Indian Institute of Information Technology Surat

Lab Report on
Machine Learning (CS 601) Practical

Submitted by

[RAHUL KUMAR SINGH] (UI21CS44)

Course Faculty

Dr. Pradeep Kumar Roy

Dr. Rajesh K. Ahir

Department of Computer Science and Engineering

Indian Institute of Information Technology Surat

Gujarat-394190, India

Jan-2024

Lab No: 1

Aim: Data Collection from E-Commerce, Twitter and Similar Platforms

Description: Write a Python script for:

(a) Collecting tweets that may incorporate owner, date of post, number of re-tweet, number of followers, no of followees , and other associated information from Twitter and store it into a .csv file. (The size of collected tweets >5000)

(b) To scrap users reviews from any E-commerce or similar portals (Ex- Amazon, Flipkart, Yelp) and store it into a csv file that may incorporate date of post, number of likes/dislikes, reviews, location, and other associated fields (The size of collected reviews>5000).

Source Code:

For Task (a):

```
# ------------------------------------------------------------------------------
from selenium import webdriver
from selenium.webdriver.common.by import By
from fake_useragent import UserAgent
from webdriver_manager.firefox import GeckoDriverManager
import time
import json
import os
from selenium.webdriver.common.keys import Keys



MY_USERNAME_VAR = os.getenv('USERNAME')
MY_PASS_VAR = os.getenv('PASS')

#
------------------------------------------------------------------------------
def wait_for_window(self, timeout = 2):
    time.sleep(round(timeout / 1000))
```

```
    wh_now = self.driver.window_handles
    wh_then = self.vars["window_handles"]
    if len(wh_now) > len(wh_then):
        return set(wh_now).difference(set(wh_then)).pop()



keywords = ["WWE","Rock","RomanReigns"]

ulrs = []
#
----------------------------------------------------------------------------------
-
options = webdriver.FirefoxOptions()
options.headless = False
ua = UserAgent()
userAgent = ua.random
options.add_argument(f'user-agent={userAgent}')
#
----------------------------------------------------------------------------------
-

driver =
webdriver.Firefox(executable_path=GeckoDriverManager().install(),options=options)
driver.get("https://twitter.com/i/flow/login")
driver.maximize_window()
time.sleep(10)
try:
    input_element = driver.find_element(By.CSS_SELECTOR,
'.r-30o5oe.r-1niwhzg.r-17gur6a.r-1yadl64.r-deolkf.r-homxoj.r-poiln3')
    input_element.click()
    time.sleep(5)
    input_element.send_keys(MY_USERNAME_VAR)
    time.sleep(5)
    next_btn =
driver.find_element(By.CSS_SELECTOR,".css-901oao.r-1awozwy.r-6koalj.r-18u37iz.r-16y2
uox.r-37j5jr.r-a023e6.r-b88u0q.r-1777fci.r-rjixqe.r-bcqeeo.r-q4m81j.r-qvutc0")
    next_btn.click()
    print("click send")
    time.sleep(5)
    login_btn =
driver.find_element(By.CSS_SELECTOR,".css-901oao.r-1awozwy.r-1cvl2hr.r-6koalj.r-18u3
```

```
7iz.r-16y2uox.r-37j5jr.r-a023e6.r-b88u0q.r-1777fci.r-rjixqe.r-bcqeeo.r-q4m81j.r-qvutc0")
    login_btn.click()
    print("click send")
    time.sleep(10)
    email_x = driver.find_element(By.CSS_SELECTOR,
'.r-30o5oe.r-1niwhzg.r-17gur6a.r-1yadl64.r-deolkf.r-homxoj.r-poiln3.r-7cikom.r-1ny4l3l.r-t60dpp.r-1dz5y72.r-fdjqy7.r-13qz1uu')
    email_x.click()
    email_x.send_keys(MY_USERNAME_VAR)
    send_x = driver.find_element(By.CSS_SELECTOR,
'.css-901oao.r-1awozwy.r-jwli3a.r-6koalj.r-18u37iz.r-16y2uox.r-37j5jr.r-a023e6.r-b88u0q.r-1777fci.r-rjixqe.r-bcqeeo.r-q4m81j.r-qvutc0')
    send_x.click()
    time.sleep(5)
    password_x = driver.find_element(By.CSS_SELECTOR,
'.r-30o5oe.r-1niwhzg.r-17gur6a.r-1yadl64.r-deolkf.r-homxoj.r-poiln3.r-7cikom.r-1ny4l3l.r-t60dpp.r-1dz5y72.r-fdjqy7.r-13qz1uu')
    password_x.click()
    password_x.send_keys(MY_PASS_VAR)
    time.sleep(5)
    login2_btn = driver.find_element(By.CSS_SELECTOR,
'.css-18t94o4.css-1dbjc4n.r-1m3jxhj.r-sdzlij.r-1phboty.r-rs99b7.r-19yznuf.r-64el8z.r-1ny4l3l.r-1dye5f7.r-o7ynqc.r-6416eg.r-lrvibr')
    login2_btn.click()
    time.sleep(10)
    session_cookies = driver.get_cookies()
    try:
        closr_btn = driver.find_element(By.CSS_SELECTOR,
'.r-30o5oe.r-1niwhzg.r-17gur6a.r-1yadl64.r-deolkf.r-homxoj.r-poiln3.r-7cikom.r-1ny4l3l.r-xyw6el.r-y0fyvk.r-1dz5y72.r-fdjqy7.r-13qz1uu')
        closr_btn.click()
    except:
        print("no close btn")
    with open('session_cookies.txt', 'w') as file:
        file.write(str(session_cookies))
    unique_numbers = set()
    keyword_numbers = {}
    for keyword in keywords:
        paths = []
        driver.find_element(By.TAG_NAME, 'body').send_keys(Keys.CONTROL + 't')
        driver.switch_to.window(driver.window_handles[-1])
```

```
driver.get("https://twitter.com/search?q={}&src=typed_query".format(keyword))
        time.sleep(10)
        # elements = driver.find_elements_by_css_selector('[data-testid="tweet"]')
        tags =
driver.find_elements(By.CSS_SELECTOR,"a.css-4rbku5.css-18t94o4.css-1dbjc4n.r-1loqt21
.r-1777fci.r-bt1l66.r-1ny4l3l.r-bztko3.r-lrvibr")

        for tag in tags:
            href = tag.get_attribute("href")
            start_index = href.find("/status/") + len("/status/")
            number = href[start_index:].split('/')[0]
            unique_numbers.add(number)

        unique_numbers_list = list(unique_numbers)
        keyword_numbers[keyword] = unique_numbers_list
        for number in unique_numbers_list:
            print(number)

    with open('keyword_numbers.json', 'w') as file:
        json.dump(keyword_numbers, file)

except Exception as e:
    print(ulrs)
    print("An error occurred:", str(e))


# driver.quit()
```

