

Data mining як засіб структурування
великих обсягів даних (Big Data) та
їх обробки

Зміст

- Big Data & Рекомендаційні системи
- Соціальні зв'язки – джерело даних для РС
- Постановка гіпотези
- Експериментальні дані
- Результати

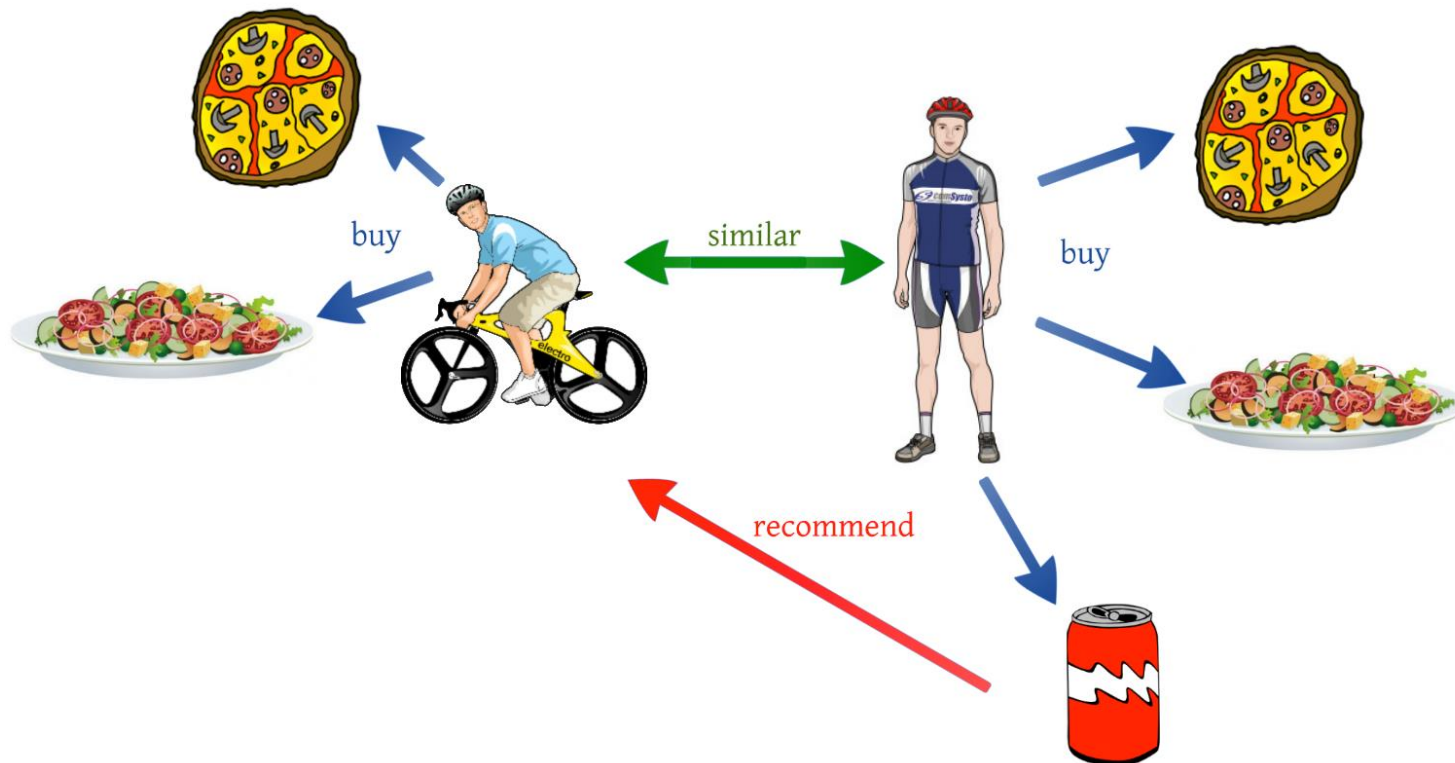
Big Data & Рекомендаційні СИСТЕМИ

- **Рекомендаційна система (РС)** — підклас системи фільтрації інформації, яка будує рейтинговий перелік об'єктів (фільми, музика, книги, новини, веб-сайти, події), яким користувач може надати перевагу.
- **Типи РС:**
 - Фільтрація вмісту
 - Колаборативна фільтрація



Фільтрація на основі вмісту

- Цей метод фільтрації ґрунтується на описі речей, що рекомендуються, і вподобань користувача за допомогою ключових слів.

