МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ НАЦІОНАЛЬНИЙ

УНІВЕРСИТЕТ “ЛЬВІВСЬКА ПОЛІТЕХНІКА”

ІНСТИТУТ КОМП'ЮТЕРНИХ НАУК ТА ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ

Кафедра ІСМ

Звіт

до лабораторної роботи №3

з дисципліни “Інновації в інформаційних системах та технологіях”

“Кластеризація”

Виконали ст. гр. ІТ-42

ТРОЦЬКО О.М.

ФЕДИК В. А.

Прийняв к.т.н., доцент

ПРОДАНЮК М.М

Львів - 2025

**Мета:** Написати на Python програму для виділення кластерів. Провести оцінку результатів.

**Вхідні дані:** набір новинних статей, що містять як правдиві, так і фейкові новин, таблиця у форматі Excel з джерелами новин, що використовувалися для dataset.

**Результати виконаної роботи**

У ході виконання лабораторної роботи ми завантажили та попередньо опрацювали датасет, що містить інформацію про текстові повідомлення з різних джерел (Telegram, Facebook, Twitter) із зазначенням їх типу (“фейк” або “правдиве повідомлення”), кількістю лайків, поширень, а також сайтом-джерелом.

За допомогою цього датасету було виконано кілька кластеризацій:

• кластеризація за ознакою **типу повідомлення** (“фейк”/“правдиве”);

• кластеризація за **джерелом розміщення** (Telegram, Facebook, Twitter);

• кластеризація за **кількістю лайків та поширень** (із поділом на кластери);

• кластеризація за **сайтами-джерелами** (на основі кількості повідомлень).

Код програми:

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

from scipy.cluster.hierarchy import dendrogram, linkage, fcluster, maxdists

from sklearn.preprocessing import StandardScaler, LabelEncoder

# Завантаження датасету

df = pd.read\_excel("dataset.xlsx")

# Вибираємо корисні стовпці для кластеризації

columns\_to\_use = ["мітка", "Like", "Поширення", "Джерело"]

df = df.dropna(subset=columns\_to\_use) # Видаляємо порожні значення

# Конвертуємо числові колонки, якщо є помилки, замінюємо їх на NaN

df["Like"] = pd.to\_numeric(df["Like"], errors="coerce")

df["Поширення"] = pd.to\_numeric(df["Поширення"], errors="coerce")

df["мітка"] = pd.to\_numeric(df["мітка"], errors="coerce")

# Видаляємо рядки, які містять NaN після конвертації

df = df.dropna(subset=["Like", "Поширення", "мітка"])

# Перетворення текстових даних у числові

df["Джерело"] = df["Джерело"].astype(str)

label\_encoder = LabelEncoder()

df["Джерело"] = label\_encoder.fit\_transform(df["Джерело"])

# Масштабування даних

scaler = StandardScaler()

data\_scaled = scaler.fit\_transform(df[columns\_to\_use])

# Вибираємо підмножину для коректної візуалізації

df\_sample = df.iloc[:100] # Беремо лише перші 100 записів

data\_scaled\_sample = data\_scaled[:100]

# Виконуємо ієрархічну кластеризацію

linkage\_matrix = linkage(data\_scaled\_sample, method="ward")

# Візуалізація дендрограми без зайвих підписів

plt.figure(figsize=(12, 6))

dendrogram(linkage\_matrix, no\_labels=True, leaf\_rotation=90, leaf\_font\_size=8)

plt.title("Дендрограма ієрархічної кластеризації")

plt.xlabel("Новини")

plt.ylabel("Відстань")

plt.show()

# Визначаємо оптимальну Мкількість кластерів

threshold\_value = maxdists(linkage\_matrix)[-1] # Використовуємо останній

df\_sample["Cluster"] = fcluster(linkage\_matrix, threshold\_value,

criterion="distance")

# Візуалізація розподілу кластерів

sns.scatterplot(x=df\_sample["Like"], y=df\_sample["Поширення"],

hue=df\_sample["Cluster"], palette="viridis", s=100,

edgecolor="k")

plt.title("Кластеризація новин за лайками та поширеннями")

plt.xlabel("Кількість лайків")

plt.ylabel("Кількість поширень")

plt.show()

Спочатку зчитуються новини з Excel файлу. Для кластеризації вибираються важливі стовпці, такі як кількість лайків, поширень, мітка новини та джерело. Некоректні або порожні значення видаляються, а текстові дані (наприклад, джерела новин) перетворюються в числові за допомогою методу кодування.

Текстові дані в стовпці "Джерело" кодуються в числа за допомогою LabelEncoder. Це дозволяє зберегти інформацію про джерела повідомлень (Telegram, Facebook, Twitter) у числовому вигляді, оскільки алгоритми кластеризації працюють тільки з числовими даними.

Далі всі дані, що використовуються для кластеризації, нормалізуються. Це робиться за допомогою StandardScaler, щоб всі ознаки (лайки, поширення, мітки) мали однаковий масштаб. Без цього кроку одна ознака (наприклад, кількість лайків) може домінувати над іншими (наприклад, джерелом повідомлення), що спотворює результати кластеризації.

Для того, щоб зробити аналіз більш зрозумілим та зручним для візуалізації, вибираються лише перші 100 записів даних. Це дозволяє зменшити обсяг даних, з якими працюватимемо при побудові графіків, а також спростити інтерпретацію результатів.

Далі виконується кластеризація за допомогою ієрархічного методу "ward". Цей метод дозволяє створити ієрархічну структуру, де кожне повідомлення можна розглядати як частину певного кластера, а потім ці кластери з'єднуються між собою.

За допомогою дендрограми візуалізується, як ці кластери формуються на кожному кроці. Дендрограма дозволяє побачити, на якому етапі класифікації кластери об'єднуються, і допомагає визначити, скільки кластерів оптимально вибрати для подальшого аналізу.

Кількість кластерів визначається на основі максимальних відстаней між групами в дендрограмі. Це дозволяє автоматично обрати найбільш відповідну кількість кластерів, яка найкраще відображає структуру даних.

На завершення будується графік, де кожен кластер відображається різними кольорами. Графік показує, як повідомлення групуються за кількістю лайків і поширень. Це дозволяє наочно побачити, як різні повідомлення класифікуються в різні групи на основі їхніх характеристик.

В результаті отримуємо такі графіки. Результати зображені на рис.1-2.



*Рис. 1. Розподіл кластерів у площині “лайки — поширення”*



*Рис. 2. Дендрограма ієрархічної кластеризації новин*

Висновок: У результаті виконання лабораторної роботи було успішно реалізовано алгоритм кластеризації даних методом k-середніх. Проведено багатовимірну кластеризацію текстових повідомлень за кількома ознаками, зокрема за типом новини, джерелом, кількістю лайків та сайтом-джерелом. Отримані результати підтвердили ефективність використання кластерного аналізу для групування схожих об'єктів без попереднього маркування. Крім того, нами було отримано практичні навички побудови кластерної моделі, вибору кількості кластерів та візуалізації результатів кластеризації, що є важливим етапом в аналізі даних та створенні систем підтримки прийняття рішень.