Хмельницький національний університет
Факультет інформаційних технологій
Кафедра комп'ютерних наук

КВАЛІФІКАЦІЙНА РОБОТА БАКАЛАВРА

на тему *Спосіб формування рекомендацій поліграфічних товарів за допомогою методу колаборативної фільтрації для електронної комерції*

Галузь знань *12 – Інформаційні технології*
Шифр і назва галузі знань

Спеціальність *122 – Комп'ютерні науки*
Шифр і назва спеціальності

Освітня програма *Комп'ютерні науки*
Назва освітньої програми

Виконав: *студент 4 курсу, група КН-19-1* — Курс, група виконавця — Підпис — *М.Б. Машталяр* — Ініціали, прізвище

Керівник: *к.т.н., доцент кафедри КН* — Науковий ступінь, посада — Підпис — *Р.О. Багрій* — Ініціали, прізвище

Нормоконтроль: *к.т.н., доцент кафедри КН* — Науковий ступінь, посада — Підпис — *Р.О. Багрій* — Ініціали, прізвище

До захисту допускаю:
Зав. кафедри КН, д.т.н., професор — Підпис — *О.В. Бармак* — Ініціали, прізвище
05 06 2023 р.

Хмельницький 2023

ХМЕЛЬНИЦЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
Факультет *інформаційних технологій*
Кафедра *комп'ютерних наук*
Освітній ступінь *бакалавр*
Галузь знань *12 – Інформаційні технології*
Спеціальність *122 – Комп'ютерні науки*
Освітня програма *освітньо-професійна програма підготовки бакалавра*

ЗАТВЕРДЖУЮ
Завідувач кафедри *комп'ютерних наук*

(підпис)
д.т.н., професор О.В. Бармак
«*06*» *03* *2023* року

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ БАКАЛАВРА

1. Тема кваліфікаційної роботи бакалавра: *«Спосіб формування рекомендацій поліграфічних товарів за допомогою методу колаборативної фільтрації для електронної комерції»*

2. Завдання видано студенту *Машталяру Матвію Богдановичу*
(прізвище, ім'я, по батькові)

3. Керівник роботи *доцент кафедри КН Багрій Руслан Олександрович*
(посада, прізвище, ім'я, по батькові)

4. Затверджено наказом університету від «*01*» *03* *2023* р. № *5*

5. Дата видачі завдання студенту: «*03*» *03* *2023* р.

6. Зміст пояснювальної записки (перелік задач) та вихідні дані:

Провести аналіз предметної області, огляд методів побудови рекомендаційних систем, існуючих програмних рішень та сформувати постановку задачі. Розробити спосіб формування рекомендацій поліграфічних товарів за допомогою методу колаборативної фільтрації. Провести тестування та верифікацію розробленого способу формування рекомендацій поліграфічних товарів. Вихідними даними є інформація про оцінки користувачів з реальних веб-систем предметної області.

7. Календарний план виконання кваліфікаційної роботи бакалавра:

№	Назва етапів (розділів) кваліфікаційної роботи бакалавра	Термін виконання	Примітка
1	*Вибір напряму дослідження та узгодження тематики кваліфікаційної роботи бакалавра з керівником*	*грудень 2022*	*виконано*
2	*Ознайомлення з предметною областю, формулювання мети та задач дослідження, визначення об'єкта та предмета дослідження*	*січень 2023*	*виконано*
3	*Робота над розділом 1 – Характеристика предметної області та постановка задачі*	*січень 2023*	*виконано*
4	*Робота над розділом 2 – Спосіб формування рекомендацій пропозицій товарів для інформаційної системи електронної комерції*	*березень 2023*	*виконано*
5	*Робота над розділом 3 – Програмна реалізація інформаційної системи з використанням запропонованого способу формування рекомендацій*	*квітень 2023*	*виконано*
6	*Оформлення пояснювальної записки згідно вимог*	*травень 2023*	*виконано*
7	*Підготовка статті до журналу, попередній захист кваліфікаційної роботи бакалавра*	*травень 2023*	*виконано*
8	*Захист кваліфікаційної роботи бакалавра на засіданні Екзаменаційної комісії*	*червень 2023*	

Виконавець: *студент 4 курсу, група КН-19-1* (Курс, група виконавця) _____ (Підпис) *М.Б. Машталяр* (Ініціали, прізвище)

Керівник: *к.т.н., доцент кафедри КН* (Науковий ступінь, посада) _____ (Підпис) *Р.О. Багрій* (Ініціали, прізвище)

Анотація

Тема кваліфікаційної роботи бакалавра: *Спосіб формування рекомендацій поліграфічних товарів за допомогою методу колаборативної фільтрації для електронної комерції*

Виконавець кваліфікаційної роботи бакалавра: *студент групи КН-19-1 Машталяр Матвій Богданович*

Керівник кваліфікаційної роботи бакалавра: *к.т.н., доцент кафедри КН Багрій Руслан Олександрович*

Кваліфікаційна робота бакалавра містить:

Пояснювальна записка				Кількість додатків
Сторінок	Рисунків	Таблиць	Джерел інформації	
60	30	2	25	2

Метою кваліфікаційної роботи бакалавра є розробка способу формування рекомендації пропозицій товарів поліграфічної продукції з використанням методу колаборативної фільтрації, що може бути використано у електронної комерції.

Результатом виконання кваліфікаційної роботи бакалавра є реалізація способу формування рекомендацій пропозицій товарів для використання у інформаційних системах електронної комерції поліграфічної продукції.

Ключові слова: колаборативна фільтрація, рекомендаційна система, пропозиції товарів.

Виконавець: *студент 4 курсу, група КН-19-1* (Курс, група виконавця) _____ (Підпис) *М.Б. Машталяр* (Ініціали, прізвище)

Зміст

Перелік скорочень

Скорочення, термін, позначення	Пояснення
КРБ	Кваліфікаційна робота бакалавра
КН	Комп’ютерні науки

Вступ

Кваліфікаційна робота бакалавра присвячена розробці способу формування рекомендацій поліграфічних товарів за допомогою методу колаборативної фільтрації для електронної комерції.

Актуальність теми.

Об'єм даних в Інтернеті з кожним роком невпинно зростає і знаходження потрібної інформації займає все більше часу. В багатьох сферах, в тому числі в електронній комерції, назріла необхідність використання ефективних способів фільтрації інформації. Колаборативна фільтрація є одним із можливих методів, що дозволяє сформувати перелік рекомендацій товарів для веб-систем електронної комерції.

Колаборативна фільтрація – це метод побудови прогнозів (рекомендацій) у рекомендаційних системах, який використовує відомі інтереси (вподобання) групи користувачів для прогнозування невідомих інтересів іншого користувача.

Мета кваліфікаційної роботи бакалавра полягає у розробці способу формування рекомендації пропозицій товарів поліграфічної продукції з використанням методу колаборативної фільтрації для веб-систем електронної комерції.

Об’єкт дослідження – процес формування рекомендації з використанням методу колаборативної фільтрації.

Предмет дослідження – методи збору та аналізу інформації, методи оцінки точності прогнозних моделей.

Завдання кваліфікаційної роботи бакалавра – отримати набори даних про оцінки користувачів поліграфічної продукції з відкритих джерел; розробити спосіб формування рекомендацій пропозицій товарів з використанням методу колаборативної фільтрації; реалізувати спосіб формування рекомендацій пропозицій товарів для використання у інформаційних системах електронної комерції поліграфічної продукції; провести тестування та верифікацію розробленого способу формування рекомендацій поліграфічних товарів.

Розділ 1 Характеристика предметної області та постановка задачі

1.1 Аналіз предметної області

Поліграфія – це сукупність діяльності та технологій, що пов'язані з друкуванням та видавництвом різноманітних матеріалів на папері або інших матеріалах. Чотири основні види поліграфії включають представницьку, рекламну, книжково-журнальну та календарну.

Представницька поліграфія включає в себе друк різних матеріалів, таких як візитні картки (рисунок 1.1), листівки, запрошення, обкладинки для документів тощо. Ці матеріали призначені для представлення компанії або бренду, розповіді про їхній діяльності та послуги, а також для підвищення їхнього професійного іміджу.

Рисунок 1.1 – Приклад продукту представницької поліграфії (візитні картки)

Рекламна поліграфія включає в себе створення та друк рекламних матеріалів, таких як брошури, листівки, плакати, банери (рисунок 1.2) та інші матеріали. Рекламна поліграфія використовується для просування товарів і послуг, підвищення уваги клієнтів та залучення нових покупців.

Рисунок 1.2 – Приклад продукту рекламної поліграфії (банер)

Книжково-журнальна поліграфія включає в себе друк книг (рисунок 1.3), журналів, буклетів та інших видань. Ці матеріали призначені для розповсюдження інформації, навчання, розваг та іншого. Книжково-журнальна поліграфія використовується в освітніх, наукових, культурних та інших сферах.

Рисунок 1.3 – Приклад продукту книжково-журнальної поліграфії (книги)

Календарна поліграфія включає в себе друк календарів (рисунок 1.4), які призначені для використання в побуті та в бізнесі. Календарі можуть бути різних типів, таких як настільні, стінові, кишенькові тощо. Вони можуть бути розроблені як з метою реклами певного бренду або компанії, так і з метою передачі корисної інформації, такої як дати свят та відпусток, фази місяця, національні свята та інші події.

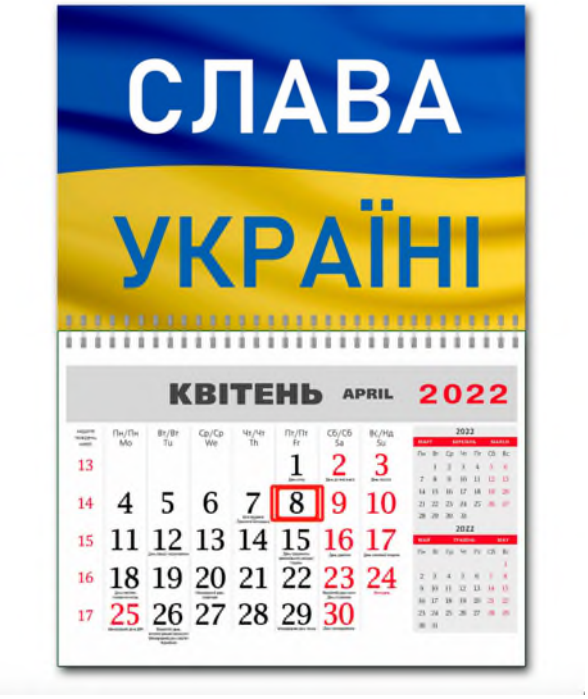


Рисунок 1.4 – Продукт календарної поліграфії (календар)

Усі види поліграфії вимагають відповідних технологій та екіпування для їх виготовлення. Це може включати в себе різноманітні типи друкарських машин (рисунок 1.5), папери, фарби та інші матеріали. Важливою складовою процесу поліграфії є дизайн та верстка матеріалів, що передують друкуванню, тому що це дозволяє виготовляти матеріали, що відповідають вимогам клієнтів та споживачів.

Поліграфічна продукція має велике значення для бізнесу, культури та інших сфер життя людей, тому що вона є ефективним засобом комунікації та реклами. Професіонали у галузі поліграфії використовують свої знання та

досвід, щоб створювати якісну та ефективну продукцію, яка задовольняє потреби різних клієнтів та споживачів.

Рисунок 1.5 – цифрова офсетна друкарська машина

Рекомендації пропозицій товарів можуть бути потрібні для таких задач, як:

1. Продаж товару або послуги.
2. Рекламування товару або послуги.
3. Підвищення свідомості про товар або послугу серед споживачів.
4. Залучення нових клієнтів до компанії або бренду.
5. Підтримання лояльності клієнтів до компанії або бренду.

У поліграфічній частині цих задач особливу увагу звертають на дизайн та виготовлення матеріалів, які будуть використовуватися для рекламування товарів або послуг. Наприклад, дизайнери можуть розробляти логотипи, брошури, листівки, плакати, каталоги, пакування тощо. Також, важливим аспектом є вибір правильних матеріалів, фарб, типів друкарських машин, які відповідатимуть вимогам клієнтів та забезпечать високу якість продукції.

При автоматизації поліграфічних задач важливо враховувати такі параметри:

1. Формат виробу, який має бути виготовлений.
2. Тираж, тобто кількість виробів, які потрібно виготовити.
3. Вид паперу, який буде використовуватися.
4. Кількість кольорів, які мають бути використані.
5. Тип друкарської машини, яка буде використана для друку.
6. Терміни виконання замовлення.

При автоматизації поліграфічних задач використовують спеціальне програмне забезпечення, яке дозволяє оптимізувати процес виготовлення продукції та зменшити ризик помилок. Також важливим є контроль якості, який дозволяє виявляти та виправляти помилки під час виробництва та забезпечувати високу якість продукції. Контроль якості може включати в себе перевірку якості вхідних матеріалів, контроль якості виробничого процесу та остаточної продукції. Для цього можуть використовуватися різні методи контролю якості, такі як візуальна оцінка, вимірювання розмірів та параметрів виробу, аналіз якості друку та кольору, перевірка на відповідність стандартам якості та інші.

Крім того, важливим етапом виробництва поліграфічної продукції є підготовка до друку. Цей етап включає в себе створення макету, підготовку файлу для друку, підготовку друкарської форми та налаштування друкарської машини. При автоматизації цих процесів важливими параметрами є точність та швидкість підготовки файлу до друку, відповідність кольору та деталей макету підготовці друкарської форми та налаштуванні друкарської машини. Загалом, виробництво поліграфічної продукції є складним та багатоетапним процесом, який вимагає високої кваліфікації працівників та використання сучасного обладнання та технологій. Втім, завдяки автоматизації та використанню стандартів якості та безпеки, можна забезпечити високу якість продукції та задовольнення потреб клієнтів.

1.2 Огляд методів побудови рекомендаційних систем

Рекомендаційні системи – це програмні комплекси, які забезпечують автоматичне рекомендування користувачам певних товарів, послуг або інформації на основі їхніх попередніх дій або інших даних. Рекомендаційні системи можуть бути розроблені для різних цілей, таких як електронна комерція, медіа, соціальні мережі, музика та інші.

Основні типи рекомендаційних систем:

1. Контентно-орієнтовані системи – використовують аналіз контенту, що був спожитий користувачем, щоб рекомендувати подібний контент. Наприклад, якщо користувач переглядає фільми жахів, система може запропонувати подібні фільми жанру жахів.

2. Колаборативні системи – використовують інформацію про попередні дії користувачів для знаходження спільних інтересів та рекомендації подібних товарів або послуг. Наприклад, якщо двоє користувачів купили схожі товари в інтернет-магазині, система може запропонувати одному з них придбати подібний товар.

3. Гібридні системи – комбінують методи колаборативного та контентно-орієнтованого підходів для забезпечення більш точних рекомендацій.

4. Демографічні системи – використовують інформацію про вік, стать, рівень освіти, заробітну плату тощо, щоб зробити рекомендації, які відповідають індивідуальним потребам та інтересам користувача.

Рекомендаційні системи можуть мати різні алгоритми, які використовуються для аналізу та обробки даних, такі як факторизація матриці, байєсовські методи, методи кластеризації, нейронні мережі та інші. Крім того, рекомендаційні системи можуть використовувати різні типи даних для аналізу та рекомендацій, такі як історії покупок, переглянутих веб-сторінок, соціальних мереж, відгуки та рейтинги товарів, демографічні дані користувачів та інші.

Рекомендаційні системи можуть бути використані для різних завдань, наприклад:

1. Підвищення продажів в інтернет-магазинах, рекомендація товарів, які можуть зацікавити користувача.

2. Рекомендації в соціальних мережах, наприклад, відображення друзів та груп, які можуть зацікавити користувача.

3. Рекомендації відео та музичного контенту на платформах стрімінгу, які відповідають інтересам та смакам користувачів.

4. Рекомендації статей, книг та інших інформаційних ресурсів на платформах онлайн-медіа.

У поліграфічній галузі рекомендаційні системи можуть використовуватися для рекомендацій друкованих продуктів, таких як книги, журнали, газети, а також для рекомендацій рекламних продуктів, наприклад, відображення рекламних банерів, які відповідають інтересам користувача.

При автоматизації поліграфічних задач, інформація про продукти може бути подана у вигляді характеристик, таких як формат, тип паперу, кількість сторінок, кількість кольорів, тип друку та інші параметри, які використовуються для автоматичної генерації замовлень та виготовлення друкованих продуктів.

Рекомендаційні системи для поліграфічних товарів можуть бути реалізовані на основі різних алгоритмів та методів, що використовуються для аналізу та обробки даних. Основними типами рекомендаційних систем для поліграфічних товарів є:

1. Контент-базована рекомендаційна система – використовує характеристики та описи товарів для визначення рекомендованих продуктів. Наприклад, якщо користувач шукає книгу певного автора, рекомендаційна система може запропонувати інші книги того ж автора, або книги з подібною тематикою.

2. Спільні фільтри – використовують інформацію про покупки користувачів для рекомендацій товарів. Наприклад, якщо користувач купує

книги про психологію, рекомендаційна система може запропонувати інші книги з цієї тематики.

3. Колаборативна фільтрація – використовує інформацію про інтереси та дії користувачів для визначення рекомендованих продуктів. Наприклад, якщо два користувачі купують книги про історію, рекомендаційна система може запропонувати інші книги з історії цим користувачам.

4. Рекомендаційні системи на основі інтерактивної реклами – використовуються для відображення рекламних банерів та інших рекламних матеріалів, які відповідають інтересам користувача. Наприклад, якщо користувач часто шукає книги про психологію, рекомендаційна система може відображати рекламні банери книг з цієї тематики.

