ДЕРЖАВНИЙ УНІВЕРСИТЕТ ТЕЛЕКОМУНІКАЦІЙ

НАВЧАЛЬНО–НАУКОВИЙ ІНСТИТУТ ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ

Кафедра інженерії програмного забезпечення

Пояснювальна записка

до магістерської роботи
на ступінь вищої освіти магістр

на тему: «**ПІДВИЩЕННЯ ЕФЕКТИВНОСТІ СИСТЕМ ПОШУКУ СХОЖИХ АКАУНТІВ В СОЦІАЛЬНИХ МЕРЕЖАХ НА ОСНОВІ КЛАСТЕРНОГО АНАЛІЗУ**»

Виконав: студент 6 курсу, групи ПДМ-61
спеціальності:
121 Інженерія програмного забезпечення
(шифр і назва спеціальності)
Дібрій Д.А.
(прізвище та ініціали)
Керівник Аверічєв І. М.
(прізвище та ініціали)
Рецензент_____
(прізвище та ініціали)
Нормоконтроль Трінтіна Н.А.
(прізвище та ініціали)

Київ –2022

ДЕРЖАВНИЙ УНІВЕРСИТЕТ ТЕЛЕКОМУНІКАЦІЙ

НАВЧАЛЬНО-НАУКОВИЙ ІНСТИТУТ ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ

Кафедра Інженерії програмного забезпечення
Ступінь вищої освіти -«Магістр»
Спеціальність підготовки – 121 «Інженерія програмного забезпечення»

ЗАТВЕРДЖУЮ
Завідувач кафедри
Інженерії програмного забезпечення
_____ Негоденко О.В.
"____" _____2022 року

ЗАВДАННЯ
НА МАГІСТЕРСЬКУ РОБОТУ СТУДЕНТУ

Дібрій Данило Андрійович
(прізвище, ім'я, по батькові)

1. Тема роботи: «Підвищення ефективності систем пошуку схожих акаунтів в соціальних мережах на основі кластерного аналізу»

Керівник роботи: Аверічев І.М., доцент кафедри ІПЗ
(прізвище, ім'я, по батькові, науковий ступінь, вчене звання)

Затверджені наказом вищого навчального закладу від «12» жовтня 2022 року №112.
2. Строк подання студентом роботи «31» червня 2022 року
3. Вхідні дані до роботи

Науково-технічна література з кластерного аналізу, машинного навчання;
Науково-технічна література з питань, пов'язаних з використанням методів кластерного аналізу в соціальних мережах;

4. Зміст розрахунково-пояснювальної записки (перелік питань, які потрібно розробити).

4.1 Огляд існуючих соціальних мереж.
4.2 Аналіз методів кластерного аналізу та методів нормалізації даних користувачів.
4.3 Розробка методу пошуку схожих акаунтів в соціальних мережах.
4.4 Проектування програмного забезпечення з методу пошуку схожих акаунтів в соціальних мережах.

5. Перелік демонстраційного матеріалу (назва основних слайдів)
   1. Мета, об’єкт та предмет дослідження
   2. Аналіз існуючих IT-рішень та їх моделей
   3. Методи оцінки схожості акаунтів
   4. Метод пошуку схожих акаунтів в соціальних мережах
   5. Порівняльний аналіз k-means та k-means mini batch

6. Дата видачі завдання «14» жовтня 2022

КАЛЕНДАРНИЙ ПЛАН

№ з/п	Назва етапів бакалаврської роботи	Строк виконання етапів роботи	Примітка
1	Отримання завдання на магістерську роботу	12.10.2022	Виконано
2	Підбір науково-технічної літератури	16.10.2022	Виконано
3	Аналіз алгоритмів та методів кластерного аналізу	20.10.2022	Виконано
4	Розробка методу	8.11.2022	Виконано
5	Проектування програмного забезпечення	20.11.2022	Виконано
6	Аналіз результатів	3.12.2022	Виконано
7	Розробка обов‘язкових демонстраційних матеріалів	10.12.2022	Виконано
8	Попередній захист роботи		
9	Здача роботи		

Студент _____ _____
(підпис) (прізвище та ініціали)

Керівник роботи _____ _____
(підпис) (прізвище та ініціали)

РЕФЕРАТ

Текстова частина магістерської роботи 57 с., 27 рис., 1 табл., 25 джерел.

КЛАСТЕРНИЙ АНАЛІЗ, СОЦІАЛЬНІ МЕРЕЖІ, НОРМАЛІЗАЦІЯ ДАНИХ, МАШИННЕ НАВЧАННЯ, КЛАСИФІКАЦІЯ, ОБРОБКА ДАНИХ.

Об'єкт дослідження – процес проектування методу для пошуку схожих акантів в соціальних мережах на основі кластерного аналізу.

Предмет дослідження – методи та алгоритми пошуку схожих акаунтів в соціальних мережах.

Мета роботи покращення пошуку груп схожих користувачів соціальних мереж за допомогою розробленого методу на основі кластерного аналізу.

Методи дослідження – у роботі були використані теоретичні методи дослідження для аналізу інформації отриманої з науково-технічної літератури, емпіричні методи для нагляду, порівняння та постановки експериментів.

Разом з активним зростанням кількості користувачів мережі інтернет збільшується і кількість даних необхідних для обробки. Використання звичайних статистичних методів для аналізу настільки великих масивів даних стає неефективним, що вимагає пошуку нових підходів до обробки та аналізу даних.

Одним з способів є аналіз за допомогою кластерного аналізу. Кластерний аналіз – це метод обробки великої кількості даних для кластеризації, тобто групування певного набору об'єктів за спільними рисами, так що утворюються окремі кластери, тобто групи схожих об'єктів які не пов'язані з об'єктами з інших кластерів.

При виконанні роботи було створено метод пошуку схожих акаунтів в соціальних мережах за допомогою алгоритму кластеризації mini batch k-means та спроектовано програмне забезпечення яке отримує, нормалізує та поділяє дані користувачів на окремі групи.

Отже, розроблено та описано метод пошуку схожих акаунтів в соціальних мережах на основі кластерного аналізу.

Даний метод та програмне забезпечення може бути використано на підприємствах яким необхідно обробляти та групувати велику кількість користувацьких даних які надходять з соціальних мереж.

Галузь використання – системи обробки великих даних отриманих з соціальних мереж.

ЗМІСТ

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ

API (Application Programming Interface) – програмний інтерфейс програми;

БД – база даних;

REST (Representational State Transfer) – передача репрезентативного стану;

HTTP (HyperText Transfer Protocol) – протокол передачі гіпертексту.

ВСТУП

Актуальність дослідження. В останні десятиліття суттєво зростає кількість користувачів мережі інтернет. Разом з сим зростає кількість людей які використовують соціальні мережі для зв'язку та отримання інформації. Активний розвиток соціальних мереж також суттєво вплинув на ринок реклами, адже завдяки інформації про користувачів яку збирають ці мережі рекламодавці мають змогу таргетувати рекламу з величезною точністю, що значно впливу на зріст продажів товарів та послуг. Адже за допомогою визначення схожих акаунтів можна легко визначити що товар який сподобався одному користувачу швидше за все привабить і іншого.

Але існує значна проблема – кількість інформації про користувачів настільки велика, що стає неможливим обробляти її звичайними методами. Через це виникла необхідність у аналізі великих даних.

Великі Дані набули широкого поширення в багатьох сферах нашого життя, хоча ми цього і не помічаємо. Їх використовують не тільки для соціальних мереж, а й медицині, телекомунікаціях та фінансових компаніях, а також в державному управлінні. За допомогою технологій Big Data підприємства отримують змогу обробляти великі масиви даних і виявляти корисні закономірності, що дають їм конкурентні переваги.

Для пошуку таких закономірностей широко використовується кластерний аналіз, адже він надає змогу ділити користувачів на групи по спільним параметрам, що дозволяє оптимізувати обробку даних для подальшого аналізу та використання. Сьогодні найбільших компаній світу: Amazon, Google, Meta активно вкладають ресурси у дослідження способів аналізу користувачів за допомогою кластеризації.

Виходячи з вищенаведеного, дослідження особливостей розробки та реалізації методу аналізу великих даних для пошуку схожих акаунтів в соціальних мережах на основі кластерного аналізу є актуальним.

Мета дослідження – покращення пошуку груп схожих користувачів соціальних мереж за допомогою розробленого методу на основі кластерного аналізу.

Завдання дослідження:

- розглянути теоретичні засади функціонування методів аналізу даних для пошуку користувачів соціальних мереж;
- дослідити методи та засоби розробки алгоритму кластерного аналізу для пошуку схожих акантів в соціальних мережах;
- розробити модель програмного забезпечення яка дозволить використовувати кластерний аналіз для обробки даних користувачів.

Об'єкт дослідження – процес проектування методу для пошуку схожих акантів в соціальних мережах на основі кластерного аналізу.

Предмет дослідження – методи та алгоритми пошуку схожих акаунтів в соціальних мережах.

Методи дослідження: у роботі були використані теоретичні методи дослідження для аналізу інформації отриманої з науково-технічної літератури, емпіричні методи для нагляду, порівняння та постановки експериментів.

Наукова новизна одержаних результатів. Результати дослідження підтверджують ефективність методів кластеризації для для пошуку схожих акаунтів в соціальних мережах.

Практична значущість результатів полягає в тому, що дослідження ґрунтується на результатах поглибленого вивчення особливостей розробки та реалізації методу аналізу великих даних для пошуку схожих акаунтів в соціальних мережах на основі кластерного аналізу.

Структура роботи. Дипломна робота складається зі вступу, трьох розділів, висновків, списку джерел посилання, додатків. Основний зміст роботи викладено на 57 сторінках. Список використаних джерел складається з 25 найменувань

1. ТЕОРЕТИЧНА ЧАСТИНА

1.1 Огляд сучасних соціальних мереж

На сьогоднішній день власниками найбільшими соціальними мережа є Facebook, Twitter та LinkedIn. Виникає необхідність розглянути їх більш детально, щоб визначити їх спільні риси та які самі дані користувачів вони збирають.

Twitter – це соціальна мережа мікроблогів, тобто коротких текстових наборів даних (280 знаків) які публікують користувачі, також відомих як «твіти». Twitter був створений у 2007 році Бізом Стоном та Еваном Вільямсом, і на сьогоднішній день має 330 мільйонів активних користувачів щомісяця. Соціальна мережа має велике значення для сучасного світу і стала одним з важливих джерел інформації, її використовують у тому числі державні посадові України для донесення інформації до населення та навколишнього світу. Для таких акаунтів використовується спеціальна відзнака – синя галочка, що підтверджує акаунт як офіційний.

Зареєстровані користувачі цієї соціальної мережі мають можливість створювати, реагувати та ділитися з іншими користувачами цими твітами, тим самим поширюючи інформацію про теми якими вони цікавляться.

Також соціальна мережа надає можливість публікувати короткі відео та фото, але більшою частиною користувачі надають перевагу саме невеликим текстам.

Крім текстового змісту у твіті присутні два додаткові блоки метаданих, що мають особливе значення: сутності та розташування.

Сутності – це згадки користувачів, хештеги, посилання URL та медіафайли, пов'язані з твітом, а розташування – це місця у реальному світі, які можна приєднати до твіту.

Публікуються твіти у стрічках повідомлень – це хронологічно впорядковані колекції твітів. Стрічка повідомлень — це будь-яка конкретна колекція твітів, що відображаються в хронологічному порядку; Однак частіше відображається пара

стрічок повідомлень. З точки зору довільного користувача Twitter, домашня стрічка це те, що він бачить відразу після входу в акаунт. У ній відображаються всі твіти користувачів, за якими він слідує. Крім домашньої стрічки є також стрічка користувача, що містить твіти, написані певним користувачем.

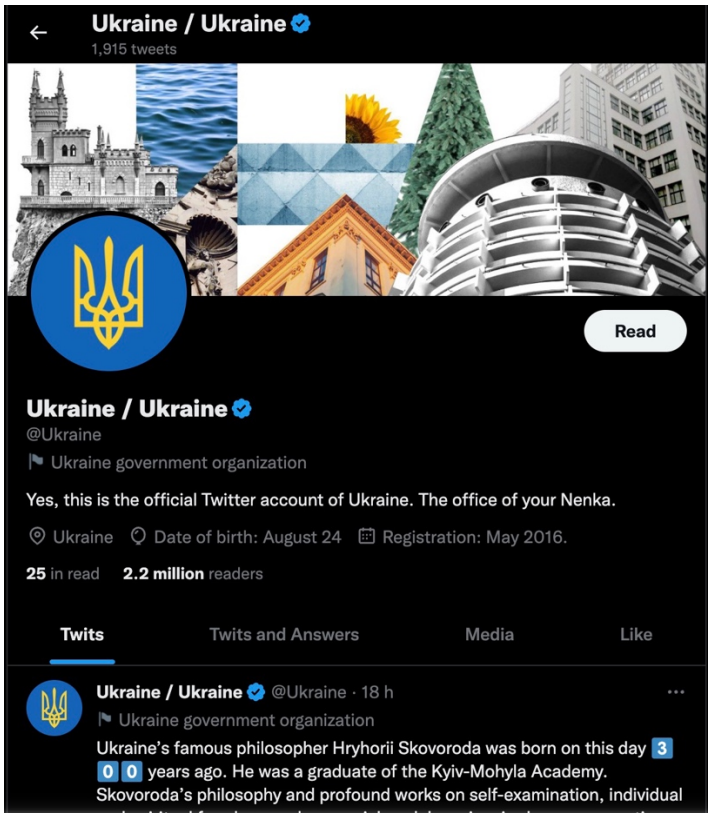


Рисунок 1.1 – Приклад акаунту Twitter

Facebook – це одна з найбільших та найстаріших соціальних мереж соціальна мережа яка нараховує два мільярди користувачів.

Facebook був заснований у 2004 році Марком Цукербергом, Едуардо Саверіном, Дастіном Московіцем і Крісом Хьюзом, які були студентами Гарвардського університету.

Facebook надає можливість розміщувати фотографії, відео та статуси. Також користувачі мають можливість створювати групи по інтересам та додавати інших користувачів до «друзів», що дозволяє вилучати велику кількість даних для подальшого аналізу. На відміну від Twitter, де використовується асиметрична модель дружби, яка відкрита і заснована на дотриманні інших користувачів без

будь-якої згоди, у Facebook модель дружби симетрична і вимагає взаємної згоди користувачів, щоб вони могли бачити взаємодії та дії один одного. Завдяки такій великій кількості даних ця соціальна мережа має змогу заробляти гроші за допомогою продажу користувацьких даних рекламним компаніям які своєю чергою аналізують дані та розділюють користувачів на рекламні групи.

Рисунок 1.2 – Сторінка користувача Facebook

LinkedIn – це соціальна мережа, орієнтованої на професійні та ділові відносини та був заснований Рейдом Гоффманом, Еріком Лі та Жан-Люком Вайлантом у 2002 році зі штаб-квартирою в Маунтін-В'ю, Каліфорнія. На сьогоднішній день ця соціальна мережа нараховує триста мільйонів користувачів по всьому світу.

На перший погляд, LinkedIn може здатися схожою на будь-яку іншу соціальну мережу, але дані, які завантажують користувачі та отримують аналітики мають зовсім іншу природу.

Якщо Twitter призначений для відправки коротких текстових повідомлень і його можна порівняти з багатолюдним форумом, Facebook — для відображення власного життя для друзів та родичів та ведення з ними листування, майже як у кімнаті наповненою друзями та родичами, які ведуть розмови, то LinkedIn можна порівняти з інтерактивним резюме та портфоліо яке надає можливість відобразити та підтвердити свої професійні навички та контакти, знайти нове місце роботи.

Ця соціальна мережа надає користувачам можливість повідомити про власний стаж, місце роботи та завершені проекти на персональній сторінці, що дозволяє працівникам з відділу ресурсів відстежувати та надавати вакансії для відповідних користувачів.

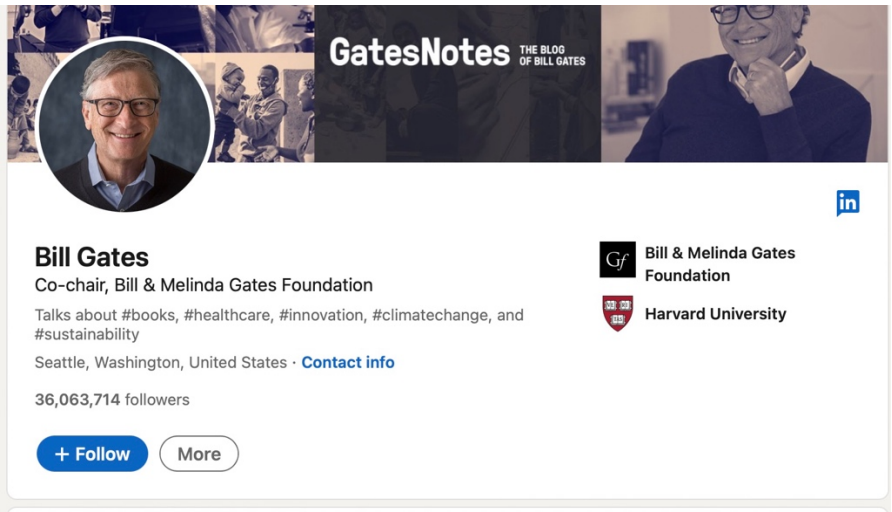


Рисунок 1.3 – Сторінка користувача LinkedIn

Виходячи з вищеописаного, можна визначити наступні спільні риси соціальних мереж:

- можливість відправки текстових повідомлень;
- можливість об’єднання в спільні соціальні групи по інтересам за допомогою певної відстежування;
- можливість за допомогою тексту надати персональну інформацію про себе.

