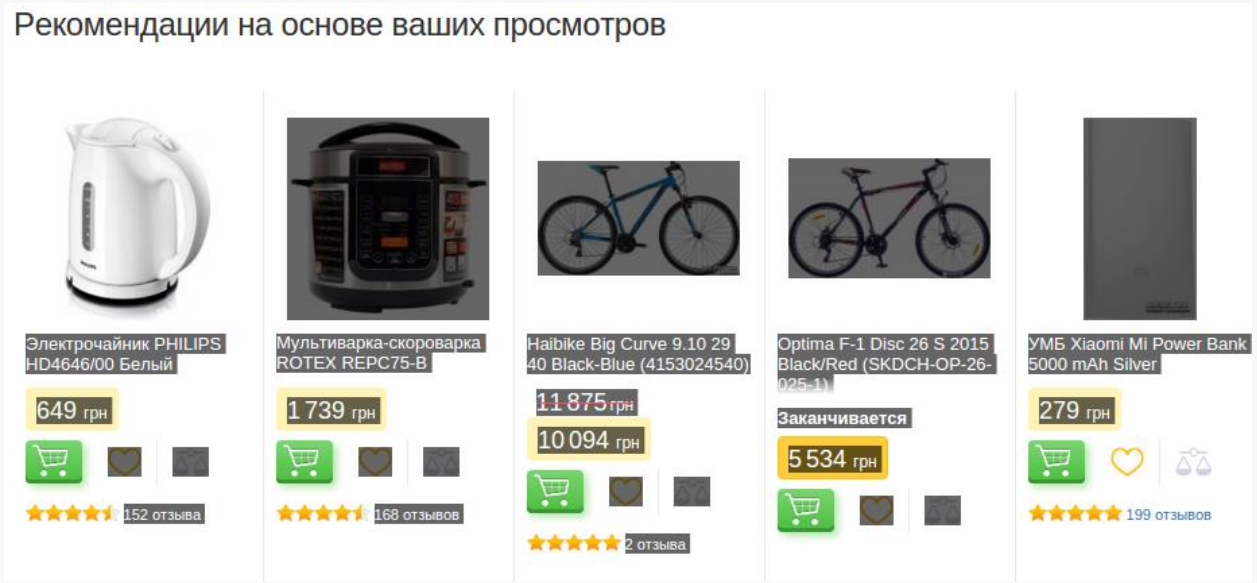
РЕКОМЕНДАЦІЙНА СИСТЕМА ДЛЯ ВИБОРУ ФІЛЬМІВ

Автор: Яковець Михайло Вікторович
Науковий керівник: Дідковська Марина Віталіївна

Актуальність

- Стрімке зростання кількості даних в мережі Інтернет
- Зростання кількості користувачів
- Переміщення комерційної сфери в Інтернет-простір



- **Мета роботи:**

1. Дослідження та розробка алгоритмів формування рекомендацій на осонові структурованих та неструктурованих даних.
2. Опис принципів роботи рекомендаційної системи для вибору фільмів.
3. Практична реалізація рекомендаційної системи для вибору фільмів на основі розробленого алгоритму.

- **Об’єкт дослідження:** рекомендаційні системи
- **Предмет дослідення:** методи та алгоритми формування списку рекомендацій фільму для користувача

Постановка задачі

- На основі
  - *вподобань конкретного користувача;*
  - *вподобань інших користувачів;*
  - *інформації про фільм*

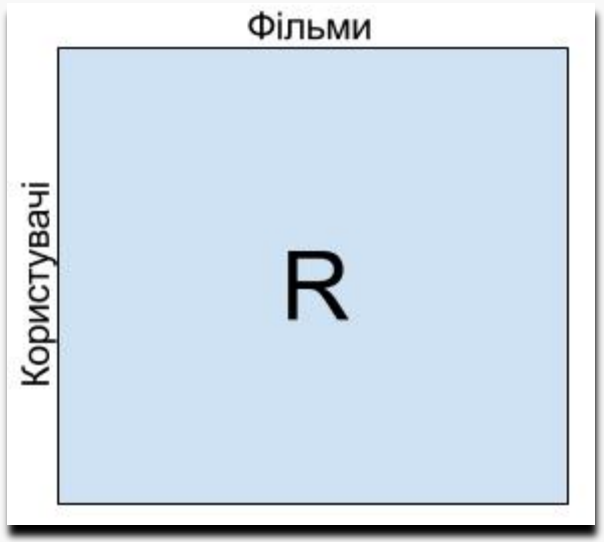
  видати релевантну рекомендацію.
- Перевірити якість рекомендацій.