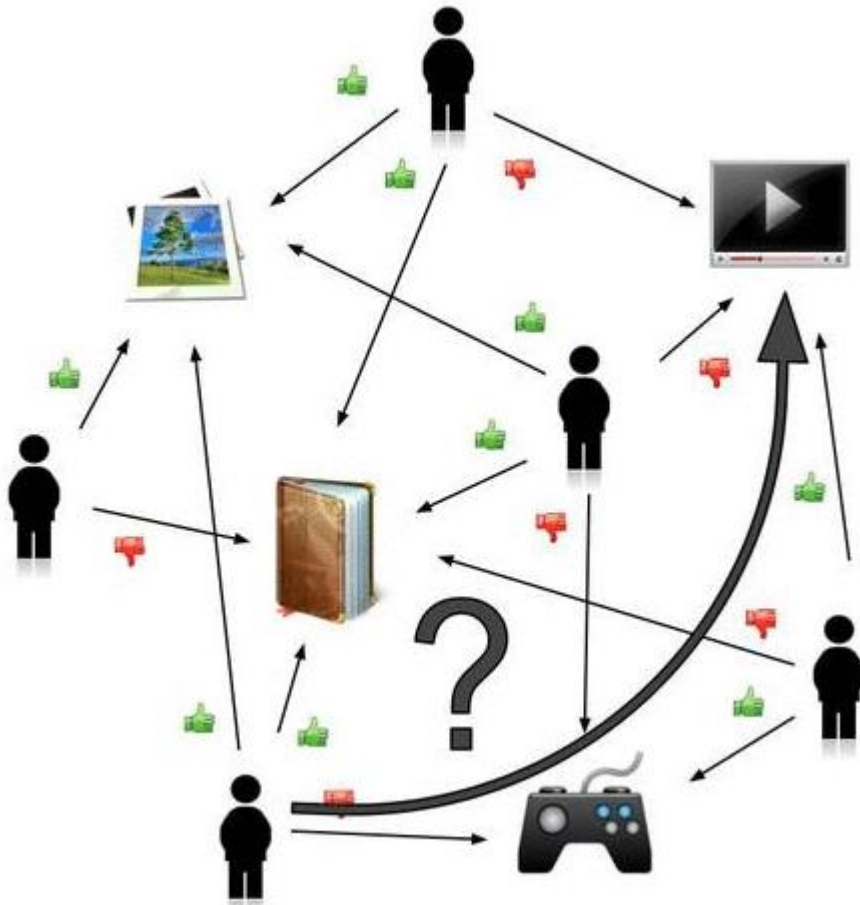



























Колаборативна фільтрація

- Це один з методів побудови прогнозу, який використовує відомі оцінки групи користувачів для прогнозування невідомих оцінок іншого користувача.

$$\text{simil}(x, y) = \cos(\vec{x}, \vec{y}) = \frac{\vec{x} \cdot \vec{y}}{\|\vec{x}\| \times \|\vec{y}\|} = \frac{\sum_{i \in I_{xy}} r_{x,i} r_{y,i}}{\sqrt{\sum_{i \in I_x} r_{x,i}^2} \sqrt{\sum_{i \in I_y} r_{y,i}^2}}$$

Приклади Колаборативної фільтрації



Netflix prize

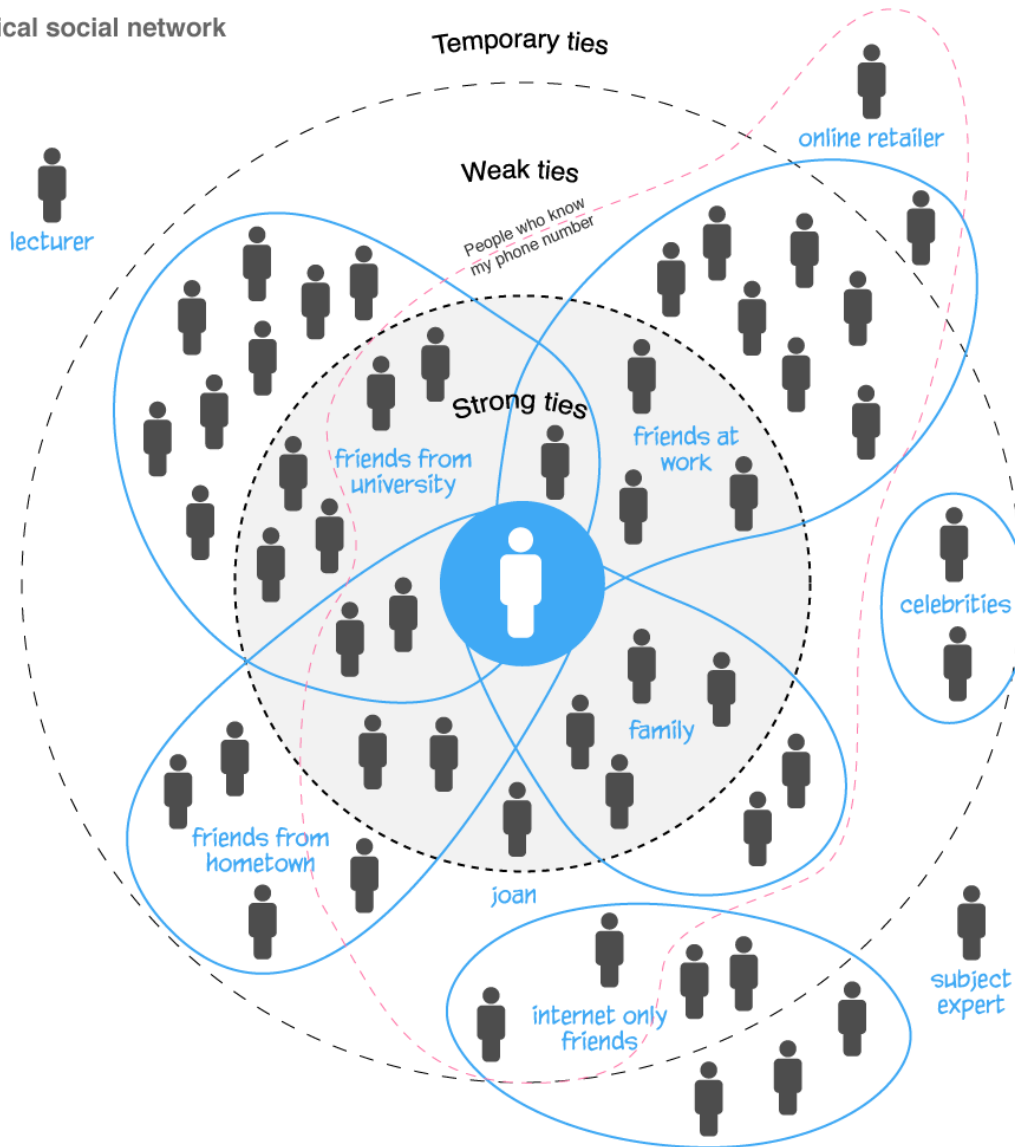
- 100,480,507 оцінок фільмів
- 480,189 користувачів
- 17,770 фільмів
- 15% покращення
- 1 billion USD