У поліграфічній галузі рекомендаційні системи можуть використовуватися для рекомендацій таких поліграфічних товарів, як книги, журнали, паперові вироби, рекламні матеріали, календарі та інше. Наприклад, рекомендаційні системи можуть пропонувати користувачам певні книги на основі їхніх попередніх виборів, рекомендувати нові книги від авторів, які сподобалися користувачеві, або запропонувати товари зі спільною тематикою.

Поліграфічна частина задач, пов'язаних з рекомендаційними системами, включає підготовку та обробку даних про товари, такі як описи, характеристики, фотографії та інші матеріали. Для книжок та інших друкованих матеріалів це можуть бути метадані, такі як ISBN, автор, видавництво, рік видання тощо. Для рекламних матеріалів можуть використовуватися характеристики, такі як розмір, матеріал, тип друку тощо.

1.3 Метод колаборативної фільтрації

Колаборативна фільтрація – це метод рекомендаційних систем, який базується на зіставленні вибірок користувачів та їхніх відгуків про товари. За допомогою цього методу, система може рекомендувати користувачам товари, які сподобалися іншим користувачам зі схожими інтересами. [1]

Існує кілька методів колаборативної фільтрації, серед яких найбільш популярні:

Метод User-Based Collaborative Filtering – заснований на схожості між користувачами. Система знаходить користувачів, які мають подібні відгуки про товари, і рекомендує нові товари на основі вибірок цих користувачів.

Метод Item-Based Collaborative Filtering – заснований на схожості між товарами. Система знаходить схожі товари на основі відгуків користувачів та рекомендує нові товари на основі відгуків на схожі товари. [2]

На перший погляд може здатись що методи User-Based Collaborative Filtering та Item-Based Collaborative Filtering виконують майже однакову функцію. Вони використовують інформацію про користувачів та предмети, щоб зробити рекомендації. Обидва методи можуть: використовувати подібність, наприклад косинусну подібність, щоб визначити, наскільки близькі користувачі або предмети; страждати від "проблеми холодного старту", коли система не має достатньої інформації про нових користувачів або нові предмети. Але, попри схожість цих методів, в них існують відмінності, наприклад:

– User-Based Collaborative Filtering використовує історію дій користувачів, тоді як Item-Based Collaborative Filtering використовує характеристики предметів;

– User-Based Collaborative Filtering може стати менш ефективним, якщо немає достатньої кількості користувачів, які взаємодіють з однаковими предметами. Item-Based Collaborative Filtering може стати менш ефективним, якщо предмети занадто різні один від одного;

– User-Based Collaborative Filtering може бути повільним, якщо мається велика кількість користувачів або предметів, оскільки потрібно порівняти кожного користувача з кожним. Item-Based Collaborative Filtering може бути швидким, оскільки порівнюється характеристики предметів, а не користувачів, але може бути менш точним, якщо два предмети занадто відмінні один від одного.

Стає очевидно, що обидва методи мають свої переваги та недоліки, і найкращий вибір залежить від конкретної ситуації та наявної інформації.

Метод Matrix Factorization (рисунок 1.6) – заснований на зменшенні розмірності вибірок користувачів та товарів, щоб знайти схожі відгуки та зробити прогнози на майбутнє. [3]

Одною з основних переваг методу MF полягає в тому, що він може працювати з дуже розрідженими матрицями, де багато елементів не має значення. Крім того, MF може вирішувати проблему "проблеми холодного старту", коли немає достатньої інформації про нових користувачів або нові предмети. Наприклад, коли користувач зареєструвався на сайті, але ще не переглядав жодного предмету.

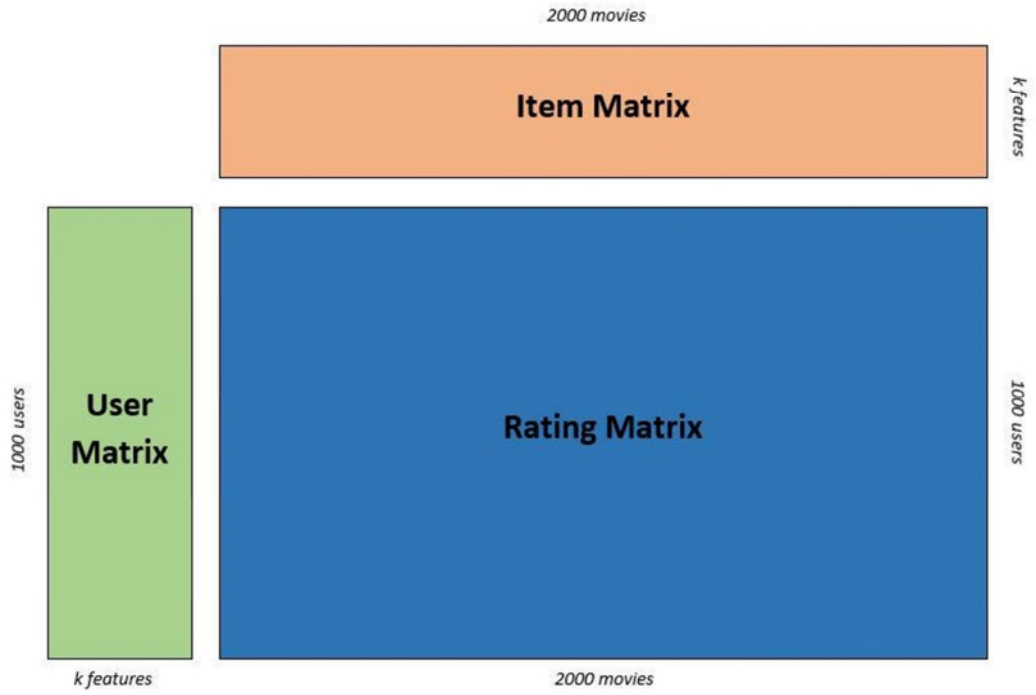


Рисунок 1.6 – візуалізація методу Matrix Factorization

Однак, MF може мати певні обмеження. Наприклад, якщо рейтинги мають шум або аномалії, це може призвести до неточних рекомендацій. Також може виникнути проблема перенавчання, коли MF дуже точно вивчає вже наявні рейтинги, але не може зробити вірні прогнози для нових користувачів або предметів.

Для розв'язання цих проблем, можуть бути використані різні модифікації методу MF, наприклад, додавання регуляризації, що дозволяє контролювати складність моделі, або використання альтернативних методів оптимізації.

Метод Hybrid Collaborative Filtering – комбінує два або більше з вищеназваних методів для отримання більш точних рекомендацій. [4] У порівнянні з іншими методами рекомендацій, гібридна колаборативна фільтрація може бути ефективнішою, оскільки вона дозволяє поєднувати переваги різних методів та компенсувати їхні недоліки. Наприклад, гібридна колаборативна фільтрація може поєднувати User-Based Collaborative Filtering та Item-Based Collaborative Filtering (рисунок 1.7), що дозволяє враховувати як взаємодію між користувачами, так і характеристики предметів. Крім того, до складу гібридної системи можуть входити інші методи, такі як Content-Based Filtering або Demographic-Based Filtering, які доповнюють рекомендації на основі інформації про контент або демографічні характеристики користувачів.

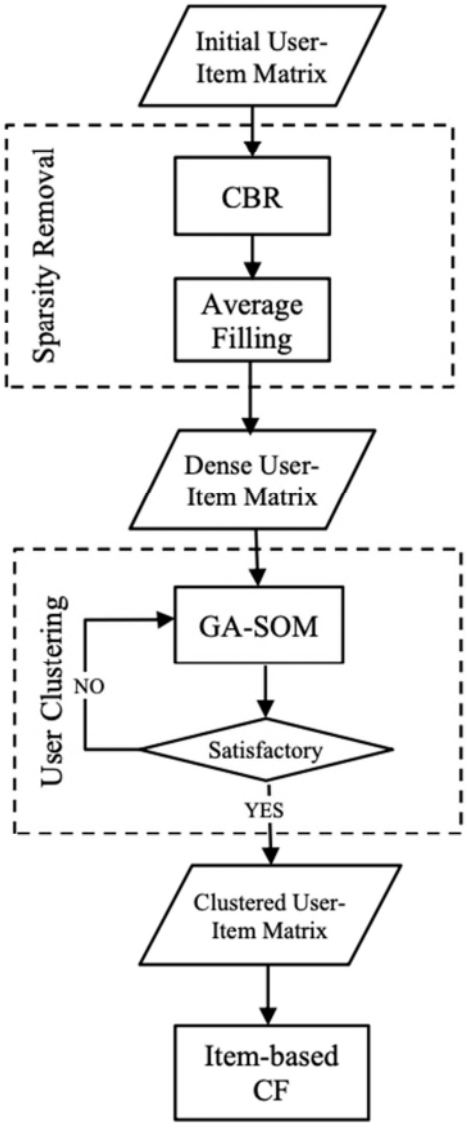


Рисунок 1.7 – Фреймворк Гібридного методу на основі User-Item Collaborative Filtering

Переваги гібридної колаборативної фільтрації полягають у тому, що вона може забезпечити більш точні та персоналізовані рекомендації для користувачів, що збільшує шанс того, що вони придбають предмети, які їм сподобаються. Однак, реалізація гібридної системи може бути складною та часоємною, і вона може потребувати значного об’єму даних та обчислювальних ресурсів для ефективної роботи.

1.4 Аналіз існуючих рішень для подібних систем

Відкриті (Open Source) рішення:

Mahout: Apache Mahout – це бібліотека машинного навчання для Java та Scala. Mahout містить реалізації алгоритмів рекомендаційних систем, таких як колаборативна фільтрація, матричний розклад та інші. Mahout можна використовувати в якості стандартної бібліотеки для розробки рекомендаційних систем, а також включає в себе інструменти для візуалізації даних та взаємодії з ними. [5]

PredictionIO: PredictionIO – це відкрита платформа для розробки рекомендаційних систем та інших систем машинного навчання. PredictionIO має вбудовані алгоритми рекомендацій, такі як колаборативна фільтрація та матричний розклад, а також має інструменти для візуалізації даних та налаштування моделей. [6]

Закриті (Closed Source) рішення:

Amazon Personalize: Amazon Personalize – це послуга рекомендаційних систем від Amazon, яка дозволяє створювати та налаштовувати рекомендаційні системи з використанням різноманітних алгоритмів, таких як колаборативна фільтрація, матричний розклад та інші. Amazon Personalize також надає інструменти для візуалізації даних та моніторингу результатів. [7]

Google Cloud Recommendations AI: Google Cloud Recommendations AI – це послуга рекомендаційних систем від Google, яка дозволяє створювати та налаштовувати рекомендаційні системи з використанням різноманітних алгоритмів, таких як колаборативна фільтрація, матричний розклад та інші. Google Cloud Recommendations AI також надає можливість використовувати гібридний підхід, поєднуючи колаборативну та контентну фільтрацію. Особливість цієї системи полягає в тому, що вона використовує глибоке навчання для покращення рекомендацій. Recommendations AI може підібрати найкращий метод рекомендації для конкретної задачі на основі доступної інформації про продукти та користувачів. [8]

1.5 Мета, завдання та вимоги до реалізації інформаційної системи

Метою кваліфікаційної роботи бакалавра є розробка способу формування рекомендацій пропозицій товарів поліграфічної продукції з використанням методу колаборативної фільтрації, що може бути використано у електронної комерції.

Для досягнення поставленої мети необхідно реалізувати виконання наступних задач:

– *отримати набори даних про оцінки користувачів поліграфічної продукції з відкритих джерел;*

– *розробити спосіб формування рекомендацій пропозицій товарів з використанням методу колаборативної фільтрації;*

– *реалізувати спосіб формування рекомендацій пропозицій товарів для використання у інформаційних системах електронної комерції поліграфічної продукції;*

– *провести тестування та верифікацію розробленого способу формування рекомендацій поліграфічних товарів.*

Розділ 2 Спосіб формування рекомендацій пропозицій товарів для інформаційної системи електронної комерції

2.1 Спосіб формування рекомендацій поліграфічних товарів за допомогою методу колаборативної фільтрації

Колаборативна фільтрація – це метод виявлення шаблонів в поведінці користувачів, що дозволяє робити рекомендації на основі взаємодій між користувачами та товарами. Цей метод можна поділити на дві основні категорії: колаборативну фільтрацію на основі користувачів (user-based) та колаборативну фільтрацію на основі товарів (item-based).

Колаборативна фільтрація на основі користувачів полягає в пошуку користувачів, що мають подібні інтереси, та рекомендації товарів, які ці користувачі оцінили високо. З іншого боку, колаборативна фільтрація на основі товарів виявляє товари, які є подібними за оцінками користувачів, та рекомендує ці товари користувачам, які вже оцінили подібні товари. У контексті колаборативної фільтрації, "сусіди" – це термін, який використовується для опису користувачів або елементів (товарів), які є подібними за певними характеристиками. Сусіди – це користувачі, які мають схожі інтереси або вподобання. Наприклад, якщо двоє користувачів оцінили набір однакових книг або фільмів подібним чином, вони будуть вважатися "сусідами". У такому контексті, система може рекомендувати користувачу товари, які його "сусіди" оцінили високо.

Розглянемо нижче схему методу колаборативної фільтрації, яка зображена на рис. 2.1.

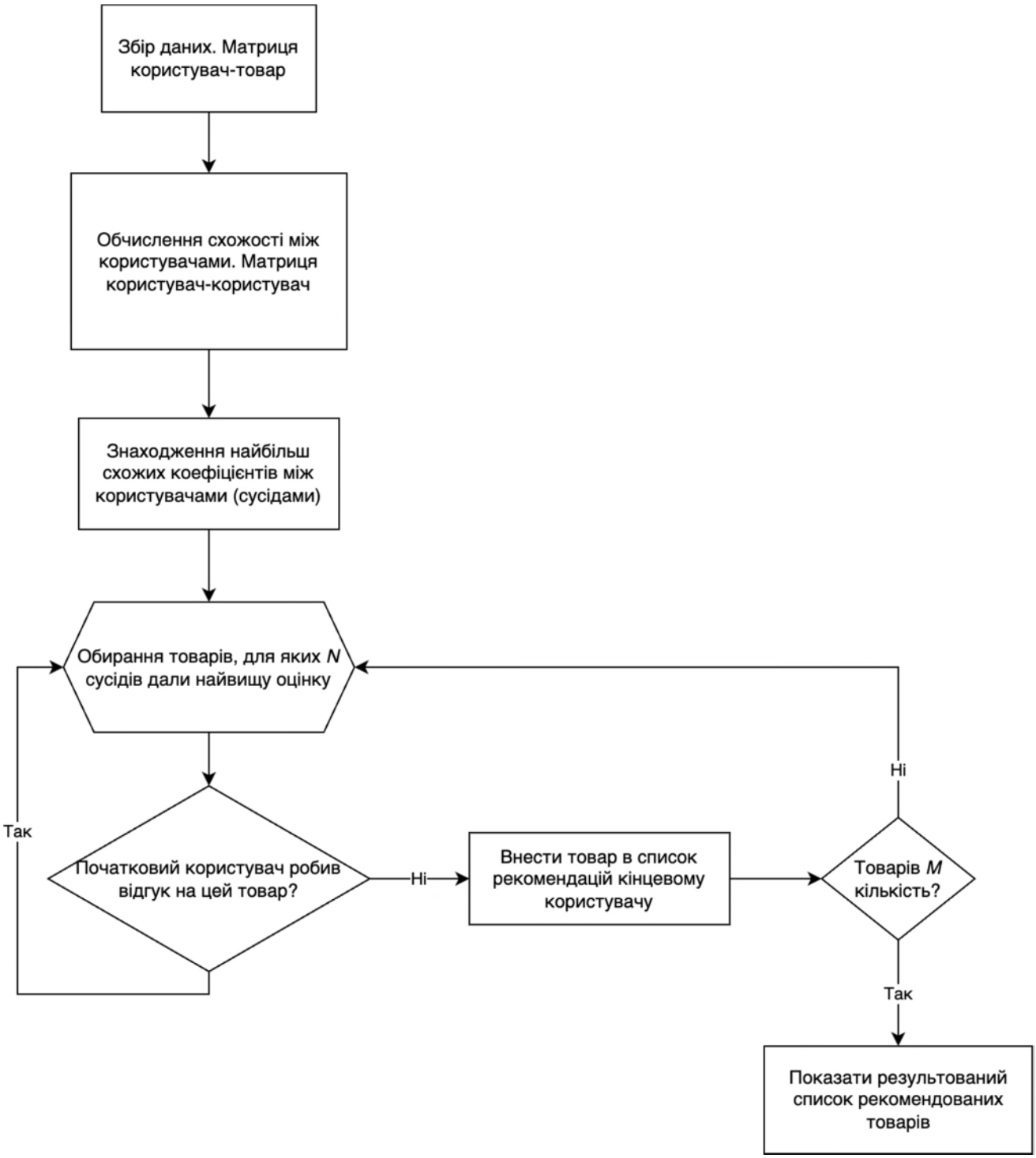


Рисунок 2.1 – Схема роботи методу колаборативної фільтрації

Для формування рекомендацій поліграфічних товарів було вирішено використати метод колаборативної фільтрації на основі користувачів (user-based). Сам спосіб є сукупністю кроків:

1. Збір даних про користувачів. Зібрати дані про взаємодію користувачів з поліграфічними товарами, такі як оцінки, відгуки, покупки, перегляди сторінок

тощо. Буде створено матрицю користувач-товар, де рядки відповідають користувачам, стовпці – товарам, а значення в комірках відображають взаємодію користувачів з товарами (наприклад, оцінки).