Задача формування рекомендацій як практична системна задача

- Принципові невизначеності;
- наявність ризиків;
- суперечності численних цілей;
- наявність елементів різної природи та інформації різного характеру (кількісного і якісного), що циркулює між цими елементами.

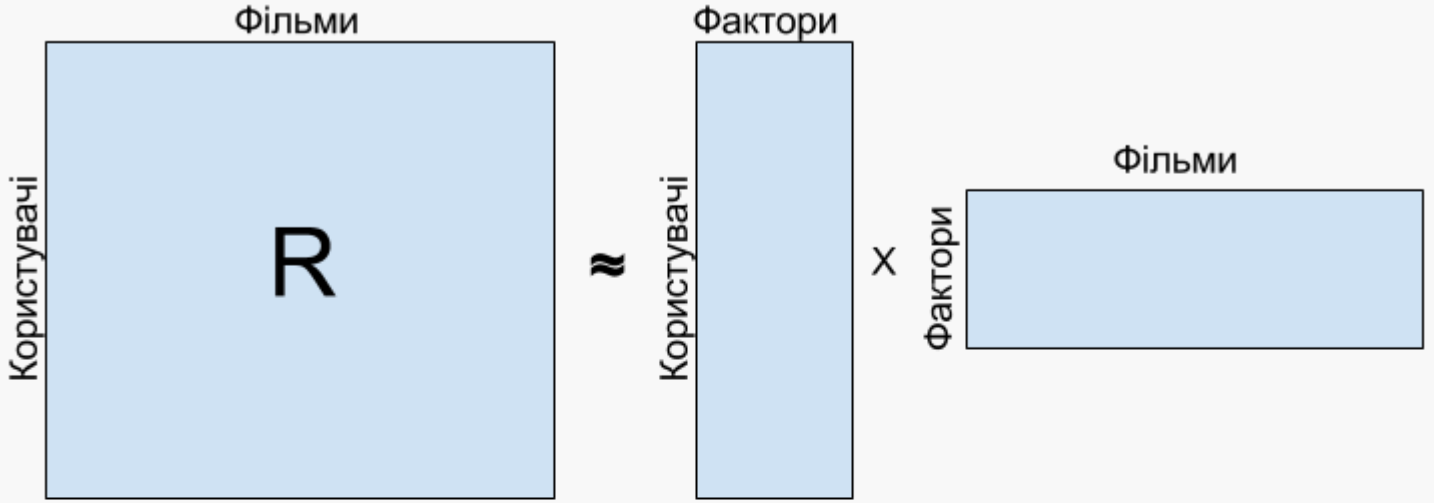
Декомпозиція предметної області

- *U* — множина користувачів, які оцінили фільми;
- *I* — множина фільмів, яка була оцінена;
- $r_{u,i}$— рейтинг *(від 1 до 10)*, якийкористувач*u* поставив фільму*i*;



Існуючі підходи до розв’язку та їхні недоліки

- Фільтрація на основі вмісту (Content filtering)
- Колаборативна фільтрація (Collaborative filtering)
  - *User-based, item-based*
  - *Виявлення прихованих факторів*



Пошук прихованих факторів

- Задача: за відомимиоцінкамикористувачівпередбачитиневідоміоцінки.

- Модель представлення оцінок:

$$r_{ui} = \mu + b_i + b_u + \mathbf{q}_i^* \mathbf{p}_u$$

де $r_{u,i}$ - оцінка з матриці рейтингів R, μ- середній рейтинг по вибірці, b_i- базовий рейтинг користувача, b_u- базовий рейтинг фільму, q_i- - вектор факторів фільму, p_u- - вектор факторів користувача.

