Лабораторна 3

«WEB-аналітика»

ФБ-31мп Синчук Владислав

Тема: 3. Збір великих даних із соціальних мереж. Автентифікація користувача. Збір інформації із джерел Deep Web.

1. Використовуючи API Twitter, написати скрипт, що буде автентифікуватися в соціальній мережі та забирати N>1000 останніх записів певного акаунту (акаунт має бути публічним), та створити датасет. Опціонально – сформувати систему сповіщення (пошта, телефон, будь-що) про знаходження певного ключового слова.

Замість твіттера буде реддіт.

Створимо візьмемо необхідні токени для того щоб взаємодіяти з Reddit API.   


Функція автентифікації в Reddit API:

*def* reddit\_auth(*client\_id*: str, *secret*: str, *user\_agent*: str) -> praw.Reddit:

    try:

        reddit\_client = praw.Reddit(*client\_id*=*client\_id*, *client\_secret*=*secret*, *user\_agent*=*user\_agent*)

        print("Auth success")

    except Exception as e:

        print("Auth failure", e)

        exit()

    return reddit\_client

Функція збирання інформації з деякого сабреддіту (не приватний акаунт як в завданні):

1) Логгер

*def* csv\_logger(*func*):

*def* wrapper(\**args*, \*\**kwargs*) -> pd.DataFrame:

        t1 = time.time()

        data = *func*(\**args*, \*\**kwargs*)

        t2 = time.time()

        try:

            df = pd.DataFrame(data)

            df.to\_csv(*f*'{*func*.\_\_name\_\_}\_out.csv')

            print(*f*'{*func*.\_\_name\_\_} run time: {t2 - t1}')

            return copy.deepcopy(df)

        except TypeError as e:

            print(e)

            return None

    return wrapper

2) Функція збирання постів на сабреддіті

@csv\_logger

*def* reddit\_scan\_subreddit(*reddit\_client*: praw.Reddit, *subreddit\_name*: str, *n\_posts*: int) -> None:

    subreddit: praw.SubredditHelper = *reddit\_client*.subreddit(*subreddit\_name*)

    print("Subreddit Display Name:", subreddit.display\_name)

    # Uncomment the line below if you want to print the subreddit description

    print("Subreddit Description:", subreddit.description)

    col\_submission\_title = []

    col\_submission\_selftext = []

    col\_submission\_created\_utc = []

    col\_submission\_over\_18 = []

    col\_submission\_num\_comments = []

    for submission in subreddit.new(*limit*=*n\_posts*):  # You can set limit=None for all posts, but be cautious with large subreddits

        col\_submission\_title.append(submission.title)

        col\_submission\_selftext.append(submission.selftext)

        col\_submission\_created\_utc.append(datetime.fromtimestamp(submission.created\_utc).strftime('%Y-%m-%dT%H:%M:%SZ'))

        col\_submission\_over\_18.append(submission.over\_18)

        col\_submission\_num\_comments.append(submission.num\_comments)

    data\_dict = {

        'title': col\_submission\_title,

        'text': col\_submission\_selftext,

        'datetime': col\_submission\_created\_utc,

        'nsfw': col\_submission\_over\_18,

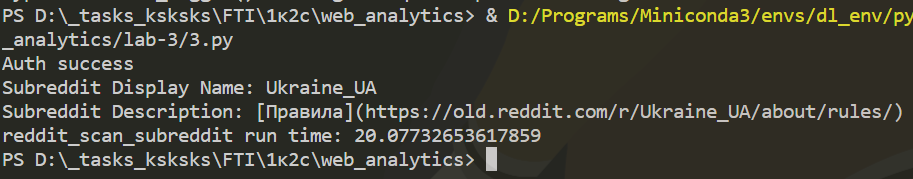
        'num\_comments': col\_submission\_num\_comments

    }

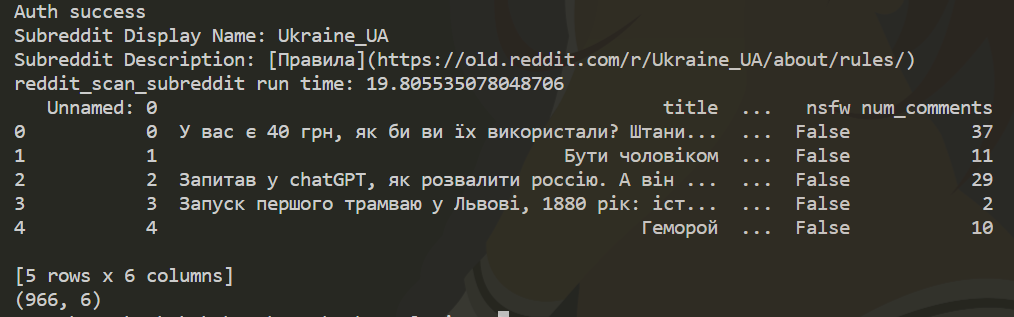
    # print(data\_dict)

    return data\_dict

Отримані результати:



Читаємо отриманий файл:



Хоча було поставлено n=1001