

**МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ ЕКОНОМІЧНИЙ УНІВЕРСИТЕТ
ІМЕНІ ВАДИМА ГЕТЬМАНА
Навчально-науковий інститут
«Інститут інформаційних технологій в економіці»
кафедра інформаційних систем в економіці**

галузь знань	12 Інформаційні технології
спеціальність	122 Комп'ютерні науки
освітня програма	Системи штучного інтелекту

форма навчання: денна

КВАЛІФІКАЦІЙНА МАГІСТЕРСЬКА РОБОТА

на тему:

**РОЗРОБКА ІНТЕЛЕКТУАЛЬНОЇ СИСТЕМИ ДЛЯ ІНТЕРНЕТ-
МАГАЗИНУ МОЛОДІЖНОГО ОДЯГУ**

здобувача **Стаєцького Миколи Олеговича**

Науковий керівник: к.е.н.
доцент

Лозовик Ю.М.

**Кваліфікаційна магістерська робота
допущена до захисту в
Екзаменаційній комісії з атестації
здобувачів вищої освіти
в.о. завідувача кафедри: к.е.н., доцент
Тішков Б.О.**

**КИЇВ 2023
Міністерство освіти і науки України**

Київський національний економічний університет імені Вадима Гетьмана
Навчально-науковий інститут
«Інститут інформаційних технологій в економіці»
кафедра інформаційних систем в економіці

галузь знань 12 «Інформаційні технології»
спеціальність 122 «Комп'ютерні науки»
Освітня програма «Системи штучного інтелекту»

ПОГОДЖЕНО:

Керівник проектної групи (гарант)
освітньо-професійної програми
«Системи штучного інтелекту»

_____ Рамазанов С.К.
“ ____ ” _____ 2023 р.

ЗАТВЕРДЖУЮ:

В.о. завідувача кафедри
_____ Тішков Б.О.

“ ____ ” _____ 2023 р.

ІНДИВІДУАЛЬНЕ ЗАВДАННЯ

Стаєцького Миколи Олеговича

денної форми навчання
на підготовку кваліфікаційної магістерської роботи
на тему: **«Розробка інтелектуальної системи для інтернет-магазину
молодіжного одягу»**

Тему затверджено наказом ректора від «20» лютого 2023 р. № 333-ст.

Кваліфікаційна магістерська робота виконується на матеріалах

наукових публікацій, інтернет-ресурсів, технічної документації

План кваліфікаційної магістерської роботи
аналіз існуючих підходів щодо розробки і впровадження інтелектуальних
модулів в системи інтернет-комерції
характеристика об'єкта дослідження та постановка задачі

Розділ III розробка проектних рішень для предметної області
Об'єкт дослідження процеси забезпечення ефективної роботи інтернет-
магазину

Предмет дослідження сукупність теоретичних, методичних і прикладних
аспектів з моделювання, проектування та програмування інтелектуальних
модулів для покращення роботи інтернет-магазину
систематизація і узагальнення сучасного досвіду з розробки нових рішень з
покращення роботи інтернет-магазину на основі модулів штучного
інтелекту та розробка власного прототипу інтелектуальної системи для
інтернет-магазину молодіжного одягу.

**Конкретні завдання, які студент повинен виконати для досягнення
поставленої мети:**

У розділі I необхідно провести аналіз існуючих підходів щодо розробки і впровадження інтелектуальних модулів в системи інтернет-комерції; навести перелік існуючих інструментів і програмних засобів для розробки систем штучного інтелекту, необхідних для покращення роботи інтернет-магазину.

У розділі II слід навести характеристика об'єкта дослідження; розглянути структуру інтелектуальної системи інтернет-магазину молодіжного одягу; описати методи, моделі і технології оброблення даних щодо вибору супутніх, взаємодоповнюючих і взаємозалежних товарів; розробити рекомендації щодо удосконалення існуючих або створення нових актуальних інтелектуальних модулів для покращення роботи інтернет-магазину

У розділі III здійснити проектування інформаційної системи; проаналізувати інформаційну систему; розробити технічні та організаційно-інформаційні механізми для впровадження запропонованих модулів інтелектуальної системи в загальну інформаційну систему інтернет-магазину

Завдання підготував

науковий керівник _____

Лозовик Юрій Миколаєвич

“ _____ ” _____ 2023 р.

Завдання одержав

студент _____

Стасцький Микола Олегович

“ _____ ” _____ 2023 р.

АНОТАЦІЯ

кваліфікаційної магістерської роботи

студента 6 курсу *Стаєцького Миколи Олеговича*, виконану на тему:
**Розробка інтелектуальної системи для інтернет-магазину молодіжного
одягу»**

Кваліфікаційна магістерська робота присвячений актуальній проблемі створення рекомендаційної системи для інтернет-магазину молодіжного одягу.

Кваліфікаційна магістерська робота складається з трьох розділів, логічно пов'язаних між собою.

Перший розділ є теоретичним. В ньому наведено аналіз існуючих підходів до оброблення природної мови та видачі рекомендацій, наводиться перелік існуючих інструментів і програмних засобів для формування рекомендованих товарів в різних сферах.

Другий розділ є аналітичним, в ньому надається характеристика об'єкта дослідження, розглядається структура інтелектуальної системи інтернет-магазину молодіжного одягу, описуються методи, моделі і технології оброблення природної мови і формування рекомендацій.

Третій розділ – конструктивний. Змодельовано процес формування рекомендацій на основі подібного опису, запропоновано користувацький інтерфейс для роботи із системою, наведено інформаційне, програмне та технічне забезпечення для системи, зроблена оцінка ефективності впровадження системи.

Висновки містять підведення підсумків з виконання, можливі шляхи розширення та покращення розробленої підсистеми та рекомендації щодо подальшого розвитку та впровадження системи.

ABSTRACT

The qualifying master's thesis is devoted to the actual problem of creating a recommendation system for an online store of youth clothing.

The qualifying master's thesis consists of three logically connected sections.

The first chapter is theoretical. It provides an analysis of existing approaches to natural language processing and issuing recommendations, lists existing tools and software for generating recommended products in various areas.

The second section is analytical, it provides a description of the research object, examines the structure of the intellectual system of the youth clothing online store, describes the methods, models and technologies of natural language processing and the formation of recommendations.

The third section is constructive. The process of forming recommendations based on such a description is modeled, a user interface for working with the system is proposed, information, software and technical support for the system are provided, and an assessment of the effectiveness of the system implementation is made.

The conclusions contain a summary of the implementation, possible ways to expand and improve the developed subsystem, and recommendations for further development and implementation of the system.

РЕФЕРАТ

Кваліфікаційна магістерська робота містить 60 сторінок, 3 таблиці, 26 рисунків, список літератури з 33 найменувань, 1 додаток.

РОЗРОБКА ІНТЕЛЕКТУАЛЬНОЇ СИСТЕМИ ІНТЕРНЕТ- МАГАЗИНУ МОЛОДІЖНОГО ОДЯГУ

Об'єкт дослідження: видача рекомендованих товарів у інтернет-магазині.

Предметом дослідження є система штучного інтернет-магазину молодіжного одягу.

Мета і завдання. Проєктування та розробка системи штучного інтернет-магазину молодіжного одягу.

Завданнями кваліфікаційної магістерської роботи є наступні: дослідити основні відомості щодо технології оброблення природної мови та підходів до видачі рекомендованих товарів; здійснити аналіз найбільш поширених інструментів; розробити модель процесу формування рекомендацій; розробити рішення щодо забезпечувальних підсистем системи штучного інтелекту для інтернет-магазину одягу.

Теоретична, методична та практична значущість отриманих результатів. В роботі досліджено предметну область щодо застосування технології оброблення природної мови для видачі рекомендацій щодо схожих товарів у інтернет-магазині одягу. Систематизовано методи і моделі, які застосовують для видачі рекомендацій, дана характеристика аналогічним інтелектуальним системам. Змодельовано процес формування рекомендацій, запропоновано користувацький інтерфейс для роботи із системою, наведено інформаційне, програмне та технічне забезпечення для інтелектуальної системи інтернет-магазину молодіжного одягу, зроблена оцінка ефективності впровадження системи.

Одержані результати можуть бути використані для видачі рекомендованих товарів в e-commerce.

Практичне значення полягає у розробленні рекомендаційної системи, призначеної для інтернет-магазину молодіжного одягу. Система може бути корисною для будь-якої аналогічних предметної області.

обка природної мови, User-Based Filtering, Content-Based Filtering, векторизація, класифікація, Natural Language Processing

Рік захисту кваліфікаційної магістерської роботи 2023 р.

[illegible]

ВСТУП

Актуальність проблеми

Видача рекомендацій є активною сферою, що швидко розвивається, зі значними досягненнями та підвищеним інтересом як з боку наукових кіл, так і підприємців. Ключовою технологією для розуміння що саме може зацікавити користувача є обробка природньої мови – Natural Language Processing (NLP).

Технологія NLP значно розширила можливості розуміння природньої мови в різних предметних галузях. Однією основних галузей застосування технології є електронна комерція. Застосування рекомендаційних систем дозволяє підвищувати ефективність торгівлі, збільшувати конверсію, та допомагає в просуванні товарів

Наразі, ринок переповнений подібними системами, проте адаптування існуючих систем не завжди призводить до бажано ефекту. Особливо в умовах “холодного старту” , коли статистичних відомостей про поведінку користувачів недостатньо для прогнозування їх поведінки.

Саме тому створення інтелектуальної системи для інтернет-магазину молодіжного одягу є досить актуальним.

Об’єкт дослідження: видача рекомендованих товарів.

Предметом дослідження є система штучного інтелекту для видачі рекомендованих товарів в інтернет-магазині молодіжного одягу.

Мета і завдання. Проєктування та розробка системи штучного інтелекту інтернет-магазину молодіжного одягу.

Для досягнення мети необхідно виконати такі завдання.

. Дослідити основні відомості щодо технології оброблення природньої мови та підходів до видачі рекомендованих товарів.

2. Здійснити аналіз найбільш поширених інструментів.

. Розробити модель процесу формування рекомендацій.

. Розробити рішення щодо забезпечувальних підсистем системи штучного інтелекту для інтернет-магазину одягу.

Матеріал дослідження. Робота виконана на основі наукових публікацій, та публікацій з мережі інтернет

Наукова новизна. Запропонована інтелектуальна система дозволить забезпечувати відвідувачів сайту якісними пропозиціями щодо схожих за описом товарів, при цьому не потребуючи великої кількості вхідних даних, що є суттєвою перевагою при “холодному старті”.

Практичне значення. Практичне значення полягає у розробленні рекомендаційної системи молодіжного одягу. Вона може бути корисною для будь-якого інтернет-магазину аналогічної предметної області.

Структура роботи. Робота складається за вступу, трьох розділів, висновків, списку використаних джерел і додатків. Перший розділ є теоретичним. В ньому наведено аналіз задачі, що вирішується, а також наводиться перелік існуючих програмних засобів для підтримки прийняття рішень в даній сфері. Другий розділ є аналітичним і присвячений дослідженню об’єкта дослідження і обґрунтування методів та технологій для вирішення поставлених задач, а також обґрунтуванню. Третій розділ – конструктивний. Наведено інформаційне, програмне та технічне забезпечення для системи Розглянуті питання оброблення текстів, проєктування інтерфейсу користувача та зроблена оцінка ефективності застосування системи.

РОЗДІЛ 1. ТЕОРЕТИЧНИЙ РОЗДІЛ. ДОСЛІДЖЕННЯ ТА АНАЛІЗ ПІДХОДІВ ДО СТВОРЕННЯ ПРЕДМЕТНОЇ ОБЛАСТІ СШІ

Дослідження предметної області. Збір інформації та вивчення матеріалів з теми

Стрімкий розвиток інтернет-продажів останні роки створив проблему перевантаження інформацією, яка перешкоджає своєчасному доступу до цікавих елементів в мережі. Хоча інформаційно-пошукові системи, такі як Google, є частковим рішенням, ці системи не можуть зіставити доступний вміст з інтересами та вподобаннями користувача

Результат цієї відсутності пріоритезації та персоналізації призвело до збільшення попиту на системи рекомендацій. Рекомендація схожих товарів є важливою частиною інтернет-продажів, оскільки допомагає покупцям знайти інші товари, які можуть їх зацікавити.

Хоча було надано кілька визначень систем рекомендацій, ми могли б визначити їх як системи фільтрації інформації, які вирішують проблему перевантаження інформацією. Вони здатні фільтрувати інформацію відповідно до опису елементів або спостережуваної поведінки самого користувача.

Враховуючи тенденцію до зростання e-commerce, актуальною є задача створення інтелектуальної системи, яка відслідковуватиме, що саме може бути цікавим користувачам інтернет-магазину

Особливо критичним для успішності бізнесу є розміщення товару у мережі інтернет, якщо вашою цільовою аудиторією є так зване “Покоління Зет” (в розмовній мові більш відомі як «зумери»). До цього покоління відносять людей, які народилися між другою половиною дев'яностих та другою половиною двохтисячних (1996—2010 роки). Те ще попередні покоління вважали “технологіями майбутнього”, зумери вважають невід’ємною частиною свого життя. Зважаючи на останні дослідження (рис. 1.1), з кожним поколінням

все більше і більше людей схилиються до онлайн-торгівлі товарами.

