

Ангеліна Хрип'як, Юрій Грицюк*

¹ НЛТУ України, Львів, Україна, 23khrupiak.a@nltu.lviv.ua

² НЛТУ України, Львів, Україна, ORCID: 0009-0003-5409-2043,
yurii.i.hrytsiuk@lpnu.ua

Система створення контенту та рекомендацій користувачу засобами машинного навчання

Анотація. Здійснено опис загальних характеристик системи створення контенту та рекомендацій користувачу засобами машинного навчання. Було розроблено систему, яка надає рекомендації на підставі схожості ключових слів елемента з іншими елементами обраної тематики. Таку систему можна застосувати в різних сферах – література, кіно, кулінарія і т.д. для формування контенту та надання рекомендацій, що відповідають інтересам користувача, створювати персоналізовані рекомендації та здійснювати пошук схожих елементів за ключовими словами.

Ключові слова – машинне навчання, персоналізований контент, ключові слова, рекомендації, косинусна подібність.

Історія виникнення рекомендаційних систем веде свій початок з розвитком інформаційних технологій та зростанням обсягу доступної інформації в мережі Інтернет. Застосування цих систем пов'язане з потребою у знаходженні затребуваних обсягів інформації, їхній фільтрації та видачі користувачу у потрібному йому вигляді, а також із зростанням їхнього інтересу до персоналізації такого виду діяльності.

Прогрес у галузі рекомендаційних систем пов'язаний з розвитком і широким використанням методів машинного навчання та алгоритмів оброблення великих обсягів даних. Актуальність застосування такої системи полягає в її здатності до адаптації та вдосконалення з часом. Машинне навчання дає змогу аналізувати значні обсяги інформації про предметну область знань, їхню взаємодію з запитом певних користувачів, їхні відгуки та якість попередніх відборів. На підставі такої інформації систему можна навчати розуміти уподобання відповідного користувачів і надавати йому рекомендації, які максимально відповідають його потребам [1].

Абстрактна модель системи. Косинусна схожість – міра подібності між двома векторами в просторі, яку використовують для порівняння текстових документів, зображень або будь-яких інших об'єктів, які можна подавати у вигляді векторів. У контексті реалізації проекту – зазначений показник є корисним для визначення ступеня схожості між різними елементами контенту [2].

У розроблюваній системі косинусна схожість та оброблення природної мови використано для створення ключових слів та обчислення подібності між відповідними елементами на їхній підставі згідно з таким порядком:

1. токенизація за словами, де спочатку текст розділяють на окремі слова або токени, що дає змогу системі розуміти текст на рівні окремих слів;
2. лематизація та стемінг тексту, коли зводять певні слова до їхньої базової форми або спільного кореня, що допомагає уніфікувати слова, які можуть мати різні форми подання;
3. стоп-слова, такі як "і", "або", "не" тощо вилучають, позаяк вони не мають змістовного значення, що допомагає зосередити роботу системи тільки на важливих словах;
4. TF-IDF – метод, який відображає важливість слова в конкретному документі порівняно з його вагомістю в усьому наборі документів, що допомагає виділити ключові слова, які найбільше характеризують кожен елемент;
5. косинусна подібність, де після отримання TF-IDF векторів для кожного елемента їх порівнюють між собою, що дає змогу виміряти кут між векторами та визначити ступінь їх подібності;
6. рекомендації на підставі схожості ключових слів, коли система порівнює кожен елемент з іншими елементами обраної тематики і виводить найбільш схожі елементи на підставі косинусної подібності таких слів;
7. пошук за власними ключовими словами користувача, коли система знайде найбільш схожий елемент серед тих, що вже мають визначені ключові слова, використовуючи їхню косинусну подібність.

Програмна реалізація. Найбільш відповідними технологіями для реалізації проекту було обрано методи машинного навчання, зокрема оброблення природної мови для аналізу текстового контенту та косинусну подібність для обчислення схожості між елементами. Ці методи було імплементовано з використанням мови програмування Python.

Розглянемо основні бібліотеки, які використовуються в коді:

- NLTK: використовується для попередньої оброблення текстових даних, як: токенізації, вилучення стоп-слів, лематизації та стемінгу описів елементів;
- Scipy: використовується для роботи з розрідженими матрицями при комбінуванні матриць TF-IDF та обчислення подібностей;
- NumPy: використовується для маніпуляцій з масивами та числових операцій, зокрема в операціях сортування та індексування;
- Sklearn: використовується для векторизації TF-IDF описів елементів та обчислення косинусної подібності між документами [3].