2. Обчислення схожості між користувачами. Використовуючи матрицю користувач-товар, схожість між користувачами буде обчислена за допомогою відповідних метрик, таких як коефіцієнт кореляції Пірсона, косинусна схожість або мірою Жаккара. Це допоможе визначити схожість між користувачами на основі їхніх вподобань.

3. Знаходження найбільш схожих користувачів за коефіцієнтом. Для кожного користувача буде знайдено найбільш схожих до нього сусідів (користувачів) шляхом сортування їх коефіцієнтів.

4. Формування рекомендацій. Для кожного користувача будуть враховуватись товари, які його сусіди оцінили високо або з якими вони взаємодіяли, але користувач ще не мав досвіду з ними. Вибиратимуться товари з найбільшою вагою, що враховує коефіцієнт схожості між користувачами та їх оцінки товарів. Ці товари будуть рекомендовані користувачу.

5. Ранжування рекомендацій. Список рекомендованих товарів ранжуватиметься, враховуючи агреговані оцінки, отримані від сусідів, та ваги схожості між користувачами. Буде використано зважену середню для формування кінцевого списку рекомендацій.

2.2 Збір даних про користувачів

Процес збору даних про користувачів є критично важливим для будь-якої системи рекомендацій. Дані користувачів можуть включати інформацію про поведінку користувачів, їхні вподобання та взаємодії з товаром.

Перш за все, потрібно визначити, які дані збираються. Це може включати демографічну інформацію про користувачів, таку як вік, стать, географічне розташування, професію тощо. Також можна збирати дані про поведінку

користувачів на веб-сайті, такі як історія перегляду, продукти, які вони купили, відгуки, які вони залишили, та товари, які вони додали до свого списку бажань.

Наступним кроком є визначення методів збору даних. Це може включати використання веб-скрапінгу для збору даних з інтернету, API для збору даних від третіх сторін, таких як соціальні мережі, або використання серверних логів для збору даних про поведінку користувачів на веб-сайті.

Після збору даних необхідно їх обробити та очистити. Це може включати видалення дублікатів, заповнення пропущених значень, конвертацію типів даних та інше. Цей етап важливий для забезпечення якості та точності даних.

В рамках цієї кваліфікаційної роботи, дані про користувачів збираються як таблиця, в якій кожен рядок – це відгук користувача на конкретну книгу. Кожен користувач має унікальний ідентифікатор. Інформація про відгуки, книги, та користувачів була взята з двох готових dataset-ів:

1. «Book Details» – інформація про книги,
2. «Reviews» – інформація про відгуки користувачів на книги.

Dataset «Book Details» має наступні поля та їх значення:

Поле	**Значення**
title	Назва товару, в даному випадку – назва книги. Це поле використовується для ідентифікації продукту та його представлення користувачам.
description	Опис товару. У випадку книги це може включати короткий зміст, контекст або резюме книги. Це поле важливе для надання користувачам додаткової інформації про товар.
authors	Автори книги. Ця інформація може бути важливою для користувачів, які

	віддають перевагу творам певних авторів.
image	Посилання на зображення обкладинки книги. Зображення часто використовуються в системах рекомендацій, оскільки вони допомагають користувачам швидко визначити, чи цікавить їх продукт.
previewLink	Посилання на попередній перегляд книги, якщо такий існує. Це може бути корисним для користувачів, які хочуть дізнатися більше про книгу перед тим, як придбати її.
publisher	Видавець книги. Ця інформація може бути корисною для користувачів, які віддають перевагу книгам від певних видавництв.
publishedDate	Дата публікації книги. Це може бути важливою інформацією, особливо якщо користувач шукає найновіші видання або книги, видані в певний період часу.
infoLink	Посилання на додаткову інформацію про книгу. Це може бути сторінка продукту на веб-сайті видавництва або інша веб-сторінка з детальною інформацією про книгу.
categories	Категорії, до яких належить книга. Це можуть бути жанри, такі як «фантастика», «історична», «науково-популярна» тощо. Ця інформація

	допомагає класифікувати книги та робити більш точні рекомендації.
ratingsCount	Кількість оцінок, які отримала книга. Це може бути корисною інформацією при ранжуванні рекомендацій, оскільки товари з високим рейтингом та великою кількістю оцінок часто вважаються більш популярними та якісними.

Dataset «Reviews» має наступні поля та їх значення:

Поле	**Значення**
id	Ідентифікатор відгуку. Це унікальне значення, яке допомагає ідентифікувати кожен окремий відгук у наборі даних.
title	Назва товару, для якого було залишено відгук. Це може допомогти у розумінні, до якого товару відноситься відгук.
price	Ціна товару на момент написання відгуку. Це може бути корисною інформацією при аналізі впливу ціни на оцінки користувачів.
user_id	Ідентифікатор користувача, який залишив відгук. Це важливе поле для відслідковування відгуків конкретного користувача.
profileName	Ім'я профілю користувача, який залишив відгук. Це може бути використано для візуалізації або

	представлення даних.
review/helpfulness	Це відношення кількості людей, які вважали відгук корисним, до загальної кількості людей, які проголосували за відгук. Це може бути корисним метриком якості відгуку.
review/score	Оцінка товару користувачем. Це числове значення, яке відображає думку користувача про товар.
review/time	Час, коли було залишено відгук. Ця інформація може бути важливою при аналізі тенденцій у часі.
review/summary	Коротке резюме або заголовок відгуку. Це може бути корисним для швидкого огляду відгуку.
review/text	Повний текст відгуку, залишеного користувачем. Це поле містить детальну інформацію про думки та досвід користувача з товаром.

Схема усіх dataset-ів і зв’язків між ними (рис 2.2):

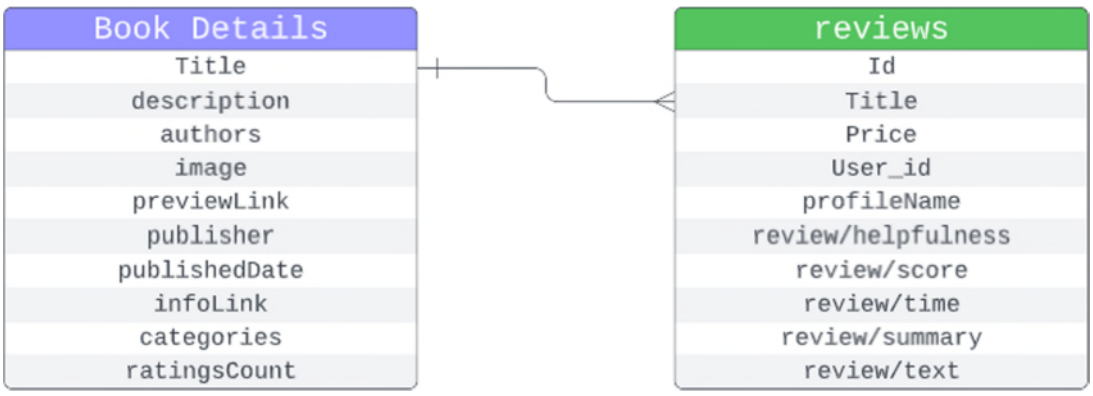


Рисунок 2.2 – Блок-схема dataset-ів

2.3 Обчислення схожості між користувачами

Матриця відповідності, яка була описана у пункті 1.3, отримана в результаті, має рядки, що відповідають користувачам, стовпці – товарам, а значення в комірках відображають взаємодію користувачів з товарами (в даному випадку – оцінки). Якщо користувач не взаємодіяв з товаром, відповідна комірка матриці буде містити 0. Розмір такої матриці може стати дуже великим, якщо існує багато користувачів та товарів, і більшість комірок матриці можуть бути порожніми (це називається "розрідженою" матрицею). У такому випадку може знадобитися використовувати спеціальні структури даних або методи для ефективної роботи з розрідженими матрицями. Наприклад, у матриці на рис. 2.3, показані відношення користувачів до предметів. У користувача під номером «1» немає відгуку про предмет з номером «1», тому в цій комірці лежить «0». Але для предмету з номером «2», у користувача з номером «1» є відгук з оцінкою «3», тому в комірці маємо відповідне число.

	Item 1	Item 2	Item 3	Item 4	Item 5
User 1	**0**	**3**	**0**	**3**	**0**
User 2	**4**	**0**	**0**	**2**	**0**
User 3	**0**	**0**	**3**	**0**	**0**
User 4	**3**	**0**	**4**	**0**	**3**
User 5	**4**	**3**	**0**	**4**	**0**

Рисунок 2.3 – Матриця «користувач-товар»

Обчислення схожості між користувачами може бути обчислена за допомогою метрик, таких як коефіцієнт кореляції Пірсона, косинусна схожість або мірою Жаккара.

Косинусна схожість – це метрика, що використовується для вимірювання того, наскільки два вектори подібні між собою [9]. Вона обчислює косинус кута між двома векторами, що дорівнює добутку векторів, поділеному на добуток їхніх довжин. Косинусна схожість варіює від -1 до 1, де 1 означає повну схожість, а -1 – повну відмінність. Якщо користувач не вказав оцінку для якогось товару, відповідне значення матриці буде дорівнювати 0. Вона зазвичай швидше обчислюється, особливо при використанні оптимізованих бібліотек. Розглянемо безпосередньо механізм роботи алгоритму [9] (форм. 2.4):

$$sim(u,a) = \frac{\sum_{i=1}^{m} r_{a,i} \cdot r_{u,i}}{\sqrt{\sum_{i=1}^{m} r_{a,i}} \cdot \sqrt{\sum_{i=1}^{m} r_{u,i}}}, \quad (2.4)$$

де, *sim(u,a)* – міра схожості користувачів a та u, $r_{u,i}$ – значення матриці R: u – рядок, i – стовпець, *sim(u,a)* приймає значення з відрізка [0, 1]. Якщо користувач не вказав оцінку для якогось предмету, відповідне значення матриці буде дорівнювати «0».

Нижче, на рис. 2.5, наведено приклад матриці схожості між користувачами, в якій u – це ідентифікатор користувача, а числа в комірках – коефіцієнти схожості [10].

	u_1	u_2	u_3	u_4	u_5	u_6	u_7	u_8	u_9	u_{10}	u_{11}	u_{12}
u_1	0	0.1	-0.2	0.4 (①)	0	0	0	0.1	0.3 (②)	-0.2	0.1	-0.1
u_2	0.1	0	0	-0.1	0	0	0	0.2	0.6	0	0	0
u_3	-0.2	0	0	-0.2	0	0	-0.1	0.4	-0.2	-0.1	0	-0.1
u_4	0.4	-0.1	-0.2	0	0.2	0	0.2	0	0	-0.1	0.6	-0.4
u_5	0	0	0	0.2	0	0	0.3	0	0.2	0	0.3	0
u_6	0	0	0	0	0	0	0	0	0	0.3	0	0
u_7	0	0	-0.1	0.2	0.3	0	0	-0.2	0.5	-0.1	0.2	0.1
u_8	0.1	0.2	0.4	0.0	0	0	-0.2	0	-0.2	-0.2	0.1	-0.2
u_9	0.3	0.6	-0.2	0	0.2	0	0.5	-0.2	0	0	0	-0.2
u_{10}	-0.2	0	-0.1	-0.1	0	0.3	-0.1	-0.2	0	0	-0.1	0.4
u_{11}	0.1	0	0	0.6	0.3	0	0.2	0.1	0	-0.1	0	0
u_{12}	-0.1	0	-0.1	-0.4	0	0	0.1	-0.2	-0.2	0.4	0	0

Рисунок 2.5 – Матриця схожості між користувачами

На матриці зеленим кольором відмічені користувачі, які мають найбільші коефіцієнти подібності. Наприклад: користувач «*u1*» найбільше подібний до користувача «*u2*», так як коефіцієнт їх схожості має значення «0.4», тобто найбільший серед усіх «сусідів» *u1*. Далі, в порядку спадання, найближчий до користувача «*u1*» є користувач «*u9*». Їх коефіцієнт схожості рівний «0.3».

2.4 Алгоритм k-найближчих сусідів

Алгоритм k-найближчих сусідів (k-Nearest Neighbors, або k-NN) є одним з найпростіших і найбільш популярних алгоритмів машинного навчання. Він належить до категорії лінійних класифікаторів і використовується в задачах класифікації та регресії [11].

Основна ідея k-NN полягає в тому, що подібні об'єкти знаходяться поруч у просторі ознак. Це означає, що новий об'єкт (або точка даних), який потрібно

класифікувати або для якого потрібно виконати прогноз, буде подібний до тих об'єктів, які знаходяться найближче до нього у просторі ознак.

k-NN використовує просту ідею для передбачення невідомих точок на основі відомих. Коли потрібно класифікувати новий об'єкт, k-NN шукає в датасеті k найближчих сусідів до цього об'єкта, і засновуючись на класах цих сусідів, новому об'єкту присвоюється клас, який є найбільш поширеним серед цих сусідів.

У випадку регресії, замість голосування за клас, k-NN бере середнє або медіану значення цілі сусідів, щоб отримати передбачуване значення для нового об'єкта. Однією з ключових властивостей k-NN є те, що він є «lazy learner» [12], тобто він не використовує тренувальні дані для навчання моделі. Замість цього, він зберігає всі тренувальні дані і використовує їх під час фази передбачення для знаходження найближчих сусідів.

Основними параметрами для k-NN є k, кількість сусідів, що враховуються, і метрика відстані, використовувана для вимірювання близькості між точками. Найчастіше використовуються Евклідова або Манхеттенська відстань [13]. Вибір правильного k та метрики відстані є критичними для ефективності алгоритму k-NN.

На рисунку 2.6, зображено тестовий зразок (зелене коло), який повинен бути класифікований як синій квадрат або як червоний трикутник. Якщо k = 3, то класифікується як червоний трикутник, тому що всередині меншого кола 2 трикутника і тільки 1 квадрат. Якщо k = 5, то він буде класифікований як синій квадрат (3 квадрата проти 2-ох трикутників всередині більшого кола).

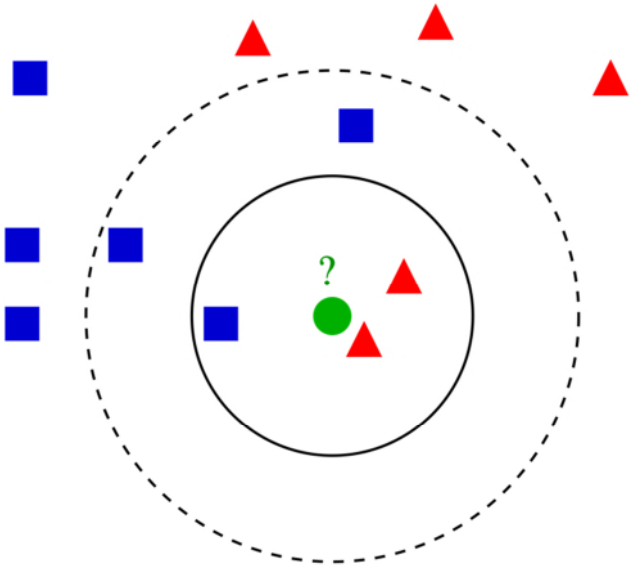


Рисунок 2.6 – Приклад k-NN класифікації. [14]

Розглянемо алгоритм «KNNWithMeans», який є базовим алгоритмом колаборативної фільтрації, що враховує середні оцінки кожного користувача. Він належить до сімейства алгоритмів k-NN (k-Nearest Neighbors) [11]. У випадку колаборативної фільтрації основаної на даних користувачів, формула алгоритму «KNNWithMeans» виглядатиме наступним чином (форм. 2.6), –

$$\hat{r}_{ui} = \mu_u + \frac{\sum_{v \in N_i^k(u)} sim(u,v) \cdot (r_{vi} - \mu_v)}{\sum_{v \in N_i^k(u)} sim(u,v)}, \tag{2.6}$$

де, k – максимальна кількість сусідів, яку треба враховувати для агрегації, $\hat{r}_{ui}$ – оцінка користувача u на елемент i, яка розраховується на основі середньої оцінки відповідного користувача μ_u доданої до вагового середнього оцінок найближчих сусідів, нормалізованого на суму подібностей $\left(\sum_{v \in N_i^k(u)} sim(u,v)\right)$.

Вагове середнє оцінок враховує різницю між оцінкою сусіда на елемент r_{vi} та середньою оцінкою сусіда μ_v, помножену на схожість між користувачем та сусідом $sim(u,v)$.

2.5 Оцінка влучності та повноти алгоритму

В даному розділі досліджено алгоритм k-найближчих сусідів (k-NN) з огляду на такі ключові метрики, як точність (precision) та повнота (recall). Для алгоритму k-NN, як і для всіх методів машинного навчання, оцінка цих метрик є критично важливою для визначення його ефективності [15].

Метрика точності показує, скільки з передбачених позитивних випадків є дійсно позитивними, тоді як метрика повноти визначає, скільки від дійсно позитивних випадків було правильно класифіковано.