Тобто основним способом надання і відповідно отримання інформації про користувача є саме тексти.

1.2 Причини використання кластеризації для аналізу даних соціальних мереж

З кожним роком до соціальних мереж приєднуються більше і більше користувачів, що вимагає значно збільшувати витрати на підтримку власних сервісів [7].

Компанія Cisco приводить наступні дані росту обсягу інтернет-трафіку з 1992 по 2017 роки: 100 ГБ – 1992, 45000 ГБ – 2017. На думку аналітиків які вивчають мережу інтернет кількість даних продовжить активно збільшуватись і надалі.

І хоча кожен рік створюється нове обладнання для передачі більшої кількості інформації, операторам телекомунікацій стає все дорожче переходити на них. Так для 44% компаній США перехід на технологію 5-G, яка надає змогу краще передавати дані через мережу, є проблемою через значні витрати на її імплементацію [8].

Це вимагає шукати нові способи монетизації за використання цих соціальних мереж, і оскільки основним ресурсом соціальних мереж є саме дані користувачів, найприбутковішим типом монетизації є продаж таргетованої реклами для користувачів. Але оскільки IT-компанії зазвичай не займаються рекламою напряму, дані користувачів групуються певним чином та продаються маркетинговим компаніям для подальшого аналізу, а маркетингові компанії своєю чергою приваблюють нових рекламодавців до соціальної мережі. Відповідно сильніший буде ефект від реклами, тим більше грошей зможуть заробляти компанії які володіють цими соціальними мережами.

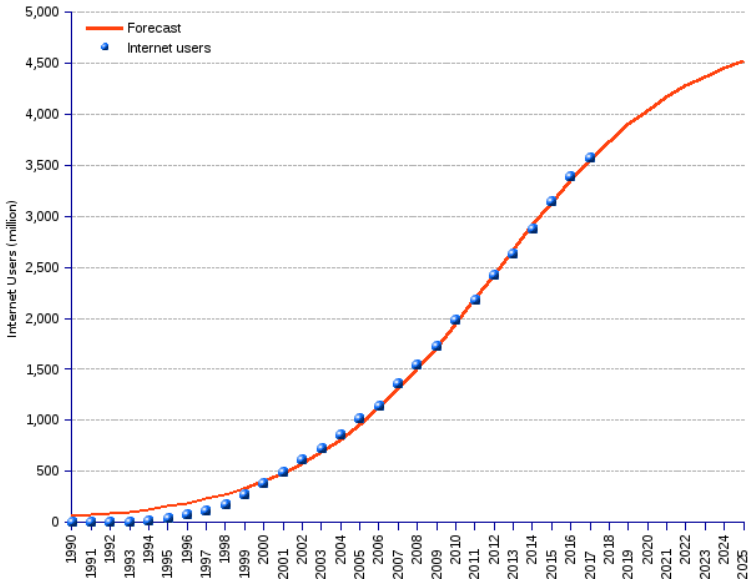


Рисунок 1.4 – Тренд на зріст користувачів мережі інтернет

Маркетинг – це певний процес направлений на визначення потенційних покупців певної послуги, їх кількості, аналізу їх інтересів для визначення оптимальної кількості товарів або послуг яку необхідно створити підприємцю для максимізації прибутку та мінімізації витрат, відповідно маркетингові компанії це окремі підприємства які надають послуги іншим підприємствам для аналізу стану ринку відповідно до приводу певного товару або послуги.

Основними продуктами діяльності маркетингових кампаній є розробка методів для продажу та просування продукції, ефективний розвиток бізнесу, вироблення маркетингових стратегій, складання бізнес-планів. На сьогоднішній день маркетингові компанії надають перевагу для просування продукції та аналізу ринків збуту саме соціальним мережам, оскільки вони містять велику кількість інформації про користувачів та надають змогу рекламувати первні продукти відповідно до вподобань користувача. Такий метод просування продукту відомий як «таргетована реклама».

Таргетована реклама – це реклама яка використовує заздалегідь відомі факти про користувача та на основі цих фактів рекомендує певний продукт. Інформацію про користувача збирають та розповсюджують сайти які користувач відвідував. Це може бути як просто і дані про те який тип сайту користувач відвідав, так і більш

детальні, наприклад оцінки фільмів які ставив користувач, товари які він придбав. Найчастіше для збирання подібної інформації використовують cookie.

Cookie – це маленький фрагмент даних про активність коритсувача на сайті який зберігає веб-браузер користувача для подальшого використання. Спочатку cookie використовували лише для пришвидшення роботи сайтів, так частина даних яку користувач бачив на сайті зберігалась локально на стороні користувача у cookie, а потім використовувалась для завантаження сторінки сайту. Через те що повторне завантаження не вимагало використання інтернету, швидкість повторного відкриття сайтів значно зростала. Наразі cookie також широко використовують для таргетованої реклами.

До переваг маркетингу саме в соціальних мережах можна віднести наступні [9]:

- тісна комунікація в соціальних мережах вимушує покупців та компанії слухати один одного, оскільки система рейтингів надає змогу переглянути якість послуг та надавати персональну інформацію про використання продуктами;
- величезна кількість трафіку та більш молода та імпульсивна аудиторія соціальних мереж є прибутковим ринком збуту;
- можливість отримати інформацію про користувачів відповідно до груп інтересів;
- рекламні компанії в соціальних мережах значно дешевші рекламних компаній у реальному світі, оскільки не вимагають розміщення реклами на білбордах, реклами по телевізору та ін. ;
- придбані персональні дані користувачів можна використовувати для подальших маркетингових компаній у найближчі роки, оскільки звички людей повільно змінюються.

Так згідно з дослідженням [10] було виявлено, що продаж товарів у соціальних мережах з можливістю залишити відгук та надати оцінку товару щодо корисності та достовірності збільшують степінь довіри до продукту або послуги, відповідно підвищуючи намір споживачів купувати продукти та кількість продажів товару.

Кластерний аналіз також використовується в інших сферах:

- Маркетинг та дослідження ринків збуту – за допомогою кластеризації маркетологи поділяють споживачів на однорідні групи. За допомогою цього дослідник ринку може визначити спільні інтереси групи користувачів які використовують продукти компанії, для визначення нових покупців. Або ринковий аналітик може бути зацікавлений у групуванні фінансових характеристик компаній, щоб мати можливість пов'язати їх із показниками їх фондового ринку.

  Першого разу кластерний аналіз у маркетингу був використаний для дослідження міст, які можна було використовувати як тестові ринки. Він використовувався для розділення міст на невелику окремі групи на основі 14 змінних, включаючи розмір міста, тираж газет і дохід на душу населення. Оскільки можна очікувати, що міста в групі будуть дуже схожими за певними ознаками, для них можуть бути використані однакові підходи у просуванні продукції.
- Астрономія – величезна кількість зірок вимагає визначення окремих груп зірок які відрізняються за певними ознаками, а кластерний аналіз можна використовувати саме для класифікації космічних об'єктів, так у дослідженні М. Фаундез-Абанса [11] за допомогою кластерного аналізу було проаналізовано дані про хімічний склад планетарних туманностей та за допомогою кластерного аналізу виявлено шість окремих груп об'єктів на які можна розділити зірки.
- Психіатрія – хвороби пов'язані з метальним здоров'ям та їх особливості досить складно чітко визначити через аналіз стану лише одного пацієнту, тому кластерний аналіз часто стає в пригоді для визначення або вдосконалення відомих діагнозів. У дослідженні [12] за допомогою кластерного аналізу даних двохсот пацієнтів з клінічною депресією були ідентифікований новий підтип депресії – ендогенна депресія.
- Біоінформатика – це наука на тісно пов'язана з біологією, інформатики та статистики, яка вивчає біологічні проблеми вирішення яких потребує обробки великої кількості даних, а сьогоденний активний розвиток геномних

і протогеномних баз даних, дозволив отримати достатню кількість людських генів для аналізу. Одним з напрямків досліджень є ДНК, але через тисячі генів які необхідно дослідити досить важко використовувати звичайні статистичні методи. Через це для аналізу генів ДНК використовують кластерний аналіз, який дозволяє визначити окремі групи генів які відповідають певним спадковим захворюванням та ін.

Виходячи з вищеописаного, можна зробити висновок про необхідність використання певних методів групування користувачів соціальних мереж за певними схожими ознаками, оскільки маркетинг в соціальних мережах є великим та активно зростаючим з кожним роком ринком для збуту товарів та послуг. Разом з активним зростанням кількості користувачів мережі інтернет збільшується і кількість даних необхідних для обробки. Використання звичайних статистичних методів для аналізу настільки великих масивів даних стає неефективним, через це виникає необхідність у розвитку методів для аналізу груп користувачів за певними спільними ознаками, що надасть можливість надавати користувачам саме ту рекламу яка їх зацікавить. У якості методів для аналізу великої кількості даних для створення гомогенних груп за певними ознаками використовуються кластерний аналіз.

1.3 Огляд наукових робіт та статей за напрямком дослідження

В ході дослідження наукових джерел було визначено, що алгоритми кластеризації широко використовуються для аналізу даних користувачів соціальних мереж. Найчастіше у якості соціальної мережі розглядають Twitter [2, 6], соціальну мережу у якій основну інформацію про користувача можна отримати з невеликих текстових повідомлень, також відомих як твіти. Кластеризацію даних користувачів відносять до окремого етапу аналізу соціальних мереж у якому використовуються методи математичної статистики та методи дата майнінгу соціальних мереж. Так в роботі [1] розглянуто типи соціальних мереж, підходи до вимірювання подібності наборів даних які отримують з соціальних мереж для

подальшого аналізу. Розглядається концепція транзитивності даних та її обчислення за допомогою коефіцієнта кластеризації. Автори встановили, що оскільки у соціальних мережах пов’язані між собою користувачі мають високу ймовірність бути «друзями» тобто мати велику кількість подібних даних, відповідно друзі цих друзів також можуть бути друзями, тобто виконується правило транзитивності даних.

Необхідно зазначити, що не існує універсального алгоритму кластеризації для обробки даних, оскільки якість роботи алгоритмів залежить від початкових даних. Так використання попередньо якісно не оброблених даних може призвести до зниженої точності алгоритму. І навпаки, правильно оброблені вхідні дані покращують якість кластеризації. У роботі [2] розглядаються методи кластеризації документів соціальних мереж Twitter Reddit, у яких основним способом передачі інформації про користувачів є невеликі блоки тексту. Розглянуто методи кластеризації k-means, k-medoids, ієрархічна кластеризація.

Результати показують, що методи кластеризації, у першу чергу k-means який дав найкращі результати, застосовані до попередньо нейронно оброблених даних забезпечують точну оцінку схожості текстових документів, відповідно і думок користувачів соціальної мережі, при умові використання правильних зовнішніх методів оцінки.

Також кластеризацію проводять за допомогою методу щільності, так у роботі [3] авторів цікавило чи можливо встановити за допомогою схожих даних з різних користувацьких акаунтів виникнення стихійного лиха в окремому регіоні земного шару. У якості методу кластеризації використовувався метод кластеризації на основі щільності, який групує дані з високим ступенем схожості. На основі текстового аналізу визначаються набори слів які з високою ймовірністю позначають тип та місце катастрофи. Проведене тестування отриманих після кластеризації даних показало ефективність кластеризації для аналізу текстових даних з акаунтів користувачів. До недоліків роботи можна віднести те, що кластеризація проводилася лише за допомогою одного методу, що не надає достатньої інформації про його ефективність порівняно з іншими.

Своєю чергою у дослідженні [4] основним методом ієрархічної кластеризації, хоча у ній також використовуються дані користувачів соціальної мережі Twitter. В статті пропонується підхід, заснований на аналітиці великих даних, який враховує дані соціальної мережі Twitter для виявлення проблем управління ланцюгом постачання у харчовій промисловості через отримання релевантних даних на основі ключових слів, їх попередню обробку необроблених і аналіз тексту за допомогою ієрархічної кластеризації. Тобто кластеризація використовувалась для аналізу акаунтів які висловлюють певне невдоволення постачею/якістю харчових продуктів через свої відгуки. За допомогою кластерного аналізу було проаналізовано дослідження ланцюжка поставок яловичини, де використовувалися тритижневі дані з Twitter.

Результат дослідження показав, що запропонований підхід текстової аналітики може бути корисним для ефективного виявлення та узагальнення важливих відгуків користувачів соціальних мереж по певній темі, у даному випадку постачання яловичини.

Наявні дослідження методів кластеризації для пошуку схожих акаунтів було детально досліджено в роботі [5]. Розглянуто використання різних алгоритмів кластеризації для аналізу даних користувачів соціальної мережі Twitter та ідентифікації неочевидних спільних патернів в текстах акаунтів які, на перший погляд не мають певної структури та їх ефективності.

Проводиться порівняльний аналіз підходів до неконтрольованого навчання, до яких відноситься кластерний аналіз, для визначення чи підтверджуються емпіричні результати аналізу даних за допомогою кластеризації. Було проведено порівняння, включно з методами кластеризації, алгоритмами, кількістю кластерів, розміром наборів даних, вимірюванням відстані, характеристиками кластеризації, методами оцінки та результатами. У висновку повідомляється, що в використання неконтрольованого навчання для аналізу даних соціальних медіа має кілька недоліків.

Були виділені наступні основні методи кластеризації:

- k-means;
- c-means;
- ієрархічна кластеризація;
- кластеризація на основі щільності.

В роботі [6] було проведено огляд наявної літератури з аналізу соціальних мереж, визначено, що проблема аналізу даних користувачів є активно обговорюваною, що робить розробку методу пошуку схожих акаунтів в соціальних мережах актуальним. Робота пропонує свої метрики аналізу, які засновані на чотирьох особливостях аналізу даних в соціальних мережах, а саме: виявлення шаблонів, злиття та інтеграція інформації, масштабованість інформації і візуалізація, які використовуються для визначення набору нових показників даних. Також дано оцінку існуючим сервісам та інструментам для аналізу соціальних мереж.

З дослідження літературних джерел випливає, що основними методами для обробки даних користувачів для розділення на гомогенні групи є методи: k-means, c-means, методи ієрархічної кластеризації, методи кластеризації на основі щільності.

Кластеризація на основі зв'язності, також відома як ієрархічна кластеризація, заснована на основній ідеї про те, що об'єкти більше пов'язані з прилеглими об'єктами, ніж об'єктами, що знаходяться далі. Ця кластеризація охоплює ціле сімейство методів, що відрізняються способом обчислення відстаней.

Алгоритми з'єднують «об'єкти» в «кластери» залежно від відстані. Кластер можна описати в основному максимальною відстанню, необхідною для з'єднання частин кластера. На різних відстанях утворюватимуться різні кластери, які можна уявити за допомогою дендрограми, що пояснює, звідки з'явилася загальна назва «ієрархічна кластеризація».

Тобто ці алгоритми не забезпечують єдине розбиття набору даних, а натомість забезпечують велику ієрархію кластерів, які зливаються один з одним на певних відстанях. Крім звичайного вибору функцій відстані, також необхідно

вибрати критерій зв’язку (оскільки кластер складається з декількох об’єктів, є кілька кандидатів для обчислення відстані) для використання.

Ця кластеризація може бути використана для пошуку спільних друзів та аналізу певних груп людей за ієрархічною ознакою, оскільки ієрархія надає змогу легко знайти зв’язаних з вами людей. Але оскільки для ефективного пошуку схожих акаунтів необхідно брати до уваги певний набір даних, та кількість даних необхідних для кластеризації не повинна бути великою, він не буде достатньо ефективним для аналізу користувачів соціальних мереж.

Кластеризація k-means представляє кожен кластер у вигляді центрального вектора, який не є членом набору даних. Коли число кластерів фіксоване рівним k, кластеризація k-means дає формальне визначення задачі оптимізації: знайти центри k кластерів і присвоїти об’єкти найближчому центру кластера, щоб квадрати відстаней від кластера були мінімальними.

Тобто необхідно визначити початкову кількість кластерів, далі алгоритм випадково обирає центри кластерів і кожен найближчий елемент приписується до цього центру. Коли усі елементи приписані до центру, визначається новий центр з ознаками які є середніми арифметичними серед елементів кластера. Однак він знаходить лише локальний оптимум і зазвичай запускається кілька разів із різними випадковими ініціалізаціями.

K-means створює кластери приблизно однакового розміру, оскільки вони завжди будуть призначати об’єкт найближчому центроїду. Це часто призводить до неправильного вирізування меж кластерів (що не дивно, оскільки алгоритм оптимізує центри кластерів, а не межі кластерів). Але також необхідно зауважити, що на відміну від методів c-means, k-means чітко призначає елементи до конкретних кластерів, що дозволяє чітко визначити схожі за певними рисами дані. Відносна простота швидкість роботи та чіткість віднесення об’єктів до кластерів робить k-means одним з найчастіше використовуваних методів кластеризації для роботи з даними з соціальних мереж, водночас необхідність конкретного задання кількості кластерів та чітке визначення кожного об’єкта може буди проблемним у разі роботи з даними які містять велику кількість шумів.

c-means підходить для випадку відсутності чіткого групування в наборі даних та належить до нечітких алгоритмів кластеризації. На відміну жорсткої кластеризації такої як k-means, нечіткі алгоритми призначають неперервні значення [0, 1] для забезпечення прийнятної кластеризації. c-means надає критерії для групування точок даних у різні кластери різного ступеня, які визначаються ступенем в залежності від схожості даних. Він містить функцію належності, яка представляє нечіткість його поведінки. За допомогою цієї функції дані прив'язуються до кожного кластера.