The image shows the Netflix logo, which consists of the word "NETFLIX" in a bold, white, sans-serif font. The letters are set against a solid red background. The logo is positioned on the right side of the slide, partially overlapping the list of statistics.

Соціальні зв'язки – джерело даних
для РС

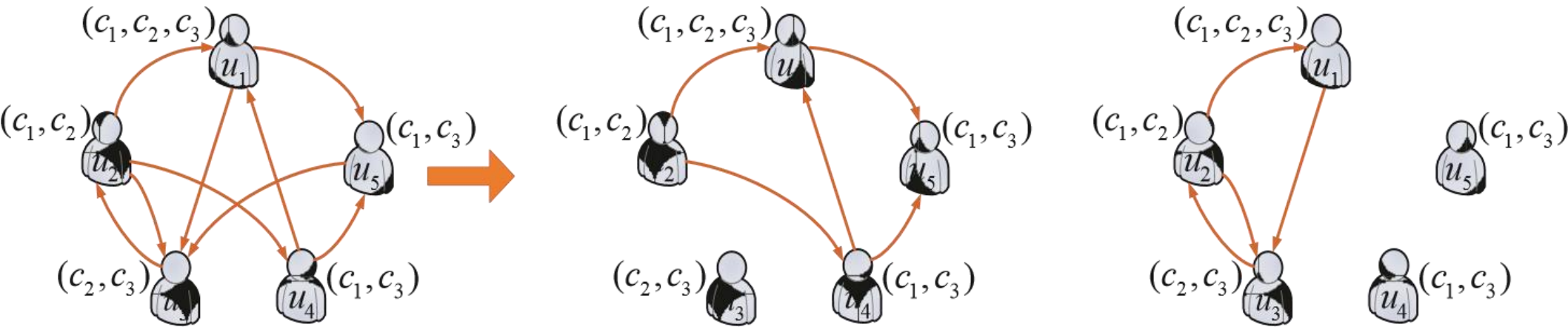
Друзі в соціальних мережах

Typical social network



Алгоритми аналізу друзів

- Категоризація друзів
- Оцінка схожості



Проблема “холодного старту”

Рішення:

- Випадкова рекомендація
- Найбільш популярна рекомендація
- Прогнозована на базі соціальних мереж

Постановка гіпотези

Профіль – додаткове джерело даних

The image shows a screenshot of a Facebook profile for Roman Kyslyi. The profile header includes a profile picture, the name "Roman Kyslyi", and an "Update Info" button. Below the header are navigation tabs for "Timeline", "About", "Friends 207", "Photos", and "More".

The "About" section contains the following information:

- Where have you worked in the past? (with a close button)
- 2 more pending items
- Studied IASA at Kiev Polytechnic Institute (Graduated in 2014)
- Lives in Kyiv, Ukraine
- From Kyiv, Ukraine (Born on March 27, 1992 (23 years old))
- Followed by 18 people

The "FRIENDS" section shows 207 friends, with three profile pictures visible: Alled Luviette, Andrii Kapranov, and Алексей Мась.