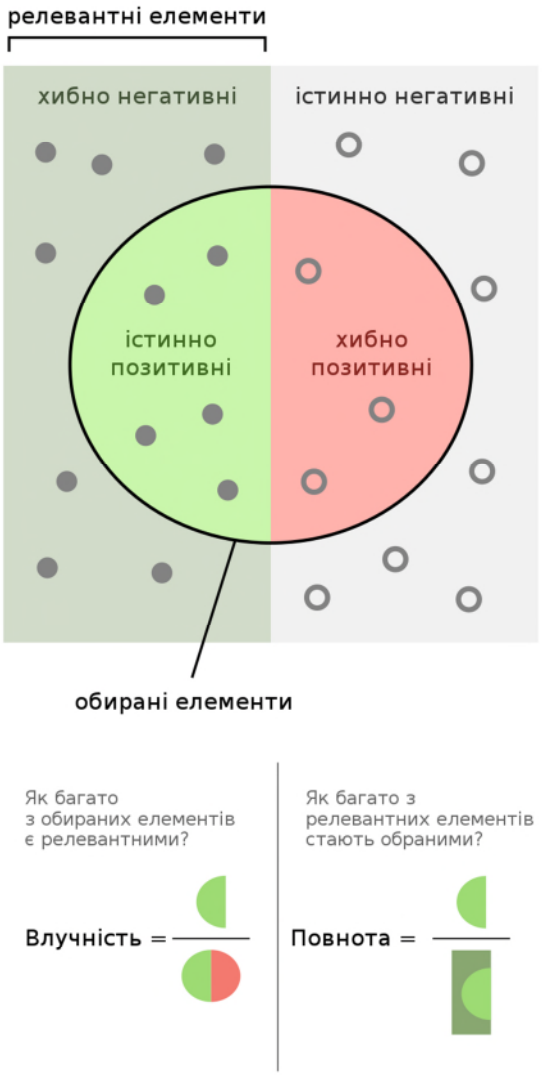


Рисунок 2.7 – Схематичний опис влучності та повноти [16]

В галузі поліграфічних товарів, влучність є часткою знайдених книг, що є релевантними запитові (формула 2.8):

$$P = \frac{B_r \cap B_f}{B_f}, \tag{2.8}$$

де, P – влучність, B_r – релевантні книги, B_f – знайдені книги.

Наприклад, для рекомендації на множині книг, влучність є числом правильних результатів, поділеним на число всіх повернених результатів.

Влучність бере до уваги всі знайдені документи, але її також можливо оцінювати на заданому рівні відсікання, враховуючи лише розташовані найвище результати, що повертає система. Таку міру називають «N-влучністю».

Влучність використовують разом із повнотою, відсотком всіх релевантних документів, який повертає пошук. Ці дві міри іноді використовують разом в оцінці F1 (або F-мірі) [17], щоби забезпечити єдине вимірювання для системи.

В галузі поліграфічних товарів повнота є часткою релевантних книг, яку вдається успішно знайти (формула 2.9).

$$C = \frac{B_r \cap B_f}{B_r}, \tag{2.9}$$

де, C – повнота, B_r – релевантні книги, B_f – знайдені книги.

В бінарній класифікації повноту називають чутливістю [18]. Її можливо розглядати як імовірність того, що релевантний документ буде знайдено за запитом.

Досягти повноти 100 % тривіально, якщо повертати у відповідь на запит всі книги. Отже, повнота сама по собі не є достатньою, й потрібно також вимірювати й число нерелевантних книг, наприклад, обчислюючи також і влучність.

Для вимірювання точності та повноти алгоритму k-NN буде проведено серію тестів на вибірці даних Y. В ході тестування застосовуватиметься стратегія k-fold cross-validation, яка отримати більш надійну оцінку результатів.

2.6 Формування рекомендацій

Формування рекомендацій відбувається в декілька етапів (рис. 2.10):

1. Спочатку вибираються «сусіди» для кожного користувача. Це ті користувачі, які найбільш схожі на поточного користувача за допомогою косинусної схожості або кореляції Пірсона. Можна обмежити кількість сусідів, щоб скоротити обчислювальні витрати.

2. Для кожного користувача знаходяться товари, які його сусіди оцінили високо або з якими вони взаємодіяли, але користувач ще не мав досвіду з ними.

3. Враховується вага схожості між користувачем та його сусідами при оцінці цих товарів. Товари, які були оцінені високо сусідами, з якими користувач сильно схожий, повинні мати більшу вагу.

4. Вибираються товари з найбільшою загальною вагою для рекомендації.

Рисунок 2.10 – Схема способу формування рекомендацій

1. Знаходження найбільш схожих користувачів за коефіцієнтом.

Для пошуку найбільш схожих користувачів за коефіцієнтом можна використати обчислену матрицю схожості, яку було отримано на попередньому етапі. За допомогою цієї матриці для кожного користувача можна визначити інших користувачів, які мають найвищий коефіцієнт схожості.

Отже, для кожного користувача можна переглянути рядок у матриці схожості, що відповідає цьому користувачу, відсортувати значення в цьому рядку у порядку спадання та взяти перші N користувачів.

2. Обрання товарів з найбільшою загальною вагою для рекомендації

З попередніх таблиць відфільтровуються ті товари, які користувач ще не оцінював, але вони були високо оцінені «сусідами» цього користувача. Цим самим вибираються потенційно «цікаві» для користувача товари, адже скоріш за все він інколи або ніколи ними не користувався. До того ж, у цих товарів хороші оцінки від інших користувачів, очевидно, що такі екземпляри користуються попитом, популярні і з великою ймовірністю сподобаються користувачеві.

При розрахунку рекомендації для товару, враховуються оцінки, які сусіди дали цьому товару, проте замість простого середнього, використовується ваговане середнє, де ваги – це схожості між користувачем і кожним сусідом.

Після того, як отримано ваговані оцінки для всіх товарів, можна обчислити суму вагованих оцінок для кожного товару і рекомендувати товари з найбільшою сумою оцінок.

Розділ 3 Програмна реалізація інформаційної системи з використанням запропонованого способу формування рекомендацій

3.1 Програмні бібліотеки та середовища розробки

Під час розробки було використано мову програмування Python а також бібліотеки: numpy, pandas, scikit-surprise, nltk. Середовищем розробки було обрано середовище розробки Visual Studio Code з плагінами для підтримки Jupyter Notebook.

Numpy – розширення мови Python, що додає підтримку великих багатовимірних масивів і матриць, разом з великою бібліотекою високорівневих математичних функцій для операцій з цими масивами [19].

Pandas – програмна бібліотека, написана для мови програмування Python для маніпулювання даними та їхнього аналізу [20]. Вона, зокрема, пропонує структури даних та операції для маніпулювання чисельними таблицями та часовими рядами. pandas є вільним програмним забезпеченням, що випускається за трипунктовою ліцензією BSD. Ця назва походить від терміну «панельні дані» (англ. panel data), який в економетрії позначає багатовимірні структуровані набори даних. Pandas дозволяє здійснювати різні операції з обробкою даних, такі як об'єднання, зміна форми, вибір, а також очищення даних та функції перегляду даних.

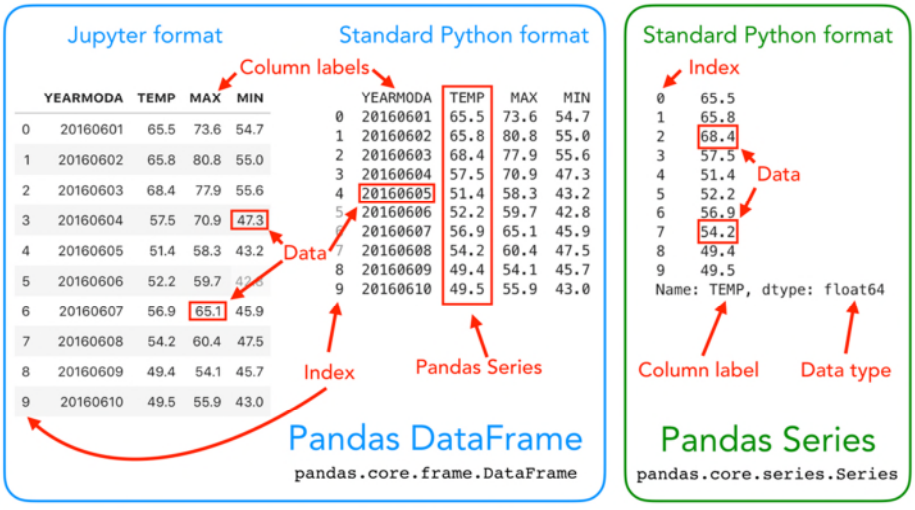


Рисунок 3.1 – Схема роботи бібліотеки «pandas» [21]

Scikit-Surprise – це бібліотека Python, призначена для простого і ефективного створення та аналізу систем рекомендацій [22]. Вона містить різні алгоритми рекомендацій, включаючи колаборативну фільтрацію, основану на користувачах та товарах, та інші. Використовується для створення моделі рекомендацій.

NLTK (Natural Language Toolkit) – це набір бібліотек, модулів та програмних інтерфейсів для обробки мови. Він надає зручні засоби для роботи з текстами на людській мові. Використовується для аналізу відгуків користувачів та визначення їх настрою. З цього набору було використано клас SentimentIntensityAnalyzer, який використовується для визначення та кількісної оцінки сентименту (тобто емоційного забарвлення) тексту. Цей аналізатор сентименту базується на моделі VADER (Valence Aware Dictionary for sEntiment Reasoning), яка спеціально розроблена для аналізу текстів з соціальних медіа. VADER використовує список слів та фраз, які асоціюються з позитивними, негативними та нейтральними емоціями.

Рисунок 3.2 – Опис функцій NLTK [23]

SentimentIntensityAnalyzer аналізує текст і видає чотири оцінки:

– pos (позитивність): ймовірність того, що текст є позитивним.
– neu (нейтральність): ймовірність того, що текст є нейтральним.
– neg (негативність): ймовірність того, що текст є негативним.
– compound (композитний): сукупний показник сентименту тексту, який враховує всі вищезгадані фактори.

Visual Studio Code (VS Code) – це вільний відкритий редактор коду, розроблений компанією Microsoft [24]. VS Code є універсальним інструментом, що підтримує велику кількість мов програмування, і є особливо популярним серед розробників на JavaScript, TypeScript і Node.js, також Python.

Jupyter Notebook – це веб-застосунок відкритого коду, який дозволяє створювати та розповсюджувати документи, що містять інтерактивний код, рівняння, візуалізації та пояснювальний текст [25]. Він широко використовується в наукових дослідженнях, обробці даних, статистичному моделюванні, машинному навчанні та інших сферах. Розширення Jupyter для Visual Studio Code дозволяє використовувати Jupyter Notebooks прямо в редакторі VS Code.

3.2 Особливості реалізації способу формування рекомендацій поліграфічних товарів

3.2.1 Первинна обробка і підготовка датасету

На початку розробки, було імпортовано бібліотеки, за допомогою яких відбувається завантаження і обробка даних. А саме: numpy та pandas. У даній реалізації, бібліотеки numpy та pandas використовуються на кількох ключових етапах.

Numpy використовується в основному для обробки числових даних у вигляді масивів. В рамках поточної реалізації, це можуть бути оцінки користувачів для поліграфічних товарів, які потім аналізуються для визначення схожості між різними продуктами або користувачами.

Pandas ж використовується для завантаження, обробки та аналізу даних. Завдяки pandas, з’являється можливість завантажувати великі набори даних про оцінки користувачів з файлів чи відкритих джерел, очищувати ці дані від пропусків або помилок, та трансформувати їх для подальшого аналізу.

Наприклад, pandas використано для визначення середніх оцінок товарів, відфільтрування товарів за певними критеріями, або ж для групування даних по користувачах чи товарах. Також з допомогою pandas виконуються різні види агрегації даних, що є важливим для статистичного аналізу та візуалізації. Таким чином, numpy та pandas відіграють ключову роль у виконанні різних задач цього проекту, пов'язаних з обробкою та аналізом даних.

Наступним кроком є формування DataFrame з даних датасетів. (рис. 3.3) Початковий датасет з усіма полями описаний у розділі 2.2.

	Id	Title	Price	User_id	profileName	review/helpfulness	review/score	review/time	review/summary	review/text
0	1882931173	Its Only Art If Its Well Hung!	NaN	AVCGYZL8FQQTD	Jim of Oz "jim-of-oz"	7/7	4.0	940636800	Nice collection of Julie Strain images	This is only for Julie Strain fans. It's a col...
1	0826414346	Dr. Seuss: American Icon	NaN	A30TK6U7DNS82R	Kevin Killian	10/10	5.0	1095724800	Really Enjoyed It	I don't care much for Dr. Seuss but after read...
2	0826414346	Dr. Seuss: American Icon	NaN	A3UH4UZ4RSVO82	John Granger	10/11	5.0	1078790400	Essential for every personal and Public Library	If people become the books they read and if "t...
3	0826414346	Dr. Seuss: American Icon	NaN	A2MVUWT453QH61	Roy E. Perry "amateur philosopher"	7/7	4.0	1090713600	Phlip Nel gives silly Seuss a serious treatment	Theodore Seuss Geisel (1904-1991), aka "D...
4	0826414346	Dr. Seuss: American Icon	NaN	A22X4XUPKF66MR	D. H. Richards "ninthwavestore"	3/3	4.0	1107993600	Good academic overview	Philip Nel - Dr. Seuss: American IconThis is b...

Рисунок 3.3 – Формування DataFrame з початкового датасету

Попередній крок надає можливість маніпулювати даними зручно, за допомогою бібліотеки pandas. Для подальшого полегшення роботи з даними,

колонки перейменовуються та вибираються з них тільки ті, які будуть використані у майбутньому аналізі. (рис. 3.4)

	book_id	user_id	review	title	rating
0	1882931173	AVCGYZL8FQQTD	This is only for Julie Strain fans. It's a col...	Its Only Art If Its Well Hung!	4.0
1	0826414346	A30TK6U7DNS82R	I don't care much for Dr. Seuss but after read...	Dr. Seuss: American Icon	5.0
2	0826414346	A3UH4UZ4RSVO82	If people become the books they read and if "t...	Dr. Seuss: American Icon	5.0
3	0826414346	A2MVUWT453QH61	Theodore Seuss Geisel (1904-1991), aka "D...	Dr. Seuss: American Icon	4.0
4	0826414346	A22X4XUPKF66MR	Philip Nel - Dr. Seuss: American IconThis is b...	Dr. Seuss: American Icon	4.0
...	...	...	...	...	...
95	B000NKGYMK	A2UMP9TJTJ6A6B	As a former Alaskan, I didn't want to have to ...	Alaska Sourdough	1.0
96	B000NKGYMK	AC2TK7NHKB5C0	For those of us who would prefer to use sourdo...	Alaska Sourdough	5.0
97	B000NKGYMK	A22T74YNRM8NTK	Make the most sublime waffles - crispy outside...	Alaska Sourdough	5.0
98	B000NKGYMK	A2E0GB5QZRX2JZ	I got this book because of all the interesting...	Alaska Sourdough	5.0
99	B000NKGYMK	A7VSVB6Z0JHOV	It's quaint, I'll give it that. The handwritte...	Alaska Sourdough	2.0

Рисунок 3.4 – Перейменування та фільтрування колонок

В результаті отримано відфільтрований набір даних, але він є досить «брудним». Завеликий розмір і велика кількість непотрібної інформації негативно впливатиме на швидкість обробки цих даних. Також, деякі відгуки, наприклад, не досвідчених користувачів, або «хейтерів» у спотворюють точність, з якою формуються рекомендації. Тому, датасет очищається і готується до аналізу. Користувачі, які надали понад двісті оглядів книг, залишаються. Розглядається, що такі користувачі мають великий досвід і їхні відгуки є актуальними. Також, зберігаються лише книги, для яких користувачі написали щонайменше п'ятдесят відгуків. (рис. 3.5)

```
                user_id  count_review  \
2564065  A3SU3TXON36T0X         201.0
1943236   A1ZNGCNKCHW11         201.0
2441685  A26BWRBPP4V2WF         201.0
482499   A2SJ11EMIOBLHH         201.0
1924280  A1DTOHMM2Y5KY0         202.0
...                 ...           ...
667567   A1G56KHOUOFWDW         243.0
153670   A3DXVYD2TQHWGD         244.0
509292    ALDRY40BPNY09         244.0
1942693  A1D5RCOILPC9LX         244.0
189875   A13O359A2KX7YY         244.0

                                                     title  title_review
2564065  Acid Dreams: The Complete Social History of LS...          51.0
1943236                                Orphans of the sky          52.0
2441685                            Run Silent, Run Deep            52.0
482499                                      Soul On Ice            53.0
1924280                    Why government doesn't work             51.0
...                                                   ...           ...
667567   Secrets of the Code: The Unauthorized Guide to...          51.0
153670                                 Fat Ollie's Book            51.0
509292              On the duty of civil disobedience              51.0
1942693       King of Foxes (Conclave of Shadows, Book 2)          51.0
189875                                    The Glass Key            52.0
```

Рисунок 3.5 – Очищення і підготовка датасету

Фільтрація даних, як описано в наведеному вище прикладі, використовується з двома основними цілями:

1. Покращення якості рекомендацій: Узагальнення часто працює краще, коли виключається "шум" від користувачів або предметів з дуже малою кількістю оцінок. В даному випадку, фільтруванням користувачів, які відгукнулися більше ніж 200 разів, та книг, які мають 50 і більше відгуків, концентрація відбувається на найбільш активних користувачах і найпопулярніших книгах, що покращує якість рекомендацій.

2. Ефективність обчислень: Робота з меншим обсягом даних значно зменшує час обчислень та ресурси, потрібні для розробки моделі. Це особливо

важливо для великих наборів даних, де навіть найефективніші алгоритми можуть виявитися повільними.

Таким чином, цей процес фільтрації допомагає зменшити розмір даних, працюючи лише з користувачами, які відгукнулися достатньо часто, і книгами, які отримали достатню кількість відгуків. Це допомагає зменшити шум і зробити модель рекомендацій більш точною і ефективною.

На попередньому рисунку (рис. 3.5) видно, що у даних змінилась нумерація індексів. В процесі обробки даних, одним з важливих етапів є оперативна і коректна робота з індексами у Pandas DataFrame. Нерідко під час маніпуляцій з даними, зокрема видалення окремих рядків, індекси можуть втратити свою послідовність або узгодженість.