Пошук прихованих факторів

- За відомим значенням оцінок знайдемо такі значення μ , b_i, b_u, q_i, p_u для усіх фільмів та користувачів, які найкраще наближають значення $r_{u,i}$до реальноїоцінки, з точки зору квадрату похибки:

$$\underbrace{\sum_{(u,i)\in\mathcal{D}} \left(\hat{r}_{ui}(\Theta) - r_{ui}\right)^2}_{\text{Якість на навчальній вибірці}} + \underbrace{\lambda \sum_{\theta\in\Theta} \theta^2}_{\text{регуляризація (λ - параметр регуляризації)}} \to \min_{\Theta}$$

- Для пошукуневідомихпараметріввикористано метод градієнтного спуску.

$$b_i = b_i + \gamma\left(e_{i,a} - \lambda b_i\right),$$
$$b_a = b_a + \gamma\left(e_{i,a} - \lambda b_a\right),$$
$$q_{a,j} = q_{a,j} + \gamma\left(e_{i,a}p_{i,j} - \lambda q_{a,j}\right)$$
$$p_{i,j} = p_{i,j} = \left(e_{i,a}p_{i,j} - \lambda p_{i,j}\right)$$

Перевірка якості моделі

- Середнє квадратичне відхилення:

$$RMSE = \sqrt{\frac{1}{|\mathcal{D}|} \sum_{(u,i) \in \mathcal{D}} (\hat{r}_{ui} - r_{ui})^2}$$

- Середня абсолютна похибка

$$MAE = \frac{1}{|\mathcal{T}|} \sum_{(u,i) \in \mathcal{T}} |\hat{r}_{ui} - r_{ui}|$$

Наївний баєсів класифікатор

- Задача: класифікуватівідгуки до фільму: позитивні, негативні, нейтральні.
- Теорема Баєса:

$$p(C|F_1,\ldots,F_n) = \frac{p(C)\ p(F_1,\ldots,F_n|C)}{p(F_1,\ldots,F_n)}.$$

- Модель наївного баєсового класифікатора:

$$p(C|F_1,\ldots,F_n) = \frac{1}{Z}p(C)\prod_{i=1}^{n} p(F_i|C)$$

Де *Z* - масштабний множник, що залежить від F_i(const);

P(C) =$\frac{\text{(кількість коментарів з класу } C)}{\text{(кількість усіх коментарів)}}$;

P(Fi/C) = $\frac{\text{(кількість входжень признаку } F_i \text{ до класу } C)}{\text{(кількість усіх признаків в класі } C)}$.

Перевірка якості моделі

- Для перевірки якості моделі використовується крос-валідація:

1. Вибірка ділиться N пособами на навчальну і перевірочну.

2. Класифікатор навчається на тестовій вибірці і випробовується на перевірочній.