The main content area shows a status update from Roman Kyslyi, posted 19 hours ago. The text of the status is "Саймон як завжди, шикарна журналістика". Below the text is a video player showing a man in a military-style vest with a "PRESS" sign, holding a "NEWS" sign. The video title is "Selfie Soldiers: Ukraine" and the description is "As the conflict in Ukraine President Vladimir Putin' involvement. But a recer". The video source is "YOUTUBE.COM". Below the video are the interaction options "Like · Comment · Share" and a "1" like count.

At the bottom, there is a partial view of another post: "Roman Kyslyi shared a link."



Roman Kyslyi

Update Info

View Activity Log 2



Timeline

About

Friends 207

Photos

More ▾

Likes

+ Add Likes



All Likes 176

Movies

TV Shows

Music

Books

Sports Teams

Athletes

People

Restaurants

Apps and Games



SPIEGEL ONLINE ✓

News/Media Website

✓ Liked ▾

✓ Following



Süddeutsche Zeitung ✓

Newspaper

✓ Liked ▾

✓ Following



Tourclub "Globus"

Sports Team

✓ Liked ▾

✓ Following



Серіал "Гвардія"

TV Show

✓ Liked ▾

✓ Following



Upwork Ukraine

Company

✓ Liked ▾

✓ Following



site.ua

Media/News/Publishing

✓ Liked ▾

✓ Following

Профіль користувача

```
{
  first_name: "Roman",
  last_name: "Kyslyi",
  relationship_status: "Single",
  name: "Roman Kyslyi",
  locale: "en_US",
  gender: "male",
  verified: true,
  email: "kvrware@gmail.com",
  religion: "belive in Jedi Forse",
  birthday: "03/27/1992",
  link: "http://www.facebook.com/100001128635849",
  - location: {
    id: "111227078906045",
    name: "Kyiv, Ukraine"
  },
  + favorite_athletes: [...],
  + hometown: {...},
  timezone: 3,
  - education: [
    - {
      - school: {
        id: "109195292432741",
        name: "Polytechnic High School"
      },
      type: "High School",
      - year: {
        id: "136328419721520",
        name: "2009"
      }
    }
  ],
  + {...}
],
  updated_time: "2014-09-24T20:02:14+0000",
  id: "100001128635849"
}
```

```
{
  first_name: "Roman",
  last_name: "Kyslyi",
  relationship_status: "Single",
  name: "Roman Kyslyi",
  locale: "en_US",
  gender: "male",
  verified: true,
  email: "kvrware@gmail.com",
  religion: "belive in Jedi Forse",
  birthday: "03/27/1992",
  link: "http://www.facebook.com/100001128635849",
  - location: {
    id: "111227078906045",
    name: "Kyiv, Ukraine"
  },
  + favorite_athletes: [...],
  + hometown: {...},
  timezone: 3,
  - education: [
    - {
      - school: {
        id: "109195292432741",
        name: "Polytechnic High School"
      },
      type: "High School",
      - year: {
        id: "136328419721520",
        name: "2009"
      }
    }
  ],
  + {...}
],
  updated_time: "2014-09-24T20:02:14+0000",
  id: "100001128635849"
}
```