Після фільтрації даних використовується метод скидання індексів рядків. Він дозволяє скинути поточні індекси DataFrame і згенерувати нові, які будуть правильно відповідати актуальному порядку рядків. В результаті, отримується структура даних, яка є більш зручною і надійною для подальшої роботи. Результат методу скидання індексів зображено на рис. 3.6 нижче.

	index	book_id	user_id	review	title	rating	count_review	title_review
0	1648451	0451519582	ACPAI5CVQKKWK	In such situations, it is customary to start o...	Wuthering Heights (Signet classics)	5.0	209.0	2160.0
1	1648452	0451519582	ACPAI5CVQKKWK	In such situations, it is customary to start o...	Wuthering Heights (Signet classics)	5.0	209.0	2160.0
2	1648453	0451519582	ACPAI5CVQKKWK	In such situations, it is customary to start o...	Wuthering Heights (Signet classics)	5.0	209.0	2160.0
3	1648454	0451519582	ACPAI5CVQKKWK	In such situations, it is customary to start o...	Wuthering Heights (Signet classics)	5.0	209.0	2160.0
4	1648455	0451519582	ACPAI5CVQKKWK	In such situations, it is customary to start o...	Wuthering Heights (Signet classics)	5.0	209.0	2160.0

Рисунок 3.6 – Дані з впорядкованими індексами

Ця процедура є важливою з точки зору забезпечення консистентності даних і забезпечує точність та надійність результатів подальшого аналізу.

3.2.2 Обчислення показника з урахуванням очищення даних від аномалій

1. Аналіз та нормалізація тональності тексту відгуків

Після проведення кроків очищення і підготовки датасету, наступним кроком є аналіз тональності тексту. Аналіз тональності, також відомий як аналіз емоційної тональності, включає в себе використання методів природної обробки мови, статистичних методів або машинного навчання для систематичного визначення характеру емоційного забарвлення тексту.

У контексті цієї роботи, цей крок дозволяє витягнути інформацію про емоційний стан користувача з його відгуків. Це допомагає розуміти, чи подобається користувачу продукт чи ні, а також розглянути ширші контекстуальні зв'язки у його відповідях.

Даний етап аналізу тональності включає в себе ряд конкретних задач. Початково, відгуки розбиваються на окремі слова і фрази, процес, який відомий як токенізація. Після цього, алгоритми використовуються для визначення емоційного забарвлення цих токенів, які агреговані для створення загального рейтингу тональності відгука.

В процесі аналізу тональності відгуків, одним із ключових етапів є визначення середнього значення тональності для кожного відгуку. Цей процес включає обрахунок метрики, яка відображає загальний емоційний настрій, виражений у відгуку, допомагаючи таким чином створити кількісний показник настрою відгуку.

Однак, отримані значення тональності можуть варіюватись від -1 до 1, що може ускладнювати подальший аналіз даних. Для вирішення цієї проблеми проводиться нормалізація отриманих значень. Ця процедура зміщує межі виміру значень в діапазон від 0 до 1.

Такий підхід має дві важливі переваги. По-перше, нормалізовані дані легше інтерпретувати, оскільки всі значення тепер знаходяться в однаковому діапазоні. По-друге, нормалізація спрощує проведення статистичного аналізу та

обробку даних машинним навчанням, оскільки багато алгоритмів працюють краще з нормалізованими даними.

2. Очищення даних від аномалій

Одним з ключових аспектів будь-якої рекомендаційної системи є її здатність генерувати точні й релевантні рекомендації. Основою для створення цих рекомендацій є якісні дані. Однак, не всі дані, які надходять до системи, є коректними або відповідають очікуванням. Наприклад, можливі випадки, коли оцінка товару не відповідає емоційному забарвленню відгуку на нього. Це може призвести до неправильних рекомендацій. Тому важливим етапом обробки даних є їх очищення від подібних аномалій. Отже, був проведений процес видалення таких неконсистентних даних, що мають відхилення між оцінкою товару та емоційним забарвленням відгуку, з метою поліпшення якості рекомендацій, що генеруються за допомогою методу колаборативної фільтрації.

Операції, що виконуються над датасетом, включають в себе видалення рядків, які можуть вважатися аномальними або непослідовними. Ці аномалії можуть бути спричинені суб'єктивністю оцінки користувача та відгуку.

Перша операція видаляє рядки, де оцінка, що була внесена користувачем, низька (1 або 2 з 5), але тональність тексту відгуку позитивна (більше 0,6). Це суперечливість може бути викликана, наприклад, помилкою при внесенні оцінки користувачем або ж наміром дати товару нижчу оцінку, ніж заслуговує за відгуком.

Друга операція видаляє записи, де оцінка висока (4 або 5 з 5), але тональність відгуку негативна (менше 0,6). Це може відбутися, якщо користувач високо оцінив товар, але у відгуку висловив негативні емоції.

Ці дії є необхідними для поліпшення якості даних, що використовуються для побудови моделей на основі методу колаборативної фільтрації. Вони сприяють отриманню більш точних та корисних рекомендацій, що виходять з використання цього методу.

3. Формування гібридного показника

Важливим кроком у побудові ефективної рекомендаційної системи є утворення надійного та релевантного метричного індикатора. Система повинна враховувати не лише числові оцінки, надані користувачами, а й емоційний вплив, що його вони викликають. З цією метою вводиться гібридний показник. Гібридний показник об'єднує два ключових параметри: числову оцінку товару та емоційний контекст відгуку.

Наступний крок – створення цього гібридного показника. Він множить оцінку користувача на товар на вираховану тональність відгука. Результатом є новий показник, який враховує обидва цих аспекти. Це покращує якість системи, оскільки враховує більше параметрів у рекомендаціях. Такий підхід може допомогти виявити невідповідності між оцінкою товару та емоційним забарвленням відгуку, що може вказувати на аномалії в даних або на несподівані вподобання користувачів.

3.2.3 Оцінка точності та повноти рекомендаційної моделі

1. Підготовка даних для моделі

Для моделі, яка вчиться формувати рекомендації потрібно підготувати DataFrame. Спочатку створюється об'єкт Reader, який вказує формат даних у файлі. У даному випадку, кожний рядок файлу повинен містити інформацію про користувача, предмет та рейтинг, розділену комами. Потім, датасет завантажується з DataFrame, використовуючи стовпці 'user_id', 'book_id', 'hybrid_score' як джерело даних для користувачів, предметів та рейтингів відповідно.

В процесі роботи з рекомендаційною системою важливим кроком є підготовка даних, яка включає завантаження, обробку та розбиття даних на навчальну та тестову вибірки. Цей етап має на меті забезпечити правильну підготовку даних для наступного аналізу та навчання моделі рекомендацій.

Опис дій:

1. Імпорт бібліотеки: Все починається з імпорту необхідної бібліотеки, яка надає функціонал для роботи з рекомендаційною системою на основі колаборативної фільтрації. Це дозволяє використовувати потрібні функції та методи для подальшої обробки та аналізу даних.

2. Завантаження даних: Наступним кроком є завантаження даних, які будуть використовуватись для рекомендаційної системи. Дані можуть бути отримані з різних джерел, наприклад, з бази даних або з файлу. У даному випадку використано функцію, яка дозволяє завантажити дані з відповідного джерела.

3. Розбиття даних: Після завантаження даних, вони розбиваються на навчальну та тестову вибірки. Навчальна вибірка використовується для навчання моделі рекомендацій, тоді як тестова вибірка використовується для оцінки ефективності моделі та перевірки її точності.

Ці дії є необхідними для підготовки даних перед подальшим застосуванням методу колаборативної фільтрації. Вони завантажують дані, розбивають дані на частини та готують їх для наступного етапу рекомендаційної системи. Правильна підготовка даних грає важливу роль у досягненні точності та ефективності рекомендаційної системи.

2. Побудова та навчання моделі рекомендацій

На цьому етапі створюється та навчається модель k-найближчих сусідів (KNN) з використанням середніх ваг (KNNWithMeans) на основі набору тренувальних даних. А саме, створюється нова модель KNN з використанням середніх ваг (KNNWithMeans). Перший гіперпараметр встановлений в значення 50, що означає, що модель враховує 50 найближчих сусідів при виробленні прогнозу для конкретного користувача або предмета. Інший параметр вказує на використання косинусної схожості для обчислення схожості між користувачами. Далі виконується навчання моделі на основі набору тренувальних даних.

Обчислюється схожість між користувачами та зберігається для подальшого використання при прогнозуванні оцінок.

3. Оцінка точності моделі на тестовому наборі даних

Створена та натренована модель використовується для оцінки її точності на тестовому наборі даних.

Для оцінки точності моделі рекомендацій використовується бібліотека Surprise. По-перше, за допомогою моделі, що була побудована, виконується прогнозування оцінок для тестового набору даних. По-друге, з використанням бібліотеки Surprise, викликається функція, яка обчислює середньоквадратичну помилку (RMSE) для порівняння прогнозованих оцінок з фактичними значеннями. Ця метрика вимірює відхилення прогнозів від фактичних значень та дозволяє оцінити точність моделі. Чим менше значення RMSE, тим краще модель прогнозує рейтинги користувачів.

Для генерування прогнозів на основі тестового набору даних використовується окремий метод моделі. Він повертає список прогнозів для кожного користувача та предмета в тестовому наборі.

Функція «rmse» від модулю «accuracy» бібліотеки «surprise» була використана для обчислення кореневого середньоквадратичного відхилення прогнозів моделі від реальних оцінок. RMSE – це стандартна метрика для оцінки точності прогнозувальних моделей. Значення RMSE зберігається для подальшого використання або виведення.

4. Визначення точності та повноти рекомендацій

На цьому етапі обчислюються дві важливі метрики оцінки для систем рекомендацій: точність (precision) та повноту (recall) для кожного користувача. Метрики обчислюються для топ-K рекомендацій, що були надані кожному користувачу.

В основі алгоритму є наступні кроки:

1. Для кожного користувача створюється список всіх його прогнозів.

2. Прогнози для кожного користувача сортуються в порядку зменшення прогнозованого рейтингу.

3. Обчислюється кількість релевантних елементів (тобто тих, рейтинг яких перевищує заданий поріг).

4. Обчислюється кількість рекомендованих елементів у топ-K (ті, які мають прогнозований рейтинг вище порогу).

5. Обчислюється кількість релевантних і рекомендованих елементів у топ-K.

6. Для кожного користувача обчислюється точність як частка релевантних і рекомендованих елементів до загальної кількості рекомендованих елементів. Якщо немає рекомендованих елементів, точність встановлюється рівною 0.

7. Для кожного користувача обчислюється повнота як частка релевантних і рекомендованих елементів до загальної кількості релевантних елементів. Якщо немає релевантних елементів, повнота встановлюється рівною 0.

Таким чином, цей алгоритм (рис. 3.7) оцінює якість системи рекомендацій з точки зору того, наскільки добре вона вибирає релевантні елементи для кожного користувача і наскільки повно ці елементи включені в список топ-K рекомендацій.

```python
def precision_recall_at_k(predictions, k=10, threshold=3):
    """Return precision and recall at k metrics for each user"""

    # First map the predictions to each user.
    user_est_true = defaultdict(list)
    for uid, _, true_r, est, _ in predictions:
        user_est_true[uid].append((est, true_r))

    precisions = dict()
    recalls = dict()
    for uid, user_ratings in user_est_true.items():

        # Sort user ratings by estimated value
        user_ratings.sort(key=lambda x: x[0], reverse=True)

        # Number of relevant items
        n_rel = sum((true_r >= threshold) for (_, true_r) in user_ratings)

        # Number of recommended items in top k
        n_rec_k = sum((est >= threshold) for (est, _) in user_ratings[:k])

        # Number of relevant and recommended items in top k
        n_rel_and_rec_k = sum(
            ((true_r >= threshold) and (est >= threshold))
            for (est, true_r) in user_ratings[:k]
        )

        # Precision@K: Proportion of recommended items that are relevant
        # When n_rec_k is 0, Precision is undefined. We here set it to 0.
        precisions[uid] = n_rel_and_rec_k / n_rec_k if n_rec_k != 0 else 0

        # Recall@K: Proportion of relevant items that are recommended
        # When n_rel is 0, Recall is undefined. We here set it to 0.
        recalls[uid] = n_rel_and_rec_k / n_rel if n_rel != 0 else 0

    return precisions, recalls
```

Рисунок 3.7 – Алгоритм обчислення точності та повноти оцінок кожного користувача

Таким чином, цей код дозволяє оцінити загальну ефективність моделі рекомендацій з точки зору точності і повноти її прогнозів.

Далі, цей метод використовується для обчислення точності (precision) та повноти (recall) прогнозів моделі на основі заданого порогу і кількості елементів k. Він приймає список прогнозів, значення k (кількість рекомендацій, які розглядаються), і порогове значення (у цьому випадку 2). Він повертає два

словники: precisions і recalls, які містять значення точності та повноти для кожного користувача відповідно.

В результаті, реалізується рекомендаційна система, яка передбачає оцінки, які користувач би дав різним книгам, і рекомендує ті, які мають найвищі передбачені оцінки.

Спочатку визначається ідентифікатор користувача, для якого потрібно зробити рекомендації. Потім отримується список усіх книг, які є в наборі даних.

Наступний крок – це передбачення оцінки, яку цей користувач би дав кожній книзі. Це досягається шляхом використання моделі, яка була навчена раніше. Передбачувані оцінки зберігаються в словнику, де ключем є ідентифікатор книги, а значенням – передбачувана оцінка.

Потім книги сортуються за передбачуваною оцінкою в порядку спадання і вибираються перші 5. Це будуть ті книги, які рекомендуються користувачеві. Наприкінці, виводяться назви рекомендованих книг і їх передбачені оцінки. Приклад вихідних даних вказаний на рис. 3.8.

```
1. B000JJ9ISM(The Daughter of Time) with predicted rating of 5.00
2. B000P3LVZA(Brave New World) with predicted rating of 5.00
3. 0681994851(Jane Eyre) with predicted rating of 5.00
4. 1556863047(The Sea Wolf) with predicted rating of 5.00
5. B000QABFIK(A Tree Grows in Brooklyn) with predicted rating of 5.00
6. B000NOWYR0(Tender Is the Night) with predicted rating of 5.00
7. B000J5ZQH8(The Sea Wolf) with predicted rating of 5.00
8. 0582528259(Jane Eyre (Simple English)) with predicted rating of 5.00
9. 0140351310(Jane Eyre: Complete and Unabridged (Puffin Classics)) with predicted rating of 5.00
10. B000NVLWSU(Night) with predicted rating of 5.00
```

Рисунок 3.8 – Приклад рекомендованих книг для користувача

3.3 Тестування способу формування рекомендацій поліграфічних товарів методом колаборативної фільтрації

В даному розділі описано тестування способу формування рекомендацій поліграфічних товарів за допомогою методу колаборативної фільтрації. Необхідно перевірити, наскільки ефективно цей метод може бути використаний

для підвищення задоволеності користувачів та покращення комерційних показників.

Процес тестування відбувається наступним чином:

1. Вибирається користувач, складається його «портрет»
2. Для цього користувача генеруються рекомендації
3. Рекомендації порівнюються з реальним «портретом» користувача.

Для тестування будуть розглянуті наступні тест-кейси:

Тест-кейси:

1. Обрано користувача №1. який написав близько 200 відгуків. Цей користувач, судячи з кількості відгуків, є активним читачем, особливо у сфері художньої літератури (рис. 3.9). Йому подобається занурюватися в вигадані світи та сюжети, які відображаються у його наданні переваги жанру «Fiction» – 74 відгуки свідчать про це.

['Fiction']	74
['Religion']	16
['Juvenile Fiction']	15
['Science']	6
['Biography & Autobiography']	6

Рисунок 3.9 – Жанри книг і кількість відгуків, які написав користувач №1

Він також проявляє інтерес до релігійних текстів з 16 відгуками в жанрі «Religion». Це може вказувати на його духовні пошуки або глибокий інтерес до релігійних питань.

Інтерес до жанру «Juvenile Fiction» з 15 відгуками може говорити про його пристрасть до книг, які цікаві для молоді, або ж він може вибирати книги для своїх дітей або молодших родичів.

В цілому, портрет цього користувача відображає широкий спектр інтересів з вираженою уподобанням до художньої літератури.

Рекомендаційна система згенерувала для цього користувача наступні книги (рис. 3.10):

```
1. 'The Giver' with predicted rating of 4.61 and genre ['Juvenile Fiction']
2. 'Man's Search for Meaning' with predicted rating of 4.26 and genre ['Existential psychotherapy']
3. 'The Lion, the Witch and the Wardrobe' with predicted rating of 4.20 and genre nan
4. 'The Big Sleep' with predicted rating of 4.19 and genre ['Fiction']
5. 'Manhattan Stories From the Heart of a Great City' with predicted rating of 4.18 and genre ['Manhattan (New York, N.Y.)']
6. 'Mere Christianity' with predicted rating of 4.15 and genre ['Religion']
7. 'A Connecticut Yankee in King Arthur's Court' with predicted rating of 4.14 and genre ['Fiction']
8. 'Harry Potter and The Sorcerer's Stone' with predicted rating of 4.14 and genre ['Juvenile Fiction']
9. 'The Daughter of Time' with predicted rating of 4.12 and genre nan
10. 'House of Mirth' with predicted rating of 4.11 and genre ['Fiction']
```

Рисунок 3.10 – Список рекомендацій для користувача №1

Розглянемо рекомендації:

1. «The Giver»: Ця книга належить до жанру «Juvenile Fiction», що відповідає інтересам користувача. Враховуючи високий прогнозований рейтинг (4.61), це, безумовно, вдалий вибір.