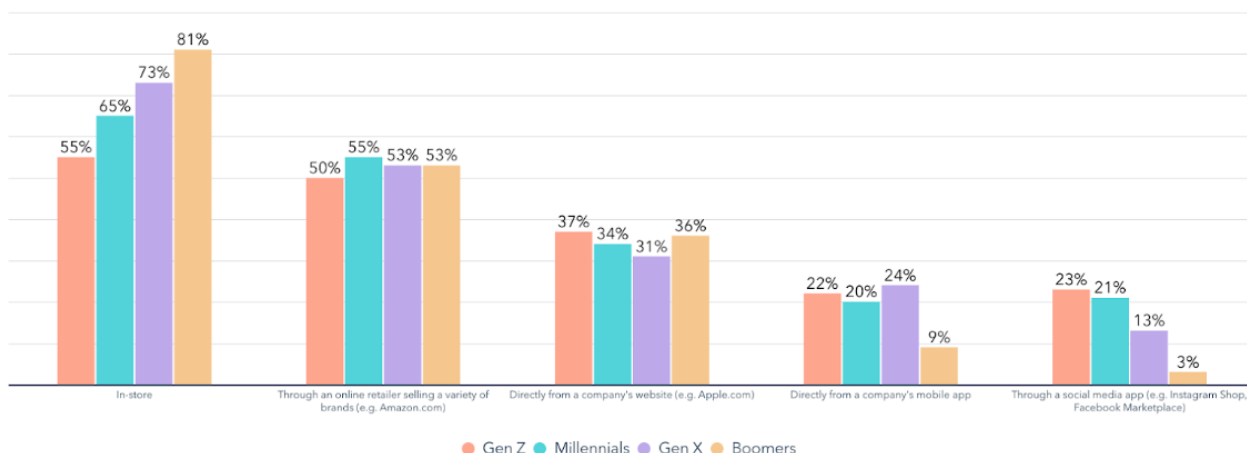


Рисунок 1.1 – Основні джерела покупки товарів для різних поколінь

Алгоритми машинного навчання в системах рекомендацій зазвичай класифікуються на дві основні категорії — методи фільтрації на основі вмісту та методи колаборативної фільтрації, хоча сучасні системи рекомендацій поєднують обидва підходи.

Фільтрування на основі вмісту: цей підхід рекомендує продукти на основі атрибутів і характеристик самих елементів. Він зосереджений на узгодженні уподобань користувача з функціями предмета. Наприклад, якщо клієнт виявив інтерес до певного бренду чи категорії, система запропонує схожі товари. Прикладом фільтрації на основі вмісту є розділ «Рекомендовано для вас» на веб-сайтах електронної комерції, як-от eBay.

Фільтрування на основі вмісту має багато переваг порівняно із колаборативним фільтруванням, зокрема:

Щоб почати давати рекомендації, не потрібні дані інших користувачів. На відміну від спільної фільтрації, фільтрації на основі вмісту не потрібні дані інших користувачів для створення рекомендацій. Після того, як користувач здійснив пошук і переглянув кілька товарів і/або зробив кілька покупок, система фільтрації на основі вмісту може почати давати відповідні рекомендації. Це робить його ідеальним для підприємств, які не мають величезної кількості

користувачів для вибірки. Це також добре працює для продавців, які мають багато користувачів, але невелику кількість взаємодій користувачів у певних категоріях або нішах.

- Рекомендації дуже актуальні для користувача. Рекомендатори на основі вмісту можуть бути повністю пристосовані до інтересів користувача, включаючи рекомендації щодо нішевих елементів, оскільки метод базується на зіставленні характеристик або атрибутів об'єкта бази даних із профілем користувача. Фільтрування на основі вмісту також є цінним для підприємств із великими бібліотеками, що містять один тип продукту, наприклад смартфони, де рекомендації мають базуватися на багатьох окремих функціях.

- Рекомендації прозорі для користувача. Високорелевантні рекомендації створюють відчуття відкритості для користувача, підвищуючи рівень його довіри до запропонованих рекомендацій. Для порівняння, із спільною фільтрацією ймовірні випадки, коли користувачі не розуміють, чому вони бачать певні рекомендації. Наприклад, скажімо, група користувачів, які придбали парасольку, також придбали пуховики. Спільна система може рекомендувати пухові пуховики іншим користувачам, які купили парасолі, але не цікавляться ними та ніколи не переглядали та не купували цей продукт.

- Ви уникаєте проблеми «холодного запуску». Колаборативна фільтрація створює потенційний сценарій «холодного запуску», коли на новому веб-сайті чи спільноті мало нових користувачів і відсутні зв'язки користувачів. Хоча фільтрація на основі вмісту потребує певних початкових вхідних даних від користувачів, щоб почати давати рекомендації, якість ранніх рекомендацій, як правило, краща, ніж система спільної роботи, яка вимагає додавання та кореляції мільйонів точок даних перед оптимізацією.

- Системи фільтрації на основі вмісту, як правило, легше створити. Наука про дані, що стоїть за системою фільтрації на основі вмісту, є відносно простою порівняно з системами спільної фільтрації, призначеними для імітації рекомендацій від користувача до користувача. Реальна робота у фільтрації на

основі вмісту полягає у призначенні атрибутів.

Колаборативна фільтрація: ця техніка аналізує поведінку та вподобання користувачів, щоб рекомендувати елементи на основі схожості з іншими користувачами. Він рекомендує продукти, які придбали або сподобалися користувачам зі схожими смаками. Прикладами систем рекомендацій спільної фільтрації є рекомендації Amazon «Клієнти, які купили це, також купили» та рекомендації фільмів Netflix.

Гібридні системи рекомендацій: ці системи поєднують численні методи рекомендацій, щоб надавати більш точні та різноманітні пропозиції. Вони використовують підходи до фільтрації на основі співпраці та контенту для підвищення якості рекомендацій. Гібридні системи зазвичай використовуються такими платформами, як Spotify, які пропонують музику на основі комбінації поведінки користувача та атрибутів елементів.

Також інколи окремо виділяють кілька альтернативних систем рекомендацій, які зазвичай використовуються в онлайн-магазинах для персоналізації досвіду покупок для клієнтів. До них належать:

Інтелектуальний аналіз правил асоціації: ця техніка визначає моделі та зв'язки між предметами, які часто купують разом. Наприклад, якщо клієнти часто купують певну сорочку та штани разом, система може порекомендувати штани тому, хто купує сорочку. Видобуток правил асоціації використовується в різних онлайн-магазинах, зокрема для цілей перехресних продажів і вищих продажів.

Фільтрування на основі демографічних показників: цей метод рекомендує продукти на основі демографічної інформації, такої як вік, стать, місцезнаходження чи інші відповідні фактори. Сегментуючи клієнтів на групи зі схожими характеристиками, система може надавати індивідуальні рекомендації. Роздрібні продавці, такі як Sephora, часто використовують фільтрацію на основі демографічних показників, щоб надавати цільові пропозиції щодо продуктів.

Системи контекстних рекомендацій: ці системи враховують контекстні

фактори, як-от час, місцезнаходження, погода чи діяльність користувача, щоб надавати рекомендації. Наприклад, програма доставки їжі може пропонувати найближчі ресторани на основі поточного місцезнаходження користувача. Контекстні рекомендації підвищують релевантність і взаємодію з користувачем, враховуючи безпосередній контекст.

Кожна система має свої сильні сторони та обмеження, і вибір того, яку використовувати, залежить від конкретних вимог і доступних даних інтернет-магазину.

2. Аналіз існуючих СШІ й інтелектуальних систем предметної області

Штучний інтелект (ШІ) використовується у онлайн магазинах для покращення користувацького досвіду та оптимізації рекомендацій. Сучасний ринок пропонує велику кількість різноманітних інструментів. Ось декілька прикладів, застосування ШІ в онлайн магазинах:

Алгоритм Amazon — це система рекомендацій, яка аналізує товари, опубліковані користувачами, і визначає переваги покупок окремих клієнтів. Він вибирає рекомендовані продукти для кожного користувача на основі їхніх попередніх покупок, взаємодії та оцінок і поєднує їх із подібними товарами, які переглядають користувачі зі схожими вподобаннями та інтересами (рис.1.2).

Customers Who Bought This Item Also Bought



Рисунок 1.2 – Рекомендаційна система “ті хто купили цей товар, також купують це”

еВау розробив модель рекомендацій, засновану на техніці обробки природної мови (NLP), зокрема на моделі BERT. Ця модель під назвою «рейтинг» використовує оцінку відстані між вставленнями як функцію та дозволяє еВау збільшити показники покупок, кліків і доходу від реклами (рис.1.3).

Модель складається з трьох етапів: отримання схожого рекламного лістингу, застосування навченого рейтингу та зміна рейтингу лістингу шляхом включення рекламної ставки продавця. Функції моделі включають рекомендовані історичні дані про товар, рекомендовану схожість елемента з початковим елементом, категорію продукту, країну та функції персоналізації користувача. Попередня версія моделей ранжирування рекомендацій оцінювала назви продуктів за допомогою TF-IDF і подібності Jaccard, але BERT, заснований на deep learning, має чудові показники щодо розуміння мови.

Microsoft CYV-00001 - 1TB Gaming Console, (Xbox One X)
 14 viewed per hour ★★★★★ 553 product ratings

Condition: **Seller refurbished**
 Quantity: **Last one**
 490 sold / [See feedback](#)

List price: ~~\$499.99~~
 You saved: \$100.99 (20% off)
 Now: **US \$399.00**
 \$36 for 12 months

[Buy It Now](#)
[Add to cart](#)
[Add to watch list](#)

1-year protection plan from SquareTrade - \$39.99

490 sold | Free delivery in 2 days | More than 98% sold

Shipping: **FREE** Standard Shipping
 Guaranteed by **Thu, Jan. 03** | [See details](#)
 Item location: Rogers, Arkansas, United States
 Ships to: United States | [See exclusions](#)

Payments: [PayPal](#) [VISA](#) [MasterCard](#) [Discover](#) [American Express](#)

PayPal CREDIT
 *\$36 for 12 months. Minimum purchase required. [Apply Now](#) | [See terms](#)

Returns: Free 30 day returns | [See details](#)

Shop with confidence
 eBay Money Back Guarantee
 Get the item you ordered or get your money back. [Learn more](#)

Seller information
 vipoutlet (168192) ★
 96.8% Positive feedback

[Save this Seller](#)
[Contact seller](#)
[Visit store](#)
[See other items](#)

THE 2018 JEEP COMPASS
 Drive with refined style in the redesigned Jeep® Compass

[Jeep](#) [BUILD & PRICE](#) [SEARCH](#)

Frequently Bought Together 1/2

Genuine Microsoft Xbox One S White Wireless Bluetooth Control...
 \$40.49
 \$69.99
 Free shipping

Forza Motorsport 7 Xbox One Brand New Factory Sealed Racing...
 \$37.99
 Free shipping

Microsoft Xbox One Media Remote Quick Easy-To-Use MYTODDLE...
 \$333.88
 + \$3.95

Xbox Kinect Sensor Adapter Connector For XBO S & XBO X & Wi...
 \$62.90
 Free shipping

AUTHENTIC Microsoft Xbox One S Wireless Controller Gamepad W...
 \$44.88
 Free shipping

Call of Duty WWII Xbox One with Zombies Brand New Sealed CO...
 \$26.79
 Free shipping

Рисунок 1.3 – Рекомендаційна система “часто купують разом”

Тож інженери eBay представили eBERT, варіант BERT, попередньо навчений на назвах товарів eBay. В автономних оцінках eBERT перевершив готові моделі BERT з результатом F1 88,9. Щоб вирішити цю проблему, eBay розробив MicroBERT, меншу версію BERT, оптимізовану для визначення ЦП. MicroBERT використовує eBERT як вчителя на етапі навчання та налаштовується за допомогою функції під назвою InfoNCE. Заголовки елементів кодуються як вектори вбудовування, а модель навчена збільшувати косинусну подібність тематичної відстані між цими векторами, одночасно зменшуючи косинусну подібність усіх інших пар заголовків елементів у міні-пакеті.

Ця нова модель рейтингу забезпечує підвищення рейтингу покупки на

3,5%, але складність цієї моделі ускладнює виконання рекомендацій у реальному часі.

Алгоритм Netflix за допомогою вимірювання схожості між елементами визначає додатковий вміст, схожий на вміст, який бачив учасник, а потім повертається до вмісту, який найбільше схожий на вміст, який споживав відвідувач.

Netflix використовує дворівневу систему ранжирування на основі рядків, де відбувається сортування (рис. 1.4):



Рисунок 1.4 – Дворівнева рекомендаційна система

У кожному рядку - найсильніші рекомендації ліворуч. У кожному стовпчику - найсильніші рекомендації вгорі.

Кожен рядок висвітлює певну тему (наприклад, 10 найпопулярніших, популярні у вашій країні, фільми жахів тощо) і зазвичай генерується за допомогою одного алгоритму. Домашня сторінка кожного учасника складається приблизно з 40 рядків із до 75 елементів, залежно від пристрою, який використовує відвідувач.

3. Постановка проблеми та формування задач

Метою роботи буде розробка рекомендаційної системи для інтернет-магазину молодіжного одягу, яка буде аналізувати опис товару і пропонувати користувачу подібні товари.

Для розробки системи штучного інтелекту був проведений аналіз теоретичних відомостей про NLP та рекомендаційні системи на основі ШІ, визначити основні підходи та методи, які застосовуються для оброблення тексту, пошуку співпадінь в ньому та видачі рекомендацій.

Результатом кваліфікаційної магістерської роботи є проєктування СШІ для видачі рекомендованих товарів користувачу, на основу пошуку схожого опису товару.

Сучасні пошукові системи сайтів використовують обробку природної мови (або NLP), щоб аналізувати пошук, розпізнавати наміри та отримувати результати, які, на їх думку, потрібні покупцям. Вони набагато досконаліші, ніж були раніше, і більше зосереджені на досягненні результатів із більшою зосередженістю та точнішим націлюванням. NLP є ключем до цього.