Можливості Facebook API

Select Permissions

User Data Permissions Extended Permissions

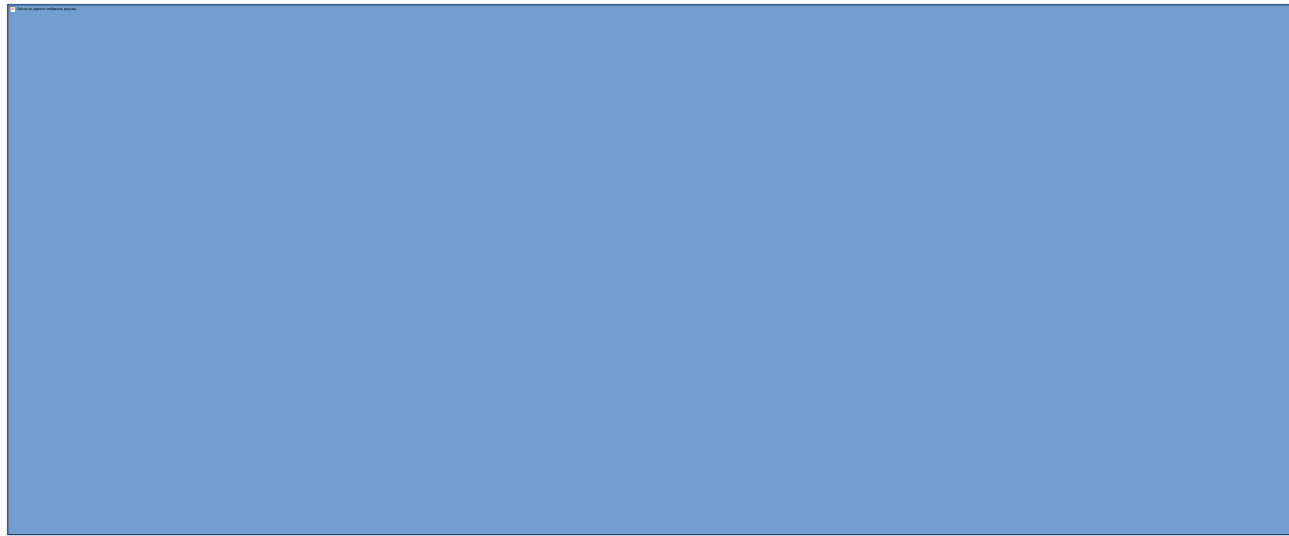
<input checked="" type="checkbox"/> user_about_me	<input checked="" type="checkbox"/> user_actions.books	<input checked="" type="checkbox"/> user_actions.fitness
<input checked="" type="checkbox"/> user_actions.music	<input checked="" type="checkbox"/> user_actions.news	<input checked="" type="checkbox"/> user_actions.video
<input checked="" type="checkbox"/> user_birthday	<input checked="" type="checkbox"/> user_education_history	<input checked="" type="checkbox"/> user_events
<input checked="" type="checkbox"/> user_friends	<input checked="" type="checkbox"/> user_games_activity	<input checked="" type="checkbox"/> user_groups
<input checked="" type="checkbox"/> user_hometown	<input checked="" type="checkbox"/> user_likes	<input checked="" type="checkbox"/> user_location
<input checked="" type="checkbox"/> user_managed_groups	<input checked="" type="checkbox"/> user_photos	<input checked="" type="checkbox"/> user_posts
<input checked="" type="checkbox"/> user_relationship_details	<input checked="" type="checkbox"/> user_relationships	<input checked="" type="checkbox"/> user_religion_politics
<input checked="" type="checkbox"/> user_status	<input checked="" type="checkbox"/> user_tagged_places	<input checked="" type="checkbox"/> user_videos
<input checked="" type="checkbox"/> user_website	<input checked="" type="checkbox"/> user_work_history	

Public profile included by default.

Get Access Token Clear Cancel

Модель SocialMF

- Ідея - профіль користувача має бути в середньому схожим на профіль його друзів, в межах похибки



$$\frac{1}{2} \sum_{(u,i)\text{obs.}} \left(R_{u,i} - \hat{R}_{u,i} \right)^2$$

Ваговий коефіцієнт

$$W = \sum_N^{i=1} (a_i * b_i) = \sum_N^{i=1} (a_i * d_i * f_i)$$

- a_i - передбачена оцінка речі(події) i в категорії c
- d_i – ваговий коефіцієнт
- f_i – індикатор, що показує чи присутня категорія в інтересах користувача

$$W = \frac{1}{2} \sum_{(u,i) \text{ obs.}} (R_{ui} R'_{ui})^2$$

Експериментальні дані

Набір даних Kaggle Event Recommendation Engine

- event
 - user
 - interested
 - not_interested
 - birthyear
 - location
 - timezone
 - likes
 - friends
- 90 000 подій
 - 140 000 користувачів

Набір даних Epinions

- user_id
- location
- gender
- birthyear
- likes
- ratings
- 74 000 користувачів
- 590 000 оцінок
- 140 000 різних предметів