2. «Man's Search for Meaning»: Ця книга може бути цікавою для користувача, оскільки вона відноситься до теми екзистенційної психотерапії, що може перетинатися з його інтересом до релігії. Прогнозований рейтинг 4.26 також позитивний.

3. «The Lion, the Witch and the Wardrobe»: Жанр цієї книги не вказаний, але це добре відома дитяча книга, що також відповідає його інтересу до «Juvenile Fiction». Прогнозований рейтинг 4.20 вказує на те, що він може насолодитися цією книгою.

4. «The Big Sleep»: Ця книга входить до жанру «Fiction», який найбільше цінується користувачем. З прогнозованим рейтингом 4.19, ця рекомендація є дуже влучною.

5. «Manhattan Stories From the Heart of a Great City»: Жанр цієї книги не відповідає безпосередньо жодному з інтересів користувача, але може бути цікавим, якщо він цікавиться географічно зорієнтованою літературою або історіями міст. Прогнозований рейтинг 4.18 вказує на те, що користувач може виявити цікавість до цієї книги.

Загалом, рекомендаційна система здебільшого згенерувала влучні рекомендації, хоча одна з них може бути менш відповідна інтересам користувача, залежно від його індивідуальних вподобань.

2. Наступним було обрано користувача №2, який написав близько 1000 відгуків. Цей користувач також виявляє сильну зацікавленість у художній літературі, оскільки відгуки на книги в жанрі «Fiction» значно переважають над іншими – 508 відгуків. Це вказує на його пристрасть до історій, що включають різноманітні сюжети та персонажів (рис 3.11).

Його інтерес до драматичних творів та детективних історій (18 і 16 відгуків відповідно) свідчить про те, що він цінує напружені, емоційні сюжети та складні загадки.

Його зацікавленість в жанрах "Juvenile Fiction" та "Young Adult Fiction" (16 і 6 відгуків відповідно) може вказувати на те, що він має дітей або підлітків, для яких він шукає книги, або просто він сам любить ці жанри.

['Fiction']	508
['Drama']	18
['Detective and mystery stories']	16
['Juvenile Fiction']	16
['England']	11
['Literary Criticism']	9
['Great Britain']	9
['Audiobooks']	8
['Biography & Autobiography']	7
['Young Adult Fiction']	6

Рисунок 3.11 – Жанри книг і кількість відгуків, які написав користувач №2

Рекомендаційна система, в свою чергу, згенерувала користувача №2 наступні книги (рис. 3.12):

```
1. 'Man's Search for Meaning' with predicted rating of 5.00 and genre ['Existential psychotherapy']
2. 'A Connecticut Yankee in King Arthur's Court' with predicted rating of 5.00 and genre ['Fiction']
3. 'The Giver' with predicted rating of 5.00 and genre ['Juvenile Fiction']
4. 'Great Gatsby (Everyman)' with predicted rating of 5.00 and genre ['American Dream']
5. 'Great Expectations' with predicted rating of 4.98 and genre ['Fiction']
6. 'Manhattan Stories From the Heart of a Great City' with predicted rating of 4.98 and genre ['Manhattan (New York, N.Y.)']
7. 'The great Gatsby (Leading English literature library)' with predicted rating of 4.93 and genre ['Fiction']
8. 'House of Mirth' with predicted rating of 4.92 and genre ['Fiction']
9. 'The Fellowship of the Rings' with predicted rating of 4.90 and genre nan
10. 'Mere Christianity' with predicted rating of 4.89 and genre ['Religion']
```

Рисунок 3.12 – Список рекомендацій для користувача №2

Розглянемо 5 перших рекомендацій:

1. «Man's Search for Meaning»: Ця книга не відповідає безпосередньо ніякому з вподобань користувача за жанрами, однак її екзистенційні теми можуть викликати інтерес. З урахуванням прогнозованого рейтингу в 5.00, ця рекомендація може бути влучною.

2. «A Connecticut Yankee in King Arthur's Court»: Ця книга відповідає головному інтересу користувача – художній літературі. Прогнозований рейтинг 5.00 свідчить про те, що він, швидше за все, насолодиться читанням цієї книги.

3. «The Giver»: Ця книга входить до жанру «Juvenile Fiction», який також представляє інтерес для користувача. Прогнозований рейтинг 5.00 говорить про високий рівень вподобань, які користувач може мати до цієї книги.

4. «Great Gatsby (Everyman)»: Ця книга може бути цікавою для користувача, оскільки вона відноситься до жанру «American Dream», що може перетинатися з його інтересом до драми та художньої літератури. Прогнозований рейтинг 5.00 також вказує на те, що користувач може виявити цікавість до цієї книги.

5. «Great Expectations»: Ця книга відповідає головному інтересу користувача – художньої літературі. З прогнозованим рейтингом 4.98, ця рекомендація є дуже влучною.

Загалом, рекомендації відповідають інтересам користувача, особливо його головним вподобанням до художньої літератури. Вони включають в себе різноманітність жанрів, що відповідає різноманітності інтересів користувача.

Прогнозовані рейтинги також високі, що вказує на те, що користувач, швидше за все, насолодиться цими книгами.

3. Для різноманітності варіантів, далі було обрано користувача №3. Цей користувач має явну зацікавленість в релігійних текстах, оскільки його відгуки на книги в жанрі «Religion» значно переважають над іншими (198 відгуків). Він цікавиться духовними питаннями, мораллю, і, можливо, богослов'ям (рис 3.13).

['Religion']	198
['History']	49
['Biography & Autobiography']	44
['Psychology']	37
['Social Science']	36
['Philosophy']	34
['Science']	23
['Business & Economics']	22
['Fiction']	19
['Political Science']	17

Рисунок 3.13 – Жанри книг і кількість відгуків, які написав користувач №3

Другим за цікавістю для цього користувача жанром є «History», з 49 відгуками. Він цікавиться минулими подіями, культурами, історичними персонажами, і, можливо, історією релігій.

Третім за цікавістю жанром є «Biography & Autobiography», з 44 відгуками. Цей користувач цікавиться реальними історіями життя людей, особливо тих, хто мав значний вплив на історію або релігію.

За ними ідуть жанри «Psychology» (37 відгуків) та «Social Science» (36 відгуків). Користувач цікавиться розумінням людського поведінка, соціальних структур та взаємодії.

«Philosophy» з 34 відгуками також є важливим жанром для цього користувача, свідчачи про його інтерес до глибоких роздумів про значення життя, істину, мораль та інші філософські питання.

Варто відмітити, що цей користувач має високий інтелектуальний інтерес до широкого спектра тем, з особливим акцентом на релігію, історію та біографії.

Розглянемо рекомендації, які система згенерувала для користувача №3 (рис. 3.14):

```
1. '1984' with predicted rating of 5.00 and genre ['Fiction']
2. 'Fahrenheit 451' with predicted rating of 5.00 and genre ['Book burning']
3. 'The Two Towers' with predicted rating of 5.00 and genre ['Fiction']
4. 'Jane Eyre' with predicted rating of 5.00 and genre ['Literary Criticism']
5. 'Jane Eyre (New Windmill)' with predicted rating of 5.00 and genre ['Charity-schools']
6. 'Pride & Prejudice (New Windmill)' with predicted rating of 5.00 and genre ['Young Adult Fiction']
7. 'Pride and Prejudice' with predicted rating of 5.00 and genre ['Fiction']
8. 'The Hobbit' with predicted rating of 5.00 and genre ['Juvenile Fiction']
9. 'Alice's Adventures in Wonderland' with predicted rating of 5.00 and genre ['Fiction']
10. 'The Fellowship of the Rings' with predicted rating of 5.00 and genre nan
```

Рисунок 3.14 – Список рекомендацій для користувача №3

Згенеровані рекомендації в основному зосереджені на художній літературі, що може бути не цілком влучним, враховуючи вподобання користувача. Хоча користувач відгукнувся на 19 книг в жанрі «Fiction», його основна зацікавленість лежить в жанрах «Religion», «History», «Biography & Autobiography», «Psychology» та «Social Science». Розглянемо аналіз кожної рекомендації:

1. «1984»: Ця книга є класикою художньої літератури, але вона не відповідає основним вподобанням користувача. Однак, оскільки вона розглядає важливі соціальні та політичні теми, користувач може її цінувати.

2. «Fahrenheit 451»: Ця книга також є художньою літературою, але зосереджена на темі цензури та контролю над інформацією, що може викликати інтерес у користувача, який цікавиться соціальними науками.

3. «The Two Towers»: Хоча це художній твір, він включає елементи міфології та релігії, що можуть викликати інтерес користувача, зацікавленого в релігії.

4. «Jane Eyre» та «Jane Eyre (New Windmill)»: Ці дві книги відображають одну й ту ж історію, але в різних виданнях. Художній твір, який включає

елементи соціальної критики, може привернути увагу користувача, але їх присутність двічі в рекомендаціях може бути не доцільною.

Загалом, хоча рекомендації можуть викликати інтерес у користувача, вони не повністю відображають його вподобання. Вони більше зосереджені на художній літературі, ніж на жанрах, що цікавлять користувача, що може вказувати на те, що система рекомендацій може бути покращена.

4. Розглянемо наступного користувача (№4), який написав більше 4 тисяч відгуків. Цей користувач є дуже активним читачем з різноманітними інтересами (рис 3.15). Ймовірніше всього – це літературний експерт. Його найпопулярнішим жанром є художня література, як свідчать 592 відгуки. З історичною літературою він також досить знайомий, залишивши 346 відгуків. Жанр «Juvenile Fiction» (270 відгуків) та «Biography & Autobiography» (268 відгуків) також є центральними в його читацьких інтересах. Можливо, він має дітей або просто любить дитячу літературу. Його інтерес до біографій і автобіографій свідчить про цікавість до життєвих історій реальних людей.

```
['Fiction']                    592
['History']                    346
['Juvenile Fiction']           270
['Biography & Autobiography']  268
['Business & Economics']       232
['Religion']                   228
['Cooking']                    165
['Computers']                  131
['Social Science']             120
['Family & Relationships']     112
```

Рисунок 3.15 – Жанри книг і кількість відгуків, які написав користувач №4

Загалом, цей користувач має дуже різноманітні інтереси, він активно читає і пише відгуки в різних жанрах, що робить його універсальним читачем.

```
1. '1984' with predicted rating of 5.00 and genre ['Fiction']
2. 'The Canterbury Tales' with predicted rating of 5.00 and genre ['Poetry']
3. 'Leaves of Grass' with predicted rating of 5.00 and genre nan
4. 'A Tree Grows in Brooklyn' with predicted rating of 5.00 and genre ['Fiction']
5. 'The Hobbitt, or there and back again; illustrated by the author.' with predicted rating of 5.00 and genre ['Fiction']
6. 'Man's Search for Meaning' with predicted rating of 5.00 and genre ['Existential psychotherapy']
7. 'A Connecticut Yankee in King Arthur's Court' with predicted rating of 5.00 and genre ['Fiction']
8. 'Jane Eyre (Everyman's Classics)' with predicted rating of 5.00 and genre ['Bildungsromans']
9. 'The Giver' with predicted rating of 5.00 and genre ['Juvenile Fiction']
10. 'The Great Gatsby' with predicted rating of 5.00 and genre ['Fiction']
```

Рисунок 3.16– Список рекомендацій для користувача №4

Враховуючи дані про користувача №4, можна припустити, що рекомендації (рис 3.16) виявилися досить влучними.

1. «1984» – художній роман, що відповідає найбільшому інтересу користувача, а також може задовольнити його зацікавленість в соціальних науках і політичній науці.

2. «The Canterbury Tales» – це класичний літературний твір, що містить елементи історії, соціальних наук та релігії, що відповідає інтересам користувача.

3. «Leaves of Grass», хоч і не має вказаного жанру, є класикою американської літератури і могла б викликати інтерес користувача, оскільки він активно читає художню літературу.

4. «A Tree Grows in Brooklyn» – художній твір, що може відповідати зацікавленості користувача в жанрі художньої літератури. Книга також містить елементи соціальних наук і історії.

5. «The Hobbit, or there and back again» – художній твір, який, крім того, що відповідає основному інтересу користувача, є популярним у різних вікових категоріях, включаючи дорослих і дітей, що може відповідати його інтересу до жанру «Juvenile Fiction».

Таким чином, усі ці книги відображають широкий діапазон інтересів користувача і, швидше за все, будуть йому цікаві.

Взявши до уваги, все вищесказане, можна зробити наступні висновки: рекомендаційна система, яка була аналізована, демонструє високу точність враховуючи інтереси і вподобання користувачів, основані на їх історії відгуків. Вона здатна враховувати широкий спектр жанрів і вибирати книги, які найкраще відповідають інтересам конкретного користувача.

Проаналізувавши кілька портретів користувачів і відповідних рекомендацій, видно, що система систематично вибирає книги, що відповідають основним жанрам, зацікавленням і темам, якими цікавиться користувач. Це свідчить про високу точність алгоритму, адже йому вдається виявити та враховувати складні і різноманітні вподобання користувачів.

В загальному підсумку, рекомендаційна система працює досить точно і ефективно враховує вподобання користувачів, що дозволяє робити влучні рекомендації.

Дана рекомендаційна система, хоча і демонструє високу точність для більшості користувачів, має обмеження, які впливають на якість рекомендацій для певних груп користувачів. Специфічно, ці обмеження стосуються користувачів, які зробили менше 200 відгуків, а також книг, на які було написано менше 50 відгуків. Це викликає складнощі в формуванні відповідних і точних рекомендацій для цих користувачів і книг.

Ці обмеження виникають через особливості датасету, що використовується для навчання моделі. Коли кількість відгуків для конкретного користувача або книги нижча за поріг, ці дані можуть відсутні в датасеті, що передається у модель. Така відсутність впливає на здатність моделі точно аналізувати і розуміти вподобання користувачів або оцінювати популярність книг.

Таким чином, для оптимізації роботи рекомендаційної системи важливо розглянути можливість розширення датасету для включення користувачів і книг з меншою кількістю відгуків. Це допоможе поліпшити якість рекомендацій для всіх користувачів і підвищити узагальнюючу здатність моделі.

Висновки

У процесі виконання роботи було успішно розроблено спосіб формування рекомендацій поліграфічних товарів за допомогою методу колаборативної фільтрації для електронної комерції. Використання методу колаборативної фільтрації дозволило створити персоналізовані рекомендації для користувачів, забезпечуючи більш високу ефективність та релевантність пропозицій товарів. Були зібрані та проаналізовані набори даних з відкритих джерел, що стосуються оцінок користувачів поліграфічної продукції, що дало можливість використати ці дані для побудови рекомендаційної системи.

Було проведено аналіз предметної області, оглянуто методи побудови рекомендаційних систем та існуючих програмних рішень. Розроблено спосіб формування рекомендацій поліграфічних товарів за допомогою методу колаборативної фільтрації. Проведено тестування та верифікацію розробленого способу.

Розроблена рекомендаційна система електронної комерції поліграфічної продукції включає функціональну та інформаційну структуру, яка забезпечує ефективну роботу на основі колаборативної фільтрації. Отримані результати підтверджують актуальність теми кваліфікаційної роботи бакалавра та можуть бути використані для подальшого розвитку рекомендаційних систем та їх впровадження в електронну комерцію.

Перелік посилань

1. Adomavicius, G., Tuzhilin, A. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. IEEE Transactions on Knowledge and Data Engineering, 17(6), 2005. С. 734-749.

2. Sarwar, B., Karypis, G., Konstan, J., Riedl, J. Item-based collaborative filtering recommendation algorithms. Proceedings of the 10th International Conference on World Wide Web, 2001. С. 285-295.

3. Koren, Y. Factorization meets the neighborhood: a multifaceted collaborative filtering model. Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2008. 426-434.

