DOI: https://doi.org/10.36910/6775-2524-0560-2024-54-15
УДК 004.657
Лавренчук Світлана Василівна, к.т.н., доцент
https://orcid.org/0000-0002-5453-3924
Христинець Наталія Анатоліївна, к.т.н., доцент
https://orcid.org/0000-0002-4836-7632
Савчук Ольга Володимирівна, здобувач
https://orcid.org/0009-0005-0629-9968
Луцький національний технічний університет, м. Луцьк, Україна

МОНІТОРИНГ ПЛАТФОРМИ YOUTUBE ЗАСОБАМИ SQL

Лавренчук С.В., Христинець Н.А., Савчук О.В. Моніторинг платформи YouTube засобами SQL. Розглядаються можливості застосування мови SQL для аналізу платформи YouTube. Детально розглядаються бібліотеки, методи та SQL-запити, створено програмний продукт, що дозволяє отримати статистику популярності відео, аналізувати зміни кількості підписників каналів та класифікувати трендові теми.
Ключові слова: платформа YouTube, аналітика, база даних, python, статистика.

Lavrenchuk S., Khrystynets, N., Savchuk O. Monitoring the YouTube Platform Using SQL. The possibilities of using the SQL language to analyze the YouTube platform are considered. Libraries, methods and SQL queries are discussed in detail, the created software product allows you to obtain statistics on video popularity, analyze changes in channel subscribers and classify trending topics.
Keywords: YouTube platform, analytics, database, python, statistics.

Постановка проблеми. У сучасному світі відеоконтент стає все більш популярним і важливим засобом спілкування. Відеохостинг YouTube вважається однією з провідних платформ у цьому сегменті, кількість користувачів і відеоконтенту зростає з кожним днем. Згідно зі статистикою Similarweb [1], сайт YouTube.com посідає друге місце за кількістю переглядів в Україні, поступаючись лише Google. Враховуючи таке зростання, дані на цій платформі необхідно ефективно відстежувати та аналізувати, щоб визначати тенденції, розуміти вподобання споживачів і вдосконалювати стратегію контенту.

Основні завдання дослідження містять створення оптимізованих запитів мовою SQL для отримання статистики про популярність відео, для аналізу змін кількості підписників каналів, визначення та класифікації трендових тем. Отримані результати будуть корисні для авторів відео контенту, маркетологів та дослідників, які бажають краще розуміти динаміку та потенціал платформи YouTube для своїх потреб. Використання SQL дасть змогу ефективно обробляти та аналізувати великі обсяги даних, забезпечуючи точні та надійні результати.

Аналіз останніх досліджень і публікацій. Багато дослідників, як академічних, так і комерційних, досліджували аналітику YouTube з різних точок зору [2-3]. Основні напрямки досліджень включають аналіз контенту, вивчення поведінки користувачів, алгоритмів рекомендацій, визначення тенденцій у відеоконтенті та аналіз впливу платформи на суспільство та культуру, а також на підприємництво в соціальних мережах [4].

В праці [5] розглядаються різні методи аналітики веб-сайтів (такі як поведінкова аналітика, конверсійна аналітика, тощо), які можна застосовувати також і до платформи YouTube.

Індійські науковці Ashwini T, Sahana LM, Mahalakshmi E, Shweta S Padti [6] розглядають можливості фреймворка Hadoop для розподіленої обробки великих даних у вигляді мультимедійного формату, зокрема на базі технології Hadoop HIVE, яка підтримує SQL-запити. Тут використано мову Java та СУБД MySQL.

Багато дослідників використовують можливості мови SQL для обробки й аналізу великої кількості даних, зібраних платформою YouTube. Наприклад, Брусенцов Юрій [7], студент Національного технічного університету «Київський політехнічний інститут імені Ігоря Сікорського» у дипломній роботі розробив програмне забезпечення для аналізу коментарів на YouTube із використанням сучасних інструментів розробки. Для створення серверної частини додатку було обрано фреймворк Fastify.js із платформою виконання Node.js, а для клієнтської частини ─ фреймворк Vue.js. Обрано систему керування базами даних MongoDB. В результаті було створено програму, яка дозволяє користувачу переглядати усі коментарі із YouTube-відео, виконувати пошук, фільтрацію, надає проаналізовані дані про текст коментарю, а саме тональність

тексту та мову, якою він написаний. Статистична інформація про аналізовані коментарі надається сервісом у вигляді графіків та діаграм [7].