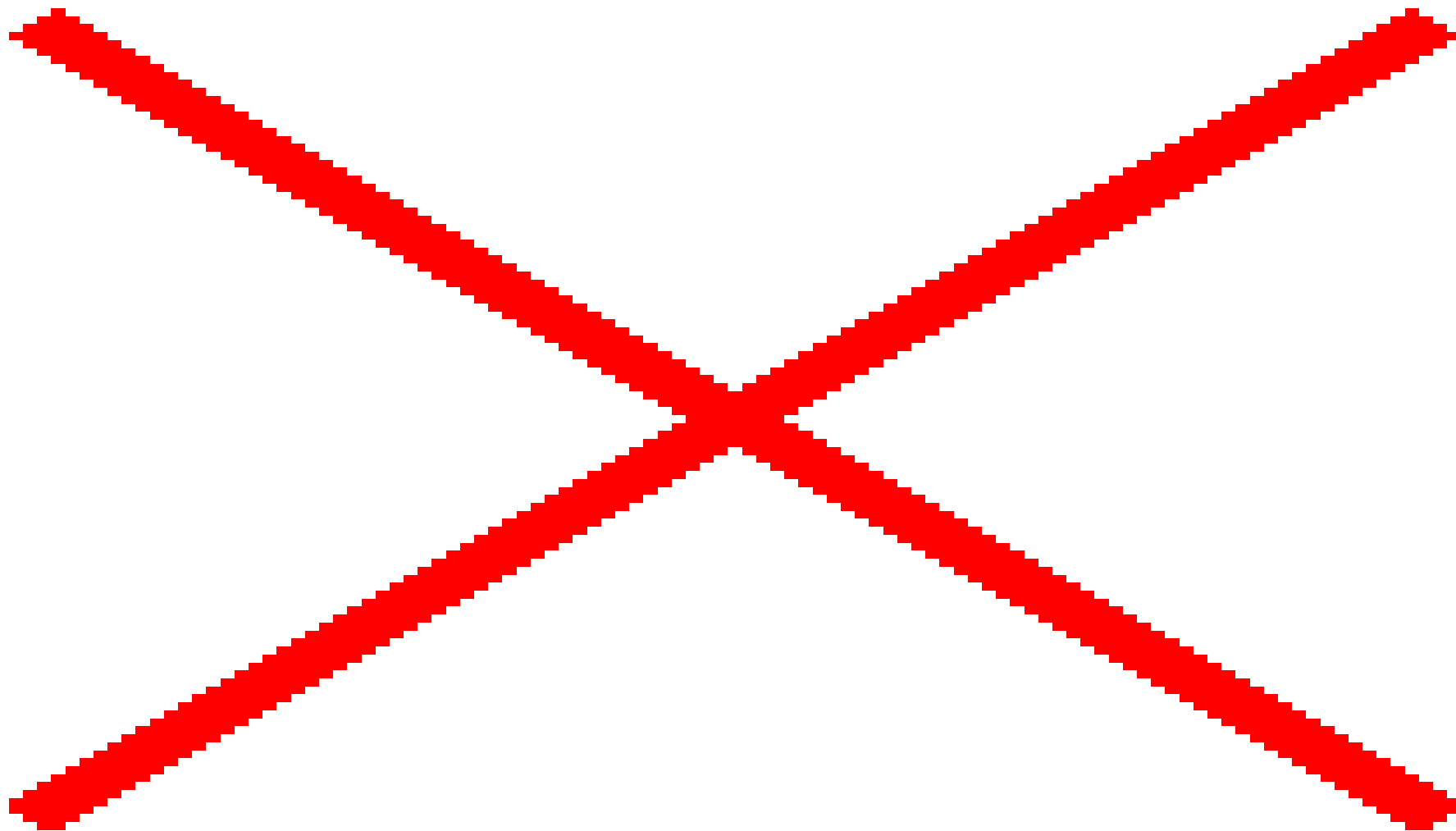
Метрика оцінювання

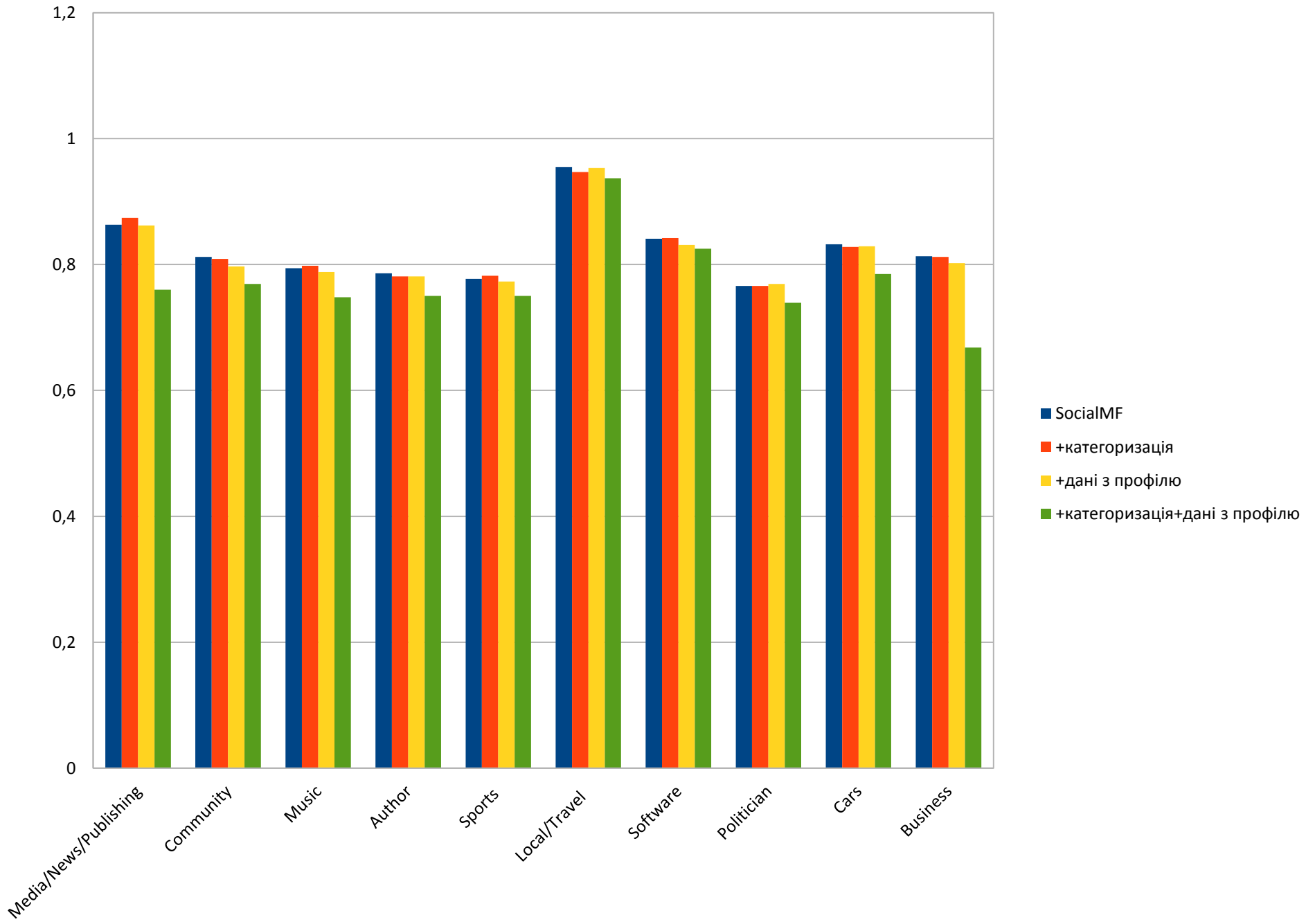
- 10 – fold cross validation
- Для оцінки точності використовувалась середньоквадратична похибка