4. Zhou, T., Kuscsik, Z., Liu, J. G., Medo, M., Wakeling, J. R., Zhang, Y. C. Solving the apparent diversity-accuracy dilemma of recommender systems 2010

5. Apache Mahout. URL: https://mahout.apache.org/

6. Apache Predictionio. URL: https://predictionio.apache.org/

7. Amazon Personalize. URL: https://aws.amazon.com/personalize/

8. Google Cloud Recommendations. URL: https://cloud.google.com/recommendations

9. Метод колаборативної фільтрації для рекомендаційних сервісів. URL: https://cyberleninka.ru/article/n/metod-kollaborativnoy-filtratsii-dlya-rekomendatelnyh-servisov

10. Матриця схожості користувачів. URL: https://www.researchgate.net/figure/An-example-of-user-similarity-matrix-of-k-RRI_fig1_351013915

11. Вікіпедія. Метод k-найближчих сусідів. URL: https://uk.wikipedia.org/wiki/Метод_k-найближчих_сусідів

12. Ліниве навчання. URL: https://www.analyticsvidhya.com/blog/2023/02/lazy-learning-vs-eager-learning-algorithms-in-machine-learning/

13. Евклідова або Манхеттенська відстань. URL: http://www.andriystav.cc.ua/Downloads/MITER/Lecture_06.pdf

14. Вікіпедія. Приклад k-NN класифікації. URL: https://upload.wikimedia.org/wikipedia/commons/thumb/e/e7/KnnClassification.svg/1920px-KnnClassification.svg.png

15. Вікіпедія. Влучність та повнота. URL: https://uk.wikipedia.org/wiki/Влучність_та_повнота

16. Вікіпедія. Схематичний опис влучності та повноти. URL: https://upload.wikimedia.org/wikipedia/commons/thumb/0/03/Precisionrecall_uk.svg/800px-Precisionrecall_uk.svg.png

17. Вікіпедія. F-міра. URL: https://uk.wikipedia.org/wiki/F-міра

18. Вікіпедія. Чутливість та специфічність. URL: https://uk.wikipedia.org/wiki/Чутливість_та_специфічність

19. Бібліотека Numpy. URL: https://numpy.org/

20. Бібліотека Pandas. URL: https://pandas.pydata.org/

21. Схема роботи бібліотеки «pandas». URL: https://raw.githubusercontent.com/geo-python/site/6dc04d7310309499435f3be2b803e752921e9f29/source/notebooks/L5/img/pandas-structures-annotated.png

22. Бібліотека Surprise. URL: https://surpriselib.com/

23. Бібліотека NLTK. URL: https://www.nltk.org/

24. Visual Studio Code. URL: https://code.visualstudio.com/

25. Visual Studio Code Jupyter Notebooks. URL: https://code.visualstudio.com/docs/datascience/jupyter-notebooks

ДОДАТКИ

Додаток А

Програмний код

Файл main.py

```
import numpy as np # лінійна алгебра
import pandas as pd # обробка даних, CSV файл I/O (напр. pd.read_csv)

import os
for dirname, _, filenames in os.walk('./kaggle/input'):
    for filename in filenames:
        print(os.path.join(dirname, filename))

ratings_df1 = pd.read_csv('./kaggle/input/amazon-books-reviews/Books_rating.csv')
bookinfo_df1 = pd.read_csv('./kaggle/input/amazon-books-reviews/books_data.csv')

ratings_df = ratings_df1
bookinfo_df = bookinfo_df1[['categories', 'authors', 'Title']]
merged_df = ratings_df1.merge(bookinfo_df1, on='Title', how='inner')
merged_df.head()

ratings_df = ratings_df[['Id','User_id','review/text','Title','review/score']]
ratings_df.rename(columns={'Id':'book_id','User_id':'user_id','review/text':'review','Title':'title','review/score':'rating'},inplace=True)



review_counts = ratings_df.groupby('user_id').count()['review']
x = (review_counts > 200) & (review_counts < 2000)
print(x[x].index)



single_reader = merged_df[merged_df['User_id'].isin(x[x].index)].groupby('User_id')['categories']
filtered_genres = single_reader.value_counts(normalize=True)
selected_genre = filtered_genres
print(selected_genre)

# Вибираємо користувачів, які залишили мінімум 200 відгуків

# ratings_df['count_review'] = ratings_df.groupby('user_id')['review'].transform('count')
# ratings_df['title_review'] = ratings_df.groupby('title')['review'].transform('count')
# ratings_df = ratings_df[ratings_df['count_review'] > 200]
# ratings_df = ratings_df[ratings_df['title_review'] > 50]
# ratings_df = ratings_df.sort_values(by=['count_review', 'title_review'], ascending=True)
# print(ratings_df.drop_duplicates(subset='user_id')[['user_id', 'count_review', 'title', 'title_review']].head(100))

x = ratings_df.groupby('user_id').count()['review'] > 200
considerable_users = x[x].index
filtered_rating = ratings_df[ratings_df['user_id'].isin(considerable_users)]



# Вибираємо книги, які мають мінімум 50 відгуків
y = filtered_rating.groupby('title').count()['review']>=50
famous_books = y[y].index
df = filtered_rating[filtered_rating['title'].isin(famous_books)]
```

```
df.head()

df = df.reset_index()
df.head()

from surprise import Dataset, Reader
from surprise import KNNWithMeans
from surprise.model_selection import train_test_split
from surprise import accuracy
import nltk
nltk.download('vader_lexicon')
nltk.download('movie_reviews')

from nltk.sentiment import SentimentIntensityAnalyzer

sia = SentimentIntensityAnalyzer()

# Вираховуємо середнє значення тональності для кожного відгуку
df['sentiment'] = df['review'].apply(lambda x: sia.polarity_scores(x)['compound'])

# В попередній матриці отримали значення від -1 до 1
# Нормалізуємо значення настрою між 0 та 1
df['sentiment'] = (df['sentiment'] - df['sentiment'].min()) / (df['sentiment'].max() - df['sentiment'].min())

df.head()

# Фільтруємо імовірно нелогічні відгуки
# 1. Оцінка 1 або 2, але настрій коментарю більше 60% задоволений
# 2. Оцінка 3 або 5, але настрій коментарю менше 60% задоволений

df = df.loc[~((df.rating.isin([1,2])) & (df['sentiment'] > 0.6))]
df = df.loc[~((df.rating.isin([4,5])) & (df['sentiment'] < 0.6))]

# Комбінуємо рейтинг та оцінку настрою в єдину метрику

df['hybrid_score'] = df['rating'] * df['sentiment']

from collections import defaultdict
def precision_recall_at_k(predictions, k=10, threshold=3):
    """Return precision and recall at k metrics for each user"""

    # First map the predictions to each user.
    user_est_true = defaultdict(list)
    for uid, _, true_r, est, _ in predictions:
        user_est_true[uid].append((est, true_r))

    precisions = dict()
    recalls = dict()
    for uid, user_ratings in user_est_true.items():

        # Sort user ratings by estimated value
        user_ratings.sort(key=lambda x: x[0], reverse=True)
```

```
        # Number of relevant items
        n_rel = sum((true_r >= threshold) for (_, true_r) in user_ratings)

        # Number of recommended items in top k
        n_rec_k = sum((est >= threshold) for (est, _) in user_ratings[:k])

        # Number of relevant and recommended items in top k
        n_rel_and_rec_k = sum(
            ((true_r >= threshold) and (est >= threshold))
            for (est, true_r) in user_ratings[:k]
        )

        # Precision@K: Proportion of recommended items that are relevant
        # When n_rec_k is 0, Precision is undefined. We here set it to 0.
        precisions[uid] = n_rel_and_rec_k / n_rec_k if n_rec_k != 0 else 0

        # Recall@K: Proportion of relevant items that are recommended
        # When n_rel is 0, Recall is undefined. We here set it to 0.
        recalls[uid] = n_rel_and_rec_k / n_rel if n_rel != 0 else 0

    return precisions, recalls

# load data from a file
reader = Reader(line_format='user item rating', sep=',')
data = Dataset.load_from_df(df[['user_id', 'title', 'hybrid_score']], reader=reader)
# Split the data into training and testing sets
trainset, testset = train_test_split(data, test_size=.25)

raw_ratings = data.raw_ratings
raw_ratings[:5]

model = KNNWithMeans(k=50, sim_options={'name': 'cosine', 'user_based': True})
model.fit(trainset)

# Evaluate the model on the test set
from surprise import accuracy
predictions = model.test(testset)
rmse = accuracy.rmse(predictions)
precisions, recalls = precision_recall_at_k(predictions, k=5, threshold=2)
print(sum(prec for prec in precisions.values()) / len(precisions))
print(sum(rec for rec in recalls.values()) / len(recalls))

# Get the user ID for whom you want to make recommendations
user_id = 'A1N1YEMTI9DJ86'

# Get the list of all items (books) in the dataset
items = df['title'].unique()

# Predict the rating the user would give to each item and store in a dictionary
item_ratings = {}
for item in items:
    predicted_rating = model.predict(user_id, item).est
```

```
    item_ratings[item] = predicted_rating

# Sort the items by predicted rating in descending order and select the top 5
top_items = sorted(item_ratings.items(), key=lambda x: x[1], reverse=True)[:10]

# print(top_items)

# Print the top 5 recommended books
for i, item in enumerate(top_items):
    print(f"{i+1}. '{item[0]}' with predicted rating of {item[1]:.2f} and genre {merged_df[merged_df['Title'] == item[0]].iloc[0].categories} {ratings_df[ratings_df['title'] == item[0]].count()['review']}")
```

КВАЛІФІКАЦІЙНА РОБОТА БАКАЛАВРА

СПОСІБ ФОРМУВАННЯ РЕКОМЕНДАЦІЙ ПОЛІГРАФІЧНИХ ТОВАРІВ ЗА ДОПОМОГОЮ МЕТОДУ КОЛАБОРАТИВНОЇ ФІЛЬТРАЦІЇ ДЛЯ ЕЛЕКТРОННОЇ КОМЕРЦІЇ

Виконав:

студент 4 курсу, групи КН-19-1

Машталяр Матвій Богданович

Керівник:

доцент кафедри КН

Багрій Руслан Олександрович

Актуальність

Об'єм даних в Інтернеті з кожним роком невпинно зростає і знаходження потрібної інформації займає все більше часу. В багатьох сферах, в тому числі в електронній комерції, назріла необхідність використання ефективних способів фільтрації інформації. Колаборативна фільтрація є одним із можливих методів, що дозволяє сформувати перелік рекомендацій товарів для веб-систем електронної комерції.

Колаборативна фільтрація – це метод побудови прогнозів (рекомендацій) у рекомендаційних системах, який використовує відомі інтереси (вподобання) групи користувачів для прогнозування невідомих інтересів іншого користувача. Вона є є потужним інструментом для компаній в сфері електронної комерції, щоб рекомендувати клієнтам друковані вироби. За допомогою цього методу компанії можуть використовувати рейтинги клієнтів, уподобання та історію покупок, щоб сформувати індивідуальні рекомендації, які сподобаються окремим особам. Ця технологія революціонізує спосіб взаємодії компаній із клієнтами та допомагає їм максимізувати потенційні продажі.

Мета і задачі роботи

Метою кваліфікаційної роботи бакалавра полягає у розробці способу формування рекомендації пропозицій товарів поліграфічної продукції з використанням методу колаборативної фільтрації для веб-систем електронної комерції, для чого слід вирішити задачі:

1. Провести аналіз предметної області.
2. Отримати набори даних про оцінки користувачів поліграфічної продукції з відкритих джерел.
3. Розробити спосіб формування рекомендацій пропозицій товарів з використанням методу колаборативної фільтрації.
4. Реалізувати спосіб формування рекомендацій пропозицій товарів для використання у інформаційних системах електронної комерції поліграфічної продукції.
5. Провести тестування та верифікацію розробленого способу формування рекомендацій поліграфічних товарів.

Розроблена програмна реалізація способу формування рекомендації пропозицій товарів поліграфічної продукції з використанням методу колаборативної фільтрації має виконувати наступні основні групи функцій:

- збір та організація даних про товари та відгуки користувачів;
- обробка даних та тренування моделі;
- генерація рекомендацій товарів для користувачів.

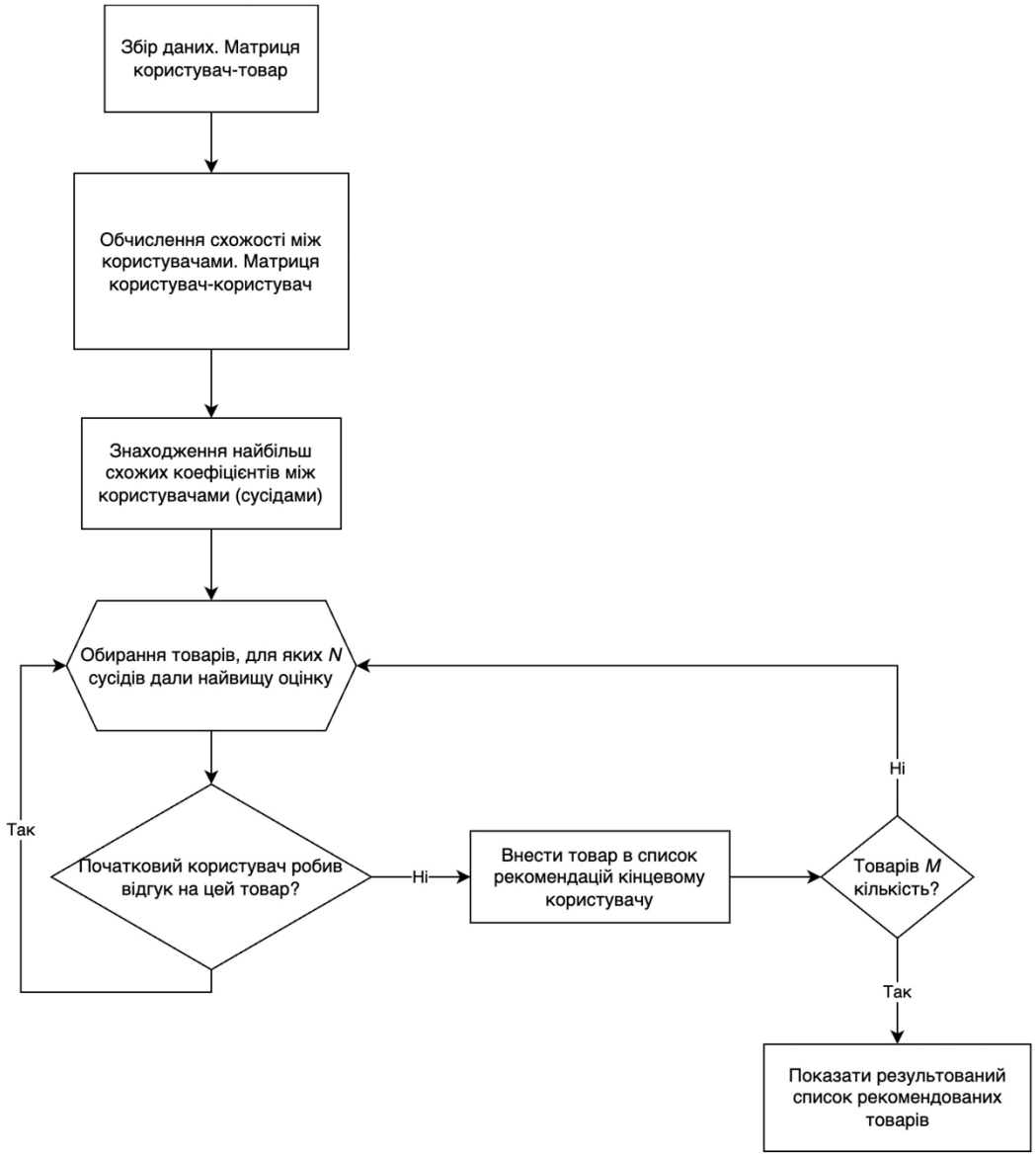


Схема роботи методу колаборативної фільтрації

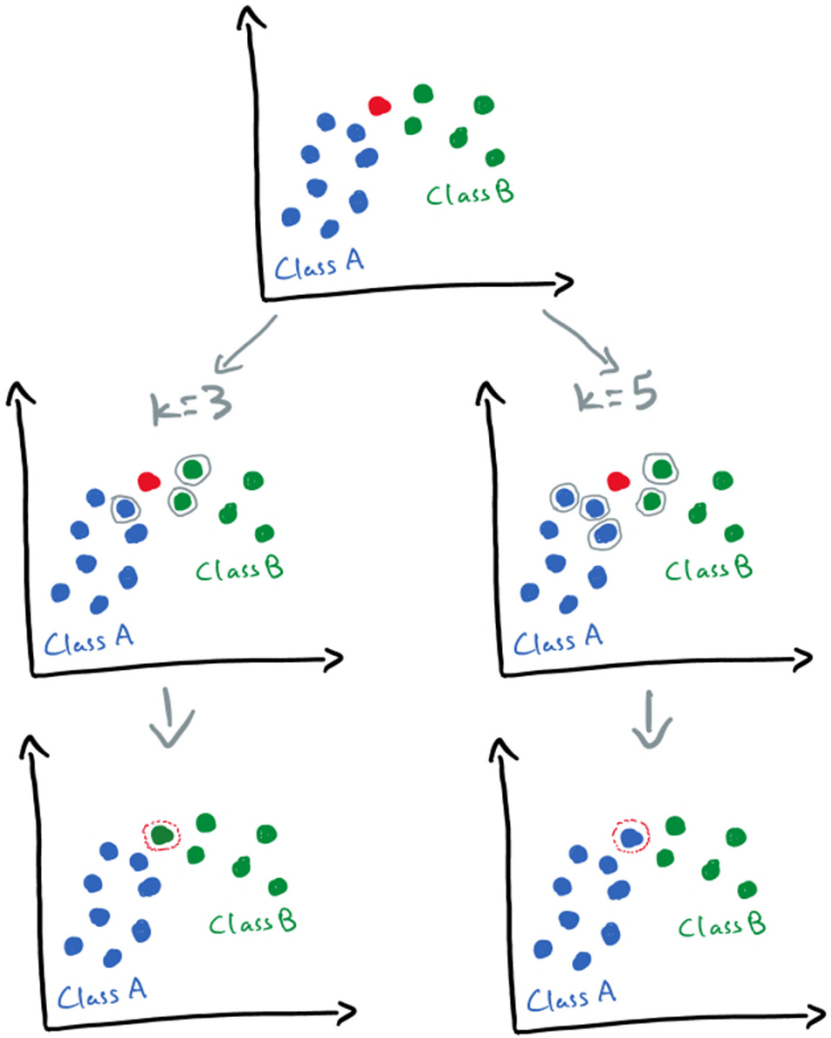


Схема роботи алгоритму k-NN

Математична модель оцінки користувачів на основі їх відгуків

Оцінка $\hat{r}_{ui}$ користувача на книгу:

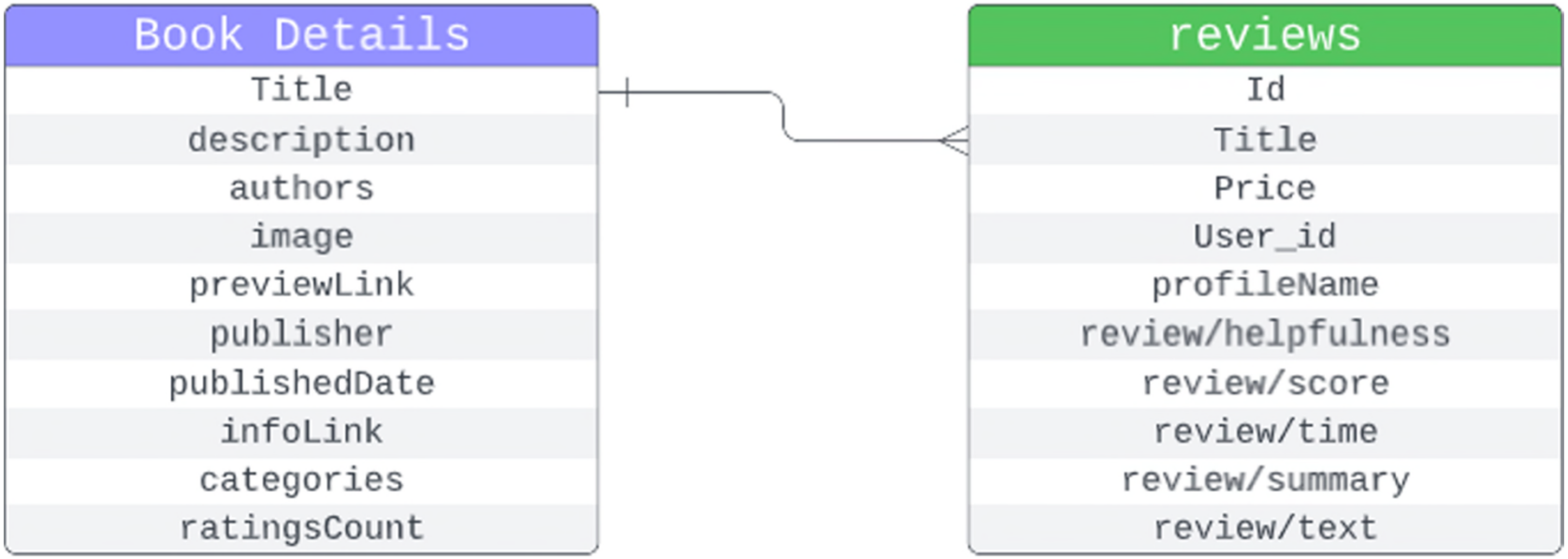
$$\hat{r}_{ui} = \mu_u + \frac{\sum_{v \in N_i^k(u)} \text{sim}(u,v) \cdot (r_{vi} - \mu_v)}{\sum_{v \in N_i^k(u)} \text{sim}(u,v)}$$

де, k – максимальна кількість сусідів, яку треба враховувати для агрегації, $\hat{r}_{ui}$ – оцінка користувача u на елемент i, яка розраховується на основі середньої оцінки відповідного користувача μ_u доданої до вагового середнього оцінок найближчих сусідів, нормалізованого на суму подібностей $\left(\sum_{v \in N_i^k(u)} sim(u,v)\right)$.

Матриця схожості користувачів

	u_1	u_2	u_3	u_4	u_5	u_6	u_7	u_8	u_9	u_{10}	u_{11}	u_{12}
u_1	0	0.1	-0.2	0.4 (①)	0	0	0	0.1	0.3 (②)	-0.2	0.1	-0.1
u_2	0.1	0	0	-0.1	0	0	0	0.2	0.6	0	0	0
u_3	-0.2	0	0	-0.2	0	0	-0.1	0.4	-0.2	-0.1	0	-0.1
u_4	0.4	-0.1	-0.2	0	0.2	0	0.2	0	0	-0.1	0.6	-0.4
u_5	0	0	0	0.2	0	0	0.3	0	0.2	0	0.3	0
u_6	0	0	0	0	0	0	0	0	0	0.3	0	0
u_7	0	0	-0.1	0.2	0.3	0	0	-0.2	0.5	-0.1	0.2	0.1
u_8	0.1	0.2	0.4	0.0	0	0	-0.2	0	-0.2	-0.2	0.1	-0.2
u_9	0.3	0.6	-0.2	0	0.2	0	0.5	-0.2	0	0	0	-0.2
u_{10}	-0.2	0	-0.1	-0.1	0	0.3	-0.1	-0.2	0	0	-0.1	0.4
u_{11}	0.1	0	0	0.6	0.3	0	0.2	0.1	0	-0.1	0	0
u_{12}	-0.1	0	-0.1	-0.4	0	0	0.1	-0.2	-0.2	0.4	0	0

Інфологічна модель датасетів



Результат програмної реалізації способу формування рекомендації пропозицій товарів поліграфічної продукції з використанням методу колаборативної фільтрації

Користувач №1

['Fiction']	74
['Religion']	16
['Juvenile Fiction']	15
['Science']	6
['Biography & Autobiography']	6

Рекомендовані книги:

1. 'The Giver' with predicted rating of 4.61 and genre ['Juvenile Fiction']
2. 'Man's Search for Meaning' with predicted rating of 4.26 and genre ['Existential psychotherapy']
3. 'The Lion, the Witch and the Wardrobe' with predicted rating of 4.20 and genre nan
4. 'The Big Sleep' with predicted rating of 4.19 and genre ['Fiction']
5. 'Manhattan Stories From the Heart of a Great City' with predicted rating of 4.18 and genre ['Manhattan (New York,

Результат програмної реалізації способу формування рекомендації пропозицій товарів поліграфічної продукції з використанням методу колаборативної фільтрації

Користувач №2

['Religion']	198	['Philosophy']	34
['History']	49	['Science']	23
['Biography & Autobiography']	44	['Business & Economics']	22
['Psychology']	37	['Fiction']	19
['Social Science']	36	['Political Science']	17

Рекомендовані книги:

1. '1984' with predicted rating of 5.00 and genre ['Fiction']
2. 'Fahrenheit 451' with predicted rating of 5.00 and genre ['Book burning']
3. 'The Two Towers' with predicted rating of 5.00 and genre ['Fiction']
4. 'Jane Eyre' with predicted rating of 5.00 and genre ['Literary Criticism']
5. 'Jane Eyre (New Windmill)' with predicted rating of 5.00 and genre ['Charity-schools']

Висновки

У рамках виконання кваліфікаційної роботи бакалавра було успішно розроблено спосіб формування рекомендацій поліграфічних товарів за допомогою методу колаборативної фільтрації для електронної комерції. Зокрема, було проведено аналіз предметної області й досліджено сучасні підходи до рекомендації поліграфічних товарів, розглянуто існуючі програмні реалізації за цим напрямком.

Розроблена програмна реалізація способу формування рекомендацій поліграфічних товарів за допомогою методу колаборативної фільтрації для електронної комерції на платформі Jupyter Notebook (Python) виконує наступні основні функції:

- Збір та організація даних про товари та відгуки користувачів.
- Обробка даних та тренування моделі.
- Генерація рекомендацій товарів для користувачів.

У якості засобів розробки було обрано платформу Jupyter Notebook, мову програмування Python, редактор програмного коду Visual Studio Code.

Thu Jun 01 19:09:34 EEST 2023, Петровський Сергій Степанович, Хмельницький національний університет, ХНУ

Anti-Plagiarism v-15.257

Максимальне співпадіння з одним документом 6.0%

Словники перевірки: en_US, ru_RU, ua_UA. **Помилок в документах: 9%**

ID: 114503 Назва: КВАЛІФІКАЦІЙНА РОБОТА БАКАЛАВРА Додано в БД: 2023-06-01 Автора: М.Б. Машталяр Керівники: Р.О. Багрій Консультанти: Опоненти:	Документ		Сумарний збіг по Базі Даних	
	Символи	Лексеми	Символи	Лексеми
	60737	929	4723 (8%)	74 (8%)

Джерело плагіату

ID	Опис	Наявність плагіату в документі	
		Символи	Лексеми

UNICHECK

Ім'я користувача:
Кафедра КН

ID перевірки:
1015376512

Дата перевірки:
01.06.2023 20:01:10 EEST

Тип перевірки:
Doc vs Internet + Library

Дата звіту:
01.06.2023 20:10:58 EEST

ID користувача:
100005671

Назва документа: **КН-19-1 Машталяр**

Кількість сторінок: **59** Кількість слів: **9867** Кількість символів: **73769** Розмір файлу: **2.60 MB** ID файлу: **1015042200**

Виявлено модифікації тексту (можуть впливати на відсоток схожості)

4.92%
Схожість

Найбільша схожість: **1.39%** з джерелом з Бібліотеки (ID файлу: **1014976252**)

4.63% Джерела з Інтернету	122	Сторінка 61
1.83% Джерела з Бібліотеки	63	Сторінка 62

0.39% Цитат

Цитати	2	Сторінка 63
Посилання	1	Сторінка 63

0%
Вилучень

Немає вилучених джерел

Модифікації

Виявлено модифікації тексту. Детальна інформація доступна в онлайн-звіті.

Замінені символи	1
Підозріле форматування	15 сторінок

РІШЕННЯ ЕКСПЕРТНОЇ КОМІСІЇ КАФЕДРИ КОМП'ЮТЕРНИХ НАУК
ПРО ДОПУСК КВАЛІФІКАЦІЙНОЇ РОБОТИ ДО ЗАХИСТУ

Підтверджуємо ознайомлення з результатом звіту подібності щодо роботи, генерованого системою виявлення текстових збігів/ідентичності/схожості:

Назва: *Спосіб формування рекомендацій поліграфічних товарів за допомогою методу колаборативної фільтрації для електронної комерції*

Автор: *студент гр. КН-19-1 Машталяр Матвій Богданович*

Спеціальність: *122 – Комп'ютерні науки*

Освітня програма: *освітньо-професійна*

Науковий керівник: *к.т.н., доц. Багрій Р.О.*

Після аналізу звіту подібності зроблено такий висновок:

№	Висновок	Позначка про відповідність
1	Запозичення, виявлені в роботі, є законними і не є плагіатом. Робота приймається до захисту.	***відповідає***
2	Виявлені запозичення не є плагіатом, розміщені в розділах, які не описують безпосередньо авторське дослідження, але кількість цитат перевищує обсяг, виправданий поставленою метою роботи Робота приймається до захисту, але має бути відкоригована. Відкоригований варіант має бути поданий на кафедру за 2 дні до захисту, разом із заявою щодо самостійності виконання письмової роботи та ідентичності друкованої та електронної версії роботи	
3	Виявлені запозичення не є плагіатом, але частково розміщені в розділах, які описують безпосередньо авторське дослідження, а кількість цитат перевищує обсяг, виправданий поставленою метою роботи. В зв'язку з цим мета роботи та поставлені завдання не були досягнені. Робота може бути допущена до захисту (наступного року) після того як буде відкоригована та допрацьована і успішно пройде повторну перевірку на академічний плагіат.	
4	Робота містить навмисні текстові спотворення, передбачувані спроби укриття запозичень або інші прояви академічного плагіату. Робота містить фабрикацію або фальсифікацію даних. Робота не допускається до захисту.	

Підтвердження:

Запозичення, виявлені в роботі, є законними і не є плагіатом, оскільки:

1) за програмою Anti-Plagiarism виявлені 6% є фрагментарними – містять поширені конструкції, загальновідомі терміни, скорочення та визначення.
2) За програмою UNICHECK виявлені 4.92%, що є запозиченнями, які розміщені в розділах аналізу існуючих технологій та прототипів, які не описують безпосередньо авторське дослідження і не стосуються результатів роботи.

Сумарний обсяг всіх запозичень, визначений системою виявлення збігів/ідентичності/схожості, складає 6% і 4.92% відповідно, що, з урахуванням наведених обґрунтувань, відповідає характеру наукового дослідження і свідчить на користь кваліфікаційної роботи.

Керівник роботи _____ Руслан Багрій

Гарант ОП _____ Олександр Мазурець

Завідувач кафедри КН _____ Олександр Бармак

ХМЕЛЬНИЦЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
МОН УКРАЇНИ

Кафедра комп'ютерних наук

ВІДГУК НАУКОВОГО КЕРІВНИКА
на кваліфікаційну роботу бакалавра

студента *гр. КН-19-1 Машталяра Матвія Богдановича*

за темою *Спосіб формування рекомендацій поліграфічних товарів за допомогою методу колаборативної фільтрації для електронної комерції*

1. Актуальність теми

Актуальність теми достатньо обґрунтована, оскільки спосіб формування рекомендацій поліграфічних товарів за допомогою методу колаборативної фільтрації дозволяє споживачам зорієнтуватися в широкому асортименті товарів і знайти ті, які відповідають їхнім інтересам і вимогам. Особливістю теми є використання одного з найбільш популярних методів рекомендаційних систем – методу колаборативної фільтрації, який базується на аналізі історичних даних про вибір користувачів.

2. Відповідність роботи предметній області Стандарту спеціальності 122 Комп'ютерні науки

Теми кваліфікаційної роботи "Спосіб формування рекомендацій поліграфічних товарів за допомогою методу колаборативної фільтрації для електронної комерції" відповідає предметній області спеціальності 122 Комп'ютерні науки та вимогам до кваліфікаційної роботи бакалавра, оскільки метою роботи є розробка рекомендаційної системи для формування пропозицій товарів поліграфічної продукції. При вирішенні поставленої задачі використано методи збору та аналізу інформації, методи машинного навчання та методи розробки рекомендаційних систем.

3. Професійні та особистісні якості бакалавра

Машталяр М. Б. під час роботи над кваліфікаційною роботою бакалавра продемонстрував розуміння теорії та практичних аспектів використання систем фільтрації інформації, що дало можливість розробити рекомендаційну систему для формування пропозицій товарів поліграфічної продукції.

4. Ступінь самостійності під час виконання кваліфікаційної роботи

Робота виконана самостійно, академічного плагіату не виявлено, стосовно всіх запозичень наведено відповідні посилання на джерела.

5. Ступінь оволодіння методами дослідження

При реалізації кваліфікаційної роботи показала достатній рівень компетентностей та володіння необхідними інструментами та обладнанням, методами, методиками та технологіями предметної області комп'ютерних наук.

6. Повнота та якість розкриття теми роботи

Тема роботи повністю розкрита, виконані усі поставлені задачі та розроблена програмна реалізація для підтвердження запропонованого способу формування рекомендацій поліграфічних товарів.

7. Логічність, послідовність, аргументованість, літературна грамотність викладення матеріалу

Викладення матеріалу логічне, послідовне та аргументоване. Мова і стиль викладення кваліфікаційної роботи відповідають стандартам, що забезпечує доступність сприймання матеріалу і відповідає вимогам до сучасних наукових робіт.

8. Можливість практичного застосування кваліфікаційної роботи бакалавра, окремих її частин

Розроблений у роботі способу формування рекомендацій поліграфічних товарів може знайти широке застосування в електронній комерції, онлайн-магазинах, друкованих виданнях або будь-якому середовищі, де присутня поліграфічна продукція і потреба в персоналізованих рекомендаціях.

9. Висновок про можливість допуску кваліфікаційної роботи бакалавра до захисту, на яку оцінку заслуговує робота

*Враховуючи високий рівень виконання та забезпечення усіх необхідних вимог, робота може бути допущена до захисту. Рекомендована оцінка «**відмінно**».*

Керівник _____ *к.т.н., доц. Руслан Багрій*

ХМЕЛЬНИЦЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
МОН УКРАЇНИ

Кафедра комп'ютерних наук

РЕЦЕНЗІЯ

на кваліфікаційну роботу бакалавра

студента *гр. КН-19-1 Машталяр Матвій Богданович*
за темою: *Спосіб формування рекомендацій поліграфічних товарів за допомогою методу колаборативної фільтрації для електронної комерції*

1. Актуальність обраної теми

Об'єм даних в Інтернеті з кожним роком невпинно зростає і знаходження потрібної інформації займає все більше часу. В багатьох сферах, в тому числі в електронній комерції, назріла необхідність використання ефективних способів фільтрації інформації. Колаборативна фільтрація є одним із можливих методів, що дозволяє сформувати перелік рекомендацій товарів для веб-систем електронної комерції.

2. Повнота розкриття мети та завдань роботи

Під час виконання кваліфікаційної роботи бакалавра було реалізовано спосіб формування рекомендацій поліграфічних товарів за допомогою методу колаборативної фільтрації, що відповідає меті та завданням кваліфікаційної роботи і розкриває їх повною мірою.

3. Зміст кожного розділу роботи

Записка кваліфікаційної роботи складається з трьох розділів. Перший розділ охоплює аналіз предметної області і формує постановку задачі. Другий розділ досліджує спосіб формування рекомендацій пропозицій товарів для інформаційної системи електронної комерції. Також в цьому розділі визначається потрібний функціонал системи, вибираються інструменти для розробки. Третій розділ зосереджується на описі програмної реалізації та тестуванні методу.

4. Оцінка розробленої інформаційної системи, її практична цінність

Розроблений спосіб формування рекомендацій поліграфічних товарів може бути застосований для накопичення інформації про поліграфічні товари та їх відповідність до кожного користувача і отримання відповідних рекомендацій.

5. Якість оформлення кваліфікаційної роботи бакалавра

Записка якісно оформлена, відповідно до встановлених вимог. Вірно та зрозуміло написана, з виразною структурою і логічною послідовністю представлення матеріалу.

6. Недоліки кваліфікаційної роботи бакалавра

Рекомендовано розглянути можливість інтеграції способу формування рекомендацій поліграфічної продукції з інформаційними системами електронної комерції.

7. Загальний висновок (допускається чи не допускається до захисту), та оцінка на яку заслуговує кваліфікаційна робота.

Враховуючи рівень виконання та забезпечення усіх необхідних вимог, робота може бути допущена до захисту. Рекомендована оцінка відмінно

Рецензент к.ф.-м.н., доц. к. ВМКЗ Ганський А.О.