Румунські дослідники здійснили багатовимірний аналіз даних [8] за допомогою Tableau Public обравши по топ-100 каналів із 17 різних категорій, вони в своєму рейтингу враховували кількість завантажених відео, кількість підписників і кількість переглядів. Вони також зробили семантичний аналіз тексту в описі каналів.

На платформі kaggle щорічно оновлюється набір даних [9], що містить статистику каналів YouTube з найбільшою кількістю підписок в форматі .csv, який і візьмемо за основу нашого дослідження.

Мета дослідження полягає у моніторингу платформи YouTube на основі статистичних даних за 2023 рік, а також для зручності роботи створенні програми у вигляді веб-сервісу з адаптивним графічним інтерфейсом, що дасть змогу здійснювати аналіз даних, таких як кількість передплатників, переглядів відео, тип каналу, популярність серед переглядачів на платформі YouTube з метою виявлення ключових тенденцій.

Основна частина дослідження.

Створено базу даних в СУБД MySQL, яка містить статистику каналів YouTube з найбільшою кількістю підписок (995 записів в таблиці), а також статистику власних каналів (рисунок 1).

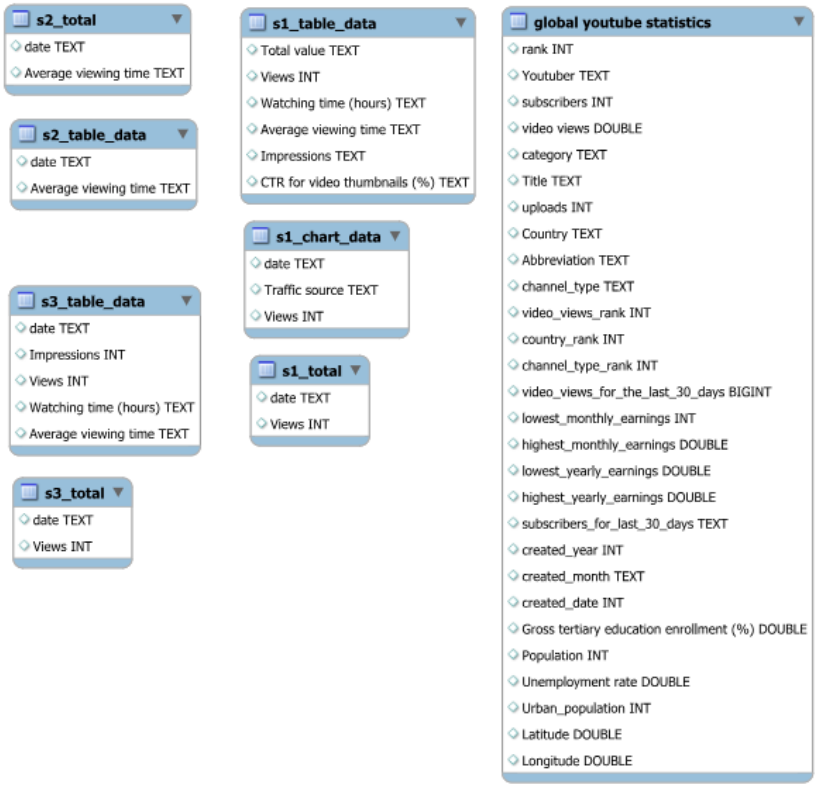

Рис. 1— Схема бази даних

Для того, щоб пересвідчитися в достовірності, порівняємо дані, отримані з [9] та дані з сайту vidIQ [10] (рисунок 2). Для цього виберемо 10 найпопулярніших каналі за кількістю переглядів:

```
USE youtubestat;
SELECT Youtuber,subscribers, `video views` FROM `global youtube statistics`
ORDER BY subscribers DESC
LIMIT 11
```

Youtuber	subscribers	video views
T-Series	245000000	228000000000
YouTube Movies	170000000	0
MrBeast	166000000	28368841870
Cocomelon - Nursery Rhymes	162000000	164000000000
SET India	159000000	148000000000
ннн Kids Diana Show	112000000	93247040539
PewDiePie	111000000	29058044447
Like Nastya	106000000	90479060027
Vlad and Niki	98900000	77180169894
Zee Music Company	96700000	57856289381
WWE	96000000	77428473662

Рис. 2— Порівняння даних з різних джерел

Як бачимо з рисунку 2, рейтинг каналів співпадає, за винятком того, що в базу даних з kaggle потрапив зайвий рядок (YouTube Movies), тому в SQL-запиті ми виводимо 11 позицій, а не 10.