Основними завданнями магістерської роботи є:
ослідження сучасних технологій обробки тексту

аналіз існуючих систем NLP та підходів до видачі рекомендованих товарів

аналіз існуючих методів штучного інтелекту для видачі рекомендацій.

розробка архітектури нової системи аналізу описи одягу і видачі найбільш подібних йому.

реалізація запроєктованих проєктних рішень.

Обґрунтування вибору підходів і технологій для проєктування та створення СШІ

СШІ являють собою комплексні комп'ютерні системи, які включають в себе технології, методи і підходи із багатьох різних галузей науки. Обрані технології та підходи повинні відповідати основним вимогам до СШІ, таким як ефективність, надійність та масштабованість.

Основними методами проєктування СШІ є:

Машинне навчання (Machine Learning) – цей метод передбачає створення моделей, які можуть навчатися на основі даних. Для реалізації машинного навчання використовуються такі технології, як нейронні мережі, метод опорних векторів та дерева рішень. Цей метод має доволі високу ефективність, але для нього потрібна якісно навчені моделі, і якщо інформація на сайті буде подана одразу кількома мовами, це суттєво ускладнить задачу із налаштуванням моделі.

Обробка зображень та відео (Computer Vision) - це метод СШІ, який передбачає аналіз та розуміння зображень та відео. Він чудово підходить для задачі класифікації одягу за категоріями, проте може виявитись не зовсім придатним для вирішення задач рекомендації, адже буде пропозиція буде складатись із візуально схожих товарів, тобто більшість пропозицій будуть одноманітними.

Обробка природної мови (Natural Language Processing) - це метод, який передбачає аналіз та розуміння людської мови. Цей підхід використовується для розробки чат-ботів, аналізу тексту та автоматичного перекладу. Для реалізації обробки природної мови використовуються такі технології, як векторні представлення слів, рекурентні нейронні мережі та моделі мови. Він буде однаково добре працювати на різних мовах і також є доволі ефективним. Саме тому у магістерській роботі зупинимось на ньому.

Для видачі рекомендацій пропонується зосередитись на пошуку схожих описів товарів. Існують два основних підходи для видачі рекомендованих товарів:

Рекомендація на основі вмісту (Content-Based Filtering) – це система

рекомендацій, яка видає користувачам відфільтровані товари на основі подібності вмісту. Нижче проілюстровано, як працює цей алгоритм (Рис. 1.5):

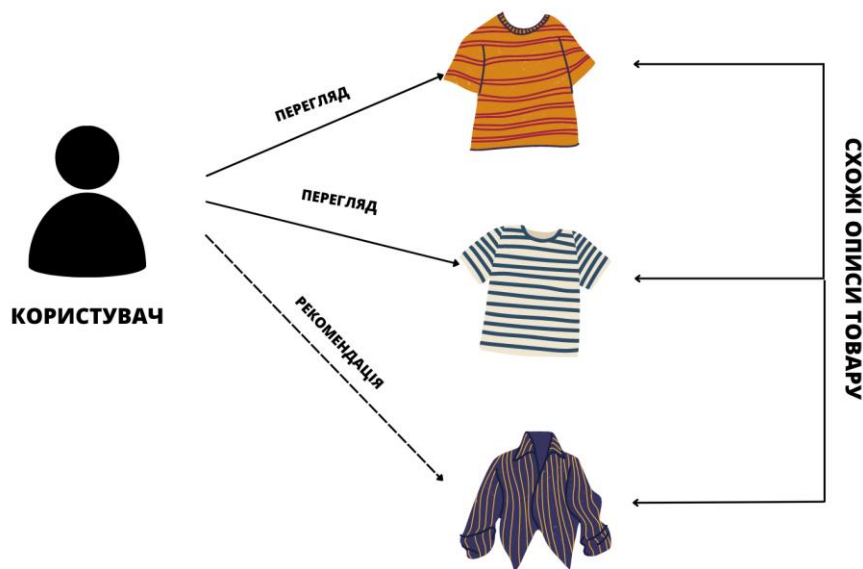


Рисунок 1.5 – Content-Based Filtering

Рекомендація на основі вподобань інших користувачів (User-Based)

Робота подібних рекомендацій заснована на створенні сегментів схожих клієнтів на основі їхніх спільних уподобань. Рекомендації надаються на основі користувачів, які згруповані разом, отже якщо клієнт має певний набір схожих характеристик разом із іншими більше досвідченішими користувачами, йому буде запропоновано перелік товарів із тих, які сподобались користувачам його сегменту.

Схематично зобразити, як працює спільна фільтрація на основі вподобань іншим користувачів можна так (рис. 1.6):

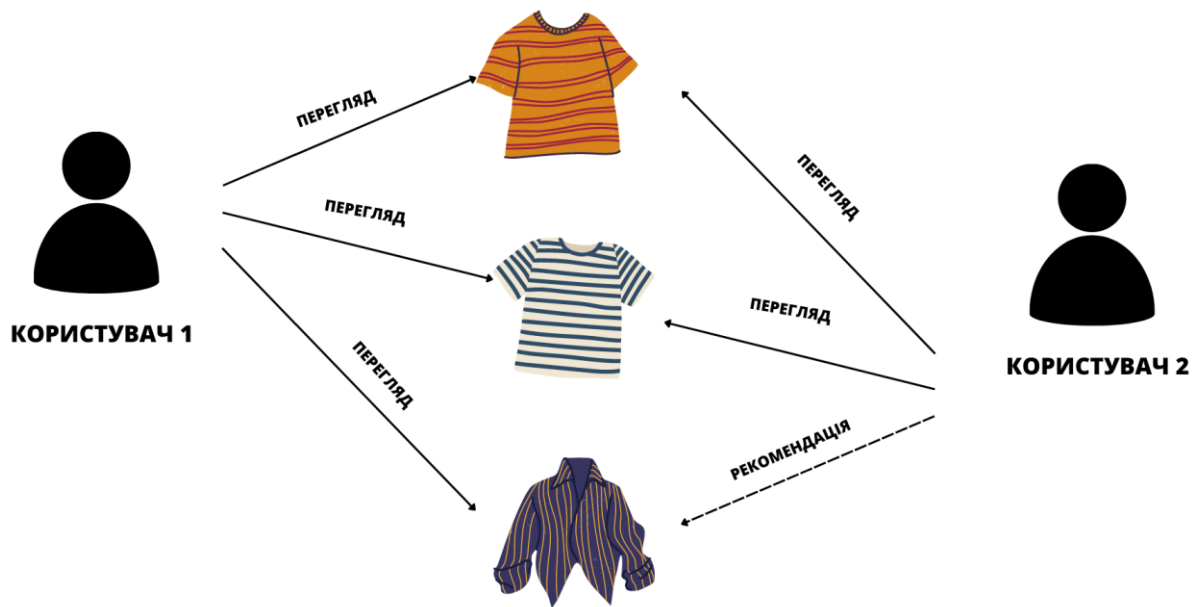


Рисунок 1.6 – User-Based Filtering

Обидва підходи мають свої переваги та недоліки, перший підхід доволі простий і ефективно працює, проте користувач ніколи не отримає рекомендацію на товар із іншої категорії. Наприклад, якщо користувачу сподобалась сорочка, йому навряд чи випаде рекомендація штанів. У другому підході немає такого недоліку, проте для його ефективної роботи потрібна солідна база користувачі, яка дозволить провести якісний кластерний аналіз. При створенні інтернет-магазину з нуля, імовірно кількість відвідувачів буде несуттєвою, а отже існуватиме величезний ризик виникнення помилок.

Тож доцільніше буде використати перший підхід.

РОЗДІЛ 2 ХАРАКТЕРИСТИКА ІНТЕЛЕКТУАЛЬНОЇ СИСТЕМИ ТА ПОСТАНОВКА ЗАДАЧІ

2.1. Характеристика об'єкта дослідження

Об'єктом дослідження є інтернет-магазин молодіжного одягу.

Інтернет-магазин молодіжного одягу повинен мати професійну команду з відповідними навичками, розділену на певні відділи (рис.2.1).

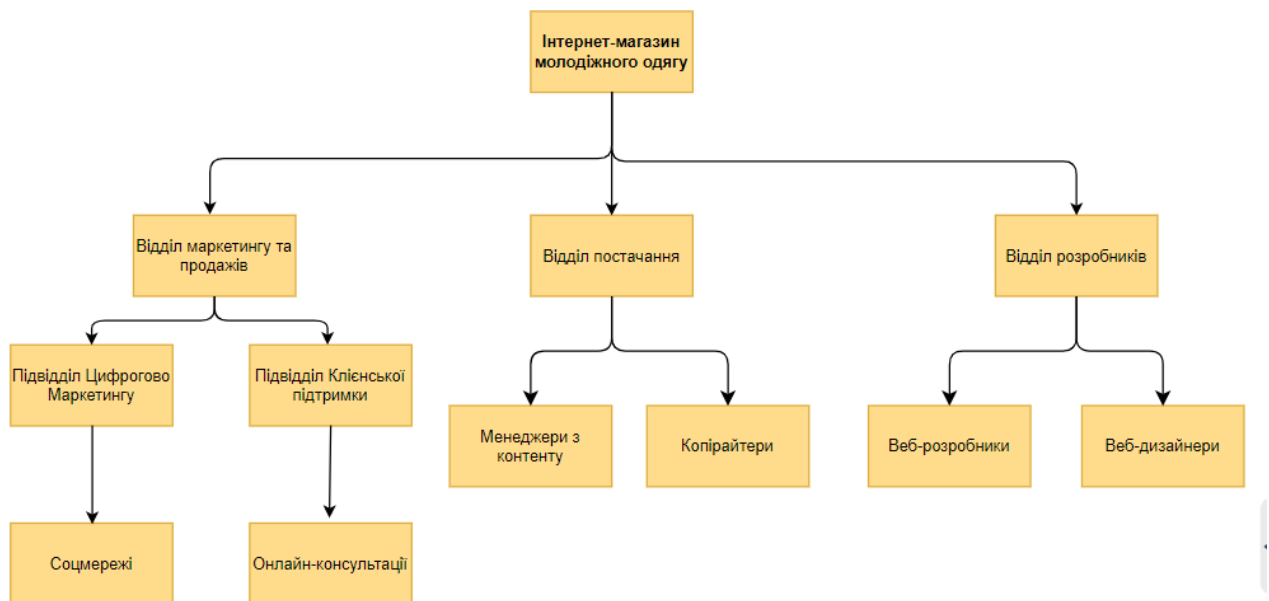


Рисунок 2.1 – Організаційна структура інтернет-магазину

На сайті можна придбати різноманітний одяг популярних у молоді брендів. Номенклатура товарів налічує тисячі одиниць. Це чоловічий, жіночий одяг усіх видів, від галантерії до верхнього одягу. Також у асортименті є аксесуари та сумки. Для постійних клієнтів пропонуються програми лояльності та знижки, регулярно проводяться акції, про які заздалегідь повідомляють клієнтів через соцмережі та електронну пошту.

Так як основна аудиторія сайту молоді люди, просуванню у соцмережах

запускається відповідна реклама та підтримується спілкування з аудиторією. До промокампаній активно залучаються лідери думок та інфлюєнсери.

Інтернет-сайт представляє собою зручний та перевірений шаблон, який містить усю необхідну інформацію.

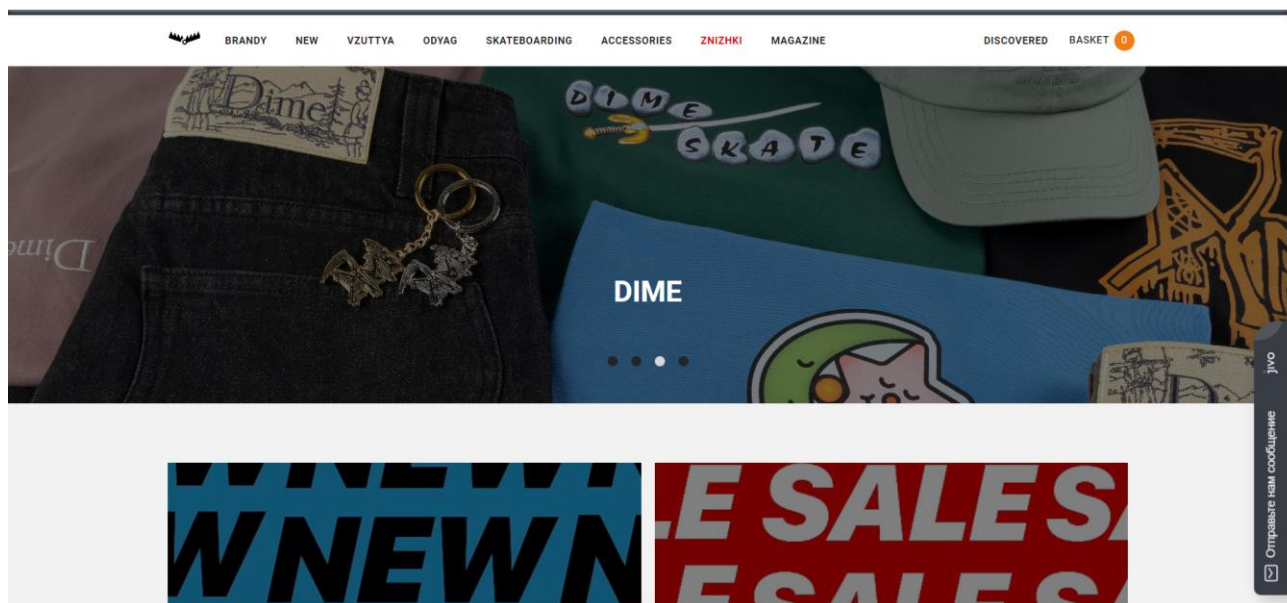


Рисунок 2.2 – Головна сторінка інтернет-магазину.