For Task (b):

```
# Importing all the required libraries
import csv
from selenium import webdriver
from selenium.webdriver.common.by import By
import time

def extract_reviews(product_url, num_reviews_to_scrape=10):
    # Calling the driver
    driver = webdriver.Chrome()
```

```
    # Requesting the Amazon product's url
    driver.get(product_url)

    time.sleep(8)

    # Extracting our review data
    reviews = []
    review_elements = driver.find_elements(By.CSS_SELECTOR, '.a-section.review')
    temp_Date = ""
    for review_element in review_elements[:num_reviews_to_scrape]:
        time.sleep(1)
        review = {}
        review['author'] = review_element.find_element(By.CSS_SELECTOR,
'.a-profile-name').text.strip()
        temp_Date = review_element.find_element(By.CSS_SELECTOR,
'.review-date').text.strip()
        review['date'] = temp_Date[temp_Date.find('on')+3:]
        review['location'] = temp_Date[12:temp_Date.find('on')-1]
        review['text'] = review_element.find_element(By.CSS_SELECTOR,
'.review-text-content').text.strip()
        review['rating'] =
review_element.find_element_by_xpath('//i[@data-hook="review-star-rating"]').text.st
rip()
        review['title'] = review_element.find_element(By.CSS_SELECTOR,
'.review-title').text.strip()
        reviews.append(review)
        print(review)

    # Terminating the WebDriver
    driver.quit()
    # Returning the reviews
    return reviews
# Product url
product_url =
'https://www.amazon.in/ZAPCASE-Compatible-Xiaomi-Covers-Carbon/product-reviews/B07GQ
Y2RN2/ref=cm_cr_arp_d_paging_btm_next_2?ie=UTF8&reviewerType=all_reviews'

# Calling the extract_reviews() function
reviews_data = []
for i in range(1,4):
    reviews_data += extract_reviews(product_url+'&pageNumber='+str(i),
num_reviews_to_scrape=10)
```