Цікаво дізнатися чи потрапили в рейтинг канали з України:

```
SELECT Youtuber,subscribers,category FROM youtubestat.`global youtube statistics`
WHERE Country='Ukraine';
```

В результаті виконання запиту в базі даних знайшлося 8 каналів, проте три з них мають не коректну назву, тому ми їх до уваги не братимемо, розглянемо детальніше інші 5 каналів (таблиця 1).

Таблиця 1 – Українські канали з найвищим рейтингом

Youtuber 2023	**Subscribers 2023**	**Subscribers 2024**	**category 2023**	**Country 2023**	**Country 2024**	**Language 2024**	**Creation data 2024**
SlivkiShow	20400000	*20700000*	Entertainment	Ukraine	*Ukraine*	Rus	11.12.2012
News 24	17700000	*6560000*	News & Politics	Ukraine	*Ukraine*	Ukr	05.02.2006
SIS vs BRO	14200000	*14100000*	Entertainment	Ukraine	*Canada*	Eng	03.03.2016
Lady Diana	13500000	*13800000*	Entertainment	Ukraine	*Ukraine*	Rus	21.03.2016
VexTrex	13300000	*13300000*	Entertainment	Ukraine	*USA*	Eng	08.09.2014

В таблиці 1 дані, які містять в шапці «2023», взято з сайту сайту [9], а дані, що в шапці містять надпис «2024» - взято з платформи YouTube. Як бачимо з таблиці 1, дані не зовсім точні, зокрема канал «SIS vs BRO» позначено як український, а насправді він є канадським, так само як канал «VexTrex» є американським, а не українським. Тому можна зробити висновок, що рейтингу потрапили лише три канали, які вказали своєю країною Україну, крім того, лише один з них («News 24») містить україномовний контент.

Також з таблиці 1 можна помітити, що 4 з 5 каналів відносяться до категорії «розваги». Тому цікаво дослідити рейтинг категорій, щоб визначити найбільш популярні, для цього напишемо запит:

```
SELECT category, count(Youtuber) FROM youtubestat.`global youtube statistics`
GROUP BY category
ORDER BY count(Youtuber) DESC;
```

Як і передбачалося, найбільші рейтинги отримала категорія «Розваги», на другому місці категорія «Музика» (рисунок 3).