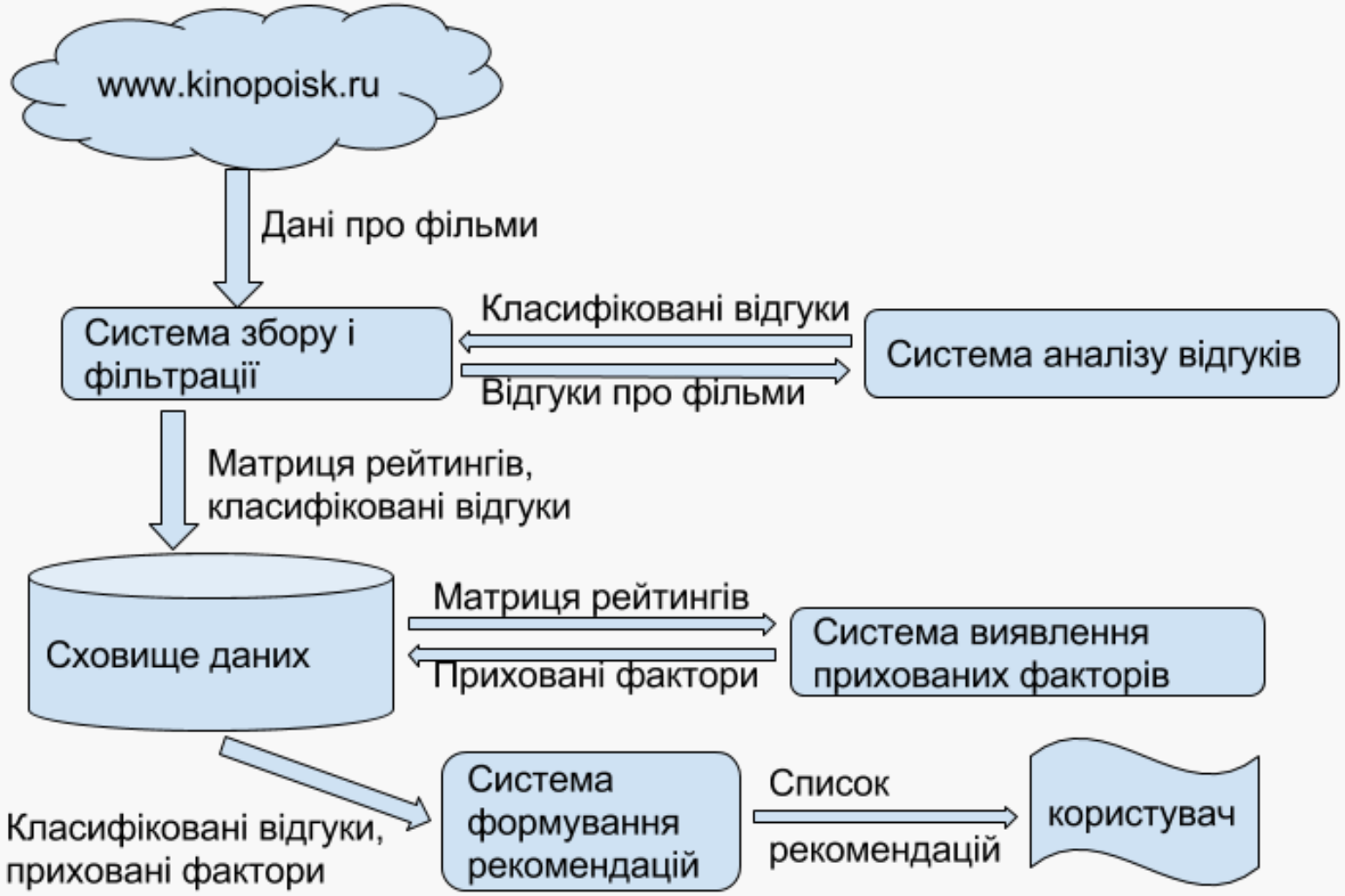
3. Для кожного експерименту рахується помилка, як відношення вдалих визначень класу, до кількості елементів в тестовій вибірці.

4. Загальна помилка рахується як середнє арифметичне помилок на кожному з розбиттів.

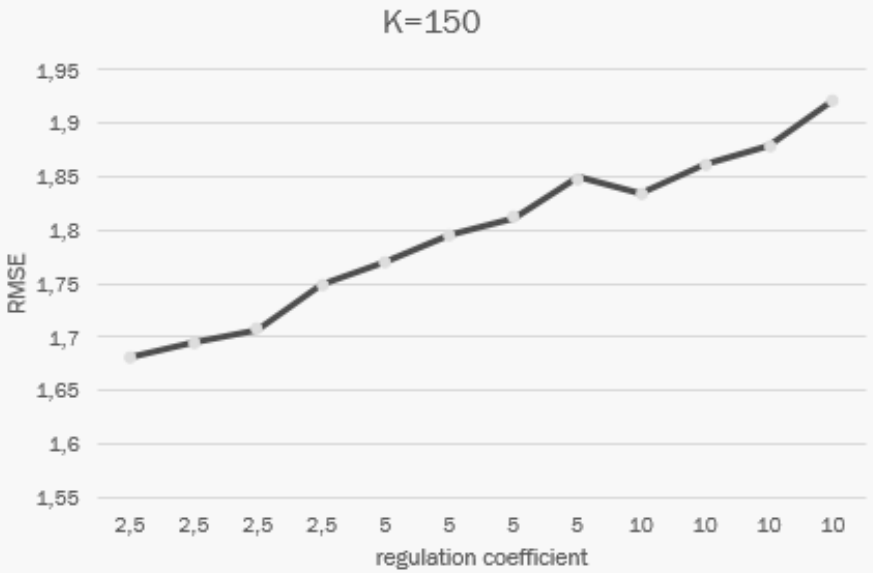
Двохетапна рекомендаційна система на основі оцінок користувачів і відгуків про фільм