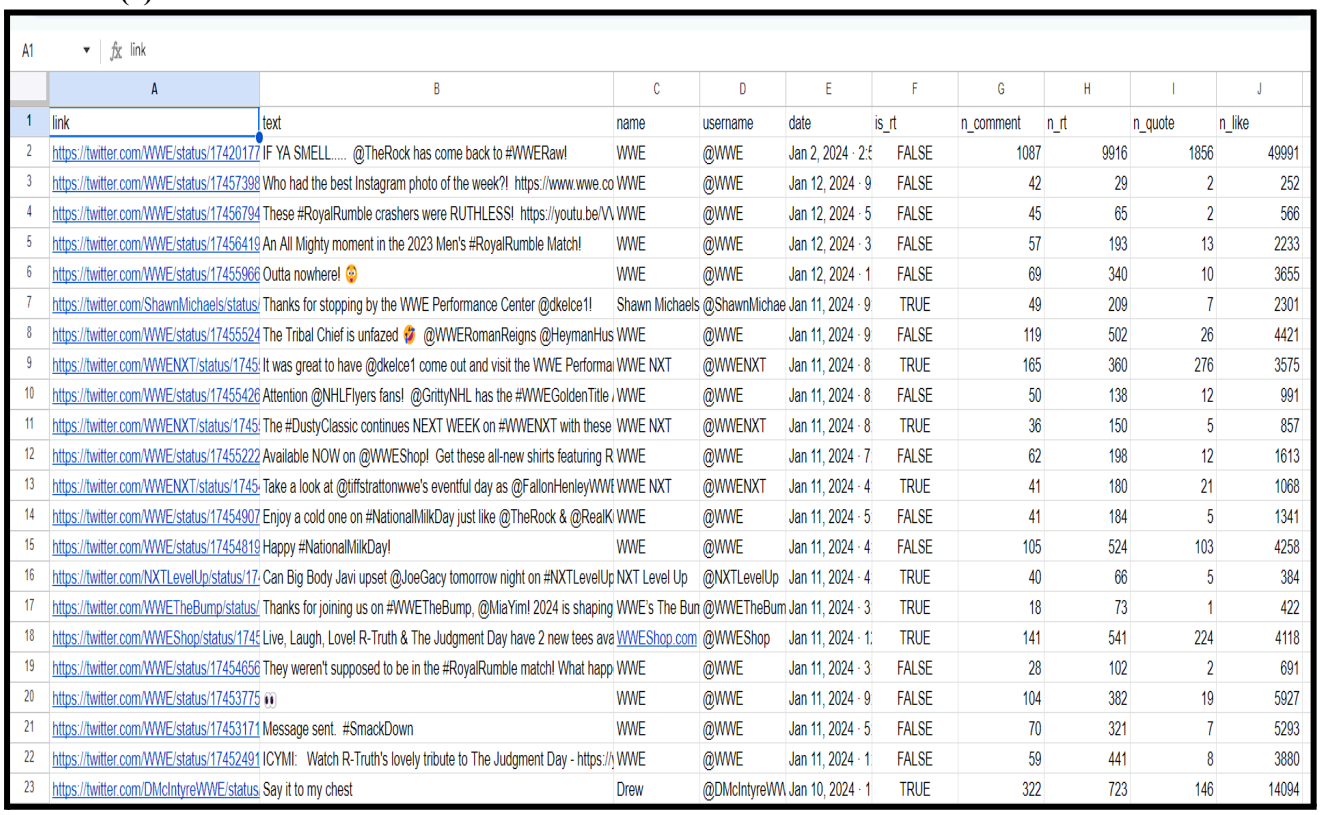
```
# Creating a function to export the data to csv
def export_csv(reviews, csv_filename='reviews_data.csv'):
    with open(csv_filename, 'w', newline='', encoding='utf-8') as csv_file:
        fieldnames = ['date','names','location','reviewtitles','ratings','reviews']
        writer = csv.DictWriter(csv_file, fieldnames=fieldnames)

        writer.writeheader()
        for review in reviews:
            writer.writerow({'date': review['date'], 'names': review['author'],
'location': review['location'], 'reviewtitles': review['title'], 'ratings':
review['rating'], 'reviews': review['text']})

# Export data to a csv file
export_csv(reviews_data)
```