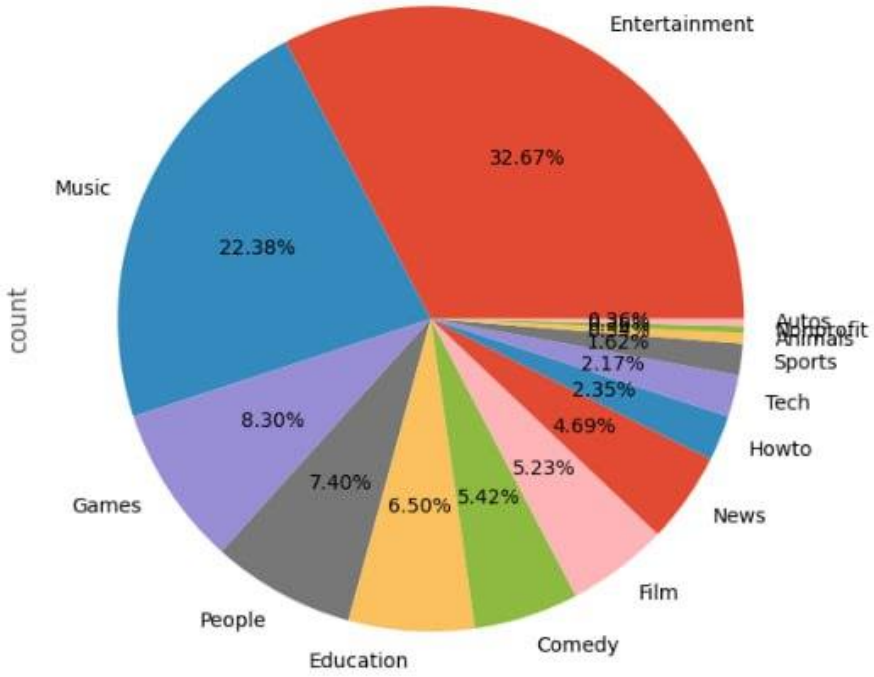

Рис. 3 – Рейтинг популярності YouTube-каналів за категоріями

Для зручності відображення аналітичних досліджень було створене програмне забезпечення у вигляді веб-сервісу для обробки даних та їх виведення у вигляді графіків. Дані зберігаються в базі даних MySQL або в форматі .csv, їх обробка здійснюється засобами мови Python з використанням бібліотек pandas, matplotlib, seaborn та Django для візуалізації та представлення результатів., веб-інтерфейс розроблений за допомогою HTML та CSS. Архітектура системи також враховує вимоги безпеки та захисту даних. Усі дані, що передаються та обробляються системою, захищені новітніми протоколами та технологіями шифрування. Безпека даних забезпечується на рівні сервера, бази даних і мережевого підключення.

Для обробки та візуалізації даних було використано різні модулі та бібліотеки, що зображено на рисунку 4.

```
from django.shortcuts import render
import io
import matplotlib.pyplot as plt
import pandas as pd
import numpy as np
import seaborn as sns
import urllib, base64
```

Рис. 4— Використані бібліотеки та модулі

from django.shortcuts import render — імпортування функції render з модуля django.shortcuts, яка використовується для рендерингу HTML-шаблонів.

Matplotlib, Pandas, NumPy, Seaborn — бібліотеки для роботи з даними та візуалізації графіків.

import urllib, base64 — бібліотеки для кодування зображення у формат base64, щоб його можна було вбудувати безпосередньо в HTML-сторінку.

os — модуль для роботи з функціями операційної системи, такими як отримання інформації

про файли та каталоги.

io — модуль для роботи з рядками та байтами у вхідних-вихідних операціях. Використовується для створення буферу для збереження графіків у вигляді байтів.

urllib — модуль для роботи з URL-адресами, використовується для кодування стрічок та формування URL для відображення графіків у веб-інтерфейсі.

base64 — модуль для роботи з кодуванням та декодуванням даних у форматі base64. Використовується для перетворення графіків у формат, який може бути відображений у HTML сторінці.

django.shortcuts — містить функції, які полегшують роботу з веб-інтерфейсом, зокрема функцію render для відображення HTML-шаблонів.

Результати статистичних дослідження платформи YouTube відображено у вигляді різних графіків, реалізація яких на веб-сторінках забезпечується фреймворком Django, приклад зображено на рисунку 5.

```
def most_popular_25(request):
    yt=load_table()

    plt.figure(figsize=(10, 10))
    yt_bp = sns.barplot(y='Youtuber',x='video views',data=yt,
                order=yt.sort_values('video views',ascending=False).Youtuber.iloc[:25],palette='Spectral') #Top 25 channels
    plt.title('25 найпопулярніших каналів') #Title for graph
    yt_bp.set_xticklabels(['0b', '50b', '100b', '150b', '200b']) #Sets x labels


    plt.plot(range=10)
    fig=plt.gcf()
    buf=io.BytesIO()
    fig.savefig(buf,format='png')
    buf.seek(0)
    string = base64.b64encode(buf.read())
    uri = urllib.parse.quote(string)
    return render(request,'most_popular_25.html', {'img': uri})
```

Рис. 5 — Код для створення графіку (стовпчаста діаграма)

Цей приклад коду визначає функцію most_popular_25 (request), яка аналізує дані та створює графік у вигляді стовпчастої діаграми для 25 найпопулярніших каналів на основі кількості переглядів відео. Далі графік зберігається у форматі PNG, конвертується в base64 та відображається на веб-сторінці «most_popular_25.html», що зображено на рисунку 6. Слід звернути увагу, що дані з рисунку 6 корелюються з даними рисунку 2.

Створений програмний продукт дозволяє робити досліджувати популярність каналів за різними критеріями, наприклад за типом або за останніми переглядами, тощо.

На початку свого розвитку YouTube намагався рекомендувати відео, які здобули найбільше кліків, в 2012 році в алгоритмі рекомендацій почали враховувати не лише кліки, але й час перегляду відео, на сьогоднішній день YouTube пропонує відео, які кожен окремий глядач, швидше за все, перегляне, а не відео, які переглядають інші люди. Персоналізовані пропозиції оцінюються відповідно до ефективності та якості відео, а також інтересу та поведінки глядачів.