$$RMSE = \sqrt{\frac{\sum_{(u,i) \in \mathcal{R}_{test}} (R_{u,i} - \hat{R}_{u,i})^2}{|\mathcal{R}_{test}|}}$$

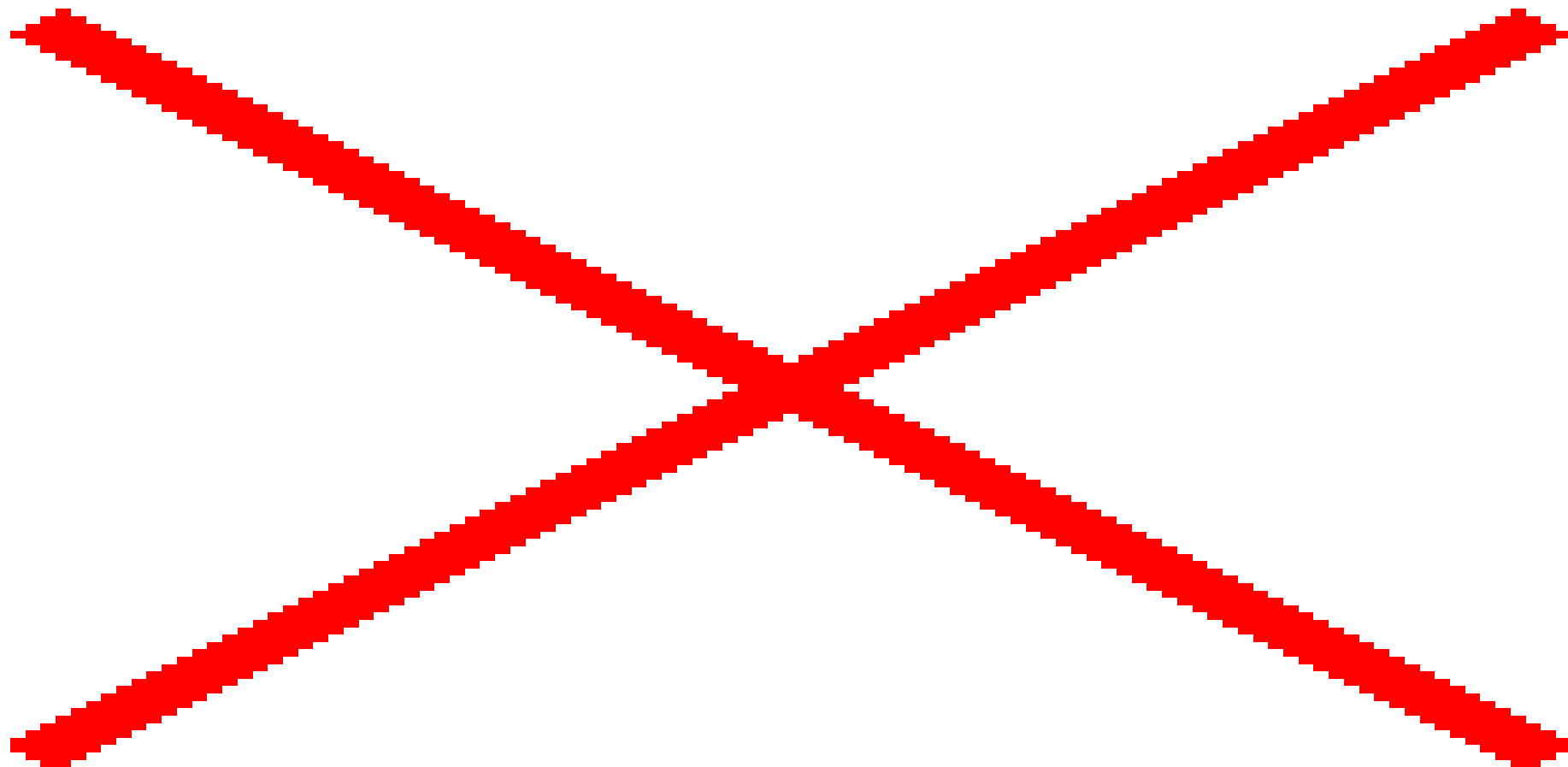
Результати