Output:

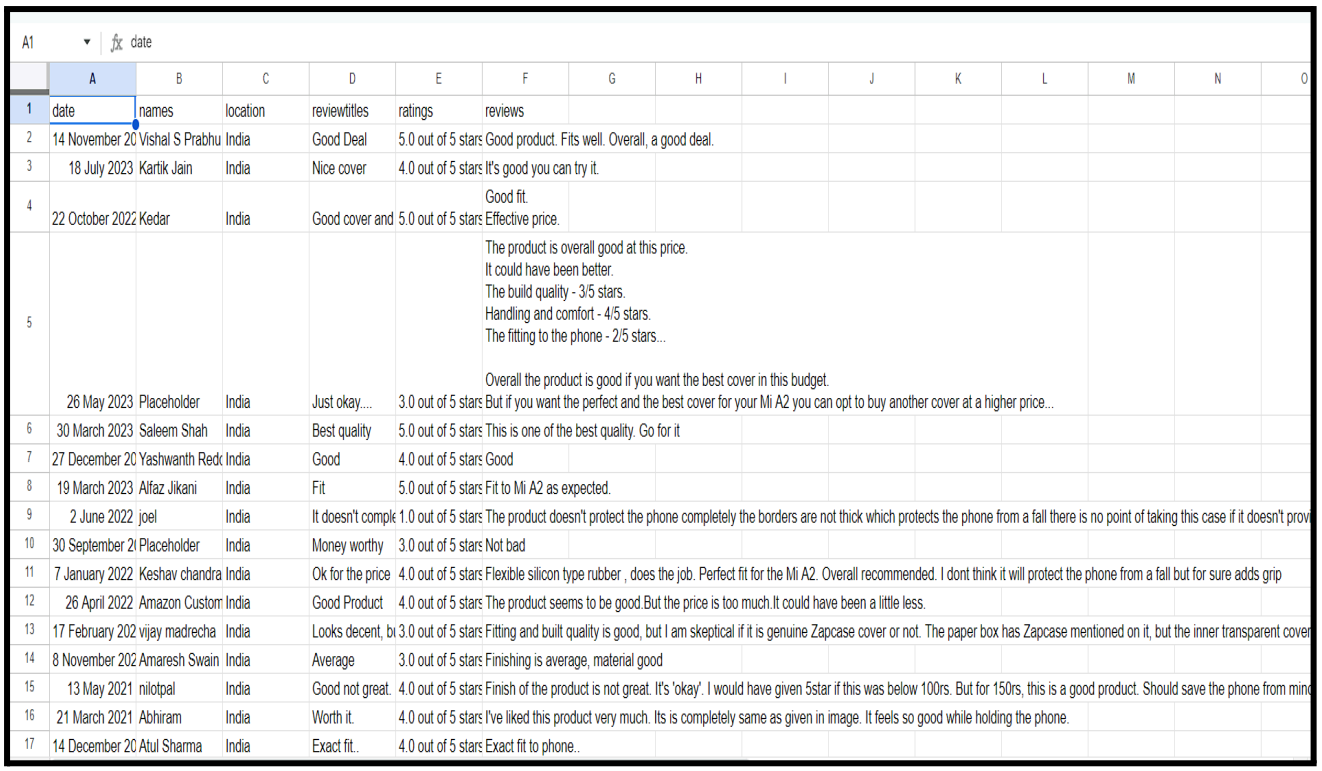
For Task (a):