Наша база даних містить окрім глобальної статистики ще й статистику власного каналу. Спробуємо дослідити джерела трафіку на своєму каналі, для цього напишемо запит:

```
SELECT `Traffic source`, sum(Views), sum(Views)/28283*100 AS '%'
FROM youtubestat.s1_chart_data
GROUP BY `Traffic source`
ORDER BY sum(Views) DESC;
```

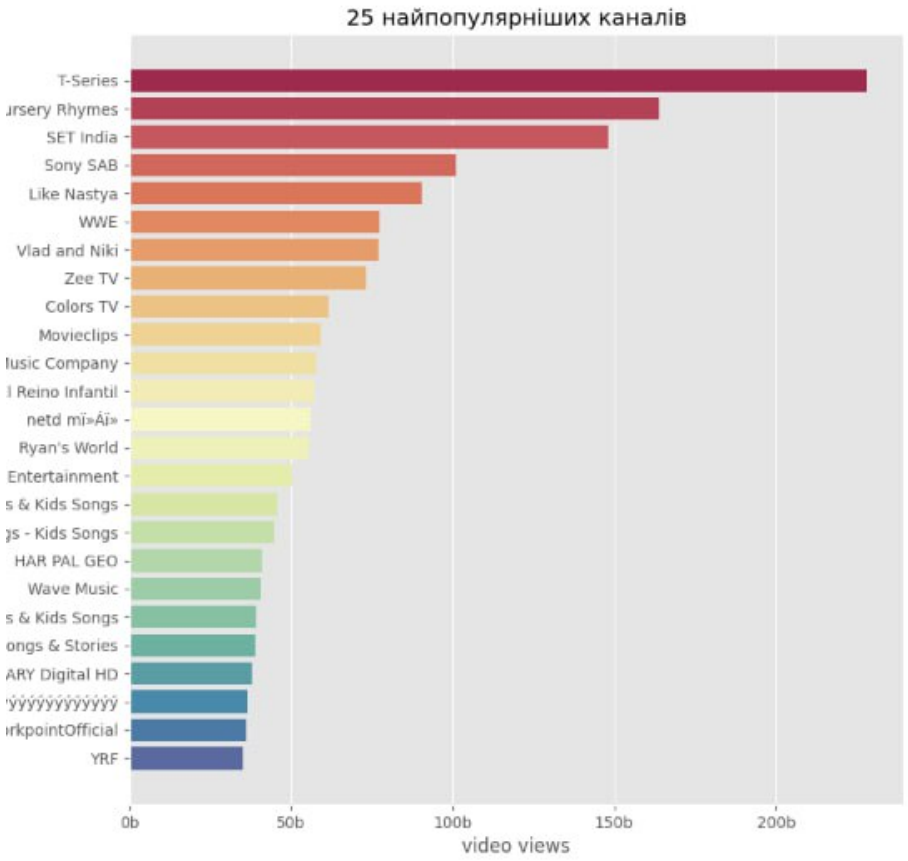

Рис. 6 — Діаграми для 25 найпопулярніших каналів

Результат виконання цього запиту наведено на рисунку 7.

Traffic source	sum(Views)	%
Рекомендовані відео	17095	60.4427
Функції вибору контенту	9307	32.9067
Сторінки каналів	575	2.0330
Адресний рядок, закладки, невідомі джерела	573	2.0260
Пошук на YouTube	566	2.0012
Інші сторінки YouTube	167	0.5905

Рисунок 7 — Джерела трафіку

Як бачимо з рисунку 7, більше 60 % трафіку займають саме рекомендовані відео, тому аналітичні дослідження платформи допоможуть краще підбирати контент та ефективніше потрапляти в списки рекомендованих відео.

Висновки. Створено базу даних в СУБД MySQL на основі csv-файлів з різних джерел (з платформи kaggle та з аналітики власного каналу). Розроблено програми у вигляді веб-сайту, що дає змогу здійснювати комплексний аналіз характеристик каналів YouTube з використанням мови SQL та інструментів візуалізації даних Python. Результати дослідження можуть бути корисними для маркетологів, дослідників та інших зацікавлених осіб, щоб ліпше розуміти платформу та розробляти стратегії взаємодії.

Список бібліографічного опису

1. Top Websites Ranking. Most Visited Websites in Ukraine. URL: https://www.similarweb.com/top-websites/ukraine/ (дата звернення: 19.02.2024)
2. Аналітика YouTube: 14 показників для відстеження ефективності вашого відеоконтенту. URL: https://wave.video/ua/blog/youtube-analytics-metrics/ (дата звернення: 20.02.2024)
3. Manko, B. A. (2023). Video advertising: Using YouTube analytics for the target audience. Journal of Information

Technology Teaching Cases, 13(1), 77-81.