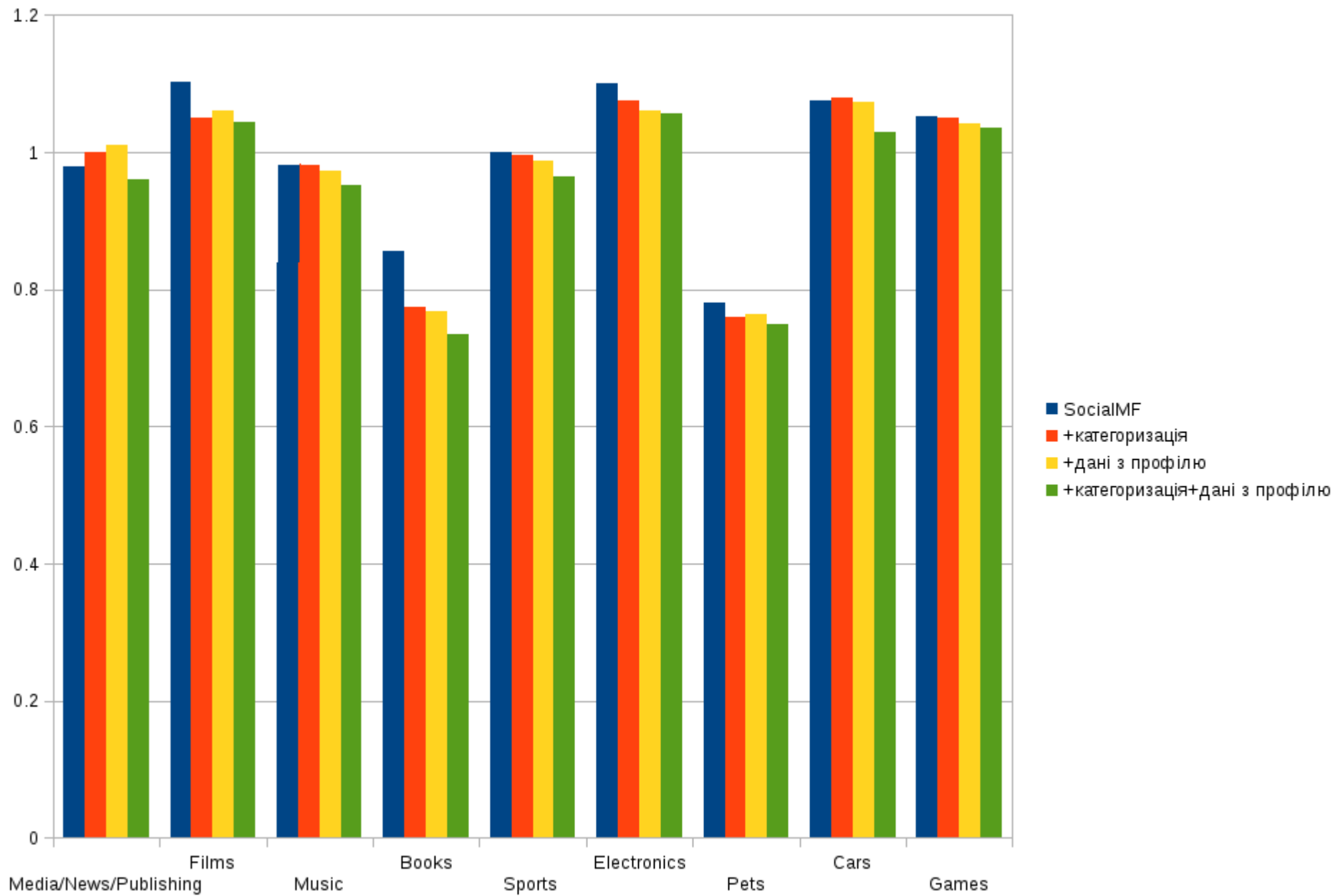
Результати тестування на наборі даних з Kaggle





Результати тестування на наборі даних з Epinions





Питання?