Ця кластеризація може ефективно використовуватись для даних які можуть відноситись одразу для декількох кластерів, оскільки до уваги береться також вірогідність потрапляння даних в кластер. Але вони мають істотні проблеми при зростанні кількості даних через зниження швидкості обробки, що вимагає використовувати їх лише з невеликими вхідними даними. Через це використання цієї кластеризації для обробки даних з соціальних мереж зазвичай є недоречним, оскільки вхідні дані мають великий розмір.

Кластеризація на основі щільності визначає об'єкти які мають відношення до кластера в залежності від простору між кластером та об'єктом, через це він може виявляти кластери довільної форми. Кластери створюються у точках в яких об'єкти максимально наближені один до одного, що дозволяє використовувати цей метод для обробки даних з великою кількістю шумів.

Водночас у разі відсутності шумів використання кластеризації на основі щільності може бути недоречною, оскільки вона вимагає великих навантажень на оперативну пам'ять комп'ютера. Через це отримати схожі результати можна за допомогою інших методів кластеризації при меншому навантаженні.

Висновок до розділу

Підсумовуючи перший розділ, можна зробити наступні висновки:

- Проаналізовані наявні компанії які займаються соціальними мережами та визначено, що через активний зріст кількості даних у мережі інтернет

актуальність пошуку схожих акаунтів в соціальних мережах за допомогою кластерного аналізу є досить великою, оскільки дозволить звичайні статистичні методи не здатні опрацьовувати настільки велику кількість даних.

- Визначено, що кластерний аналіз активно використовується для аналізу даних користувачів соціальних мереж, що підтверджує актуальність теми дослідження.
- Визначено, що у якості методів кластеризації соціальних мереж активно використовуються наступні:
    - k-means;
    - c-means;
    - ієрархічна кластеризація;
    - кластеризація на основі щільності.

Беручи це до уваги, необхідно детальніше ознайомитись з цими методами для визначення того, який буде найбільш ефективним для вирішення поставленої задачі.

2. НАУКОВО-ДОСЛІДНА ЧАСТИНА

2.1 Огляд кластерного аналізу

Однією з базових навичок живих істот є вміння розрізняти та групувати об’єкти за певними ознаками для класифікації. Класифікація є фундаментальною для більшості наук, неможливо уявити біологію без класифікації організмів, астрономію без класифікації зірок, медицину без класифікації хвороб.
Така широка необхідність у поділі та групуванні об’єктів призвела до виникнення аналізу на основі математичних методів, який називається кластерним аналізом.

Кластерний аналіз – це метод обробки великої кількості даних для кластеризації, тобто групування певного набору об’єктів за спільними рисами, так що утворюються окремі кластери, тобто групи схожих об’єктів які не пов’язані з об’єктами з інших кластерів. Методи кластеризації широко використовуються в біології, астрономії та психіатрії, а з появою соціальних мереж і в сфері ІТ. Не існує ідеального методу для розділення на кластери, усе залежить від попередньо встановлених вимог, що робить кластерний аналіз гнучким інструментом у вмілих руках. Такий аналіз належить до машинного навчання без вчителя – це один з методів машинного навчання який система самостійно навчається виконувати поставлену задачу без контролю з боку експериментатора.

Машинне навчання – це галузь дослідження штучного інтелекту яка фокусується на дослідженні методів та алгоритмів для імітації людського мислення та навчання і тісно пов’язане з обчислювальною статистикою. Алгоритми машинного навчання використовують в різних областях де важко або неможливо розробити звичайні алгоритми для виконання поставлених задач. Зазвичай проблеми пов’язані з великою кількістю даних які необхідно обробити. До таких сфер використання можна віднести медицину, штучний інтелект, маркетинг, розпізнавання мовлення, сільське господарство. Активний розвиток соціальних мереж та великих даних активно допомагає розвитку цієї галузі, оскільки надає значну кількість початкових даних подальшої для обробки та аналізу.

Машинне навчання без вчителя проводиться з великою кількістю об’єктів для визначення їх схожостей та взаємозв’язків між об’єктами, на відміну від машинного навчання з вчителем, коли система визначає закономірності по попередньо встановленим правильним відповідям.

Кластерний аналіз був розроблений в антропології Драйвером і Кребером в 1932 і введений в психологію Джозефом Зубіним в 1938 і Робертом Трайоном в 1939 і широко використовувався Кеттеллом, починаючи з 1943 для класифікації теорії ознак в психології особистості.

Поняття «кластер» не може бути точно визначено, що є однією з причин, чому існує так багато алгоритмів кластеризації.

До характеристик кластера можна віднести наступні:

- зовнішня згуртованість, тобто однорідність;
- зовнішня ізоляція, тобто відокремленість.

Мозку людини притаманно автоматично визначати кластери на основі відстаней між точками. «Кластери», присутні на малюнках, зазвичай зрозумілі для більшості людей не знайомих з терміном, навіть якщо вони приймають різні форми, що дозволяє не давати чітке формальне визначення терміну.

Оскільки кластерний аналіз працює з багатовимірних даних які зазвичай представлені у вигляді матриць, результат його роботи можна представити графічно. Графічні відображення багатовимірних даних можуть надати розуміння структури даних, і вони можуть бути корисними для аналізу наявності кластерів у даних.

Рисунок. 2.1 – Приклади кластерів

Приклад на рисунку 2.2 демонструє набір даних який не має певної кластерної структури, але під час кластеризації може метод може виявити певні окремі групи даних. Тут більшість спостерігачів дійшли б висновку, що не існує «природної» кластерної структури, просто єдина однорідна сукупність точок, але після кластеризації можуть виникнути окремі групи даних розділені по кластерах. Як наслідок у спостерігача виникає проблема інтерпретації отриманих даних, оскільки він може не мати попередньої інформації про дані й може зробити висновки про існування окремих кластерів в наборі даних який не повинен їх містити. Через це важливим етапом який передує кластерному аналізу є аналіз структури даних яка буде використана для кластеризації, які саме бізнес-процеси та правила відбору по схожості повинні бути визначені для отримання надійних результатів.

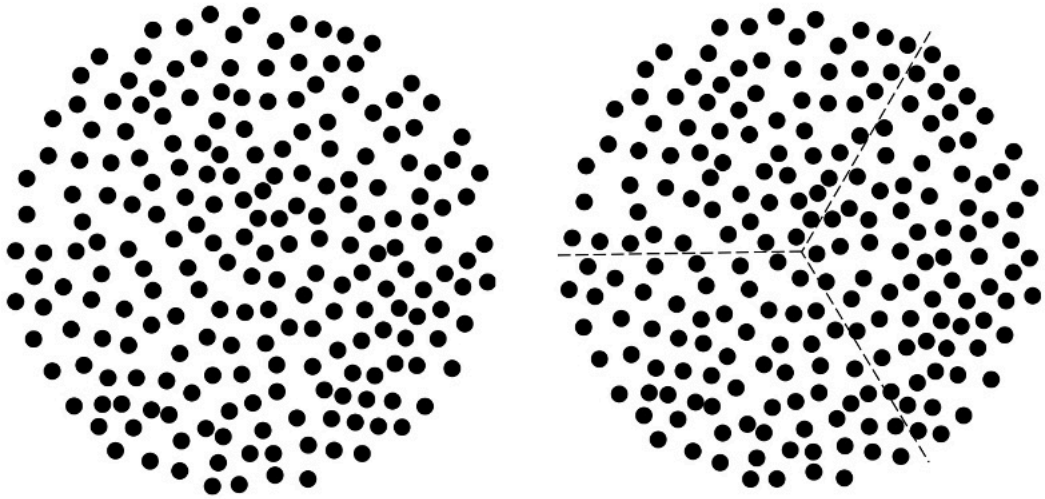


Рисунок 2.2 – Приклад однорідного набору даних до та після кластеризації

Спрощено алгоритм кластеризації можна представити наступним чином (Рис. 2.3 – 2.6):

- Завантаження початкових даних.
- Вибір початкових центрів згідно з обраним методом кластеризації.
- Віднесення кожного об’єкту до певного кластера.

- Перерахунок центрів кластерів відповідно до положення обраних об’єктів кластера.
- Повторення кроків 3-4 у разі зміни кластерів.
- Завершення кластеризації.

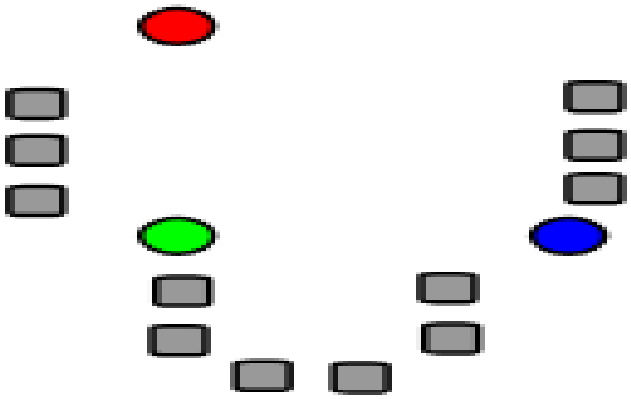


Рисунок 2.3 – Вибір початкових точок

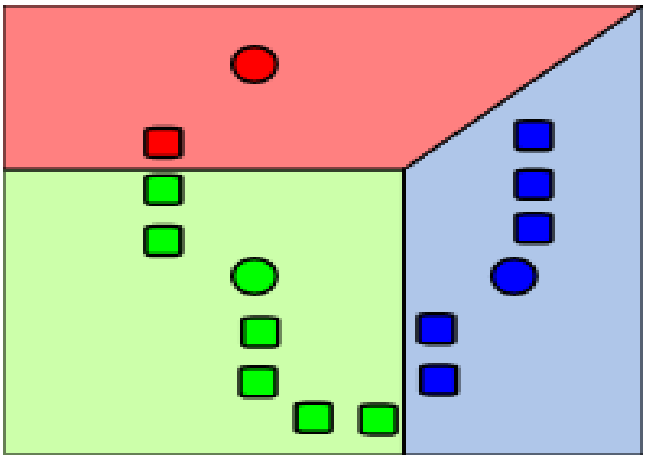


Рисунок 2.4 – Віднесення об’єктів до кластерів

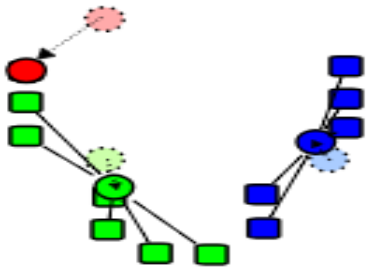


Рисунок 2.5 – Визначення нових центрів кластерів

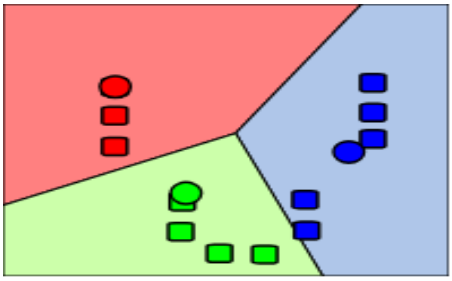


Рисунок 2.6 – Кінець кластеризації

У більшості випадків початкові дані являють собою матрицю стовпці якої є досліджуваними об’єктами, а строки – їх характеристиками (Функція 2.1).

$$X = \begin{bmatrix} x_{11} & x_{1m} \\ x_{n1} & x_{nm} \end{bmatrix}, \tag{2.1}$$

де n – кількість об’єктів для класифікації, а m – характеристики об’єкта.

Числові значення які можуть прийняти характеристики поділяються на декілька типів:

- Якісні – приймають два або більше значень до яких можна віднести певні числові значення, але вони не будуть відповідати впорядкованості цих значень та не можуть бути використані для арифметичних операцій (стать, місце роботи).
- Рангові – з цими числовими значеннями немає сенсу проводити арифметичні операції, але вони чітко відображають порядок характеристик (рівень вдоволеності сервісом).
- Кількісні значення – впорядковані та можна проводити арифметичні операції (зріст, вік).

Приведення характеристик об’єкта до певних кількісних характеристик залежить від початкових даних та поставленої задачі, тому неможливо визначити універсальний метод для приведення показників до порівняного виду. Використання таких методів називають нормалізацією або нормуванням. У разі

приведення характеристик об’єкта до числових значень або якщо характеристики початково мали такі значення, можна скористатись наступними формулами нормалізації які широко використовуються:

$$x_{ij}^{n} = \frac{x_{ij} - \overline{x_j}}{\sigma_j}, \tag{2.2}$$

де x_{ij}^{n} – це нормалізований параметр певного об’єкта, x_{ij} – це значення яке приймає показник *j* об’єкта *i,* $\overline{x_j}$ – середнє значення характеристики *j* для усіх об’єктів, σ_j – середньоквадратичне відхилення характеристики *j.*

У разі використання цієї формули характеристики об’єкта приймають середнє значення 0, а відхилення від нього дорівнює 1.

Інша формула перетворює характеристик шляхом відображення їх на інтервал [0; 1]:

$$x_{ij}^{n} = \frac{x_{ij} - x_j^{min}}{x_j^{max} - x_j^{min}}, \tag{2.3}$$

де $x_j^{min} = \min\{x_{ij}\}$, а $x_j^{max} = \max\{x_{ij}\}$.

Після проведення нормалізації та вибору початкових точок кластерів слідує етап оцінки схожості об’єктів за їх характеристиками і є найважливішим, оскільки саме на цьому етапі визначається до якого кластеру належить об’єкт.

Зазвичай схожість об’єктів визначається тим, наскільки «близько» один до одного або як далеко вони один від одного. Як вже було вказано вище, алгоритми кластеризації мають за відправну точку багатовимірну матрицю, елементи якої відображають кількісну міру близькості. Два об’єкти вважаються «близькими», коли їх відмінності або відстань невеликі, або подібність велика.

Як і при нормалізації, для цього етапу не існує універсального методу, і вибір цілком залежить від набору даних який досліджується і їх природи. Широко

використовуються наступні формули для визначення схожості, оскільки вони підходять для визначення схожості взаємопов’язаних об’єктів:

$$d_{pq} = \sqrt{\sum_{i=1}^{n} (p_i - p_q)^2}, \tag{2.4}$$

де p_i та p_q це точки в n-вимірному просторі. Ця формула також відома як формула Евкліда, і оскільки дані об’єкту представлені у вигляді точок на координатній прямій, за допомогою неї можна визначити мінімальну відстань між ними.

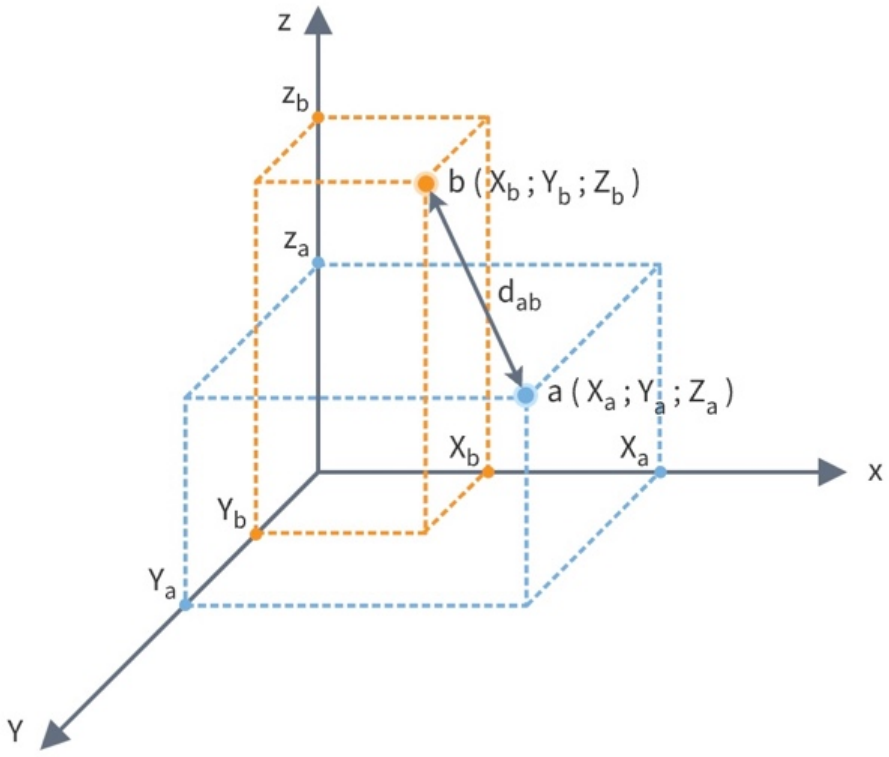


Рисунок. 2.7 – Визначення схожості двох об’єктів за допомогою формули Евкліда

При великій кількості значення n ця формула може надавати недостатньо точні значення, у цьому разі може бути використана формула відома як відстань Махаланобіса:

$$d_M(x_i, x_j) = (x_i - x_j) F^{-1} (x_i - x_j)^T , \tag{2.5}$$

Де x_i, x_j це вектори характеристик, F це матриця коваріації, у разі якщо ця матриця одинична формула зводиться до формули відстані Евкліда.