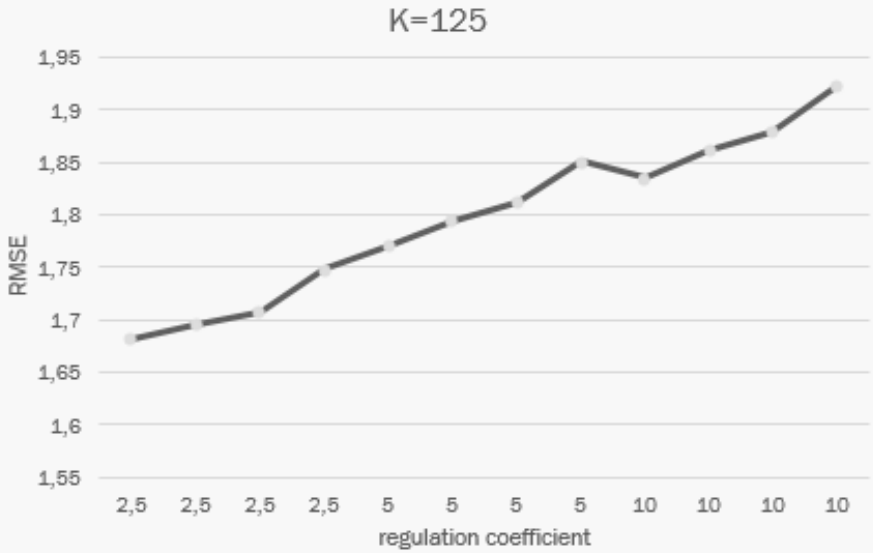
- Система збору і фільтрації даних
- Система виявлення прихованих факторів на основі оцінок користувачів
- Система формування рейтингу фільмів на основі класифікації відгуків про фільми
- Система формування списку рекомендацій