На рис. 1,*а* зображено форму для введення ключових слів, наприклад фільму, а на рис. 1,*б* – результат його пошуку за ключовими словами.

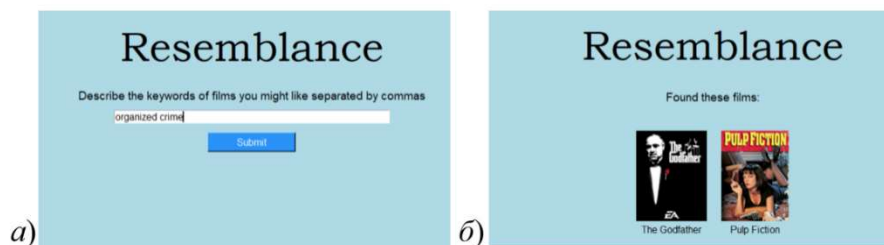


Рисунок 1. Введення ключових слів до фільмів (*а*), результати їхнього пошуку (*б*)

На рис. 2,*а* зображено введення назви схожого фільму, а на рис. 2,*б* – результат їхнього пошуку за введеною назвою.

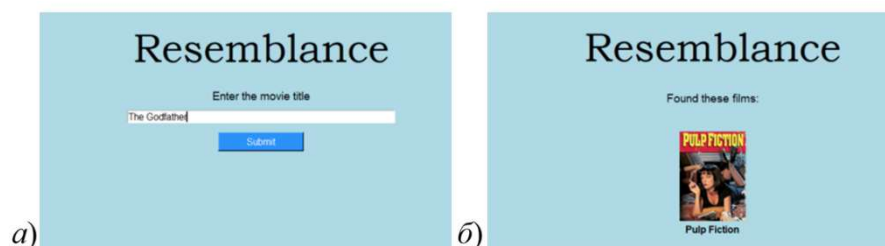


Рисунок 2. Введення назви схожого фільму (*а*), результати їхнього пошуку (*б*)

На рис. 3,а зображено введення ключових слів для різних тематик, а на рис. 3,б – результат їхнього пошуку за ключовими словами.

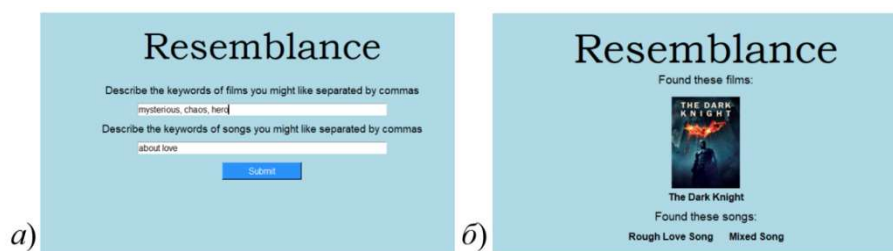


Рисунок 3. Введення ключових слів до різних тематик(а), результати їхнього пошуку (б)

Висновок. Найбільш відповідними технологіями для реалізації системи створення контенту та рекомендацій користувачу було обрано методи машинного навчання, зокрема – оброблення природної мови для аналізу текстового контенту та косинусна подібність для обчислення схожості між елементами. Внаслідок реалізації програмного забезпечення була розроблена інформаційна система, яка може аналізувати контент, створювати персоналізовані рекомендації на підставі інтересів користувачів та здійснювати пошук схожих елементів за ключовими словами. Подальші дослідження будуть спрямовані на вдосконалення методики машинного навчання системи для знаходження найбільш правдоподібних рекомендацій, затребуваних користувачем.

Список використаних літературних джерел

- [1] Rekomendatsiini systemy. [In Ukrainian]. URL: <https://skalar.ua/ua/expertise/recommender-systems>
- [2] Kosynus podibnosti. [In Ukrainian]. URL: https://uk.wikipedia.org/wiki/Косинус_подібності
- [3] 10 naikrashchykh bibliotek Python dlia mashynnoho navchannia ta ShI. [In Ukrainian]. URL: <https://www.unite.ai/uk/10-best-python-libraries-for-machine-learning-ai/>

A system for creating content and user recommendations using machine learning

Anhelina Khrypiak, Yurii Hrytsiuk

A description of the general characteristics of system for creating content and user recommendations using machine learning has been made. A system that creates recommendations based on keyword matching of elements with other elements of the selected topic has been developed. The significance of the system lies in the possibility of applying the developed system in various fields, such as literature, cinema, cooking, etc., to generate content and recommendations that meet the user's interests, while using only one system.