4. Shetty, A., Abreo, B., D'Souza, A., Kondana, A., & Karimbi, K. M. (2021, May). Video Description Based Youtube Comment Classification. In Applications of Artificial Intelligence in Engineering: Proceedings of First Global Conference on Artificial Intelligence and Applications (GCAIA 2020) (pp. 667-678). Singapore: Springer Singapore.

5. Христинець, Н., Мельник, К., Фенюк, А., & Копчук, В. (2023). Аналітика веб-ресурсів як способи рейтингування інформаційних систем. КОМП'ЮТЕРНО-ІНТЕГРОВАНІ ТЕХНОЛОГІЇ: ОСВІТА, НАУКА, ВИРОБНИЦТВО, (53), 228-232. https://doi.org/10.36910/6775-2524-0560-2023-53-34

6. Ashwini, T., Sahana, L. M., Mahalakshmi, E., & Shweta, S. P. YOUTUBE DATA ANALYSIS USING HADOOP FRAMEWORK. DOI: 10.33564/IJEAST.2021.v05i11.051

7. Брусенцов Ю. О. Веб-сервіс для аналізу коментарів на YouTube. Кваліфікаційна робота «Бакалавр». Київ, 2020. 26 с.

8. Lupșa-Tătaru, D. A., & Lixăndroiu, R. (2022). YouTube channels, subscribers, uploads and views: a multidimensional analysis of the first 1700 channels from july 2022. Sustainability, 14(20), 13112.

9. Global YouTube Statistics 2023. Kaggle: Your Machine Learning and Data Science Community. URL: https://www.kaggle.com/datasets/nelgiriyewithana/global-youtube-statistics-2023/data (дата звернення: 09.03.2024).

10. Top 50 YouTube Channels. Boost Your Views And Subscribers On YouTube - vidIQ. URL: https://vidiq.com/youtube-stats/top/50/ (date of access: 10.03.2024).

References

1. Top Websites Ranking. Most Visited Websites in Ukraine. URL: https://www.similarweb.com/top-websites/ukraine/ (дата звернення: 19.02.2024)

2. Аналітика YouTube: 14 показників для відстеження ефективності вашого відеоконтенту. URL: https://wave.video/ua/blog/youtube-analytics-metrics/ (дата звернення: 20.02.2024)

3. Manko, B. A. (2023). Video advertising: Using YouTube analytics for the target audience. Journal of Information Technology Teaching Cases, 13(1), 77-81.

4. Shetty, A., Abreo, B., D'Souza, A., Kondana, A., & Karimbi, K. M. (2021, May). Video Description Based Youtube Comment Classification. In Applications of Artificial Intelligence in Engineering: Proceedings of First Global Conference on Artificial Intelligence and Applications (GCAIA 2020) (pp. 667-678). Singapore: Springer Singapore.

5. Khrystynets, N., Melnyk, K., Fenyuk , A., & Kopchuk, V. (2023). Analytics of web resources as ways of rating information systems. COMPUTER-INTEGRATED TECHNOLOGIES: EDUCATION, SCIENCE, PRODUCTION, (53), 228-232. https://doi.org/10.36910/6775-2524-0560-2023-53-34

6. Ashwini, T., Sahana, L. M., Mahalakshmi, E., & Shweta, S. P. YOUTUBE DATA ANALYSIS USING HADOOP FRAMEWORK. DOI: 10.33564/IJEAST.2021.v05i11.051

7. Brusentsov Yu. O. Web service for analyzing comments on YouTube. Qualifying work "Bachelor". Kyiv, 2020. 26 p.

8. .Lupșa-Tătaru, D. A., & Lixăndroiu, R. (2022). YouTube channels, subscribers, uploads and views: a multidimensional analysis of the first 1700 channels from july 2022. Sustainability, 14(20), 13112.

9. Global YouTube Statistics 2023. Kaggle: Your Machine Learning and Data Science Community. URL: https://www.kaggle.com/datasets/nelgiriyewithana/global-youtube-statistics-2023/data (date of access: 09.03.2024).

10. Top 50 YouTube Channels. Boost Your Views And Subscribers On YouTube - vidIQ. URL: https://vidiq.com/youtube-stats/top/50/ (date of access: 10.03.2024).