Розділ з товарами має чимало фільтрів за брендом, розміром, вартістю, тощо.

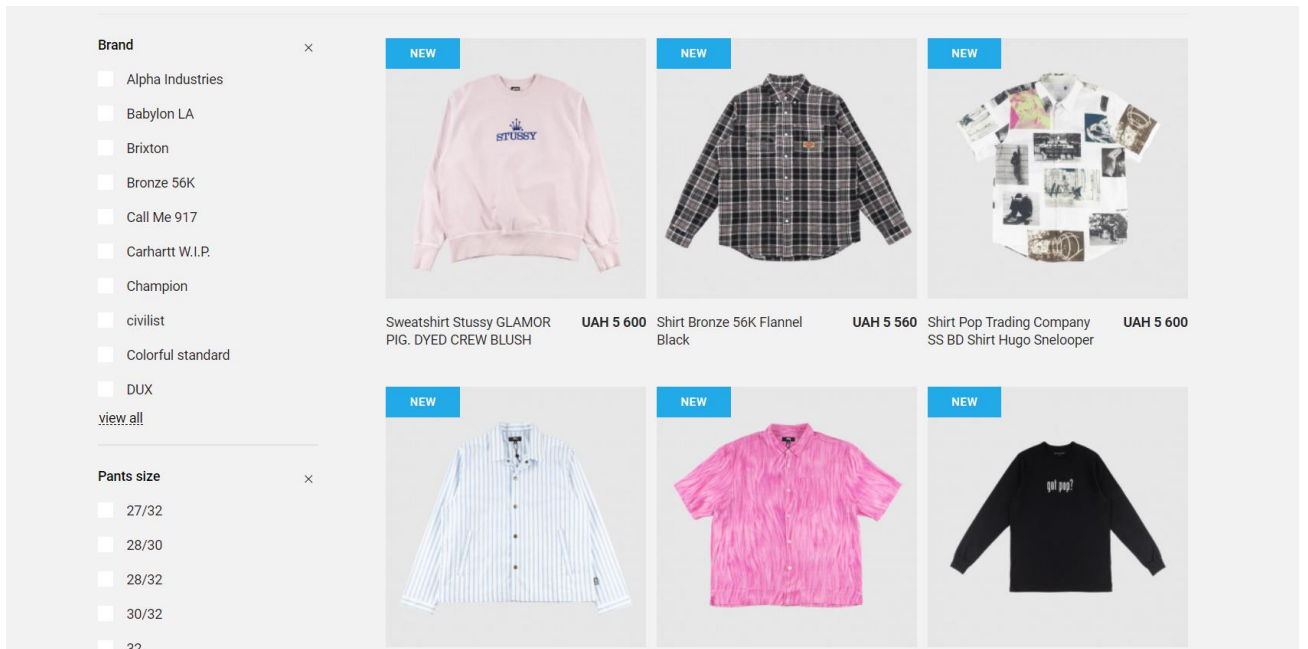


Рисунок 2.3 – Розділ з товарами

При перегляді відвідувачем сайту товарів доцільно буде використати рекомендаційну систему для підбору 4-5 найбільш схожих за описом товарів.

2.2. Структура і характеристика інтелектуальної системи та її компонентів

Рекомендаційна система призначена для швидкого аналізу опису одягу і аксесуарів та видачі схожих на них товарів, які імовірно можуть зацікавити користувача.

Система зможе працювати із різними мовами, адже заснована на NLP методі, який не потребує заздалегідь навченої моделі, яку потрібно перевчати під різні мови.

Основними функціями системи є:

- порівняння опису товару;
- формування матриці подібності;

- видача переліку найбільш схожих товарів.

Матриця подібності – це інструмент, що використовується для виведення та оцінки результатів. Вона дозволяє систематизувати та візуалізувати дані порівняння для прийняття обґрунтованих рішень. У матриці порівнянь рядки та стовпці представляють елементи, які потрібно порівняти. У кожній комірці матриці вказується оцінка або ранжування відповідно до певного критерію чи метрики. Ці оцінки можуть бути числовими значеннями, рангами, позначками або іншими відповідними мірками.

Вхідні дані подаються на обробку із бази даних. Нижче наведено приклад результату SQL запиту до бд, яким отримується необхідна інформація. Нам потрібен ідентифікатор товару та його опис (рис. 2.4).

products			
	name	description	
0	DKNY Unisex Black & Grey Printed Medium Trolle...	Black and grey printed medium trolley bag, sec...	
1	EthnoVogue Women Beige & Grey Made to Measure ...	Beige & Grey made to measure kurta with churid...	
2	SPYKAR Women Pink Alexa Super Skinny Fit High-...	Pink coloured wash 5-pocket high-rise cropped ...	
3	Raymond Men Blue Self-Design Single-Breasted B...	Blue self-design bandhgala suitBlue self-desig...	
4	Parx Men Brown & Off-White Slim Fit Printed Ca...	Brown and off-white printed casual shirt, has ...	
5	SHOWOFF Men Brown Solid Slim Fit Regular Shorts	Brown solid low-rise regular shorts, has four ...	
6	Parx Men Blue Slim Fit Checked Casual Shirt	Blue checked casual shirt, has a spread collar...	
7	SPYKAR Women Burgundy Alexa Super Skinny Fit H...	Burgundy coloured wash 5-pocket high-rise jean...	
8	Parx Men Brown Tapered Fit Solid Regular Trousers	Brown solid regular trousers regular trousers	
9	DKNY Unisex Black Large Trolley Bag	Black solid large trolley bag, secured with a ...	
10	Kenneth Cole Women Navy Blue Solid Backpack	Navy Blue backpackNon-Padded haul loop1 main c...	
11	Parx Men Green Printed Polo Collar T-shirt	Green printed T-shirt, has a polo collar, and ...	
12	Police Men To Be The King Eau De Toilette 125 ml	To Be The King Eau De Toilette	
13	Vishudh Women Maroon Printed Kurta with Palazzos	Maroon printed kurta with palazzos Maroon stra...	
14	Parx Men Red Slim Fit Checked Casual Shirt	Red checked casual shirt, has a spread collar,...	
15	Being Human Clothing Navy Printed Casual Slim ...	Navy blue printed casual shirt, has a spread c...	
16	Homesake Gold-Toned & White Solid Handcrafted ...	Type: Handcrafted Table Lamp with ShadeColour:...	
17	SEJ by Nisha Gupta Set of 6 Mustard & Blue Pri...	Set content: 6 table mats Shape: Rectangle Col...	

Рисунок 2.4 – Розділ з товарами

Для роботи в рамках кваліфікаційної магістерської роботи вхідні дані будуть завантажені у систему в CSV форматі.

Вихідні дані подаються у вигляді списку тайтлів 4 найбільш схожих за описов товарів.

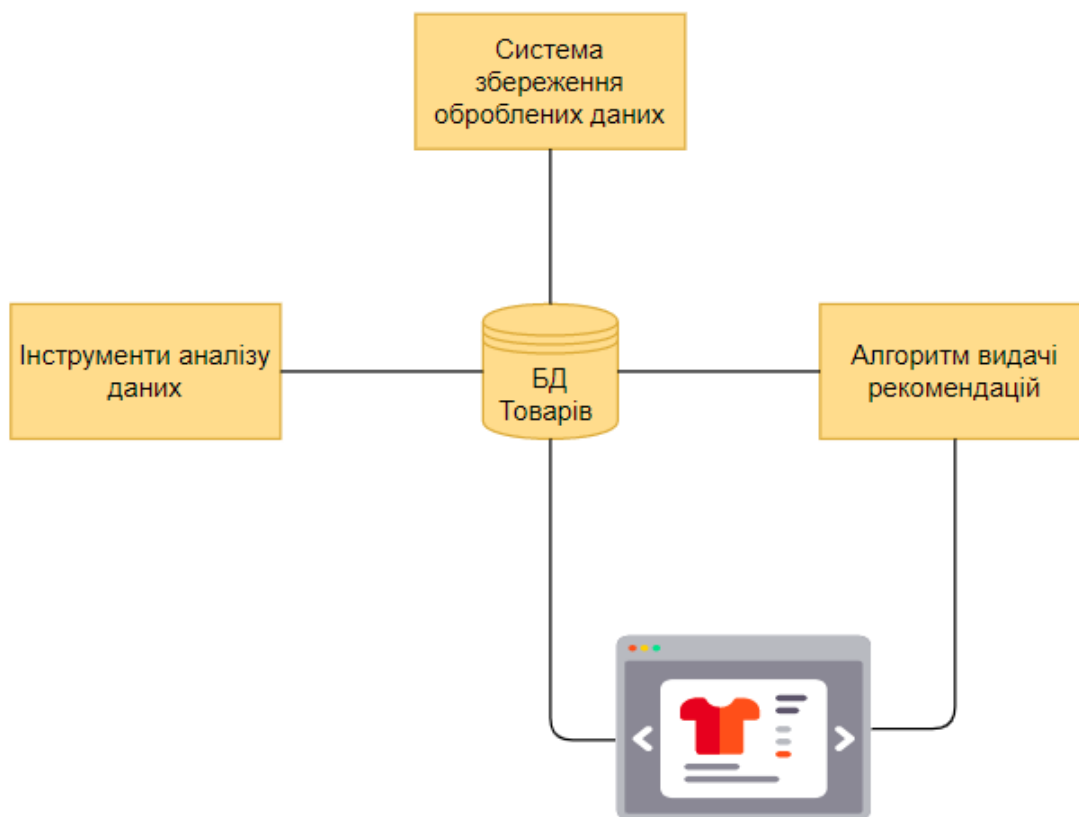


Рисунок 2.5 – Структура інтелектуальної системи рекомендацій інтернет-магазину молодіжного одягу

Інтелектуальна система для інтернет-магазину молодіжного одягу включає в себе наступні елементи:

- актуальна база даних товарів;
- інструменти аналізу даних, які дозволяють обробляти зібрані дані та формувати матрицю порівнянь;
- алгоритми видачі рекомендацій;
- інтерфейс користувача, що надає доступ до рекомендованих товарів у вигляді пропозицій на сторінці із тим товаром, який переглядає користувач;
- система збереження оброблених даних.

Інтелектуальна система є окремим інструментом аналізу, який може працювати як окремий модуль інтернет-магазину.

Методи, моделі і моделювання процесів і елементів складних систем

Методи дослідження й синтезу компонент систем штучного інтелекту

Перед початком роботи потрібно підготувати текстовий опис товарів належним чином.

Основні кроки підготовки тексту до аналізу включають:

чищення тексту. Залежно від завдання очищення включає видалення неалфавітних символів, різних тегів, URL-адрес, знаків пунктуації, пробілів та інших елементів розмітки.

єреведення тексту в нижній регістр

Застосування стоп-слів. Стоп-слова – це часто вживані слова, які не несуть в собі жодного додаткового змісту, але можуть вплинути на результат аналізу.

. Перетворення текстового вмісту у векторне представлення, яке виділяє слова, що використовуються в схожих або ідентичних контекстах.

У задачах обробки природної мови (NLP), де текст є основним видом даних, векторизація тексту використовується для того, щоб перетворити неструктурований текст у вектори, які можуть бути опрацьовані алгоритмами машинного навчання. Це дозволяє використовувати різні методи та моделі машинного навчання для аналізу та розуміння текстових даних.

Векторизація тексту — це процес перетворення тексту в числове представлення.

Існує кілька підходів до векторизації тексту. Один з найпоширеніших методів - це "Bag of Words" (мішок слів), де кожен документ розглядається як набір слів, а векторизація включає підрахунок кількості входжень кожного слова в документі. Інший популярний метод - це TF-IDF (term frequency-inverse

document frequency), який враховує як частоту входження слова в документ, так і важливість слова в контексті всього корпусу текстів.

Окрім цих базових методів, існують більш складні підходи, такі як word2vec, GloVe і BERT, які використовують нейронні мережі для векторизації слів або фраз з урахуванням семантики та контексту. Коротко опишемо кожен із них.

Word2Vec — це техніка машинного навчання, розроблена Томасом Міколовим і його групою з вивчення даних у Google. Він спирається на deep learning, щоб навчити комп'ютер вивчати мову. Вбудовані слова – це вектори, створені для представлення слів та їх контексту. Обробка природної мови є складним процесом, який вимагає розуміння структурних зв'язків між словами та контекстом. Слова рідко мають значення окремо, тому модель вбудовування слів допомагає нам зрозуміти, що означає дане слово.

Векторна методологія вбудовування слів — це сучасна техніка NLP, яка використовується для кодування значення слів. Він присвоює кожному слову значення від 0 до 1 за кількома різними вимірами та впорядковує їх за подібністю та несхожістю.

Комп'ютер призначає координати залежно від того, як часто він «бачить» спільне входження слів. Word2vec працює, відображаючи слова, які з'являються в тому самому контексті, а антоніми зазвичай знаходяться в тій самій області простору (рис. 2.4).