A1 link

	A	B	C	D	E	F	G	H	I	J
1	link	text	name	username	date	is_rt	n_comment	n_rt	n_quote	n_like
2	https://twitter.com/WWE/status/17420177	IF YA SMELL..... @TheRock has come back to #WWERaw!	WWE	@WWE	Jan 2, 2024 · 2:5	FALSE	1087	9916	1856	49991
3	https://twitter.com/WWE/status/17457398	Who had the best Instagram photo of the week?! https://www.wwe.co	WWE	@WWE	Jan 12, 2024 · 9	FALSE	42	29	2	252
4	https://twitter.com/WWE/status/17456794	These #RoyalRumble crashers were RUTHLESS! https://youtu.be/VV	WWE	@WWE	Jan 12, 2024 · 5	FALSE	45	65	2	566
5	https://twitter.com/WWE/status/17456419	An All Mighty moment in the 2023 Men's #RoyalRumble Match!	WWE	@WWE	Jan 12, 2024 · 3	FALSE	57	193	13	2233
6	https://twitter.com/WWE/status/17455966	Outta nowhere! 😳	WWE	@WWE	Jan 12, 2024 · 1	FALSE	69	340	10	3655
7	https://twitter.com/ShawnMichaels/status/	Thanks for stopping by the WWE Performance Center @dkelce1!	Shawn Michaels	@ShawnMichae	Jan 11, 2024 · 9	TRUE	49	209	7	2301
8	https://twitter.com/WWE/status/17455524	The Tribal Chief is unfazed 🤷 @WWERomanReigns @HeymanHus	WWE	@WWE	Jan 11, 2024 · 9	FALSE	119	502	26	4421
9	https://twitter.com/WWENXT/status/1745	It was great to have @dkelce1 come out and visit the WWE Performa	WWE NXT	@WWENXT	Jan 11, 2024 · 8	TRUE	165	360	276	3575
10	https://twitter.com/WWE/status/17455426	Attention @NHLFlyers fans! @GrittyNHL has the #WWEGoldenTitle	WWE	@WWE	Jan 11, 2024 · 8	FALSE	50	138	12	991
11	https://twitter.com/WWENXT/status/1745	The #DustyClassic continues NEXT WEEK on #WWENXT with these	WWE NXT	@WWENXT	Jan 11, 2024 · 8	TRUE	36	150	5	857
12	https://twitter.com/WWE/status/17455222	Available NOW on @WWEShop! Get these all-new shirts featuring R	WWE	@WWE	Jan 11, 2024 · 7	FALSE	62	198	12	1613
13	https://twitter.com/WWENXT/status/1745	Take a look at @tiffstrattonwwe's eventful day as @FallonHenleyWWE	WWE NXT	@WWENXT	Jan 11, 2024 · 4	TRUE	41	180	21	1068
14	https://twitter.com/WWE/status/17454907	Enjoy a cold one on #NationalMilkDay just like @TheRock & @RealK	WWE	@WWE	Jan 11, 2024 · 5	FALSE	41	184	5	1341
15	https://twitter.com/WWE/status/17454819	Happy #NationalMilkDay!	WWE	@WWE	Jan 11, 2024 · 4	FALSE	105	524	103	4258
16	https://twitter.com/NXTLevelUp/status/17	Can Big Body Javi upset @JoeGacy tomorrow night on #NXTLevelUp	NXT Level Up	@NXTLevelUp	Jan 11, 2024 · 4	TRUE	40	66	5	384
17	https://twitter.com/WWETheBump/status/	Thanks for joining us on #WWETheBump, @MiaYim! 2024 is shaping	WWE's The Bun	@WWETheBum	Jan 11, 2024 · 3	TRUE	18	73	1	422
18	https://twitter.com/WWEShop/status/1745	Live, Laugh, Love! R-Truth & The Judgment Day have 2 new tees ava	WWEShop.com	@WWEShop	Jan 11, 2024 · 1	TRUE	141	541	224	4118
19	https://twitter.com/WWE/status/17454656	They weren't supposed to be in the #RoyalRumble match! What happ	WWE	@WWE	Jan 11, 2024 · 3	FALSE	28	102	2	691
20	https://twitter.com/WWE/status/17453775	👀	WWE	@WWE	Jan 11, 2024 · 9	FALSE	104	382	19	5927
21	https://twitter.com/WWE/status/17453171	Message sent. #SmackDown	WWE	@WWE	Jan 11, 2024 · 5	FALSE	70	321	7	5293
22	https://twitter.com/WWE/status/17452491	ICYMI: Watch R-Truth's lovely tribute to The Judgment Day - https://	WWE	@WWE	Jan 11, 2024 · 1	FALSE	59	441	8	3880
23	https://twitter.com/DMcIntyreWWE/status	Say it to my chest	Drew	@DMcIntyreWW	Jan 10, 2024 · 1	TRUE	322	723	146	14094