Відстань Махаланобіса можна використовувати для даних у яких вектори x_i, x_j однаково важливі для подальшої класифікації, а точки у просторі розподілені відносно центру у вигляді еліпса, а для відстані Евкліда у вигляді кола (Рис. 1.12).

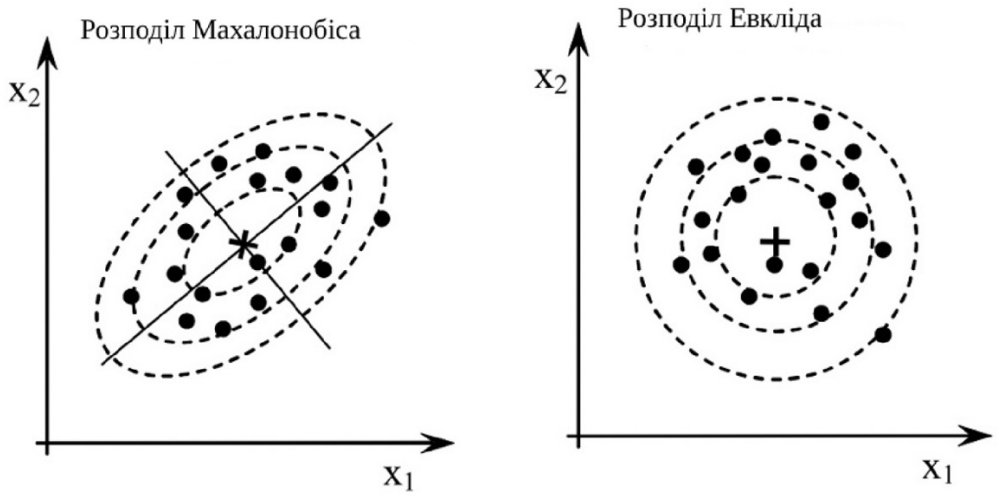


Рисунок 2.8 – Приклад розподілу точок на плоскості за допомогою формули Махалонобіса та Евкліда

2.2 Ієрархічна кластеризація

Для ієрархічної кластеризації початкові дані не поділяються на певну кількість класів або кластерів за один крок, замість цього поділ на кластери складається з поділу даних на окремі кластери з одного кластера до n кластерів, кожен з яких містить одне значення, або навпаки злиття з цих окремих кластерів даних у єдиний кластер.

Методи ієрархічної кластеризації які зливають n-кластери у єдиний кластер називаються агломераційними методи, а методами поділу називають методи які послідовно поділяють n кластерів на дрібніші. Обидва типи ієрархічної

кластеризації ставлять перед собою задачу розділити або поєднати на кожному етапі дані, за допомогою певного значення схожості.

Поділ або об’єднання в цих методах, є незворотними, тому, коли агломеративний алгоритм об’єднав два набори даних, вони не можуть бути згодом розділені, а коли роздільний алгоритм здійснив розділення, це неможливо скасувати, що є досить значним недоліком, оскільки завдяки цьому потенційні помилки в вибірці не можуть бути виправлені у подальших ітераціях. Ієрархічний метод страждає від того недоліку, що він ніколи не може виправити те, що було зроблено на попередніх етапах.

Через те що агломераційні ієрархічні методи зводять дані до єдиного кластера, а методи поділу розбивають весь набір даних на n кластерів одиничних кластерів, визначення оптимального поділу накладається на дослідника який проводить аналіз і визначає на якому саме етапі дані були поділені на достатню кількість кластерів.

Ієрархічні класифікації, створені шляхом агломерації або розподілу, можуть бути представлені двовимірною діаграмою, відомою як дендрограма, яка ілюструє злиття або поділи, зроблені на кожному етапі аналізу (Рис. 2.9). Оскільки цей алгоритм повністю розбиває або об’єднує дані, визначення достатнього рівня розбиття лежить на досліднику.

За допомогою багатовимірних даних можна побудувати дендрограми для кожної окремої змінної, але це не дуже допоможе у розкритті кластерної структури даних, оскільки граничний розподіл кожної змінної може не точно відображати розподіл повного набору змінних.

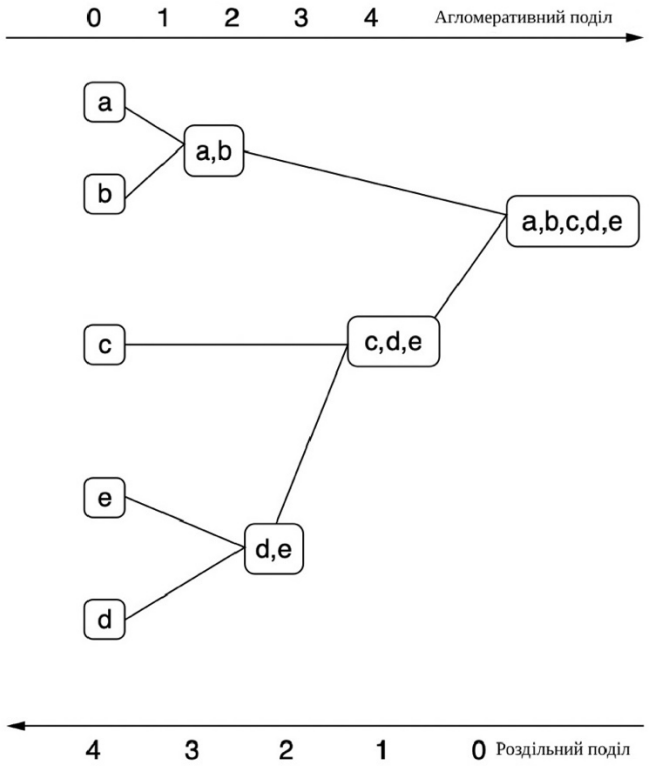


Рисунок 2.9 – Приклад агломеративного та роздільного поділу даних

Зазвичай агломераційним методам надають перевагу перед методами поділу оскільки методи поділу вимагають занадто великих ресурсів для обробки даних і через це не вважається достатньо ефективним, тому має сенс віддати перевагу саме дослідженню агломераційного методу.

Цей алгоритм поділяє дані на декілька розділів на кожному етапі, на першому дані складаються з *n*-одиничних кластерів де *n* це кількість досліджуваних об’єктів На останньому етапі кластери об’єднуються в єдиний кластер який містить усі *n* кластерів. На кожному етапі поєднання кластерів відбувається по принципу найближчої подібності, тобто визначаються та об’єднують кластери по принципу схожостей об’єктів. Для оцінки схожості об’єктів дані попередньо приводяться до матриці даних і порівнюються за допомогою формул представлених у розділі 1.

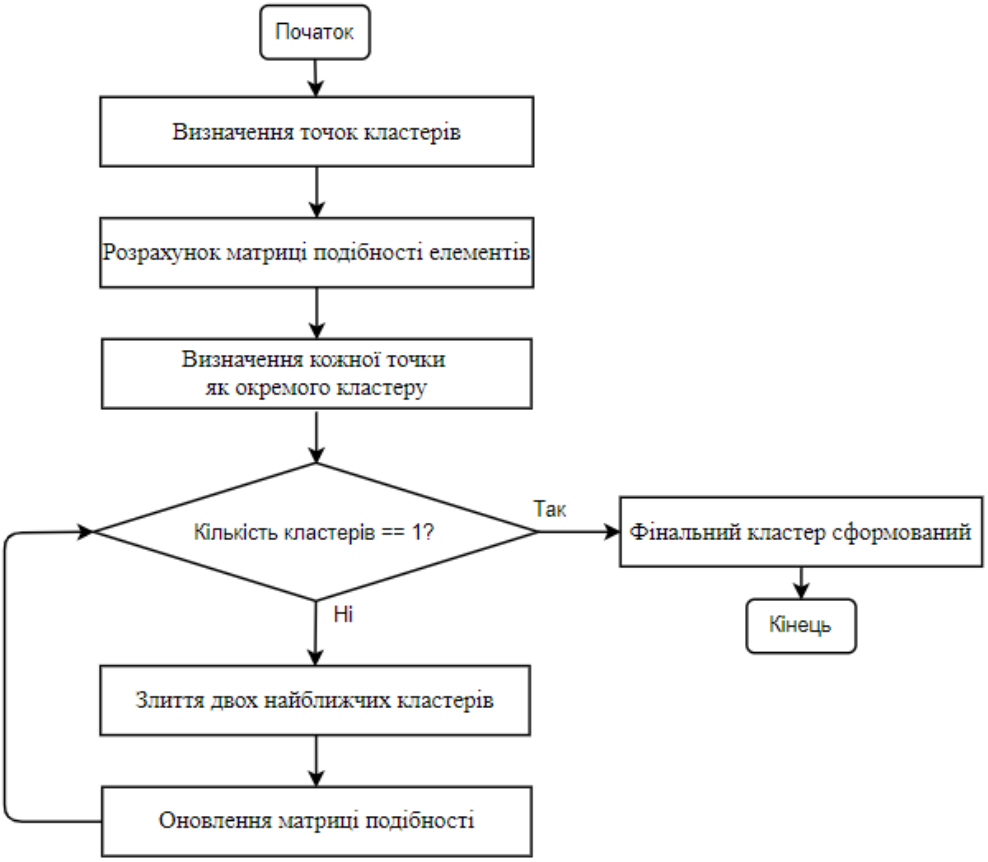


Рисунок 2.10 – Алгоритм ієрархічної кластеризації

З вищеозначеного випливає те, що дана кластеризація може бути використана для пошуку спільних друзів та аналізу певних груп людей за ієрархічною ознакою, оскільки ієрархія надає змогу легко знайти зв’язаних з вами людей. Але оскільки для пошуку схожих акаунтів використовують велику кількість акаунтів, графічне відображення кластеризації у вигляді дендрограми буде не інформативним, та вимагатиме значну кількість обчислювальних ресурсів. Через це ієрархічна кластеризація не буде достатньо ефективною вирішення для поставленої задачі.

2.3 Кластеризація k-means

k-means є найпростішим і найбільш часто використовуваним алгоритмом. Він починається з випадкового визначення початкових точок кластерів та додає до кожного кластера об’єкти на основі подібності між об’єктом та центром кластера.

Коли усі об’єкти розподілені по кластерах, визначаються центроїди кожного кластера.

Центроїд – це вектор, який обчислюється як середні значення характеристик кожного об’єкта кластера . Алгоритм k-means продовжує свою роботу доки кластери та їх центроїди не будуть змінюватись. Алгоритм k-means популярний, оскільки його легко реалізувати, а його часова складність дорівнює *O(N)*, де *N* — кількість об’єктів. Основна проблема цього алгоритму полягає в тому, що він чутливий до вибору початкових точок, більш того, кількість кластерів яка повинна бути знайдена за допомогою k-means кластеризації повинна бути визначена дослідником.

Для визначення оптимальної кількості кластерів використовується метод ліктя. Ідея методу ліктя полягає в тому, щоб виконати кластеризацію k-means на наборі даних для діапазону значень k, і для кожного значення k обчислити суму квадратів помилок. Після цього необхідно побудувати лінійну діаграму для кожного значення k, і спираючись на цю діаграму визначити «лікоть», який буде найкращим значенням для кількості кластерів. При збільшенні k кількість сума квадратів помилок буде зменшуватись, але чим більше k, тим більшу кількість кластерів ми отримаємо, що може призвести до появи кластерів які не відображають реального стану системи. Тобто необхідно визначити невелике значення k, яке все ще має низьку кількість помилок.

Формула для визначення суми квадратів помилок має наступний вигляд:

$$SSE = \sum_{i=1}^{n}(x_i - \bar{x})^2, \tag{2.6}$$

де n – це кількість об’єктів вибірки i це кількість характеристик окремого об’єкта x_i це значення яке приймає характеристика, $\bar{x}$ це середнє значення яке приймають характеристики об’єкта.

На рисунку 2.11 Можна чітко побачити, що кількість помилок при $k > 3$ зменшується значно менше ніж при менших значеннях k, що робить значення $k = 3$ оптимальним.

Протягом застосування алгоритму визначаються лише локально-оптимальні кластери, через це виникає необхідність ітеративного виконання алгоритму k - means при заданій кількості кластерів для різних для отримання оптимальних результатів кластеризації.

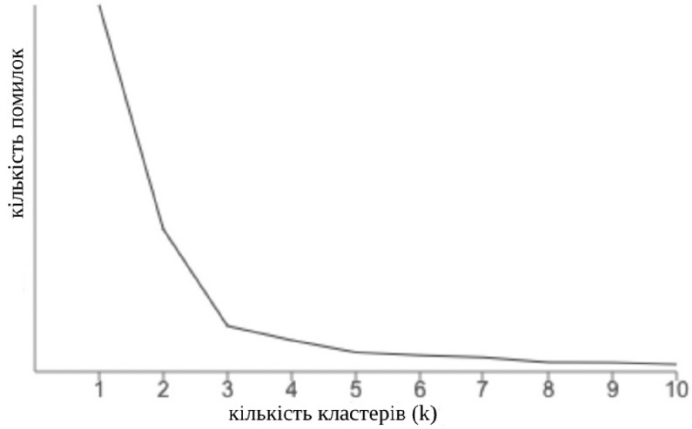


Рисунок 2.11 Визначення оптимальної кількості кластерів методом ліктя

До особливості k-means також можна віднести розбиття простору на діаграму Вороного, тобто таке розбиття при якому простір розбивається на площини максимально близькі до об’єктів.

Також необхідно зауважити, що оскільки початковий вибір центрів кластерів є випадковим, це може стати джерелом похибки. Через це існує необхідність у попередньому дослідженні початкових даних для визначення їх структури та особливостей, що дозволить виявити неточності в кластерах після завершення роботи алгоритму. Також можливий варіант самостійного визначення початкових точок кластерів для першої ітерації алгоритму, у разі якщо дослідник впевнений в оптимальності цих точок.

Існує необхідність у попередньому очищенні початкових даних від шумів, оскільки даний алгоритм, на відміну від алгоритму щільності, не здатен якісно відокремлювати їх, що може призвести до похибки у фінальному результаті.

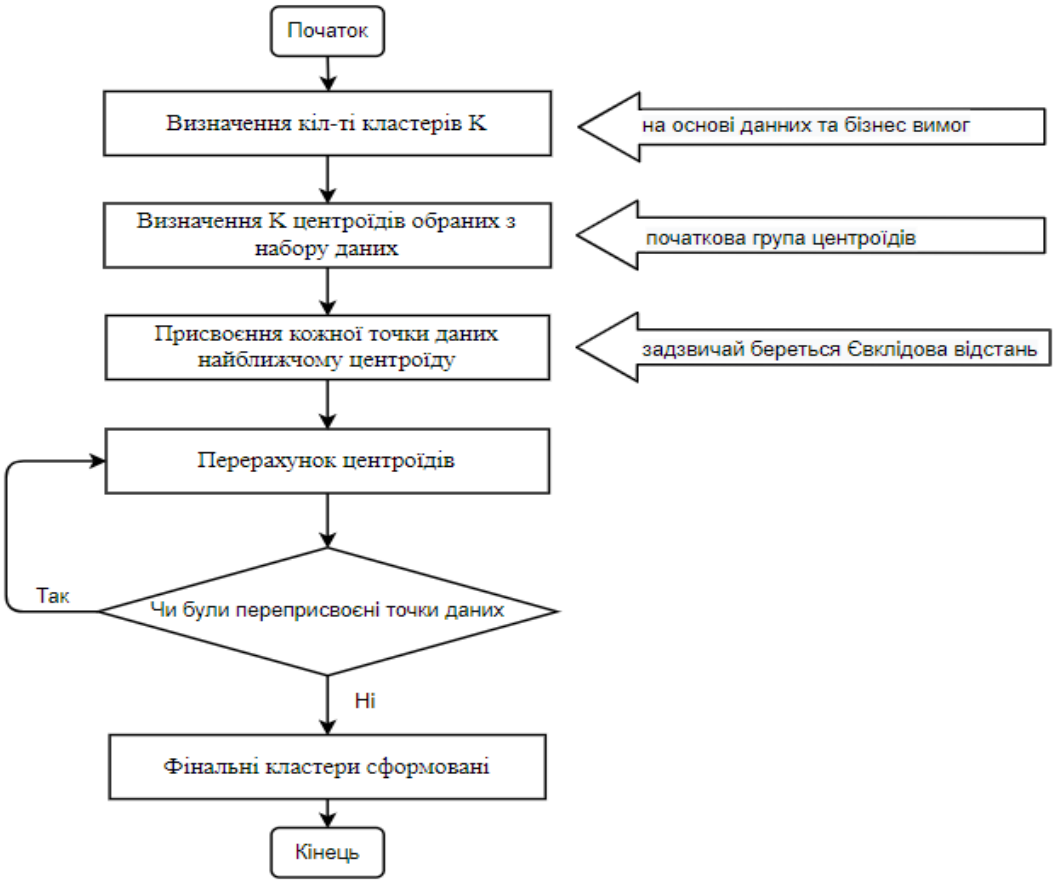


Рисунок 2.12 – Алгоритм k-means

До недоліків алгоритмів k-means часто відносять необхідність чітко визначити кількість кластерів, також отримані в результаті роботи кластери будуть приблизно одного розміру. Можна відмітити необхідність декількох запусків алгоритму для отримання кращого результату. Але також необхідно зауважити, що на відміну від методів c-means, k-means чітко призначає елементи до конкретних кластерів, що дозволяє чітко визначити схожі за певними рисами акаунти, а необхідність чітко визначити кількість кластерів перетворюється з недоліку на перевагу коли річ заходить про використання результатів кластеризації для подальшої таргетованої реклами. Це дозволяє зробити висновок про те, що алгоритми k-means добре підійдуть для обробки даних з метою пошуку схожих акаунтів у соціальних мережах.