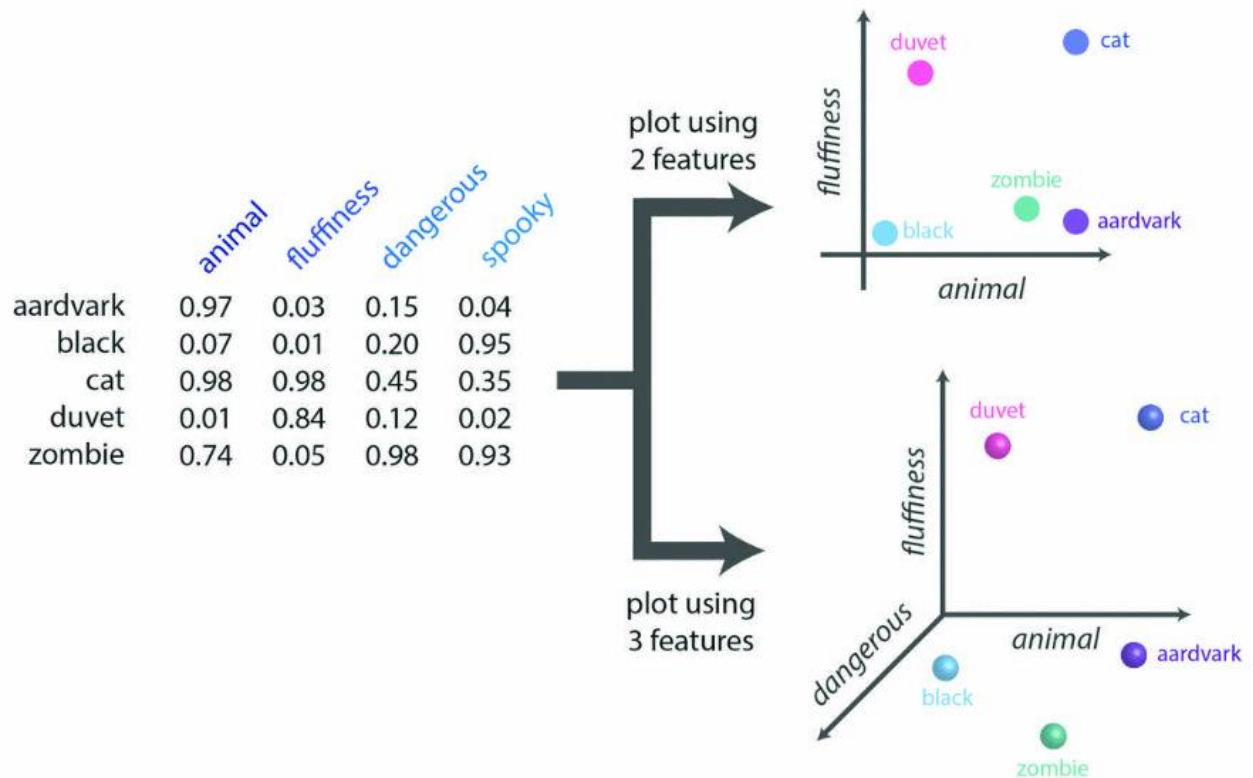


Рисунок 2.6 – Ілюстрація роботи Word2Vec

Існує два основних варіанти Word2Vec: Continuous Bag of Words (CBOW) і Skip-gram. CBOW передбачає цільове слово за його контекстом, тоді як Skip-gram передбачає контекстні слова за цільовим словом. Word2Vec широко використовувався в задачах обробки природної мови, таких як моделювання мови, класифікація документів і пошук інформації.

— ще одна популярна модель вбудовування слів, яка також представляє слова як щільні вектори. Він поєднує глобальну статистику спільного використання слів із локальною контекстною інформацією, яку використовує Word2Vec. Ключова ідея GloVe полягає в тому, щоб зафіксувати зв'язок між словами шляхом аналізу співвідношення ймовірностей їх спільного появи. GloVe вивчає вбудовування слів шляхом розкладання на множники матриці статистики спільного використання слів. Отримані вкладення демонструють значущі лінійні

підструктури та можуть вловлювати як семантичну, так і синтаксичну подібність між словами.

Показники подібності, які використовуються для оцінки найближчих сусідів, створюють єдиний скаляр, який кількісно визначає спорідненість двох слів (рис. 2.7). Ця простота може бути проблематичною, оскільки два наведених слова майже завжди демонструють більш запутані відносини, ніж це можна охопити одним числом. Наприклад, чоловіка можна вважати схожим на жінку в тому, що обидва слова описують людей; з іншого боку, ці два слова часто вважають протилежними, оскільки вони підкреслюють головну вісь, за якою люди відрізняються один від одного.

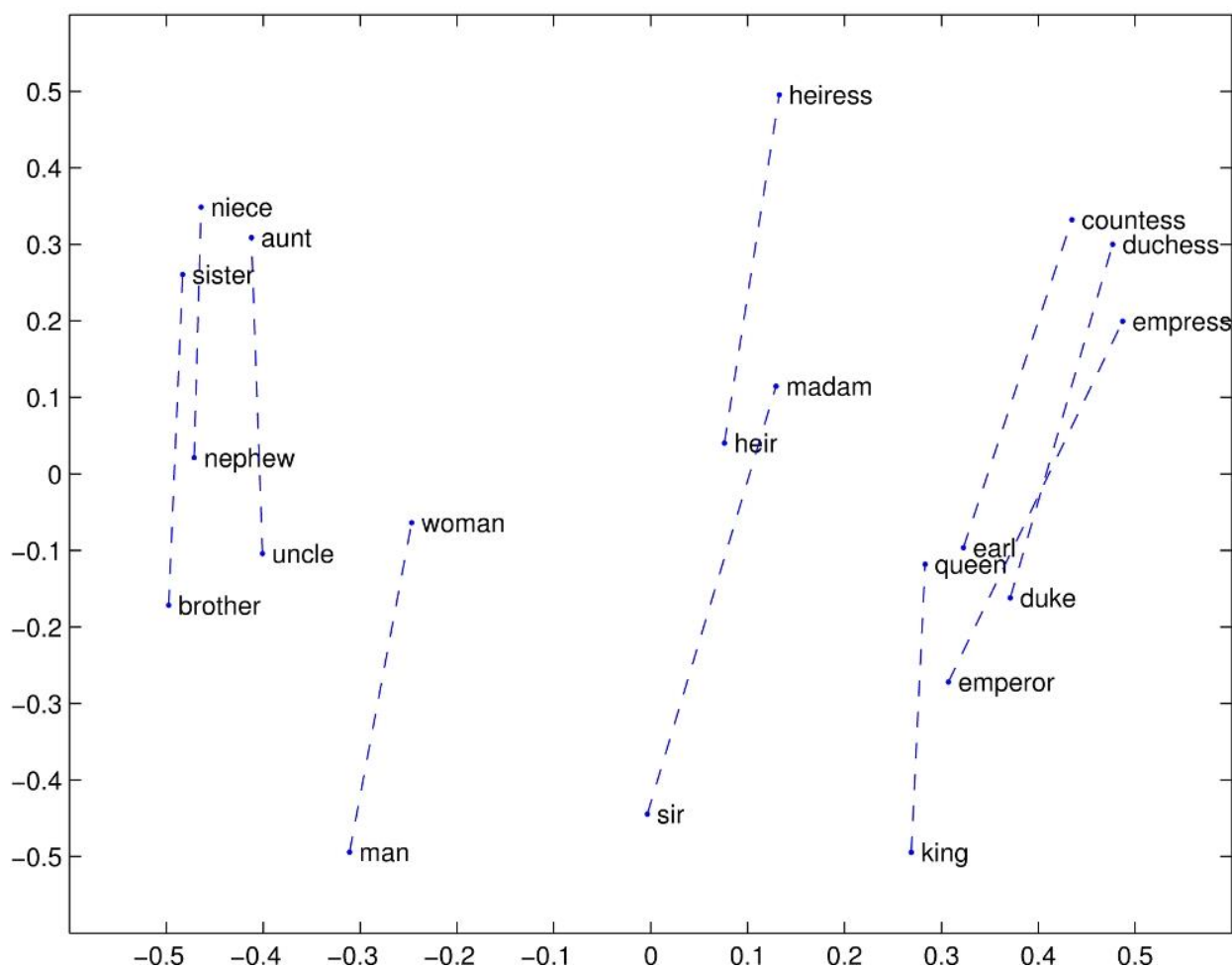


Рисунок 2.7 – Ілюстрація роботи GloVe

GloVe широко використовується в різних програмах обробки природної мови, включаючи аналіз настроїв, машинний переклад і розпізнавання об'єктів.

— це найсучасніша мовна модель, яка зробила революцію в задачах обробки

природної мови. На відміну від Word2Vec і GloVe, які генерують статичні вбудовані слова, BERT вивчає контекстуальне представлення слів. Він заснований на трансформаційній архітектурі, яка дозволяє отримувати контекстну інформацію слів, розглядаючи як лівий, так і правий контекст у двонаправленому порядку. BERT навчена на великому корпусі текстових даних за допомогою моделювання пропущених слів та завдань прогнозування наступного речення. Після попереднього навчання BERT можна налаштувати на конкретні подальші завдання, такі як класифікація тексту, розпізнавання іменованих об'єктів і відповіді на запитання. BERT досяг вражаючої продуктивності в широкому діапазоні тестів обробки природної мови та став наріжним каменем у багатьох програмах NLP.

Векторизація тексту дозволяє представити текстові дані у вигляді числових векторів, що відкриває можливості для застосування різноманітних аналітичних методів та моделей машинного навчання для обробки та аналізу тексту.

оделі та методи оптимізації в системах штучного інтелекту

Задача рекомендації подібних товарів відноситься до задач класифікації в системах штучного інтелекту. Розглянемо модель, використану у кваліфікаційній магістерській роботі.

Для оцінювання схожості використовується метод косинусної подібності. обчислює L2-нормалізований скалярний добуток векторів. Тобто, якщо x та y - вектори-рядки, їх косинус-подібність визначається за формулою:

x

y Де, $\|x\|$ та $\|y\|$ - це довжина двох векторів x та y .

=

x

Це називається косинусною подібністю, оскільки евклідова (L2) нормалізація проектує вектори на одиничну сферу, а їх скалярний добуток є косинусом кута між точками, позначеними векторами.

Це модель є популярним вибором для обчислення подібності документів, представлених як вектори tf-idf.

2.3.3. Методи та моделі управління в системах штучного інтелекту

Визначення того, що визначає гарну рекомендацію, — це проблема, з якою зіштовхується чимало підприємств. Визначення «хороших» рекомендацій допомагає оцінити продуктивність створених вами рекомендацій. Якість рекомендації можна оцінити за допомогою різних підходів, які вимірюють охоплення та точність.

Точність – це частка правильних рекомендацій із загальної кількості можливих рекомендацій, тоді як охоплення вимірює частку об'єктів у просторі пошуку, для яких система може надати рекомендації.

Метод оцінки рекомендації залежить виключно від набору даних і підходу, використаного для створення рекомендації. Системи рекомендацій мають кілька концептуальних подібностей із проблемою класифікації та регресійного моделювання.

Загальноприйнятими показниками статистичної точності для оцінки точності рекомендацій є RMSD, MAE та k-кратна перехресна перевірка.

Перехресна перевірка K Fold

K-кратна перехресна перевірка (K - Fold Crosscheck) — це популярний підхід, який використовується в машинному навчанні для оцінки продуктивності моделі та пом'якшення переналаштування. Методика включає в себе розбиття доступного позначеного набору даних на K однакових за розміром підмножин. Потім модель навчається та оцінюється K разів, щоразу

використовуючи інше згортання як набір перевірки, а решта K-1 згортань як навчальний набір. Цей процес дозволяє більш повно оцінювати продуктивність моделі, використовуючи різні підмножини даних для навчання та перевірки точності отриманих систем у порівнянні з тестовим набором.

Для оцінки навченості системи береться середнє значення балів точності. Загалом, цей метод корисний для запобігання переналаштуванню вашої моделі, однак це об'ємний процес з точки зору обчислень.

MAE (середня абсолютна похибка)

Формула визначення середньої абсолютної похибки виглядає так:

=

Де, y_i – прогнозоване значення,

n – загальна кількість значень

n

y Представляє середнє абсолютне значення кожної помилки в прогнозуванні рейтингу рекомендованого товару. Чим менше показник MAE, тим краще.

Середнє квадратичне відхилення)

Формула визначення середнього квадратичного відхилення виглядає так:

—

$\bar{x}$

Де, x_i – прогнозоване значення,

$\hat{x}_i$ – середнє значення,

N – загальна кількість значень

n

x Подібний показник до MAE, але має сильніший штраф, коли прогноз дуже далекий від справжнього значення, і слабший штраф, коли прогноз ближчий до справжнього значення

—

Зведення в квадрат різниці справжніх і прогнозованих значень замість суми абсолютних значень. Це гарантує, що результуюче значення завжди додатне і більше, коли різниця велика, і менше, коли різниця низька. Чим нижчий показник RMSD, тим краще.

Ці показники часто використовуються для оцінки якості рекомендації, але в них відсутні різні компоненти. Наявність даних користувачів, пов'язаних із рекомендаціями, є важливою для визначення справжньої якості рекомендації. Можливість відстежувати частоту відвідувань рекомендацій, залученість до платформи, чуйність тощо забезпечить чіткішу точку зору щодо якості рекомендацій.

Інші компоненти, про які слід знати, це знати, коли змінювати рекомендації, якщо користувач не взаємодіяв з ними протягом X часу, або коли перенавчати систему на основі нових оцінок або взаємодії користувачів. Ви також хотіли б звернути увагу на те, чи обмежують ці рекомендації користувача до підрозділу продуктів, як рекомедатор ставиться до новизни, різноманітності та упередженості вибору. А/В-тестування часто використовується для відстеження цих показників.

РОЗДІЛ 3 РОЗРОБЛЕННЯ ПРОЕКТНИХ РІШЕНЬ

3.1. Моделювання та проектування бази знань для інтелектуальної системи рекомендацій

Перед початком роботи, можемо виділити основні етапи, які мають місце в обробці текстової інформації:

збір даних

чистка даних

бробка даних

наліз результатів

Для підготовки та аналізу даних ми будемо використовувати мову Python із відповідними бібліотеками. Нижче, проілюстрований процес обробки інформації інтелектуальною системою (рис. 3.1).

