**Assignment on Platform Research**

**Task #1: Scrapping from TikTok**

Task 1 in is all about creating a python script to scrape tiktok profile data. Written carefully in python considering edge cases. Besides the flexibility of scraping 100 profile by changing the main url from where we want to gather data.

**Methodology:**

First step in this process was selecting a suitable library for the task. I started with the Beautifulsoup library , but faced a few problems, like the same Scrapy library, maybe just because of a few internal problems. However, I overcame this problem by using the Selenium library to get the source code of the site with the help of chrome browser and driver. And to make this happen created an account. Through which I got the urls of the suggested profile and later scraped data from those urls. And this url list can be changed to get the data of the desired profile. Further main tasks started from loading necessary libraries, setting up the web browser, driver and creating driver for my profile page ended up with suggested profile urls, file creates the python script to scrape the data storing in a csv file.

Step by step technical solutions:

Step 1:

It Started with Importing necessary library.

import os

import re

import numpy as np

import pandas as pd

from selenium import webdriver

from bs4 import BeautifulSoup

import time

import csv

import urllib.request

import selenium

from selenium import webdriver

from selenium.webdriver.common.by import By

from selenium.webdriver.common.keys import Keys

from webdriver\_manager.chrome import ChromeDriverManager

from selenium.webdriver.chrome.service import Service

from selenium.webdriver.chrome.options import Options

Step 2:

Setting up the web driver and initializes the driver(Chrome Browser)

options = webdriver.ChromeOptions()

options.add\_experimental\_option("excludeSwitches", ["enable-logging"])

options.add\_argument('--headless')

options.add\_argument('--disable-gpu')

#THIS INITIALIZES THE DRIVER (AKA THE WEB BROWSER)

path = Service(r'C:\Users\Sudhakor\Scrapy\chromedriver.exe')

driver = webdriver.Chrome(options=options, service = path)

Step 3:

Generating the source code for the main page of the site in the profile section and this URL can be changed to get the desired data. Here Green color code is defining the main page

driver.get("https://www.tiktok.com/")

source\_code = driver.page\_source

def url\_body(source\_code):

soup = BeautifulSoup(source\_code, 'html.parser')

url\_body = soup.find\_all('div', class\_ = 'tiktok-1mo2fkg-DivUserLinkContainer e797se20')

return url\_body

body = url\_body(source\_code)

Step 4.

Grabbing all the available url and from the main body part and store in url\_list.

body = url\_body(source\_code)

url\_list = []

def url\_generator(body):

for i in body:

url = 'https://www.tiktok.com' + i.find('a')['href']

url\_list.append(url)

url\_generator(body)

Step: 5

Then comes the main function which will scrape all the required data from a url. And this portion will handle all the necessary adjustments. Set the post limit will be 50 for each and set None otherwise(empty). And Converted the engagement number into thousand(K) standard, calculate average by dividing average by the total numbers of post available in the profile must be less then 50. And ended up returning the fields.

def finding\_content(url):

driver.get(url)

source\_code = driver.page\_source

soup = BeautifulSoup(source\_code, 'html.parser')

name = soup.find('h2', class\_ = "ekmpd5l5").text.strip()

about = soup.find('h2', class\_ = 'e1457k4r3').text

about = re.sub('[^a-zA-Z0-9]', ' ', about)

about = re.sub(' +', ' ', about).strip()

img\_link = soup.find('img', class\_ = 'e1e9er4e1')['src']

following = soup.find('strong', title = 'Following').text

followers = soup.find('strong', title = 'Followers').text

likes = soup.find('strong', title = 'Likes').text

all\_posts = []

engagement = 0

posts\_main = []

try:

posts\_main = soup.find\_all('div', class\_ = 'e19c29qe7')

if len(posts\_main)> 50:

posts\_main = posts\_main[:50]

else:

pass

for post in posts\_main:

post\_link = post.find('a')['href']

views = post.find('strong', class\_ = 'video-count tiktok-1nb981f-StrongVideoCount e148ts222').text

if views.endswith('M'):

views = views.split('M')[0]

try:

fract = views.split('.')[1]

le =(3-len(fract))

zero = ''

for i in np.zeros(le):

zero += '0'

views = (views + zero).replace('.', '')

except:

views = str(views) + ('000')

elif views.endswith('B'):

try:

fract = views.split('.')[1]

le =(6-len(fract))

zero = ''

for i in np.zeros(le):

zero += '0'

views = (views + zero).replace('.', '')

except:

views = str(views) + ('000000')

elif views.endswith('K'):

views = views.split('K')[0]

try:

views = views.replace('.', '')

except:

pass

else:

views = int(views)/1000

engagement+=int(views)

all\_posts.append(post\_link)

except:

all\_posts = []

average\_engagement = str(np.round(engagement/len(posts\_main),2))+'K'

return name, img\_link, about, likes, followers, following, all\_posts, average\_engagement

**Step 6:**

In this last iterate through all the url from the url list, store the scraped values onwards in a dictionary. Last but not the least, converted the dictionary into pandas dataframe and proceeded into a csv file.