2.4 Кластеризація c-means

Основним недоліком k-means, як було описано вище, є потреба у заданні числа кластерів дослідником, що не завжди надає достатньо точні результати. Також можлива ситуація особливого розташування даних - на перетині кластерів. В такому випадку метод k-means може працювати нескінченно довго, оскільки він не здатен остаточно визначити до якого саме кластеру слід віднести об’єкт.

Кластеризація c-means належить до нечіткої кластеризації, тобто такої у якій об’єкти не призначаються конкретному кластеру: вони мають лише певний ступінь приналежності, яка вказує на те наскільки сильно об’єкт належить до того чи іншого кластера, на відміну від чітких методів k-means та ієрархічної кластеризації, де значення належності об’кта може дорівнювати одиниці (належить кластеру) або нулю (не належить кластеру).

Нечітка кластеризація декілька наступних переваг перед чіткими методами:

- дані про об’єкт можна співвідносити з іншими кластерами, що надає більше інформації;
- ступінь належності явно вказує наскільки інші кластери можуть бути оптимальними, оскільки різниця між кластерами може бути незначна у межах поставленої задачі.

У нечіткому кластерному аналізі кількість підгруп вважається відомою, а належність об’єкта в кожному кластері як і в k-means оцінюється за допомогою ітераційного методу, на основі цільової функції. Концепція функції приналежності походить від нечіткої логіки, розширення булевої логіки, в якій поняття істини та хибності замінені поняттями часткової істинності.

Оскільки для нечіткого аналізу не можна примінювати функції схожості чіткої кластеризації, існують окремі функції для визначення членства об’єкта.

Аналіз ступеня приналежності призначає випадкам «ступінь приналежності» до двох або більше кластерів за допомогою ймовірності членства в кластері. Так для набору об’єктів n та кількості кластерів k, маючи вектор характеристик x_i критерій визначення відповідності можна обчислити за наступною формулою:

$$\sum_{t=1}^{k}\sum_{i=1}^{n} u_{it}^{v}\, d^2(x_i, m_t) \ , \qquad (2.7)$$

де m_t це центр кластера *t*, $u_{it}^{v} > 0$ для усіх $i = 1, \ldots, n$ та $\sum_{t=1}^{k} u_{it} = 1$, $d(x_i, m_t)$ це відстань від об'єкта до центра кластера яка може бути визначена формулами для чіткої кластеризації, наприклад відстані Евкліда, *v* це ступінь нечіткості, якому зазвичай задають значення $v = 2$, оскільки $v = 1$ надає чіткі результати.

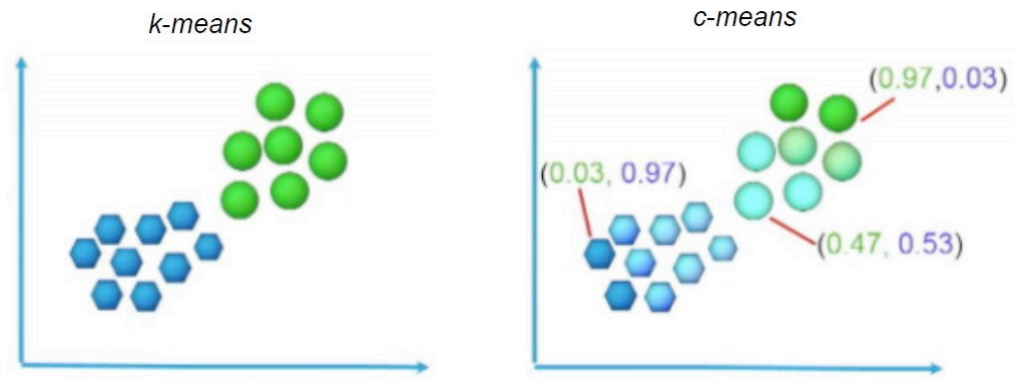


Рисунок 2.13 – Порівняння k-means та c-means

Рисунок 2.14 – Алгоритм c-means

c-means кластеризація має ряд значних недоліків [13], основним з яких є значна кількість обчислювальних ресурсів яка необхідна для проведення кластеризації. При цьому слід зауважити, що необхідність визначення приналежності користувача соціальних мереж декільком кластерам одночасно не завжди є необхідною, що робить c-means кластеризацію занадто ситуативною і через це використання цього алгоритму кластеризації для пошуку схожих акаунтів в соціальних мережах виглядає недоречним.

2.5 Кластеризація на основі щільності

Методи k-means, c-means та ієрархічна кластеризація розраховані для пошуку кластерів з вибірки яка була попередньо оброблена, а значення які належать вибірці достатньо схожі. Через це результат їх обробки у разі присутності шуму у наборі початкових даних результат кластеризації може бути досить неточним.

Через це виникає необхідність у алгоритмі який здатен опрацьовувати дані які містять певні шуми, що якраз і є сильною стороною кластеризації на основі щільності.

Суть цієї кластеризації полягає в попередньому визначенні скупчення об’єктів у просторі, оскільки шуми у цьому разі будуть розташовані дуже далеко відносно досліджуваних об’єктів. Основним методом визначення кластерів у кластеризації на основі щільності є DBSCAN (Density-Based Spatial Clustering Of Applications With Noise — просторова кластеризація додатків із шумом).

Принцип роботи DBSCAN наступний:

- Задається значення радіуса у якому між двома точками може виникнути сусідство. Якщо значення вибрано замале, велика частина даних буде вважатися викидом, якщо ж навпаки значення вказано завелике – буде створено великі кластери з не репрезентативними вибірками даних.
- Задається мінімальна кількість сусідів (об’єктів) у радіусі. Чим більший набір даних, тим більше значення необхідно задати.

- Шукаються сусідні об’єкти в зазначеного радіуса і визначаються основні точки з найбільшою кількістю сусідів.
- Для кожної основної точки, якщо вона ще не призначена кластера, створюється новий кластер.
- Кожну точку яка належить радіусу основної додають до кластера до якого належить основна точка.
- Точки a і b називаються густинно зв’язаними, якщо існує точка c, яка має достатню кількість точок серед своїх сусідів і обидві точки a і b знаходяться на попередньо заданій максимальній відстані. Це ланцюговий процес. Отже, якщо b є сусідом c, c є сусідом d, d є сусідом e, який своєю чергою, є сусідом a, означає, що b є сусідом a.
- Усі точки які не призначені до кластерів будуть вважатись шумом, оскільки вони знаходяться на занадто великій відстані.

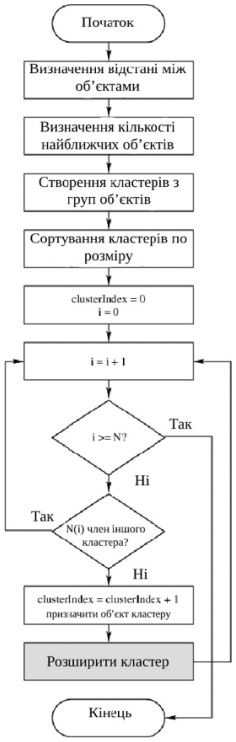


Рисунок 2.15 – Алгоритм кластеризації на основі щільності

Ключовим недоліком кластеризації на основі щільності є те, що вона очікує деякого падіння щільності для виявлення меж кластера оскільки у першу чергу

призначена для роботи з наборами даних які містять шуми. Це може призвести до низької якості кластеризації у разі використання попередньо оброблених наборів даних які не містять зайвих значень.

Через досить високу потенційну неточність та відсутність у необхідності складних умов для аналізу кластеризація на основі розподілу не підходить для аналізу користувачів соціальних мереж.

2.6 Створення методу пошуку схожих акаунтів в соціальних мережах

Попередній аналіз алгоритмів кластеризації показав, що не існує певного методу «срібної кулі» який дозволяє аналізувати будь-який масив даних однаково ефективно, вибір необхідного методу суцільно залежить від поставлених в бізнес-вимогах задач. Оскільки неможливо описати роботу усіх методів, в даній роботі буде розглянутий лише алгоритм k-means як найбільш універсальний та відповідний до аналізу акаунтів користувачів.

Але оскільки вважається, що початкові дані матимуть досить великий об’єм, через те що в деяких соціальних мережах кількість користувачів сягає мільярду, доречно буде використовувати модифікований для роботи з великою кількістю даних k-means mini batch (у звичайного k-means кількість обчислень збільшується зі збільшенням кількості елементів).

Основною відмінністю від звичайного k-means є те, що він ітеративно розбиває вхідні дані на рандомізовані батчі – тобто невеликі частини початкового масиву даних які обираються випадково [14].

Для певного набору заданих даних T = $\{x_1, x_2, .., x_p\}$, $x_i \in R^{m*n}$ де x_i представляє об’єкт, який є n-вимірним вектором, а m це кількість об’єктів які містить набір T, пошук центрів кластерів C (попередньо вказана кількість кластерів) з мінімізацією кількості даних T обчислюється за формулою:

$$min \sum_{x \in T} ||f(C, x) - x||^2, \quad (2.8)$$

де $f(C, x)$ повертає найближчий центр кластера до об'єкта x.

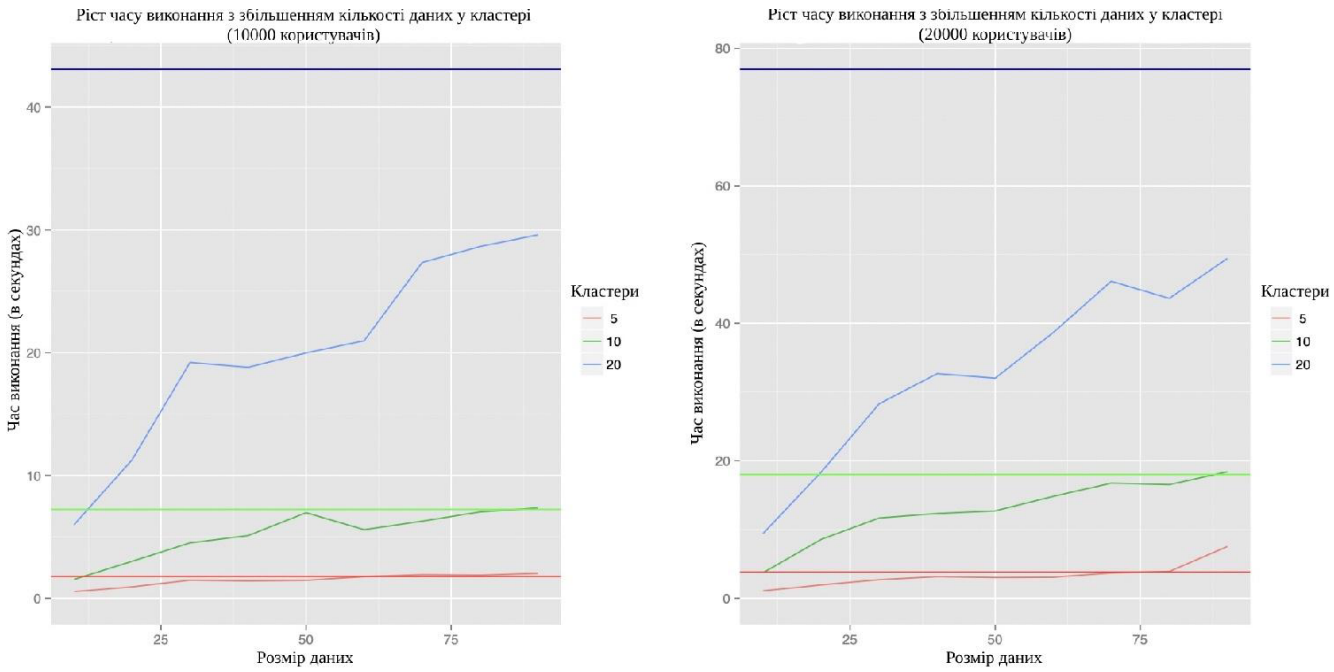


Рисунок 2.16 – Порівняння швидкість роботи k-means і mini batch k-means з ростом кількості вхідних даних

Час, необхідний для об'єднання з використанням лише вибірки набору даних, буде меншим, ніж час, необхідний для використання всіх даних. Так час кластеризації зменшується з використанням різних розмірів мінібатчів (від 10% до 90% вхідних даних) і різної кількості кластерів для фіксованого розміру кластера. На графіку горизонтальною лінією позначено час, необхідний k-means для різної кількості кластерів (2.16).

Зі збільшенням кількості кластерів відносна економія часу на обчислення також збільшується. Економія обчислювального часу більш помітна лише тоді, коли кількість кластерів дуже велика. Вплив розміру кластерів на час обчислення також більш очевидний, коли кількість кластерів більша. З графіка видно, що час обчислення лінійно збільшується зі збільшенням розміру набору даних (подвоєння кількості прикладів подвоює час, необхідний для обох алгоритмів). На це немає очевидної причини, оскільки час обчислення mini batch k-means різко скорочується, коли кількість кластерів велика.

Зі збільшенням кількості кластерів співвідношення також збільшується. Це означає, що отримати подібну функцію оцінки стає важче, коли кількість кластерів велика. Майже ідентична поведінка може спостерігатися для подібності до результівного розподілу k-means, з цього можна зробити висновок, що збільшення кількості кластерів зменшує подібність mini batch k-means до k-means.

Як було визначено вище, алгоритми кластеризації k-means сильно страждають у разі якщо надані дані містять шуми. Через це виникає необхідність у попередній нормалізації даних. Не існує певного універсального методу нормалізації, оскільки дані які надходять можуть приймати різні значення, тому метод нормалізації повинен бути обраний дослідником беручи до уваги поставлену задачу. Надалі дані можуть бути приведені до певних кількісних характеристик за допомогою формул які були описані в розділі 2.1.

Беручи до уваги усе вищеописане, метод пошуку схожих акаунтів в соціальних мережах можна описати наступними кроками:

- Отримання початкових даних користувачів соціальних мереж.
- Нормалізація даних та приведення до кількісних характеристик до значень [0, 1].
- Створення векторної матриці X, де x_i це користувач, а x_j це його характеристики отримані з API яке відповідає за відправку даних.
- Визначення кількості необхідних кластерів за допомогою методу ліктя.
- Проведення кластеризації mini batch k-means, у якості міри оцінки схожості можна використовувати відстань Евкліда.
- Отримання результівної вибірки даних.

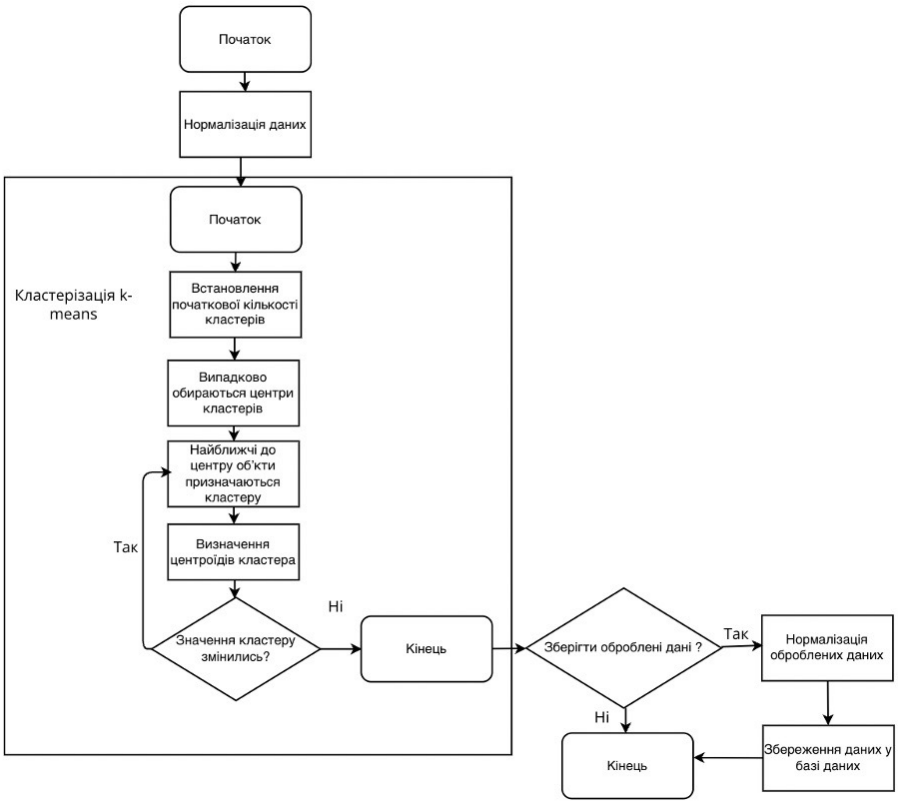


Рис. 2.18 Блок-схема методу пошуку схожих акаунтів в соціальних мережах за допомогою кластеризації

Висновок до розділу

Підсумовуючи перший розділ, можна зробити наступні висновки:

- Проаналізовано наявні методи кластерного аналізу та визначено, що метод mini batch k-means є найбільш ефективним для визначення схожих користувачів соціальних мереж, оскільки здатен обробляти велику кількість схожих даних та визначати кластери з великою точністю.
- Проаналізовані методи пошуку схожості та нормалізації даних, визначено що для k-means кластеризації можуть бути використані відстань Евкліда та відстань Махаланобіса.