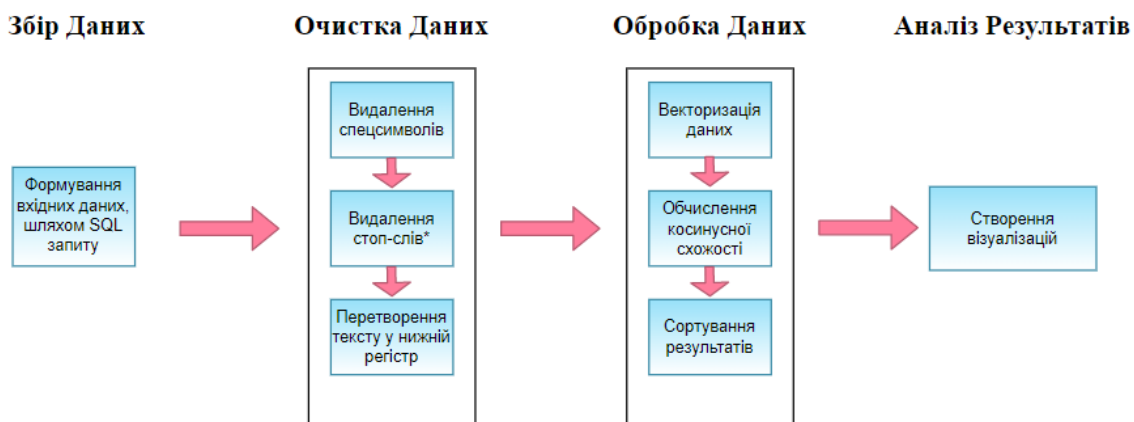


Рисунок 3.1 – Етапи формування рекомендацій

Необхідні вхідні дані для аналізу можуть здійснюватися шляхом виконання SQL запиту до бази даних, де містяться усі наявні товари інтернет-магазину. Альтернативно, можна використати набір даних у форматі CSV або

Результати SQL запиту мають містити наступну інформацію :

- ☐ ідентифікатор товару;
- ☐ тайтл товару;
- ☐ текстовий опис товару.

Отриманий результат потрібно підготувати для NLP. Процедура підготовки включає наступні кроки:

- чищення тексту від усіх спецсимволів, цифр, тощо.
- перетворення тексту на нижній регістр.
- виключення стоп-слів.

Результат запиту до БД можна також представити у вигляд CSV файлу (рис. 3.2).

A	B	C
1 ProductID name		description
2 10017413 DKNY Unisex Black & Grey Printed Medium Trolley Bag		Black and grey printed medium trolley bag, secured with a TSA lockOne handle on the top an
3 10016283 EthnoVogue Women Beige & Grey Made to Measure Custom Made Kurta Set with Jacket		Beige & Grey made to measure kurta with churidar andB dupattaBeige made to measure cal
4 10009781 SPYKAR Women Pink Alexa Super Skinny Fit High-Rise Clean Look Stretchable Cropped Jeans		Pink coloured wash 5-pocket high-rise cropped jeans, clean look, no fade, has a button and z
5 10015921 Raymond Men Blue Self-Design Single-Breasted Bandhgala Suit		Blue self-design bandhgala suitBlue self-design bandhgala blazer, has a mandarin collar, singl
6 10017833 Parx Men Brown & Off-White Slim Fit Printed Casual Shirt		Brown and off-white printed casual shirt, has a spread collar, long sleeves, button placket, ci
7 10014361 SHOWOFF Men Brown Solid Slim Fit Regular Shorts		Brown solid low-rise regular shorts, has four pockets, a button closure
8 10017869 Parx Men Blue Slim Fit Checked Casual Shirt		Blue checked casual shirt, has a spread collar, long sleeves, button placket, curved hem, one
9 10009695 SPYKAR Women Burgundy Alexa Super Skinny Fit High-Rise Clean Look Stretchable Ankle Jeans		Burgundy coloured wash 5-pocket high-rise jeans, clean look, no fade, has a button and zip c
10 10000571 Parx Men Brown Tapered Fit Solid Regular Trousers		Brown solid regular trousers regular trousers
11 10017421 DKNY Unisex Black Large Trolley Bag		Black solid large trolley bag, secured with a TSA lockOne handle on the top and one on the si
12 10000911 Kenneth Cole Women Navy Blue Solid Backpack		Navy Blue backpackNon-Padded haul loop1 main compartment with zip closurePadded back
13 10000245 Parx Men Green Printed Polo Collar T-shirt		Green printed T-shirt, has a polo collar, and short sleeves
14 10001989 Police Men To Be The King Eau De Toilette 125 ml		To Be The King Eau De ToiletteB
15 10016727 Vishudh Women Maroon Printed Kurta with Palazzos		Maroon printed kurta with palazzos Maroon straight calf length kurta, has a round neck, thre
16 10017749 Parx Men Red Slim Fit Checked Casual Shirt		Red checked casual shirt, has a spread collar, long sleeves, button placket, curved hem, one
17 1000795 Being Human Clothing Navy Printed Casual Slim Shirt		Navy blue printedB casual shirt, has a spread collar, a full button placket, two pockets,B long
18 10003803 Homesake Gold-Toned & White Solid Handcrafted Table Lamp with Shade		Type: Handcrafted Table Lamp with ShadeColour: Gold-Toned and whitePattern: SolidShape: V
19 10012765 SEJ by Nisha Gupta Set of 6 Mustard & Blue Printed Table Placemats		Set content: 6 table mats Shape: Rectangle Colour: Mustard and blue Pattern: Printed Suitab
20 10014413 YAK YAK Men Camel Brown Solid Polo Collar T-shirt		Camel brown solid T-shirt, has polo collar, short sleeves
21 10013483 PARFAIT Plus Size Black Striped Non-Wired Lightly Padded T-shirt Bra P5252		Black striped medium-coverage T-shirt bralightly Padded and Non-Wired seamless cupsRegu
22 10001265 Michael Kors Women Sexy Amber Eau de Parfum 100ml		Michael Kors Sexy Amber Eau de ParfumFragrance Notes:Top notes - amberHeart notes - sand
23 10004155 HIGHLANDER Men Mustard & Black Slim Fit Checked Casual Shirt		Mustard and black checked casual shirt, has a spread collar, long sleeves, button placket, cui
24 10012795 SEJ by Nisha Gupta Set of 6 Multicoloured Printed Table Placemats		Set content: 6 table matsShape: RectangleColour: MulticolouredPattern: PrintedSuitable For
25 10009723 SPYKAR Women Blue Adora Skinny Fit Mid-Rise Clean Look Stretchable Ankle Jeans		Blue medium wash 5-pocket mid-rise jeans, clean look, heavy fade, has a button and zip clos
26 10012853 JEWEL JUNCTION Gold-Toned Square Cufflinks		A pair of gold-toned square solid cufflinks
27 10006001 ID Men Tan Brown Fisherman Leather Sandals		A pair of tan brown fisherman sandalsLeather upper with velcro closureCushioned footbedP
28 10017377 Story@home Multicoloured Geometric 210TC Cotton 1 King Bedsheet with 2 Pillow Covers		Set content: 1 King bedsheet with 2 pillow coversQuality: Fine Thread count: 210Colour: Mul
29 10015839 Raymond Men Blue Slim Fit Printed Formal Shirt		Blue printed formal shirt, has a spread collar, long sleeves, button placket, straight hem, one

Рисунок 3.2 – Вхідні дані

Для очистки тексту в описі товару від усіх спецсимволів можна використати regular expressions. Також потрібно перевести текст в єдиний регістр. Ці дві дії можна об'єднати створивши єдину функцію для очистки (рис. 3.3). Для роботи нам потрібна стандартна бібліотека re.

```
#Функція для очистки даних
def clean_text(text):
    # Видалення спецсимволів
    text = re.sub(r'^\w\s', '', text)
    # Переведення в нижній регістр
    text = text.lower()
    return text
```

Рисунок 3.3 – Функція очистки даних

На наступному кроці потрібно провести очистку тексту від стоп-слів. З цією метою можна використати як вже готові бібліотеки так і створити власну. Слова мають виключатися з огляду на свою незначущість. Наприклад, в бібліотеці sklearn до таких слів належать “and”, “the”, “him” і т.д. Цей крок ми

об'єднаємо із векторизацією тексту.

У даній роботі для перетворення очищеного тексту на вектори використовується метод "Bag of Words" (мішок слів). Він визначає кожне слово в описі товару і створює вектор із числами, які відображають число входжень кожного слова із словника. Інформація про порядок та структуру слів ігнорується, тому метод і має назву “мішку”.

За допомогою CountVectorizer() створюється матриця векторів слів на основі описів товарів, де кожен рядок відповідає одному товару, а кожен стовпець відображає наявність кожного слова зі словника в описі товару .

Далі використовується cosine_similarity() для обчислення косинусної схожості між векторами описів товарів, що дозволяє порівнювати товари та встановлювати ступінь їхньої схожості на основі текстового опису (рис.3.4).

```
# Створення матриці векторів слів на основі описів товарів з виключенням стоп-слів і функцією очистки
vectorizer = CountVectorizer(stop_words='english').fit_transform(products['description'].apply(clean_text))

# Обчислення косинусної схожості між векторами описів товарів
similarity_scores = cosine_similarity(vectorizer)
```

Рисунок 3.4 – Код для векторного представлення та обчислення схожості між векторами описів товарів

Маючи 12491 позицію в списку товарів, в результаті обробки ми отримали матрицю розміром 12491 x 12491.

```
array([[1.          , 0.07511746, 0.13483997, ..., 0.04264014, 0.          ,
        0.05884899],
       [0.07511746, 1.          , 0.02321192, ..., 0.02936101, 0.          ,
        0.04052204],
       [0.13483997, 0.02321192, 1.          , ..., 0.15811388, 0.          ,
        0.05455447],
       ...,
       [0.04264014, 0.02936101, 0.15811388, ..., 1.          , 0.          ,
        0.          ],
       [0.          , 0.          , 0.          , ..., 0.          , 1.          ,
        0.          ],
       [0.05884899, 0.04052204, 0.05455447, ..., 0.          , 0.          ,
        1.          ]])
```

Рисунок 3.5 – Матриця з результатами

3.2. Розроблення користувацького інтерфейсу. Елементи та структура

Для зручного представлення результатів аналізу, можна використовувати різноманітні чарти. Вони потрібні нам для оцінки результатів аналізу. Нижче представлений програмний код для створення візуалізації типу heatmap (рис.

```
#побудова візуалізації типу heatmap на основі матриці порівнянь
from matplotlib import pyplot as plt
heatmap = plt.pcolor(similarity_scores)
plt.colorbar(heatmap)
```

Рисунок 3.6 – Код для візуалізації матриці порівнянь

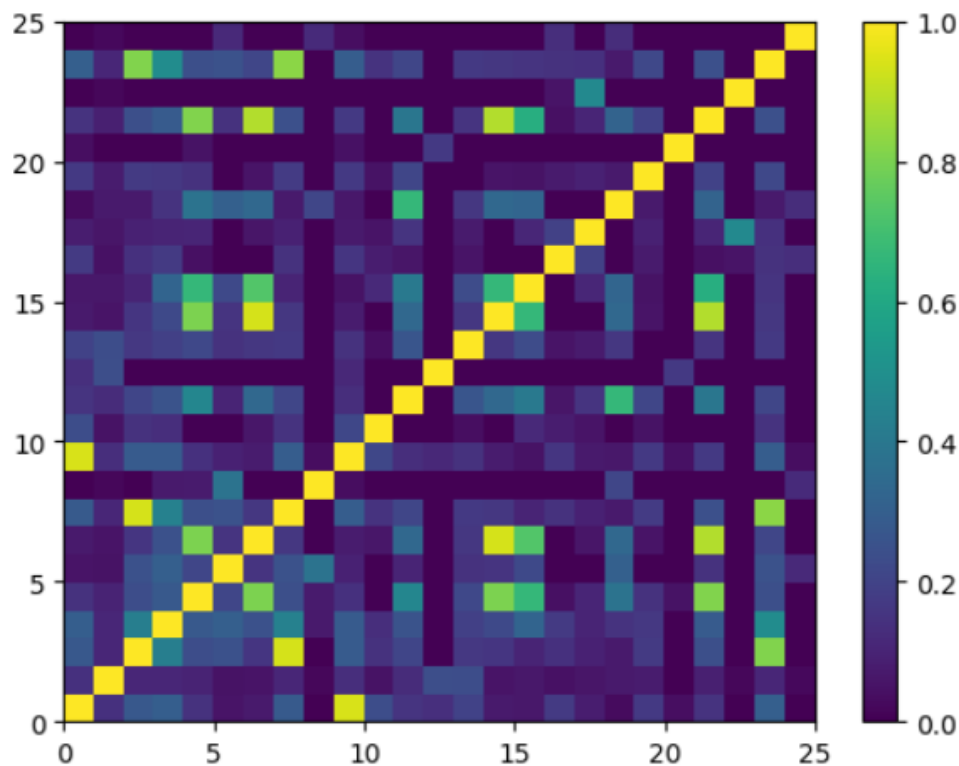


Рисунок 3.7 – Візуалізація типу heatmap на основі матриці порівнянь для 25 товарів

Це наш “рейтинг” схожості товарів. По горизонталі та вертикалі знаходиться індекс товару. На рисунку найбільш схожі товари будуть відображені як крапки жовтого на зеленому кольорі. З heatmap наглядно можна побачити, що товари з індексом “0” та “9” виглядають як добра комбінація для пропозиції. Звернувшись до вхідних даних, можемо перевірити чи дійсно це так (рис. 3.8).