For Task (b):



	A	B	C	D	E	F
1	date	names	location	reviewtitles	ratings	reviews
2	14 November 20	Vishal S Prabhu	India	Good Deal	5.0 out of 5 stars	Good product. Fits well. Overall, a good deal.
3	18 July 2023	Kartik Jain	India	Nice cover	4.0 out of 5 stars	It's good you can try it.
4	22 October 2022	Kedar	India	Good cover and	5.0 out of 5 stars	Good fit. Effective price.
5	26 May 2023	Placeholder	India	Just okay....	3.0 out of 5 stars	The product is overall good at this price. It could have been better. The build quality - 3/5 stars. Handling and comfort - 4/5 stars. The fitting to the phone - 2/5 stars... Overall the product is good if you want the best cover in this budget. But if you want the perfect and the best cover for your Mi A2 you can opt to buy another cover at a higher price...
6	30 March 2023	Saleem Shah	India	Best quality	5.0 out of 5 stars	This is one of the best quality. Go for it
7	27 December 20	Yashwanth Redd	India	Good	4.0 out of 5 stars	Good
8	19 March 2023	Alfaz Jikani	India	Fit	5.0 out of 5 stars	Fit to Mi A2 as expected.
9	2 June 2022	joel	India	It doesn't compl	1.0 out of 5 stars	The product doesn't protect the phone completely the borders are not thick which protects the phone from a fall there is no point of taking this case if it doesn't prov
10	30 September 2(	Placeholder	India	Money worthy	3.0 out of 5 stars	Not bad
11	7 January 2022	Keshav chandra	India	Ok for the price	4.0 out of 5 stars	Flexible silicon type rubber , does the job. Perfect fit for the Mi A2. Overall recommended. I dont think it will protect the phone from a fall but for sure adds grip
12	26 April 2022	Amazon Custom	India	Good Product	4.0 out of 5 stars	The product seems to be good.But the price is too much.It could have been a little less.
13	17 February 202	vijay madrecha	India	Looks decent, b	3.0 out of 5 stars	Fitting and built quality is good, but I am skeptical if it is genuine Zapcase cover or not. The paper box has Zapcase mentioned on it, but the inner transparent cover
14	8 November 202	Amaresh Swain	India	Average	3.0 out of 5 stars	Finishing is average, material good
15	13 May 2021	nilotpal	India	Good not great.	4.0 out of 5 stars	Finish of the product is not great. It's 'okay'. I would have given 5star if this was below 100rs. But for 150rs, this is a good product. Should save the phone from mino
16	21 March 2021	Abhiram	India	Worth it.	4.0 out of 5 stars	I've liked this product very much. Its is completely same as given in image. It feels so good while holding the phone.
17	14 December 20	Atul Sharma	India	Exact fit..	4.0 out of 5 stars	Exact fit to phone..

Conclusion:

- Efficient and direct access to Twitter's data through the API.
- Provides real-time data retrieval, enabling instant updates.
- Offers structured data in JSON format for easy processing.
- Overcomes API limitations for certain tasks, such as scraping dynamic content using custom scraping.
- Allows interaction with web elements, useful for user-specific searches.