3. ПРАКТИЧНА ЧАСТИНА

3.1 Розгляд наявних API для отримання даних користувачів соціальних мереж

Twitter API – це RESTful API розроблений компанією Twitter для вилучення даних користувачів з цієї соціальної мережі. Для підключення до цього API попередньо необхідно зареєструватись у Twitter у якості розробника, цей спосіб отримати доступ до API Twitter — спосіб контролю та взаємодії зі сторонніми розробниками. Це необхідно через допущені раніше зловживання платформами соціальних мереж, тепер також потрібно подати заявку на обліковий запис розробника і отримати схвалення створення нових додатків. Разом з додатком також буде створено набір струму нової аутентифікації, які дозволять вам отримати програмний доступ до платформи Twitter.

В такому випадку мається на увазі, що ви створюєте додаток, який маєте намір авторизувати для доступу до даних свого облікового запису, тому такий підхід може здатися трохи дивним: чому б просто не використовувати своє ім'я користувача та пароль для доступу до API? Річ у тім, що такий підхід може бути непоганий для вас, але інші, наприклад, ваш друг або колега, можуть почуватися не зовсім комфортно, передаючи комбінацію імені користувача та пароля, щоб отримати можливість користуватися результатами роботи вашої програми. Передача облікових даних завжди погана ідея. На щастя, цю проблему помітили та створили стандартний протокол відкритої авторизації під назвою OAuth3 (скорочено від Open Authorization), який призначений для подібних ситуацій та загалом для соціальних мереж. На цей час протокол є стандартом для соціальних мереж.

Facebook Graph Api – це API від компанії Facebook яка надає дані користувачів у вигляді графа - структури даних, що представляє соціальні зв'язки і складається з вузлів і ребер, що їх пов'язують. Graph API також надає засоби взаємодії з графом, також існує графічний інтерфейс Graph API Explorer який надає

можливість взаємодії з графом за допомогою користувацького інтерфейсу. В якості формату відповіді Facebook Graph Api повертає формат json, що дозволяє легко обробляти дані що надходять.

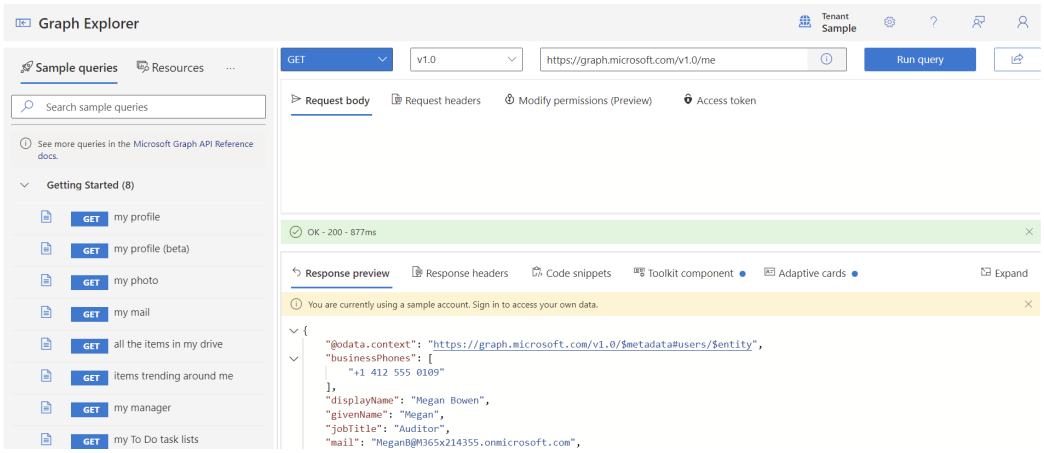


Рисунок 3.1 – Інтерфейс Graph API Explorer

Graph API Explorer зручно використовувати, коли у розробника існує токен OAuth, пов'язаний з певним набором дозволів для програми, що розробляється, і необхідно виконати кілька запити для налагодження програми. Токен OAuth можна отримати при реєстрації сторінки у мережі Facebook.

На відміну від Twitter API, який дозволяє отримати лише базову інформацію про користувача, Facebook API дозволяє проводити складний аналіз користувацьких сторінок за наступними характеристиками:

- наскільки популярна сторінка?
- наскільки активно проявляють себе підписники сторінки?
- чи є серед підписників особливо активні?
- які теми на сторінці обговорюються найактивніше?

LinkedIn API – це API соціальної мережі LinkedIn яке надає інформацію про сторінки користувачів та його колег у форматі CSV. Як і для Twitter та Facebook, для отримання даних необхідно мати акаунт у мережі LinkedIn.

Слід зазначити, що протягом багатьох років у LinkedIn постійно вносилися зміни до API, що обмежують доступ до інформації. Наприклад, спроба отримати

всі дані про контакти через API може завершитися помилкою 403 («Заборонено»). З іншого боку, LinkedIn, як і раніше, дозволяє завантажити архів з інформацією про всі ваші контакти, про що ми поговоримо в розділі «Завантаження файлу з інформацією про контакти в LinkedIn». Цей архів містить ті ж дані, що доступні користувачеві на веб-сайті LinkedIn. Через це дані отримані з API можуть містити не всю інформацію яку запитує розробник.

З розгляду наявних API можна зробити висновок, що інформація про користувача надходить у вигляді текстових даних json або CSV формату, тобто попередня нормалізація даних буде стосуватись саме текстових значень.

3.2 Нормалізація текстових даних

Через відсутність єдиного стандарту для отримання даних с соціальних мереж виникає необхідність в попередній нормалізації даних, оскільки вхідні текстові дані можуть приймати різні значення несучи при цьому однакову інформацію, що надалі може завадити кластеризації. В якості прикладу припустимо що ми намагаємось створити кластери користувачів які в якості спільної риси мають країну, а вхідні дані містять текстове значення яке відповідає країні яку встановив користувач.

Можливий набір значень який буде ідентифікувати Україну буде виглядати так:

- Україна
- УКР
- Ukraine
- UKR
- UA

Без попередньої нормалізації даних буде отримано набір кластерів які не фактично будуть відноситись до однієї країни.

У випадку текстових повідомлень популярним є метод розбиття текстових повідомлень на окремі слова, або речення, також відомий як сегментація. Алгоритм сегментації передбачає розбиття кожного повідомлення на окремі частки, та

подальше видалення шумів (слова які не несуть у собі ніякої інформації, наприклад пунктуація).

Розглянемо алгоритм сегментації на прикладі. В якості прикладу сегментації візьмемо наступне речення: «Backgammon is one of the oldest known board games.». Першочергове розбиття речення надасть нам наступний масив слів: ['Backgammon', 'is', 'one', 'of', 'the', 'oldest', 'known', 'board', 'games', '.'].

Одразу можна побачити що в масиві містяться елементи які не несуть у собі ніякої інформації, тобто є шумом. Через це наступний крок алгоритму видаляє з отриманого масиву даних елементи які не будуть використовуватись, після цього ми отримуємо масив даних який можна використовувати для аналізу акаунтів, наприклад якщо у двох акаунтах неодноразово зустрічається слово «Backgammon», це свідчить про те що акаунти мають певну схожість, можливо обидва користувачі полюбляють саме цю гру : ['Backgammon', 'one', 'oldest', 'known', 'board', 'games']

Аналіз по соціальним групам своєю чергою це просто спрощений аналіз текстових повідомлень, оскільки для нього достатньо використати назви цих груп, або їх ID чи посилання у разі якщо система зберігає інформацію про це. Для прикладу візьмемо групи на які підписана людина яка полюбляє автомобілі та створимо з них масив: ['Love Cars', 'F1', 'BMW', 'Car builds'].

Далі ці дані їх можна привести до єдиного стандарту в залежності від поставлених задач, так, наприклад у разі пошуку користувачів які зацікавлені в автомобілях марки 'BMW' значення на векторі користувача який відстежує даний автомобіль буде визначене як 1, а інші марки автомобілів на проміжку [0, 1] в залежності від відповідності інших автомобілів марці 'BMW'.

В якості прикладу оберемо користувачів соціальної мережі Facebook які відповідають характеристикам вищеописаного масиву даних. В якості еталонного користувача було обрано такого, який відповідає усім критеріям масиву, тобто відстежує усі групи.

Для користувачів x_1-x_4 з характеристиками акаунтів y_1-y_4 можна створити матрицю можна заповнити наступним чином:

Таблиця 3.1 – Векторна матриця користувачів

Векторна матриця користувачів	y_1	y_2	y_3	y_4
x_1	1	1	1	1
x_2	1	0,64	0	0,5
x_3	0	0,3	0	0,6
x_4	1	0,8	1	0,3

В якості методу визначення схожості речень та слів для попередньої нормалізації можна скористатись методом TF-IDF, одного з фундаментальних методів вилучення релевантних даних з текстових документів. Абревіатура TF-IDF розшифровується як term frequency - inverse document frequency (частота слова - зворотня частота документа) і позначає метрику, яку можна використовувати для пошуку документів у корпусі шляхом обчислення нормалізованих оцінок, що виражають відносну важливість слів у документах.

Математично TF-IDF виражається як добуток частоти слова на обернену частоту документа:

$$tf\_idf = tf * idf, \tag{3.1}$$

де *tf* представляє важливість слова в конкретному документі, а *idf* — важливість слова для всього корпусу. tf_idf, надає оцінку, що враховує обидва фактори.

Частоту слова можна визначити просто як кількість разів, яке це слово зустрілося в тексті нормалізоване з загальною кількістю слів у тексті, щоб враховувалася довжина документа. Обернена частота документа надає узагальнену можливість нормалізації оцінок документів у корпусі для позбавлення їх від шумів, також відомих як стоп-слова які були розглянуті вище.

Natural Language Toolkit являє собою бібліотеку для мови програмування Python для роботи з даними людської мови. Її можна використати для попередньої нормалізації текстових даних користувачів.

За допомогою Natural Language Toolkit можна класифікувати, токенізувати, та проводити синтаксичний аналіз текстів, що дозволить оцінити схожість користувацьких даних і обрати найбільш відповідні значення в залежності від цілі подальшої кластеризації користувацьких даних в модулі кластеризації.

Наведемо декілька методів для оцінки схожості даних які реалізує Natural Language Toolkit:

- Відстань Левенштейна – це мінімальна кількість операцій вставок, видалень та замін, необхідна для перетворення одного рядка в інший. Наприклад, для перетворення dad на bad потрібна одна операція заміни (першої літери d на b), в результаті редакційна відстань між цими двома словами дорівнює 1.
- Подібність n-грам – це спосіб оцінки тексту через представлення його у вигляді послідовності всіх можливих груп з n лексем. Цей спосіб забезпечує основу для підрахунку словосполучень.
- Відстань Жаккара – це метрика для оцінки подібності текстів через представлення їх у вигляді множин. Для обчислення відстані Жаккара необхідно визначити унікальні елементи загальні для двох множин та розділити їх на загальне число унікальних елементів.

3.3 Розгляд існуючих бібліотек реалізуючих алгоритми кластерізації

scikit-learn це набір бібліотек мови Python який реалізує методи кластеризації. scikit-learn було створено в 2007 році як проект Google Summer of Code Девідом Курнапо, з того часу бібліотеки активно розвивали і сьогодні scikit-learn є одним з найпопулярніших наборів бібліотек для машинного навчання та кластерного аналізу великих даних. Оскільки кластерний аналіз є різновидом машинного навчання без вчителя, scikit-learn також містить в собі набір методів кластерного аналізу таких як k-means та DBSCAN (Рис. 3.2).

Geopy це набір бібліотек для аналізу географічних назв для подальшого визначення координат які можуть відповідати користувачам. Цю бібліотеку

доречно використовувати у разі кластеризації користувачів з врахуванням їх геолокації та подальшого використання цих даних для побудови картограм.

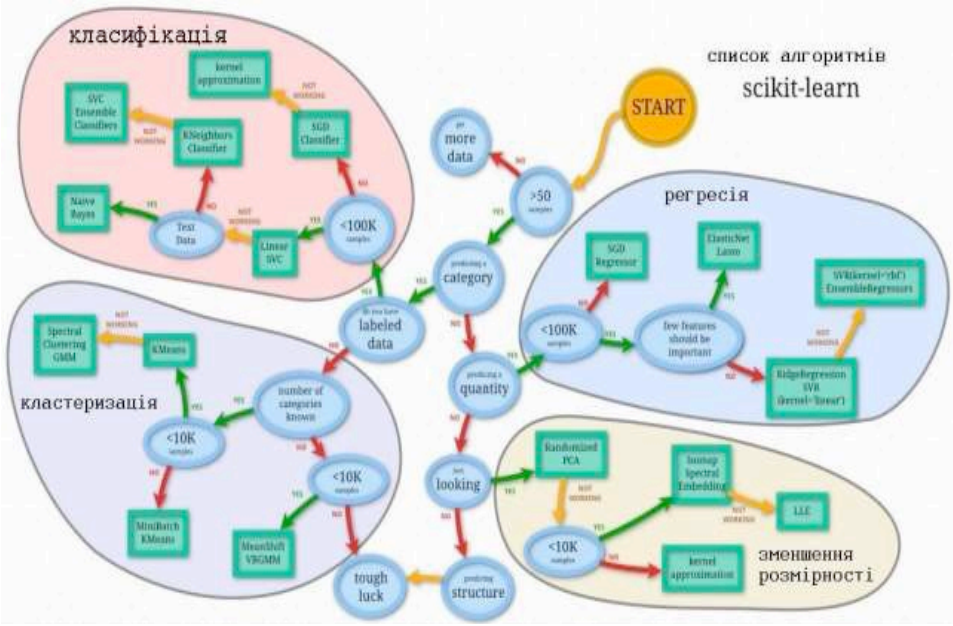


Рисунок 3.2 – Класифікація реалізованих методів бібліотеки scikit-learn

Matplotlib — це бібліотека Python для візуалізації даних. У ній можна побудувати двовимірні (плоскі) та тривимірні графіки. Використання цієї бібліотеки надасть можливість візуалізувати результат роботи кластерного аналізу.

3.4 Проектування програмного забезпечення

Кластеризація великих масивів даних зазвичай є довготривалим процесом, через це запит на кластеризацію необхідно робити асинхронним.

Асинхронні запити дозволяють користувачу використовувати програмне забезпечення після надсилання запиту навіть якщо запит знаходиться в обробці, оскільки процес відповідальний за обчислення даних не блокує інші процеси. Альтернативою таким запитам можуть бути синхронні запити, які блокують виконуються послідовно і блокують інші процеси, що призводить у тому числі до блокування користувацького інтерфейсу.

Після завершення кластеризації має сенс зберегти отримані результати, що

вимагає від модуля кластеризації зробити запит до бази даних.

Беручи до уваги необхідні вимоги для отримання, нормалізації, зберігання та кластеризації даних має сенс розділити програмне забезпечення на наступні модулі:

- Модуль отримання та зберігання даних. Містить у собі базу даних та інтерфейс для виконання CRUD (create, insert, update, delete) операцій.
- Модуль нормалізації даних. Містить методи для нормалізації вхідних даних та вихідних даних та надсилання їх до модуля кластеризації або модулю отримання та зберігання даних.
- Модуль кластеризації. Містить методи для кластеризації даних які надійшли з модуля нормалізації.
- Модуль користувацького інтерфейсу. Містить методи для взаємодії користувача та системи через графічний інтерфейс.

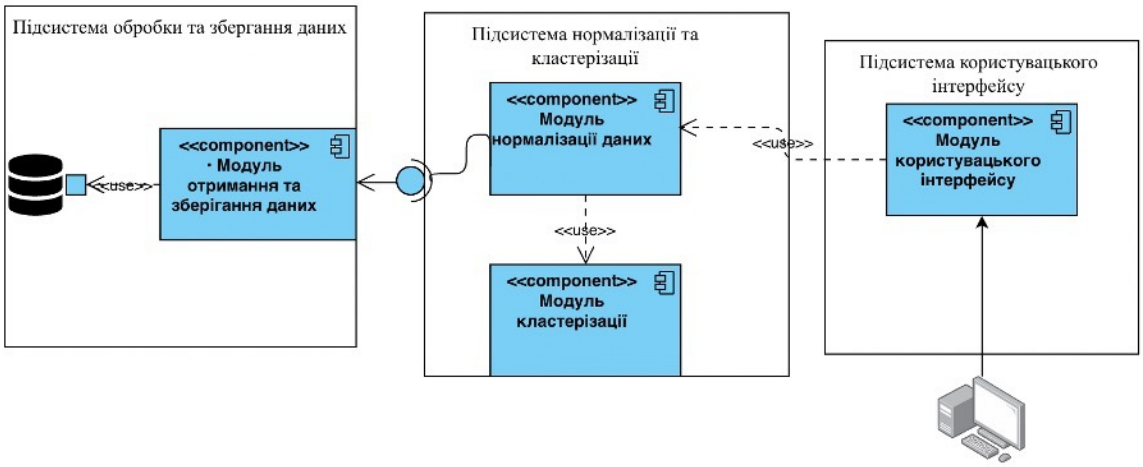


Рисунок 3.3 – Діаграма компонентів системи

Модуль користувацького інтерфейсу може являти собою як локально використовуваний модуль, тобто брати дані лише з наявної БД, так і може бути віддаленим. У локальному випадку користувачу буде достатньо прямого виклику до модуля нормалізації через інтерфейс, який своєю чергою виконає одну з необхідних команд.

