duration': duration,
})
print(movie_df.info())
movie_df
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6 entries, 0 to 5
Data columns (total 3 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   movie     6 non-null      object
 1   year      6 non-null      object
 2   duration  6 non-null      object
dtypes: object(3)
memory usage: 272.0+ bytes
None
```

	movie	year	duration
0	Logan	2017	119m
1	Logan Lucky	2017	119m
2	The Taking of Deborah Logan	2014	119m
3	The Night Logan Woke Up	SS 1	119m
4	The Two Worlds of Jennie Logan	1979	119m
5	Logan's Run	1976	119m

Next steps: View recommended plots

```
movie_df.to_csv('/content/movie_df.csv')
```

Data Preparation

- Collected data may not be compatible or formatted correctly
- Data must be prepared before it can be added to a data set
- Extract, Transform and Load (ETL)
  - process for collecting data from a variety of sources, transforming the data, and then loading the data into a database

Data preprocessing

Data Processing is a process of cleaning the raw data i.e. the data is collected in the real world and is converted to a clean data set. In other words, whenever the data is gathered from different sources it is collected in a raw format and this data isn't feasible for the analysis. Therefore, certain steps are executed to convert the data into a small clean data set, this part of the process is called as data preprocessing. Most of the real-world data is messy, some of these types of data are: 1. Missing data: Missing data can be found when it is not continuously created or due to technical issues in the application (IOT system). 2. Noisy Data This type of data is also called outliners, this can occur due to human errors (human manually gathering the data) or some technical problem of the device at the time of collection of data. 3. Inconsistent data: This type of data might be collected due to human errors (mistakes with the name or values) or duplication of data. These are some of the basic pre processing techniques that can be used to convert raw data. 1. Conversion of data: As we know that Machine Learning models can only handle numeric features, hence categorical and ordinal data must be somehow converted into numeric features. 2. Ignoring the missing values: Whenever we encounter missing data in the data set then we can remove the row or column of data depending on our need. This method is known to be efficient but it shouldn't be performed if there are a lot of missing values in the dataset. 3. Filling the missing values: Whenever we encounter missing data in the data set then we can fill the missing data manually, most commonly the mean, median or highest frequency value is used.

1. Machine learning: If we have some missing data then we can predict what data shall be present at the empty position by using the existing data. 5. Outliers detection: There are some error data that might be present in our data set that deviates drastically from other observations in a data set. [Example: human weight = 800 Kg; due to mistyping of extra 0]

```
movie_df['year'].unique()
```

```
array(['2017', '2014', 'SS 1', '1979', '1976'], dtype=object)
```

```
movie_df.dtypes
```

```
movie       object
year        object
duration    object
dtype: object
```

```
movie_df['year'] = (movie_df.year.apply(lambda x:x.replace('(I)','')))
```

```
movie_df['year'].unique()
```

```
array(['2017', '2014', 'SS 1', '1979', '1976'], dtype=object)
```

```
movie_df['year'] = (movie_df.year.apply(lambda x:x.replace('(II)','')))
```

```
movie_df['year'] = (movie_df.year.apply(lambda x:x.replace('(III)','')))
```

```
movie_df['year'].unique()
```

```
array(['2017', '2014', 'SS 1', '1979', '1976'], dtype=object)
```

```
movie_df['year'] = (movie_df.year.apply(lambda x:x.replace('(','')))
```

```
movie_df['year'].unique()
```

```
array(['2017', '2014', 'SS 1', '1979', '1976'], dtype=object)
```

```
movie_df['year'] = (movie_df.year.apply(lambda x:x.replace(')','')))
```

```
movie_df['year'].unique()
```

```
array(['2017', '2014', 'SS 1', '1979', '1976'], dtype=object)
```

Need to convert SS1 to NaN

```
movie_df['year'] = movie_df['year'].astype(int)
```

```
---------------------------------------------------------------------------
ValueError                                Traceback (most recent call last)
<ipython-input-72-6273ff180712> in <cell line: 1>()
----> 1 movie_df['year'] = movie_df['year'].astype(int)

                                 6 frames
/usr/local/lib/python3.10/dist-packages/pandas/core/dtypes/astype.py in astype_nansafe(arr, dtype, copy, skipna)
    168     if copy or is_object_dtype(arr.dtype) or is_object_dtype(dtype):
    169         # Explicit copy, or required since NumPy can't view from / to object.
--> 170         return arr.astype(dtype, copy=True)
    171
    172     return arr.astype(dtype, copy=copy)

ValueError: invalid literal for int() with base 10: 'SS 1'
```

```
print(movie_df['year'].unique())
```

```
['2017' '2014' 'SS 1' '1979' '1976']
```

```
# Replace non-numeric values with NaN
movie_df['year'] = pd.to_numeric(movie_df['year'], errors='coerce')

# Drop rows with NaN values in the 'year' column
movie_df = movie_df.dropna(subset=['year'])

# Convert 'year' column to integer type
movie_df['year'] = movie_df['year'].astype(int)
```

```
<ipython-input-74-7b367c589dc7>:8: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  movie_df['year'] = movie_df['year'].astype(int)
```

```
movie_df['year'].unique()
```

```
array([2017, 2014, 1979, 1976])
```

```
movie_df.head(10)
```

	movie	year	duration
0	Logan	2017	119m
1	Logan Lucky	2017	119m
2	The Taking of Deborah Logan	2014	119m
4	The Two Worlds of Jennie Logan	1979	119m
5	Logan's Run	1976	119m

Next steps: View recommended plots

```
movie_df.tail(10)
```

	movie	year	duration
0	Logan	2017	119m
1	Logan Lucky	2017	119m
2	The Taking of Deborah Logan	2014	119m
4	The Two Worlds of Jennie Logan	1979	119m
5	Logan's Run	1976	119m

```
movie_df
```

	movie	year	duration
0	Logan	2017	119m
1	Logan Lucky	2017	119m
2	The Taking of Deborah Logan	2014	119m
4	The Two Worlds of Jennie Logan	1979	119m
5	Logan's Run	1976	119m

Next steps: View recommended plots