index	name
0	DKNY Unisex Black & Grey Printed Medium Trolley Bag
1	EthnoVogue Women Beige & Grey Made to Measure Custom Made Kurta Set with Jacket
2	SPYKAR Women Pink Alexa Super Skinny Fit High-Rise Clean Look Stretchable Cropped Jeans
3	Raymond Men Blue Self-Design Single-Breasted Bandhgala Suit
4	Parx Men Brown & Off-White Slim Fit Printed Casual Shirt
5	SHOWOFF Men Brown Solid Slim Fit Regular Shorts
6	Parx Men Blue Slim Fit Checked Casual Shirt
7	SPYKAR Women Burgundy Alexa Super Skinny Fit High-Rise Clean Look Stretchable Ankle Jeans
8	Parx Men Brown Tapered Fit Solid Regular Trousers
9	DKNY Unisex Black Large Trolley Bag

Рисунок 3.8 – Список товарів

Для наочного представлення найбільш близьких товарів побудований граф (рис. 3.10). Кожен товар позначений числовим індексом, як і на попередній візуалізації. Код для візуалізації даного графіку наведений нижче (рис. 3.9).

```
import networkx as nx
G = nx.to_networkx_graph(similarity_scores)
nx.draw(G, with_labels=True, font_weight='bold')
```

Рисунок 3.9 – Код для візуалізації типу граф

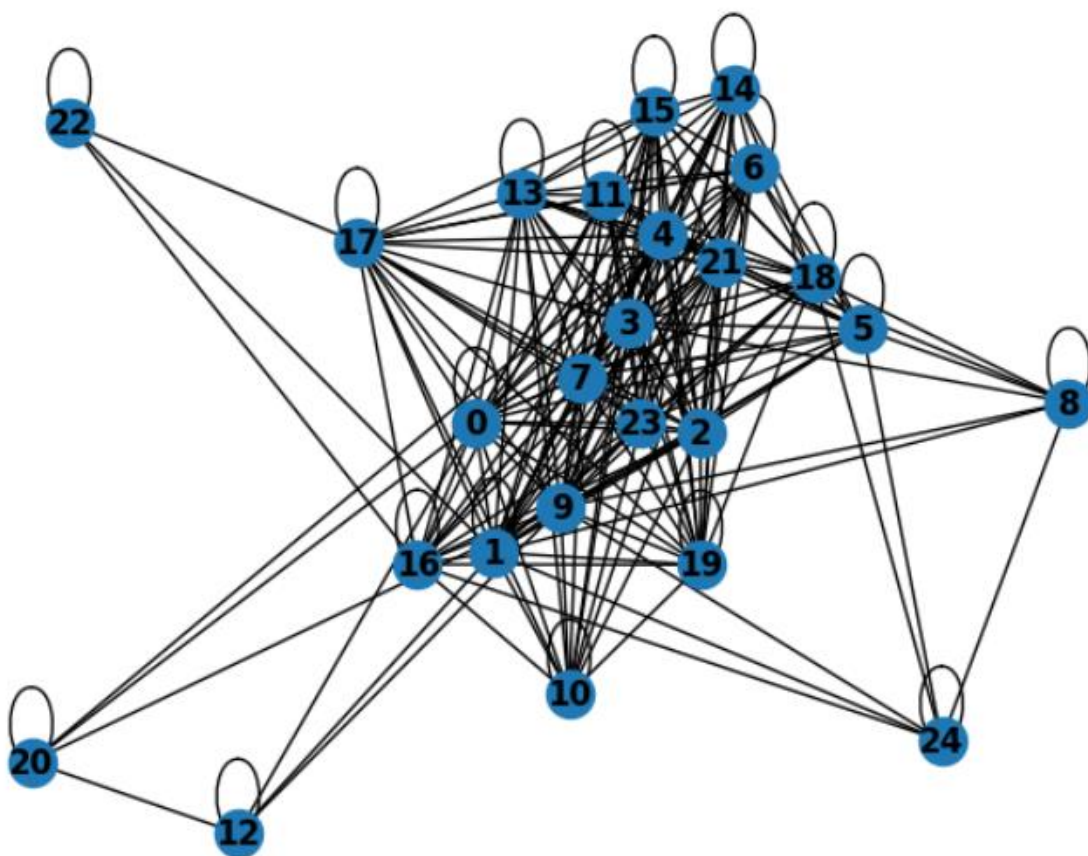


Рисунок 3.10 — Граф, який відображає відносини між товарами

Рекомендації мають доповнювати веб-сторінку із описом кожного товару інтернет-магазину. Нижче наведено прототип такої веб-сторінки із відображенням рекомендованих товарів (рис. 3.11).

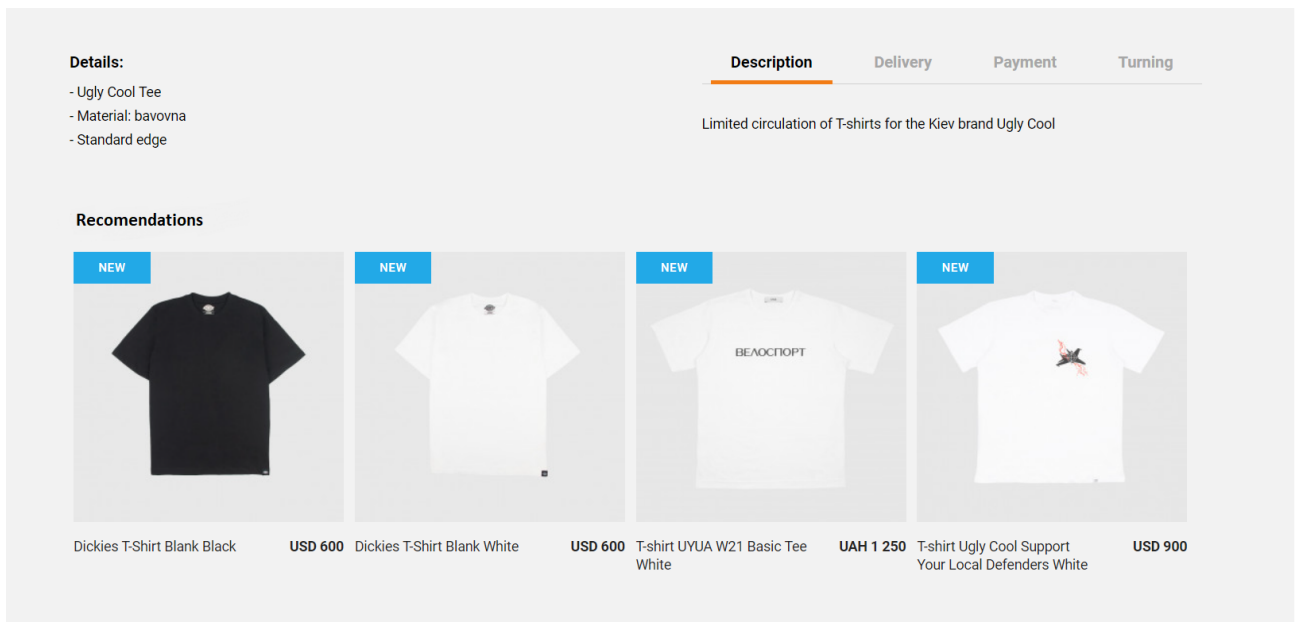


Рисунок 3.11 — Прототип веб-сторінки із рекомендованими товарами

3.3. Проектування забезпечувальних підсистем інтелектуальної системи. Реалізація системи

3.3.1. Інформаційне забезпечення

Інформаційне забезпечення інтелектуальної системи видачі рекомендацій відвідувачам інтернет-магазину молодіжного одягу формується шляхом збору інформації з БД про наявні товари та їх текстовий опис.

Загальна структура інформаційного забезпечення наведена на малюнку (рис. 3.12)

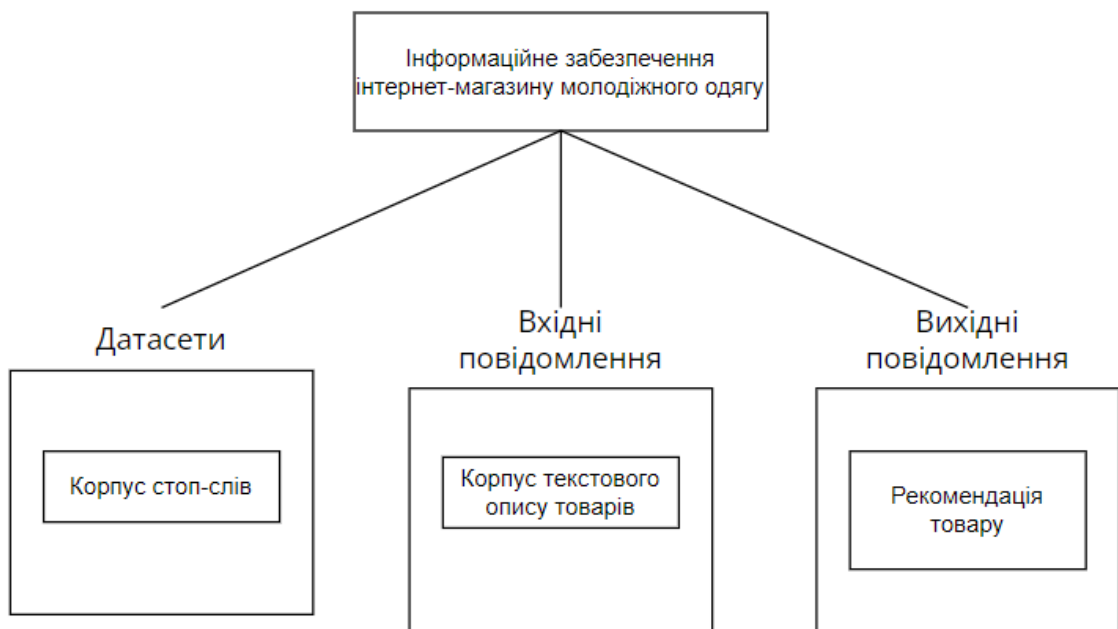


Рисунок 3.12 – Структура інформаційного забезпечення

Корпус тексту (text corpus) являє собою великий та структурований набір зчитуємих машинами та систематично зібраних текстів. Корпуси тексту є важливим інструментом для багатьох завдань у галузі обробки природної мови (NLP) та машинного навчання.

Інформаційне забезпечення включає також датасет стоп-слів для аналізу.

Таблиця 3.1 – Структура вхідних повідомлень

Назва сутності	Позначення	Тип даних	Обов'язкове поле
Ідентифікатор товару	id	String [5]	Ні
Назва товару		String [100]	Так
Опис товару		String [10000]	Так

3.3.2. Програмне забезпечення

Основою програмної реалізації системи є мова програмування Python та різноманітні бібліотеки для роботи з текстами, матрицями, побудовою візуалізацій тощо. Нижче наведено опис кожної з них

Бібліотека Pandas - це потужна бібліотека для обробки та аналізу даних у мові програмування Python. Вона надає широкий набір функцій для роботи з даними, таких як створення, зчитування, збереження, фільтрація, сортування, групування, об'єднання та редагування даних у форматі таблиць (DataFrame) та серій (Series).

Pandas забезпечує зручний та ефективний інтерфейс для роботи з даними та включає в себе такі можливості, як:

- о Зчитування та запис даних з різних форматів файлів, таких як CSV, Excel, SQL, JSON, HTML, XML та інших.
- о Робота зі статистичними даними, включаючи обчислення медіани, середнього значення, кореляції та інших метрик.
- о Фільтрація даних та виконання операцій вибору, об'єднання, групування та агрегації.
- о Робота з датами та часом, включаючи розрахунок інтервалів, перетворення форматів та інших операцій.
- о Візуалізація даних за допомогою графіків та діаграм.

Крім того, Pandas дозволяє легко інтегруватися з іншими бібліотеками Python, такими як NumPy, Matplotlib, Scikit-learn та інші, для більш ефективної роботи з даними та аналізу їх результатів.

Бібліотека NumPy (Numerical Python) - це бібліотека для наукових обчислень у мові програмування Python. Вона надає ефективні засоби для роботи з масивами даних, включаючи числові дані, які можна використовувати для розв'язання різноманітних задач, зокрема в математиці, фізиці, науці про дані та інших галузях.

Деякі з можливостей, які надає NumPy, включають в себе:

- о Масиви: NumPy надає широкий набір функцій для створення та роботи з масивами даних. Масиви в NumPy дозволяють зберігати великі об'єми даних та ефективно виконувати операції з ними.
- о Операції з масивами: NumPy надає широкий набір операцій для масивів даних, включаючи векторні операції, операції лінійної алгебри, операції статистики та ін.
- о Індекссування та зрізи: NumPy дозволяє легко доступатися до елементів масивів та виконувати зрізи масивів для вибірки певних даних.
- о Випадкові числа: NumPy містить функції для генерації випадкових чисел та розподілів випадкових величин.
- о Інтеграція з іншими бібліотеками: NumPy інтегрується з іншими бібліотеками Python, такими як SciPy, Pandas та Matplotlib, для більш ефективної роботи з науковими обчисленнями та аналізу даних.

NumPy є однією з найбільш важливих бібліотек у наукових обчисленнях у

Також відома як Sklearn - це бібліотека для машинного навчання у мові програмування Python. Вона надає реалізацію багатьох алгоритмів машинного навчання, таких як класифікація, кластеризація, регресія та ін. Scikit-learn надає такі можливості:

- о Реалізація багатьох алгоритмів машинного навчання: Scikit-learn містить реалізації багатьох класичних алгоритмів машинного навчання, таких як лінійна регресія, логістична регресія, метод опорних векторів (SVM), кластеризація K-means та багато інших.