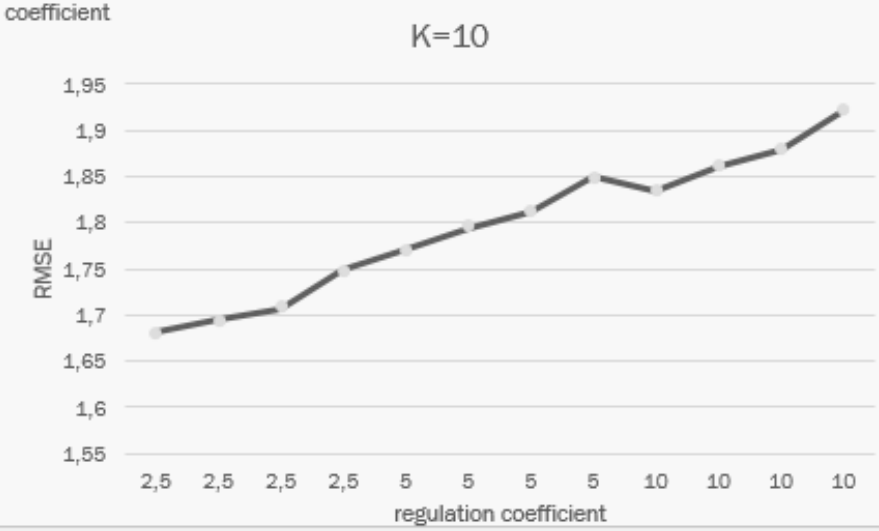
Схема рекомендаційної системи



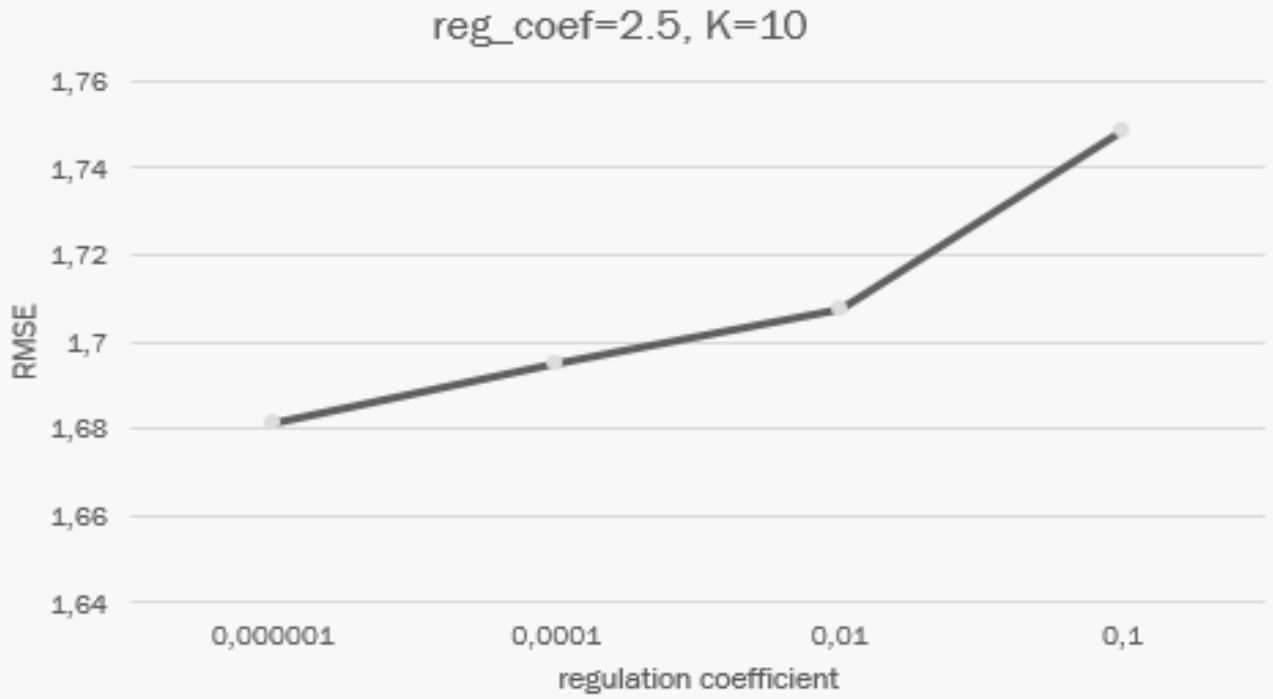
Аналіз практичних результатів







Аналіз практичних результатів



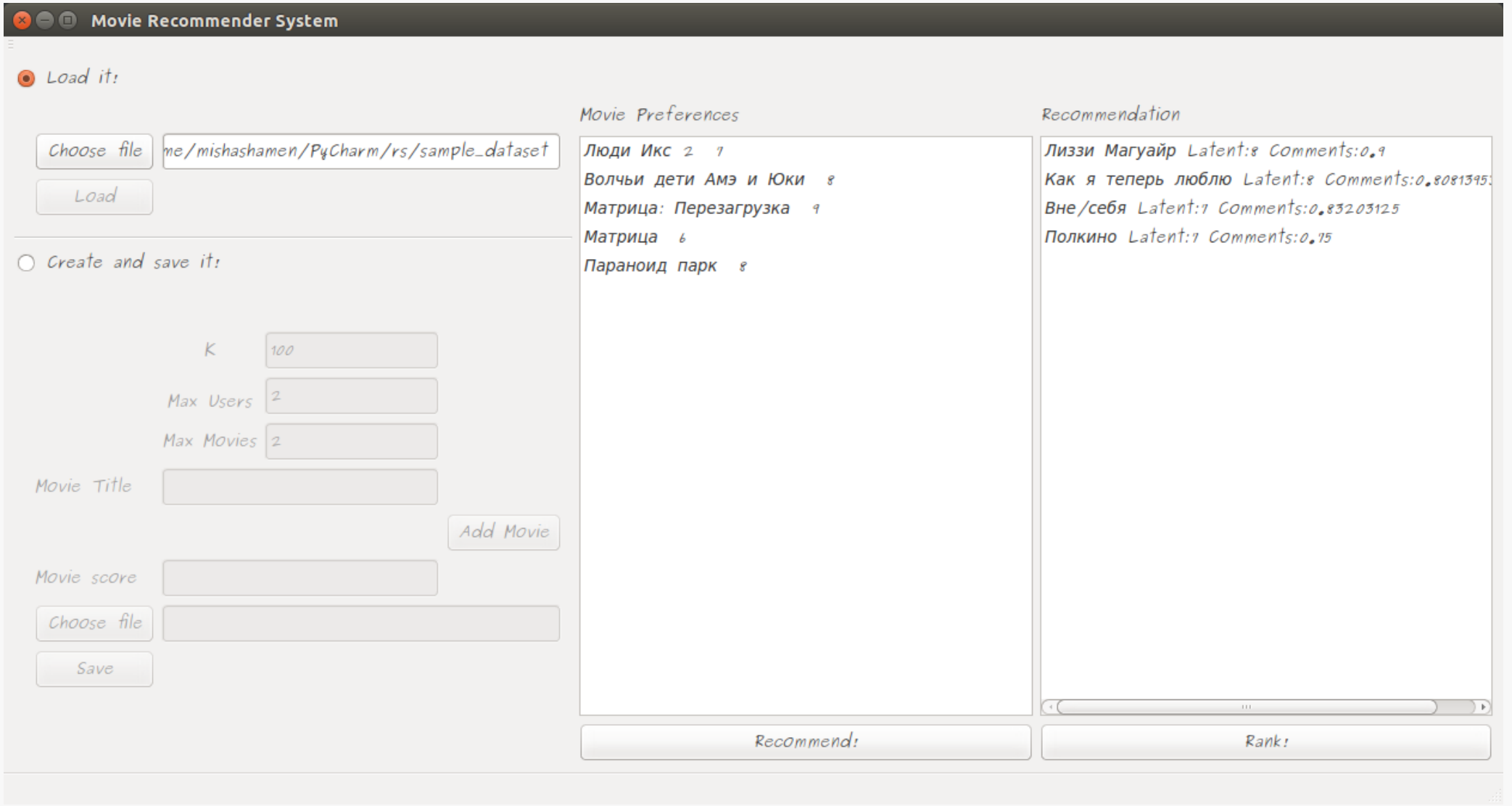
- Для зменшення похибки RMSE, потрібно зменшувати коефіцієнт регуляризації і збільшення кількості ітерацій навчання.

Аналіз практичних результатів

Признаки	Точність “Погані – Інші”	Точність “Хороші – Нейтральні”
Уніграми	81.5%	74.6%
Бігами	79.5%	74.8%
Змішані	79.8%	76.6%

- В ролі признаків належності до класу було обрано змішаний тим, тобто множину біграм і уніграм.
- Ієрархічний класифікатор є більш точним при класифікації на початковому рівні (Погані, інші відгуки)

Демо роботи продукту



Висновки

- Розроблено підсистему збору і фільтрації даних про фільми і користувачів
- Запропоновано і розроблено підсистему формування рекомендацій на основі оцінок користувачів та неструктурованих даних про фільми у вигляді відгуків про фільми
- Налаштовано параметри для алгоритму пошуку рекомендацій на основі прихованих факторів, щоб збільшити точність прогнозу за RMSE і MAE
- Налаштовано ієрархічний наївний баєсів класифікатор для аналізу відгуків

Шляхи подальшого розвитку

- Для формування рейтингу на основі даних про кількість позитивних, негативних, нейтральних відгуків, застосовується середня оцінка фільму, тобто:

$$\frac{(\text{поз.відгуки} + 0.75 * \text{нейтр.відгуки})}{(\text{кількість усіх відгуків})}$$

- Для збереженняданих про фільмивикористовуваласьфайлова система. Щобпідвищитишвидкістьроботи з даними, варто перенести ситему на реляційну БД.