profile\_list = []

for url in url\_list:

name, img\_link, about, likes, followers, following, all\_posts, average\_engagement = finding\_content(url)

profile = {'Username': name,

'Image':img\_link,

'About':about,

'Likes': likes,

'Followers' : followers,

'Following' : following,

'Posts' : all\_posts,

'Average engagement' : average\_engagement

}

print(profile)

profile\_list.append(profile)

print('Saving profile: ',name)

df = pd.DataFrame(profile\_list)

df.to\_csv('TIKTOK\_DATA.csv', index = False)

Dataset Link: <https://github.com/Sudhakordas/ActiveFence-assignmetn/blob/master/TIKTOK_DATA.csv>

**Task #6: Collect violence content from Facebook**

Finding 10 abusive Facebook links.

Graphic violence:

1.<https://www.facebook.com/1035756516/videos/1232502846125/>

2.<https://www.facebook.com/albert.longoria.56/videos/102023489824840/>

3.<https://www.facebook.com/mehmet.sahin.5621/videos/103297606350828/>

4.https://www.facebook.com/reel/5376332329155797?s=yWDuG2&fs=e

Nudity

1.<https://fb.watch/gMfXNwVBeJ/>

2.<https://www.facebook.com/reel/1822534568081686?s=yWDuG2&fs=e>

3.<https://fb.watch/gMfXNwVBeJ/>

4.<https://fb.watch/gMg6hXiprE/>

Selling illegal goods (selling drugs, prostration and more)

1.<https://fb.watch/gMdoPfABG_/>

2.<https://fb.watch/gMeb-4KHdL/>

**Why do I think these content of this links are abuse.**

Categorizing these types of contents cannot be subjective because of biases. And That’s why Facebook set some rules as a guideline for the creator.

**Rules for graphic violence.**

In many instances, when people share this type of content, they are condemning it or raising awareness about it. When people share anything on Facebook, we expect that they will share it responsibly, including carefully choosing who will see that content.

To avoid content being taken down for being too violent or graphic:

Don't

* Share graphic images for sadistic pleasure or to glorify violence.
* Share extremely graphic videos, such as those depicting a beheading.

**First link** under this category showing animal cruelty beheading a Goat. **Second link** contains the brutal data of beading a man. Last but not the least **third link** is example of total animal cruelty torturing innocent animals. And Fourth link is showing some dangerous video and people got scared and confused.

**Rules for Nudity and sexual activity**

Post content displaying genitals or focusing in on fully exposed buttocks. We also restrict some images of female breasts if they include the nipple, but our intent is to allow images that are shared for medical or health purposes. We also allow photos of women actively engaged in breastfeeding or showing breasts with post-mastectomy scarring. We also allow photographs of paintings, sculptures and other art that depicts nude figures.

Post images of sexual activity. Restrictions on the display of sexual activity also apply to illustrated and digitally created content unless the content is posted for educational, humorous or satirical purposes. Explicit images of sexual intercourse are prohibited.

Describe sexual acts in vivid detail. Descriptions of such acts that go into vivid detail may be removed.

All the **four links** in this section contain such type of 18+ contents which are beyond the terms and the regulations of Facebook.

**Content about non-medical drugs (other than alcohol or tobacco) including posts that:**

* Coordinates or encourages others to sell non-medical drugs
* Depicts, admits to, attempts purchase, or promotes sales of non-medical drugs by the poster of the content or their associates
* Promotes, encourages, coordinates, or provides instructions for use of non-medical drugs
* Admits, either in writing or verbally, to personal use of non-medical drugs unless posted in a recovery context
* It prohibits content that depicts the sale or attempt to purchase marijuana and pharmaceutical drugs. This includes content that:
* Mentions or depicts marijuana or pharmaceutical drugs

**Makes an attempt to sell or trade, by which we mean any of the following:**

* Explicitly mentioning the product is for sale or trade or delivery
* Asking the audience to buy
* Listing the price
* Encouraging contact about the product either by explicitly asking to be contacted or including any type of contact information
* Attempting to solicit the product, defined as:
* Stating interest in buying the product, or
* Asking if anyone has the product for sale/trade
* This applies to both individual pieces of content and Pages and Groups primarily dedicated to the sale of marijuana or pharmaceutical drugs

And first link in this category is showing how do they sell drugs and use and encourage young people to sell drugs. In the second link a person selling drugs openly.

**Way of collecting the data**

All of us know that social media platform is full of data of every type. And data for the specific task can be obtained in many different ways. In this project collection of abusive data is being instructed. Hence to perform this task has to done some research on the regulations on the Facebook content. As Facebook is full of wide spread groups and profile targeting right groups and profile needs some analysis. After selecting those group, then come the selecting right data which comes under the rules and regulation of Facebook.

**YES**

Browse for the data within those groups and profile

**Check data**

Store in the database

**NO**

Finding right group and profile which contain the required data

Researching Facebook terms and regulations