У разі віддаленого виклику інтерфейс повинен надіслати підсистемі нормалізації та кластеризації HTTP запити, які той оброблятиме та відповідатиме. Обов’язково до HTTP повинно додаватись:

- IP адреса, з якого відбувається запит;
- тип запиту;
- дані запиту у форматі json або xlm.

Відповідь повинна виглядати наступним чином:

- відповідь на запит з кодом 200;
- у разі внутрішніх проблем надсилання повідомлення про помилку з кодом 400;
- у разі проблем з БД надсилання повідомлення про помилку з кодом 500.

Після отримання та нормалізації даних необхідно зберегти їх у певному сховищі даних. В залежності від розміру даних у якості сховищ даних можуть виступати реляційні або NOSQL бази даних.

Реляційна база даних зберігає значення у вигляді відношень. Візуально дані відображають у вигляді таблиць, але представлення про те що реляційні бази даних застосовують таблиці в якості структури даних є не вірним[1].

Елементи в реляційних базах представлені наступним чином[2]:

- Сутності – це певні об’єкти які цікавлять користувачів бази (клієнти, країни).
- Стовпці – це окрема частина даних яка зберігається в таблиці.
- Строки – це набір стовпців які разом повністю описують об’єкт.
- Таблиці – це набір строк.
- Результівний набір – це тимчасова таблиця яка зберігається в оперативній пам’яті і є результатом SQL-запиту.
- Первинний ключ – це один чи декілька стовпців які можна використовувати у якості унікального ідентифікатору для кожної строки.
- Зовнішній ключ – це один або декілька ключів які можна використовувати для ідентифікації строк іншої таблиці.

Реляційні бази даних дозволяють зберігати значні об’єми даних в постійній пам’яті, але через це швидкість отримання даних є повільнішою в порівнянні з NOSQL сховищами даних.

NOSQL бази даних зберігають дані в оперативній пам’яті комп’ютера, що дозволяє їм швидко отримувати та записувати їх, але разом з цим вони не мають змоги зберігати дані постійно та зберігати великі масиви даних.

Такі бази даних не мають одного типу представлення даних, та можуть буди розділені наступним чином:

- Бази даних документів – об’єднують кожен ключ зі складною структурою даних, яка називається документом. Документи можуть містити пари ключ-масив або пари ключ-значення або навіть вкладені документи.
- Ключ-значення – кожен окремий елемент зберігається як пара ключ-значення.
- Зберігання з широкими стовпцями – ці типи баз даних оптимізовані для запитів до великих наборів даних, і замість рядків вони зберігають стовпці даних разом.
- Графові – зберігають інформацію яку можна представити у виді графів, наприклад соціальні зв’язки.

Виходячи з діаграми компонентів системи, можна необхідно створити наступні модулі: графічного інтерфейсу, нормалізації, кластеризації та бази даних.

inter – модуль який відповідає за графічний інтерфейс користувача та відправляє дані завантажені користувачем до модуля norm. Реалізує наступні методи:

- button_clust/1 відповідає за кнопку надсилання даних та приймає в якості аргументу необроблені дані;
- button_load/1 відповідає за отримання даних та приймає в якості аргументу числове значення > 0;
- button_save/1 відповідає за зберігання даних та приймає в якості аргументу оброблені нормалізовані дані;
- image/1 відповідає за відображення кластеризованих даних у вигляді графіку та приймає у якості аргументу оброблені нормалізовані дані;

- table/1 відповідає за відображення кластеризованих даних у вигляді таблиці та приймає у якості аргументу оброблені нормалізовані дані.

norm – це модуль який відповідає за нормалізацію даних перед надсиланням до модуля кластеризації та до користувача. Також через цей модуль записуються дані в базу даних. Реалізує наступні методи:

- clust/1 модуль для зовнішнього виклику. Викликає методи norm.normalize/1, db.insert/1, norm.clust/1. Приймає не оброблені дані в якості аргументу;
- norm.normalize/1 нормалізує отримані дані. Приймає не оброблені дані у якості аргументу;
- clust_db/1 нормалізує кластеризовані дані для зручного відображення та зберігання в базі даних.

clust_mod – це модуль для проведення кластеризації даних відповідним методом. Має єдиний метод clust/1 який кластеризує нормалізовані дані та повертає їх.

DB – це модуль для спілкування з базою даних. Реалізує наступні методи:

- db.insert/1 робить запит до бази даних типу INSERT. В якості первинного ключа використовується ключ/стовпець (в залежності від типу бази даних) автоінкрементне числове поле id. Приймає нормалізовані дані в якості аргументу;
- db.insert/2 працює аналогічно db.insert/1, але зберігає дані після кластерізації. Приймає нормалізовані кластерізовані дані та іменовану константу after_clust якості аргументу;
- db.get_by_id/1 робить запит до бази даних типу GET.

В загальному вигляді система повинна працювати наступним чином:

- Користувач робить запит на обробку даних за допомогою інтерфейсу.
- Модуль інтерфейсу викликає модуль нормалізації та кластеризації.
- Модуль нормалізації нормалізує дані та відправляє до модуля кластеризації.
- Модуль кластеризації оброблює дані за допомогою методу mini batch k-means.

- Дані знов потрапляють до модуля нормалізації та зберігаються в БД.
- Кластеризовані дані відображаються на користувацькому інтерфейсі у разі успішного виконання.
- Користувач має змогу завантажити продивитись або завантажити інші кластеризовані дані.

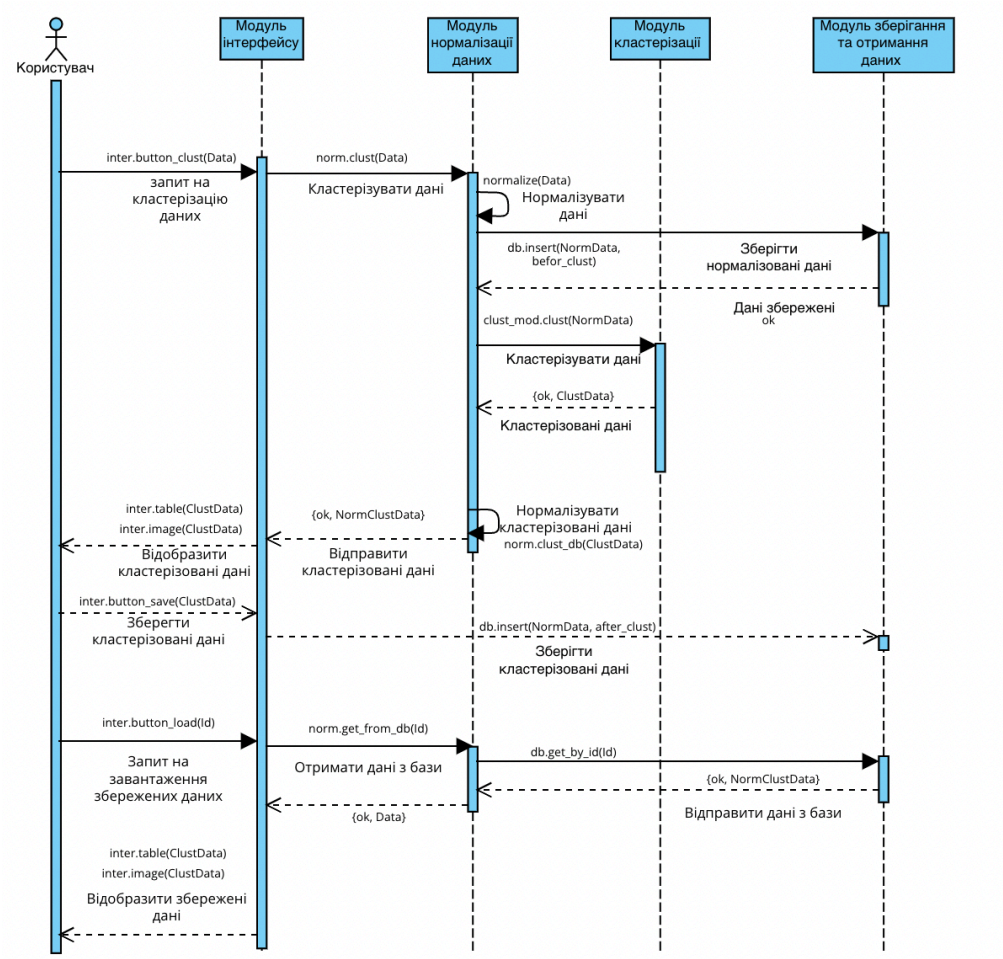


Рис. 3.4 Діаграма послідовності

3.5 Проектування користувацького інтерфейсу

Визначимо вимоги до користувацького інтерфейсу відповідно до потреб функціональних компонентів системи:

- Кнопка «опрацювати» для опрацювання акаунтів методом кластеризації.
- Кнопка «отримати» для отримання даних з БД.

- Кнопка «зберегти», для збереження опрацьованих даних користувачів.
- Панель для відображення графічних даних.
- Панель для відображення табличних даних.

В якості сервісу для проектування інтерфейсів було обрано Figma.

Figma – це графічний онлайн редактор для побудови користувацьких інтерфейсів. У ньому можна створити прототип сайту, інтерфейс програми та мобільного додатку, редагувати інтерфейс в реальному часі.

Після обрання необхідних початкових даних та натиснення кнопки «опрацювати» користувач має змогу продивитись графічний результат роботи та вибірку даних з якими вона проводилась.

Панелі відображення графічних та табличних даних виводять значення лише у разі вдалого завершення кластеризації або завантаження даних з БД, у разі помилки на панель відображення графічних даних виводиться повідомлення про відповідну помилку. Також кнопка «зберегти», активується для можливості збереження файлу результівної роботи.

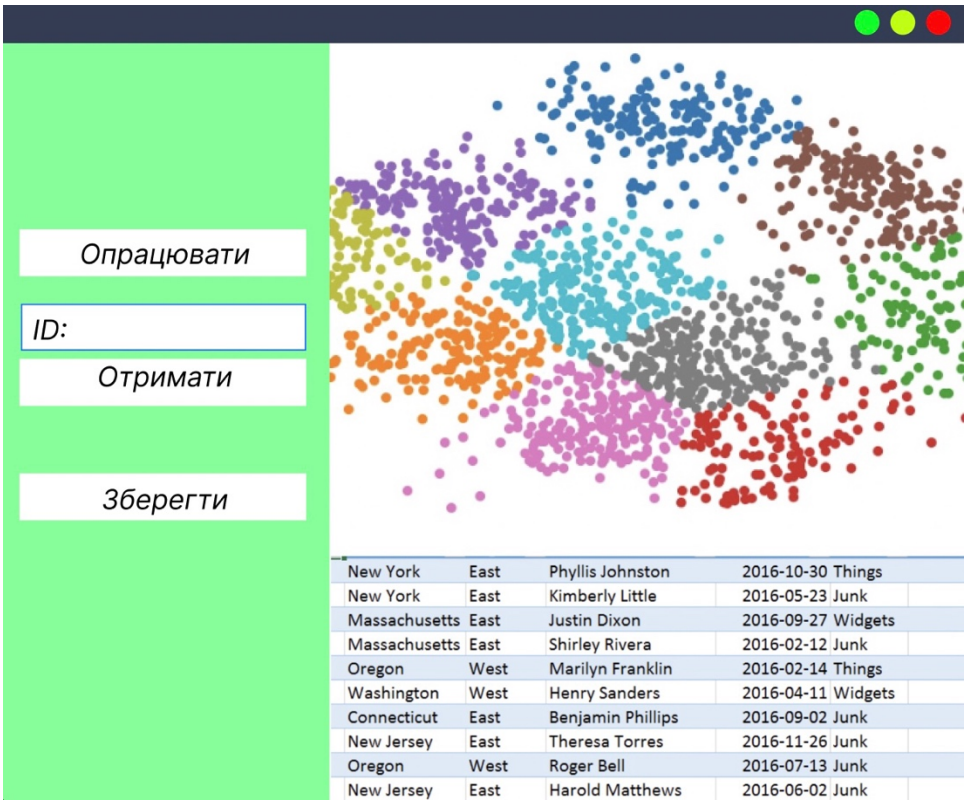


Рисунок. 3.5 Макет користувацького інтерфейсу

Висновок до розділу

Підсумовуючи третій розділ, можемо зробити такі висновки:

- Досліджено наявні бібліотеки для проведення кластеризації, нормалізації та обробки даних на мові програмування Python.
- Спроектовано модель програмного забезпечення та користувацький інтерфейс для обробки, та збереження даних для пошуку схожих акаунтів соціальних мереж за допомогою методу кластерного аналізу mini batch k-means розробленого у розділі 2.

ВИСНОВКИ

У даній роботі були проаналізовані існуючі підходи до побудови та використання кластеризації у соціальних мережах, як саме вони можуть бути використані для пошуку схожих акаунтів.

Було проведене дослідження наукових робіт присвячених використанню кластеризації у соціальних мережах та проаналізовані існуючі соціальні мережі. Визначено що через активне збільшення користувацьких даних у мережі інтернет тема обробки користувацьких даних методами кластеризації є актуальною, оскільки звичайні статистичні методи не здатні обчислити таку кількість інформації.

Проведено дослідження існуючих алгоритмів та методів кластеризації для пошуку схожих акаунтів, кластеризації таких як k-means, c-means, ієрархічна кластеризація та кластеризація на основі щільності. Також були досліджені методи нормалізації даних для подальшого перетворення у матрицю користувачів з векторів користувацьких характеристик.

Проведено порівняльний аналіз алгоритмів k-means та k-means mini batch. Встановлено що час виконання k-means mini batch значно менший для великої кількості вхідних даних.

Розроблено метод пошуку схожих акаунтів в соціальних мережах за допомогою алгоритму k-means mini batch.

Проведено аналіз існуючих бібліотек кластеризації та та нормалізації, досліджено API для отримання даних акаунтів з соціальних мереж, спроектовано програмне забезпечення яке реалізує метод пошуку схожих акаунтів в соціальних мережах.

ПЕРЕЛІК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Shazia Tabassum, João Gama. Social Network Analysis : An Overview. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 2018.
2. Stephan Curiskis. An evaluation of document clustering and topic modelling in two online social networks: Twitter and Reddit. Information Processing and Management 2019.
3. Mochammad Haldi Widianto, Ivan Diryana Sudirman, Muhammad Hanif Awaluddin. Application of Density Based Clustering of Disaster Location in Realtime Social Media. TEM Journal 2020. C.929-936.
4. Akshit Singh. Social media data analytics to improve supply chain management in food industries. Transportation Research Part E: Logistics and Transportation Review 2018. C.8-16.
5. Noufa Alnajran, Keeley Crockett, David McLean, Annabel Latham. Cluster Analysis of Twitter Data: A Review of Algorithms. Manchester Metropolitan University 2018. C.2-6.
6. David Camacho. The four dimensions of social network analysis: an overview of research methods, applications, and software tools. Universidad Rey Juan Carlo 2020. C.3-16.
7. Helpnetsecurity, Ongoing and initial costs top list of barriers to 5G implementation. [Електронний ресурс] - Режим доступу: https://www.helpnetsecurity.com/2020/09/18/barriers-5g-implementation/
8. Martin Hacks, The growth of data consumption is unsustainable. [Електронний ресурс] - Режим доступу: https://www.martinhacks.com/the-growth-of-data-consumption-is-unsustainable/
9. Victoria Bolotaeva. Marketing Opportunities with Social Networks. Northern Kentucky University.
10. Duong HanhTien. Examining the influence of customer-to-customer electronic word-of-mouth on purchase intention in social networking sites.