- о Легкий використання та налагодження: Scikit-learn є досить простою бібліотекою для використання та налагодження, оскільки вона містить прості та інтуїтивні інтерфейси для багатьох задач машинного навчання.

- о Обробка та підготовка даних: Scikit-learn містить інструменти для попередньої обробки та підготовки даних, таких як масштабування, кодування категоріальних змінних, заповнення пропущених значень та інше.

- о Оцінка моделей: Scikit-learn надає інструменти для оцінки та порівняння моделей машинного навчання, таких як крос-валідація, ROC-криві, матриці помилок та інше.

- о Візуалізація: Scikit-learn містить інструменти для візуалізації даних та моделей машинного навчання, такі як графіки, діаграми розсіювання та інше.

Scikit-learn є однією з найбільш популярних бібліотек для машинного навчання у Python.

Для роботи з модулем обробки тексту застосовуємо наступні бібліотеки (рис. 3.13):

```
import pandas as pd
import numpy as np
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.metrics.pairwise import cosine_similarity
```

Рисунок 3.13 – Бібліотеки Python для роботи з модулем обробки тексту

Для реалізації контрольного прикладу, описаного в розділі 3.3.1, використано середовище Google Colab.

, також відомий як Google Colab, - це безкоштовна послуга хмарних обчислень, яка дозволяє користувачам запускати та писати код у середовищі Jupyter Notebook, використовуючи веб-браузер.

Основні особливості Google Colab:

Є безкоштовною платформою, яка забезпечує безкоштовний доступ до GPU (графічних процесорів) та TPU (тензорних процесорів) для прискорення обчислень.

Підтримує мову програмування Python та має вбудовану підтримку для багатьох популярних бібліотек, таких як NumPy, Pandas, Matplotlib, TensorFlow, PyTorch та ін.

Дозволяє користувачам легко спільно працювати над проектами з допомогою можливості ділитися ноутбуками з іншими користувачами та працювати з ними в режимі реального часу.

Дозволяє зберігати файли в хмарі, обмінюватися ними між пристроями та користувачами.

Не потребує налаштування інфраструктури, такої як сервери, бази даних або інші інструменти, що забезпечує легку розгортку та використання.

Можна стверджувати, що Google Colab - це функціональний та зручний інструмент для розробки і виконання проектів з машинного навчання та інших областей науки та технологій.

Технічне забезпечення

Для роботи інтелектуальної системи із видачі рекомендацій у інтернет-магазині молодіжного одягу необхідне наступне обладнання:

Комп'ютер для розгортання та функціонування системи. Для роботи системи достатньо комп'ютера марки Lenovo IdeaPad 5 з наступними технічними характеристиками (таблиця 3.2).

Таблиця 3.2 – Технічні вимоги до комп'ютера, необхідні для функціонування інтелектуальної системи.

№ зп	Назва характеристики	Характеристика комп'ютера
	Розмір корпусу	Nettop
	Материнська плата	Lenovo IdeaPad 5
	Процесор	Ryzen 5 5500U
	Тактова частота процесора	2,5 ГГц
	Тип роз'єму	LGA1151
	Кількість ядер	
	Об'єм оперативної пам'яті	Гб
	Тип пам'яті	DDR4
	Кількість додаткових слотів пам'яті	Без слотів
	Відеокарта	AMD Radeon Vega 11
	Тип накопичувача	SSD
	Об'єм диска	256 Гб
	Операційна система	Win 11 Pro
	Кількість Ethernet портів	1 x RJ45
	Швидкість з'єднання	500 Мб/с
	Потужність блоку живлення	65 Вт
	Наявність Wi-Fi	Так
	Наявність Bluetooth	Так
	Інтерфейс USB 3.0	Так
	Вихід HDMI	Так
	Вихід DisplayPort	Так

Для підтримки високої швидкості обчислень пропонується використання паралельної обробки даних, розподіленої обробки даних, обчислювальних кластерів та інших технологій. З метою збереження даних під час аварійних ситуацій і збоїв, важливо забезпечити систему резервних копіюванням даних.

3.3.4. Організаційно- економічне забезпечення

Створення будь-якого програмного продукту вимагає оцінювання його економічної ефективності. Створення програмного продукту є доцільним лише за умови, що прибутки, отримані від продажу чи експлуатації системи будуть

вищими ніж витрати на розробку та підтримання. В іншому випадку, продукт буде вважатися неефективним.

При створенні рекомендаційної системи виконуються наступні види робіт:

- дослідження теоретичних відомостей;
- аналіз аналогічних на ринку інструментів;
- аналіз існуючих методів для рекомендаційних систем;
- моделювання процесу створення рекомендацій;
- формування корпусів слів;
- створення списку стоп-слів;
- препроцесинг тексту;
- здійснення аналізу;
- створення інтерфейсу системи;
- створення серверної частини;
- тестування системи;
- введення в експлуатацію.

Приблизний кошторис витрат на реалізацію проекту представлений в таблиці 3.3.

Таблиця 3.3 – Кошторис витрат на створення веб-системи

Найменування робіт	Вартість години (грн)	Кількість годин	Загальна вартість (грн)
Дослідження теоретичних відомостей			
Аналіз аналогічних на ринку інструментів			
Аналіз існуючих методів для рекомендаційних систем			
Моделювання процесу створення рекомендацій			
Формування корпусів слів			

Препроцесинг тексту			
Здійснення аналізу			
Створення інтерфейсу системи			
Створення серверної частини			
Тестування системи			
Усього			

Оцінена підсумкова вартість реалізації складає 42 500 грн.

Запуск системи в роботи дозволить значно збільшити кількість відвідувачів інтернет-магазину молодіжного одягу, підвищити конверсію, і відповідно, прибутки підприємства.

ВИСНОВКИ

За результатами виконання кваліфікаційної магістерської роботи сформовані наступні висновки.

Важливим напрямом розвитку електронної комерції рекомендаційні системи, що використовують Natural Language Processing (NLP). Це технологія, що дозволяє переводити людську мову в считуваємий комп'ютерами формат, придатний для аналізу. В рамках роботи досліджена галузь видачі рекомендацій відвідувачам інтернет-магазину (recommendation engine), що пропонує схожі товари на основі спільних атрибутів або прогнозування поведінки користувачів.

В ході дослідження було виявлено, що на сучасному ринку IT-продуктів існують величезна кількість технічних рішень, придатних для використання. Проте, вони мають різну ефективність, яка сильно залежить від предметної галузі.

Основними підходами для видачі пропозиції є: фільтрування на основі вмісту, колаборативне фільтрування, інтелектуальний аналіз правил асоціації, фільтрування на основі демографічних показників, системи контекстних рекомендацій.

У виконанні аналізу з метою пошуку схожих товарів можна визначити такі основні етапи: збір даних; очистка даних; обробка тексту засобами NLP; аналіз отриманих результатів.

В кваліфікаційній магістерській роботі реалізовано аналіз опису товарів інтернет-магазину молодіжного одягу з метою пошуку співпадінь. Шляхом обробки інформації про наявні товари, їх опис, було здійснено векторне представлення та сформована матриця, яка відображає рейтинг схожості товарів. Також було реалізоване візуальне представлення результатів.

Створено прототип інтерфейсу користувача системи, описано програмне та технічне забезпечення. Розраховано вартість створення системи, яка оцінена у 42500 грн.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

acobs W. How AI Product Recommendations Can Benefit Your Online Store [Електронний ресурс]. – Дата візиту 06.05.2023. – Режим доступу до ресурсу:

aj N. How to Build a Recommendation System [Електронний ресурс]. – Дата візиту 05.05.2023. – Режим доступу до ресурсу:

sed Recommendation System. [Електронний ресурс]. – Дата візиту 18.04.2023. – Режим доступу до ресурсу: <https://pub.towardsai.net/how-to-build-a-content-based->

окументація NumPy [Електронний ресурс]. – Дата візиту 19.04.2023. – Режим доступу до ресурсу: <https://numpy.org/project/numpy/>

окументація Pandas [Електронний ресурс]. – Дата візиту 19.04.2023. – Режим доступу до ресурсу: <https://pandas.pydata.org/>

окументація Scikit-learn [Електронний ресурс]. – Дата візиту 20.04.2023. – Режим доступу до ресурсу: <https://scikit-learn.org/>

окументація Collab [Електронний ресурс]. – Дата візиту 20.04.2023. – Режим доступу до ресурсу: <https://colab.research.google.com/>

[Електронний ресурс]. – Дата візиту 10.05.2023. – Режим доступу до ресурсу:

[Електронний ресурс]. – Дата візиту 10.05.2023. – Режим доступу до ресурсу:

окументація Scikit-learn [Електронний ресурс]. – Дата візиту 11.05.2023. –

крІнформ [Електронний ресурс]. – Дата візиту 11.05.2023. – Режим доступу до

р

е

с

и

[Електронний ресурс]. – Дата візиту 11.05.2023. – Режим доступу до ресурсу:

i

k

i

p

ing Xiong The Impact of Artificial Intelligence and Digital Economy Consumer
Online Shopping Behavior on Market Changes [Електронний ресурс]. – Дата візиту
9.05.2023. – Режим доступу до ресурсу: [https://www.researchgate.net/publication/354111111](#)

o

y

affrey Pennington, Richard Socher, Christopher D. Manning GloVe: Global
Vectors for Word Representation [Електронний ресурс]. – Дата візиту 9.05.2023. –
Режим доступу до ресурсу: <https://nlp.stanford.edu/projects/glove/>

an Berthiaume Newegg integrates ChatGPT AI into online shopping experience
[Електронний ресурс]. – Дата візиту 9.05.2023. – Режим доступу до ресурсу:

a

Blog [Електронний ресурс]. – Дата візиту 9.05.2023. – Режим доступу до
ресурсу: [https://www.newegg.com/ai-chatgpt/blog](#)

e

by Brand Recommendations are Important for your Business [Електронний ресурс].
у Дата візиту 12.05.2023. – Режим доступу до ресурсу:

p

сО. Сергеев-Горчинський, Г. В. Іщенко Інтелектуальний Аналіз Даних
[Електронний ресурс]. – Дата візиту 9.05.2023. – Режим доступу до ресурсу:

Titles [Електронний ресурс]. – Дата візиту 12.05.2023. – Режим доступу до

ерші кроки в NLP: розглядаємо Python-бібліотеку NLTK в реальному завданні [Електронний ресурс]. – 2020. – Режим доступу до ресурсу:

nference on Computing Informatics and Networks. 2021. P. 365–375

бробка природної мови (NLP) методами машинного навчання в Python [Електронний ресурс]. – 2022. – Режим доступу до ресурсу:

к вирішити 90% завдань NLP: покроковий посібник з обробки природної мови [Електронний ресурс]. – 2018. – Режим доступу до ресурсу:

снови Natural Language Processing для тексту [Електронний ресурс]. – 2019. – Режим доступу до ресурсу:

an Berthiaume Newegg integrates ChatGPT AI into online shopping experience [Електронний ресурс]. – Дата візиту 9.05.2023. – Режим доступу до ресурсу:

[Електронний ресурс]. – Дата візиту 9.05.2023. – Режим доступу до ресурсу:

ow Netflix’s Recommendation Engine Works? [Електронний ресурс]. – Дата візиту 9.05.2023. – Режим доступу до ресурсу:

artina Pranjić Consumer Behavior Analysis 101: How to Learn More About Your

p

e

[Електронний ресурс]. – Дата візиту 4.05.2023. – Режим доступу до ресурсу:

y

Similarity [Електронний ресурс]. – Дата візиту 9.05.2023. – Режим доступу до

c

ерші кроки в NLP: розглядаємо Python-бібліотеку NLTK в реальному завданні

[Електронний ресурс]. – Дата візиту 9.05.2023. – Режим доступу до ресурсу:

Додаток А.

```
import pandas as pd
import numpy as np
import re
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.metrics.pairwise import cosine_similarity

# Завантаження даних про товари та їх описи
df = pd.read_csv("products.csv")
products = df[['name', 'description']]

# Функція для очистки даних
def clean_text(text):
    # Видалення спецсимволів
    text = re.sub(r'[\^\w\s]', "", text)
    # Переведення в нижній регістр
    text = text.lower()
    return text

# Створення матриці векторів слів на основі описів товарів з виключенням стоп-слів і функцією очистки
vectorizer = CountVectorizer(stop_words='english').fit_transform(products['description'].apply(clean_text))

# Обчислення косинусної схожості між векторами описів товарів
similarity_scores = cosine_similarity(vectorizer)

# Функція для отримання рекомендацій товарів на основі назви товару
def get_recommendations(product_name, similarity_scores=similarity_scores, products=products):
    product_index = products[products['name'] == product_name].index.values[0]
    product_similarity_scores = list(enumerate(similarity_scores[product_index]))
    product_similarity_scores = sorted(product_similarity_scores, key=lambda x: x[1], reverse=True)
    product_similarity_scores = product_similarity_scores[1:6]
    product_indices = [i[0] for i in product_similarity_scores]
    return products.iloc[product_indices]['name']

# Функція для отримання рекомендацій товарів на основі назви товару
def get_recommendations(product_name, similarity_scores=similarity_scores, products=products):
    product_index = products[products['name'] == product_name].index.values[0]
    product_similarity_scores = list(enumerate(similarity_scores[product_index]))
    product_similarity_scores = sorted(product_similarity_scores, key=lambda x: x[1], reverse=True)
    product_similarity_scores = product_similarity_scores[1:6]
    product_indices = [i[0] for i in product_similarity_scores]
    return products.iloc[product_indices]['name']
```