11. M. Faúndez-Abans. Classification of planetary nebulae by cluster analysis and artificial neural networks. Astron. Astrophys 1996.
12. Toshi A Furukawa. A polydiagnostic study of depressive disorders according to DSM-IV and 23 classical diagnostic systems. PSN 2002.
13. Songyin Deng. Clustering with Fuzzy C-means and Common Challenges. Journal of Physics: Conference Series 2020.
14. Bo Xiao. SMK-means: An Improved Mini Batch K-means Algorithm Based on Mapreduce with Big Data. Tech Science Press 2018.
15. Matthew A. Russell, Mikhail Klassen. Mining the Social Web 2019. C.169-188.
16. Brian S. Everitt, Sabine Landau, Morven Leese, Daniel Stahl. Cluster Analysis, 5th Edition. C.7-13.
17. Kristina Sinaga, Miin-Shen Hang. Unsupervised K-Means Clustering Algorithm. IEEE Access 2020. C.3-11.
18. Tshepo Chris Nokeri. Data Science Solutions with Python 2021. C.125-140.
19. Frank Kane. Hands-On Data Science and Python Machine Learning 2017. C.174-181.
20. Marc Peter Deisenroth, Aldo Faisal, Cheng Soon Ong. Mathematics for Machine Learning 2020. C.348-368.
21. Xiangjie Kong, Yajie Shi. Academic Social Networks: Modeling, Analysis, Mining and Applications 2019. C.6-9.
22. Zhongying Zhao. An Incremental Method to Detect Communities in Dynamic Evolving Social Networks 2018. C.2-4.
23. Yifu Zeng, Yantao Zhou, Xu Zhou, Fei Zheng. Fuzzy clustering-based skyline query preprocessing algorithm for large-scale flow data analysis. The Journal of Supercomputing 2020. C.1323-1328.
24. Sumalee Sangamuang. A Graph-Based Algorithm For Interpersonal Ties Clustering In Signed Networks. Tehnicki glasnik 2019. C.275-278.
25. Carlo Perrotta, Ben Williamson. The social life of Learning Analytics: cluster analysis and the ‘performance’ of algorithmic education. Learning, media and technology 2018. C.9-13.

Додаток

ДЕРЖАВНИЙ УНІВЕРСИТЕТ ТЕЛЕКОМУНІКАЦІЙ

НАВЧАЛЬНО–НАУКОВИЙ ІНСТИТУТ ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ

Кафедра інженерії програмного забезпечення

МАГІСТЕРСЬКА РОБОТА

«Розробка методу пошуку схожих акаунтів в соціальних мережах на основі кластерного аналізу»

Виконав: студент групи ПДМ-61 Дібрій Данило Андрійович

Керівник: доцент кафедри ІПЗ Аверічев І.М

Київ - 2022

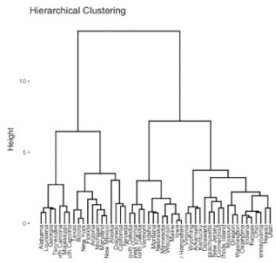
МЕТА, ОБ'ЄКТ, ПРЕДМЕТ ДОСЛІДЖЕННЯ

Мета роботи: покращення пошуку груп схожих користувачів соціальних мереж за допомогою розробленого методу на основі кластерного аналізу.

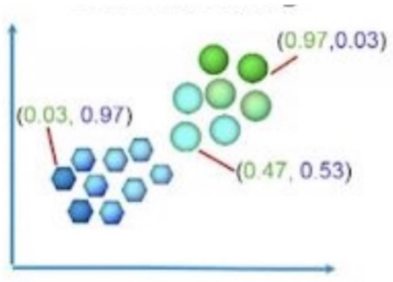
Об'єкт дослідження: процес проектування методу для пошуку схожих акаунтів в соціальних мережах на основі кластерного аналізу.

Предмет дослідження: методи та алгоритми пошуку схожих акаунтів в соціальних мережах.

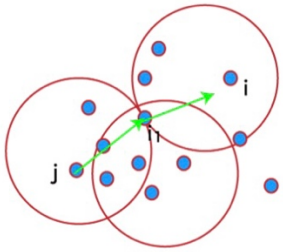
АНАЛІЗ ІСНУЮЧИХ ІТ-РІШЕНЬ ТА ЇХ МОДЕЛЕЙ



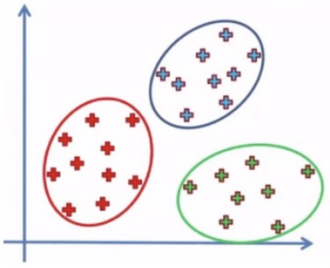
Ієрархічна кластеризація – кластеризація за допомогою дендрограм



c-means – кластеризація відносить об’єкти до кластерів з певною вірогідністю

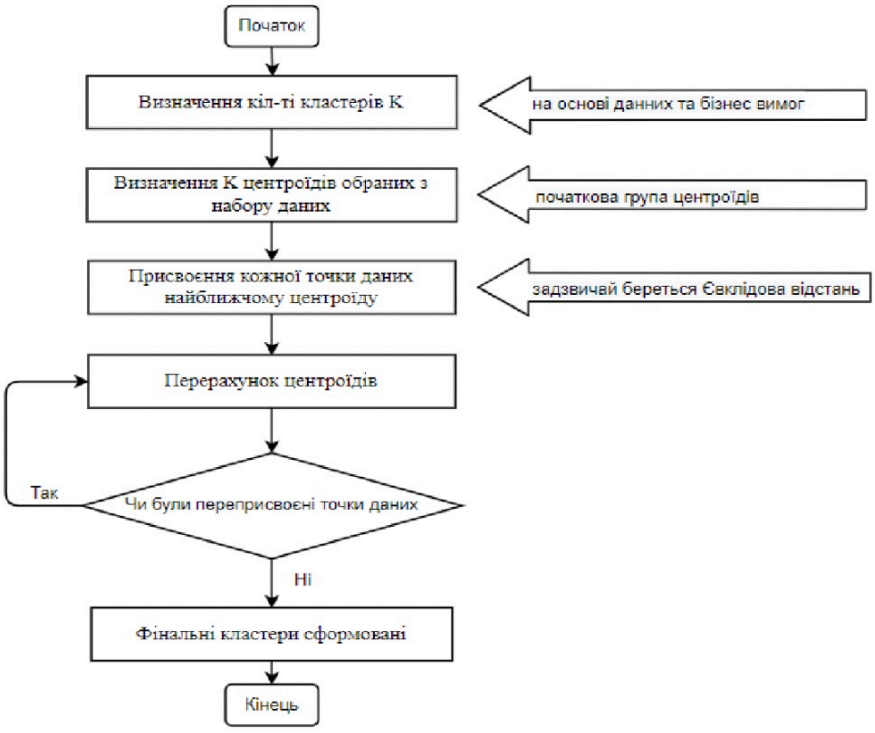


Кластеризація на основі щільності – враховує щільність розташування об’єктів



k-means – кластеризація випадково обирає початкові центри та визначає найближчі об’єкти

КЛАСТЕРИЗАЦІЯ K-MEANS



МЕТОДИ ОЦІНКИ СХОЖОСТІ АКАУНТІВ

Формула відстані Евкліда:

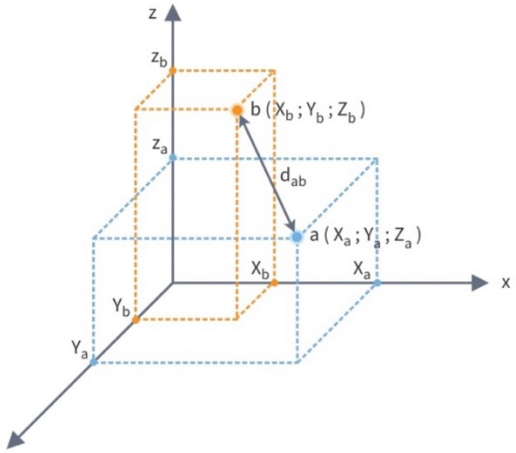
$$d_{pq} = \sqrt{\sum_{i=1}^{n}(p_i - p_q)^2},$$

де p_i та p_q це точки в n-вимірному просторі.

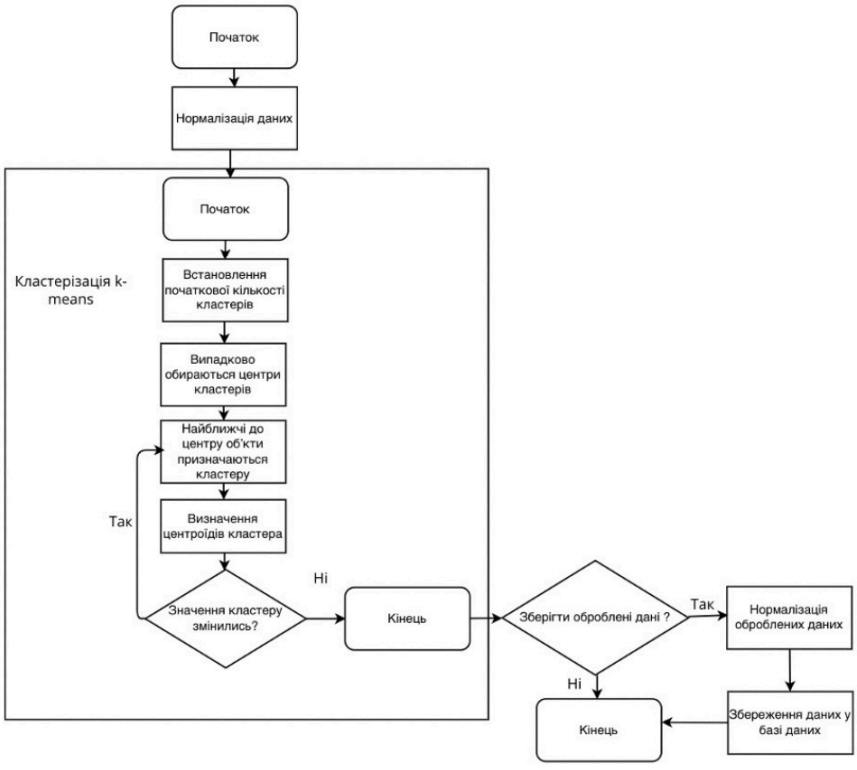
При великому значенні n ця формула може надавати недостатньо точні значення, у цьому разі може бути використана формула відома як відстань Махаланобіса:

$$d_M(x_i, x_j) = (x_i - x_j)F^{-1}(x_i - x_j)^T ,$$

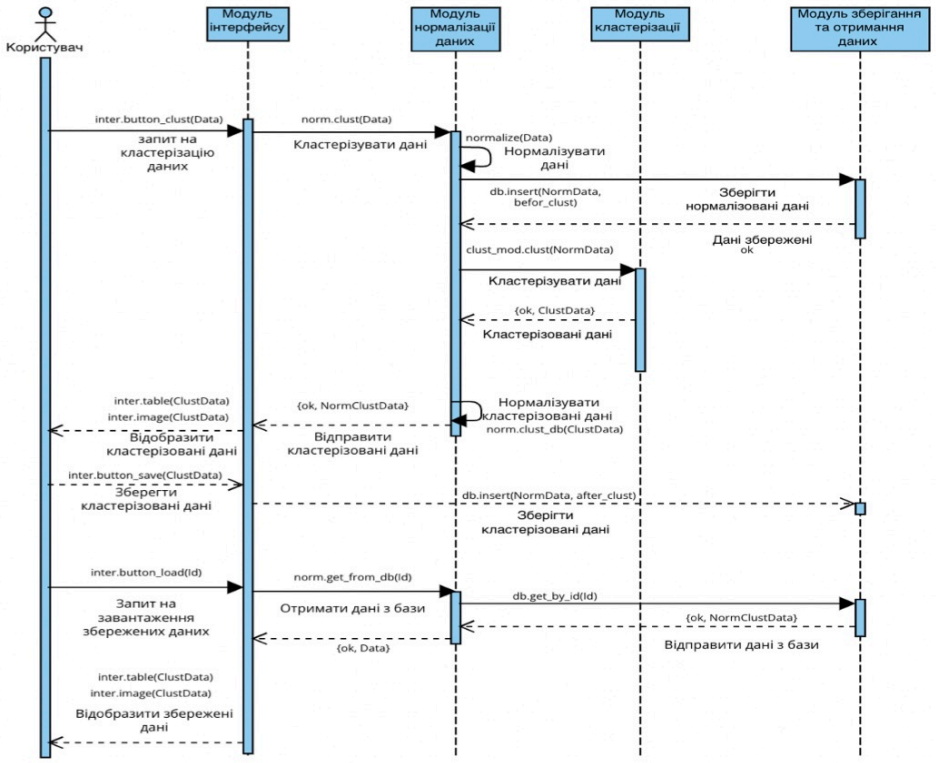
Де x_i, x_j це вектори характеристик, F це матриця коваріації, у разі якщо ця матриця одинична формула зводиться то формули відстані Евкліда.



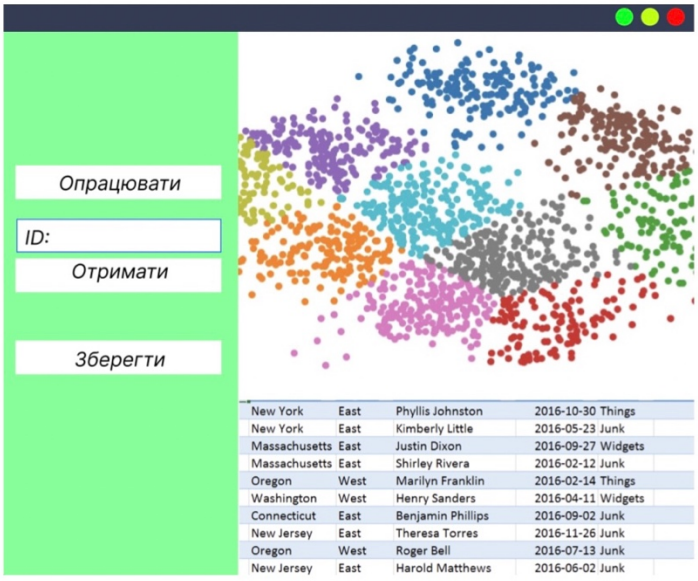
МЕТОД ПОШУКУ СХОЖИХ АКАУНТІВ В СОЦІАЛЬНИХ МЕРЕЖАХ



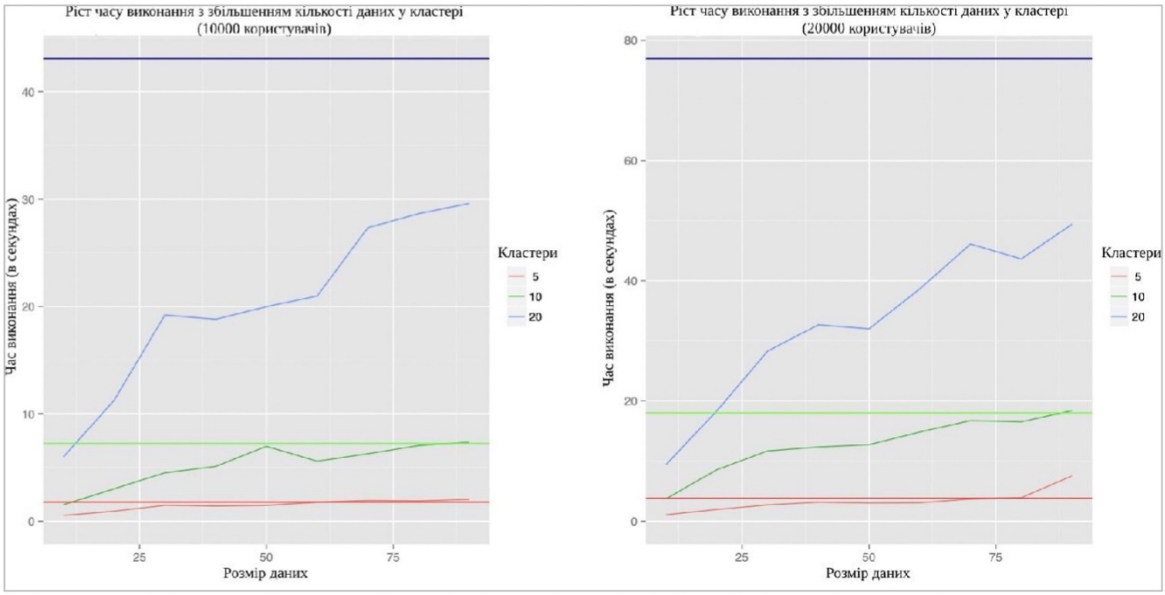
ДІАГРАМА ПОСЛІДОВНОСТІ



ІНТЕРФЕЙС ПРОГРАМИ ДЛЯ КЛАСТЕРИЗАЦІЇ СХОЖИХ АКАУНТІВ



ПОРІВНЯЛЬНИЙ АНАЛІЗ K-MEANS ТА K-MEANS MINI BATCH



ВИСНОВКИ

1 Проаналізовано існуючі підходи до побудови та використання кластеризації у соціальних мережах.

2. Проведено дослідження існуючих алгоритмів та методів кластеризації для пошуку схожих акаунтів.

3. Проведено порівняльний аналіз алгоритмів k-means та k-means mini batch. Встановлено що час виконання k-means mini batch значно менший для великої кількості вхідних даних.

4. Розроблено метод пошуку схожих акаунтів в соціальних мережах за допомогою алгоритму k-means mini batch.

5. Проведено аналіз існуючих бібліотек кластеризації та спроектовано програмне забезпечення яке реалізує метод пошуку схожих акаунтів в соціальних мережах.

ПУБЛІКАЦІЇ ТА АПРОБАЦІЯ РОБОТИ

Тези доповідей:

1. Аверічев І.М., Дібрій Д.А. Розробка методу пошуку схожих акаунтів в соціальних мережах на основі кластерного аналізу// XV Науково-технічна конференція «Сучасні інфокомунікаційні технології» – Київ: ДУТ, 2022.
2. Аверічев І.М., Дібрій Д.А. Використання кластеризації для пошуку схожих акаунтів в соціальних мережах // Науково-технічна конференція «Проблеми комп'ютерної інженерії» – Київ: ДУТ, 2022.

ДЯКУЮ ЗА УВАГУ!