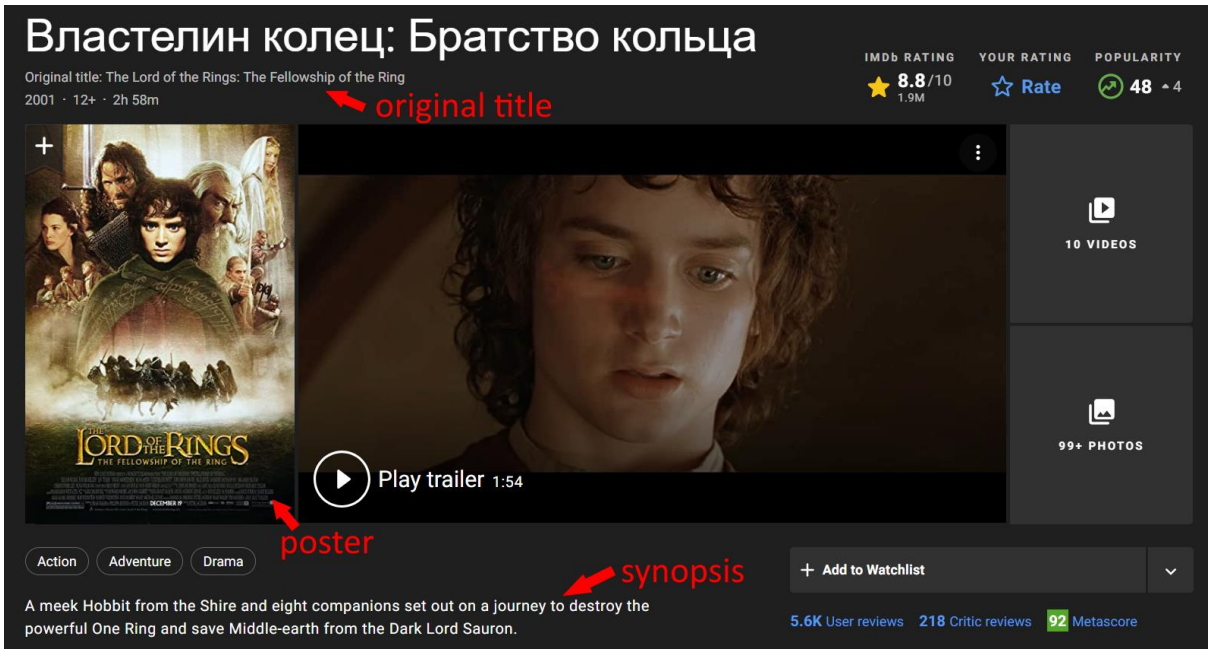
Лабораторное задание №2

Аннотация

В реальных индустриальных задачах зачастую возникает необходимость анализировать и строить предсказания с использованием данных разной модальности. Современные модели позволяют эффективно комбинировать векторные представления текста и изображений, отбирать наиболее информативные признаки и решать широкий спектр задач классификации и регрессии.

Задание

В данном задании вам предстоит по названию, постеру (изображению) и краткому описанию фильма предсказывать его жанр из заранее утвержденного списка категорий.



Для выполнения задания необходимо:

1) Собрать данные для обучения и валидации: названия, постеры и краткие описания фильмов с сайта **imdb.com**. Для этого предлагаем воспользоваться API IMDb: https://github.com/IMDb-API/IMDbApiLib.

Требование к dataset-у – не менее 200 примеров на каждую из описанных 24-х категорий:

1. Action
2. Adventure
3. Animation
4. Biography
5. Comedy
6. Crime
7. Documentary
8. Drama
9. Family
10. Fantasy
11. Film Noir
12. History
13. Horror
14. Music
15. Musical
16. Mystery
17. Romance
18. Sci-Fi
19. Short Film
20. Sport
21. Superhero
22. Thriller
23. War
24. Western

2) Выбрать алгоритмы и модели для извлечения информативных признаков из текста и изображений (на ваше усмотрение).
3) Выбрать модель для предсказания по извлеченным признакам меток жанров фильмов (на ваше усмотрение).
4) Обучить модель для решения задачи классификации фильмов, информация о которых представлена названием, постером и кратким описанием, на категории жанров.
5) Провалидировать модель на собранных данных.
6) Подготовить презентацию, в которой описать используемые алгоритмы и модели, способы учета мультимодальности данных, обоснование настройки гиперпараметров моделей, результатов на валидационном и тестовом датасете.
7) Тестовый датасет для оценки качества будет представлен за неделю до даты защиты.

Защита

Защита результатов выполнения задания будет производиться в формате очной презентации. Презентация должна содержать информацию, описанную в пункте 6. Кроме того, будет оцениваться качество классификации на примерах из тестового датасета.

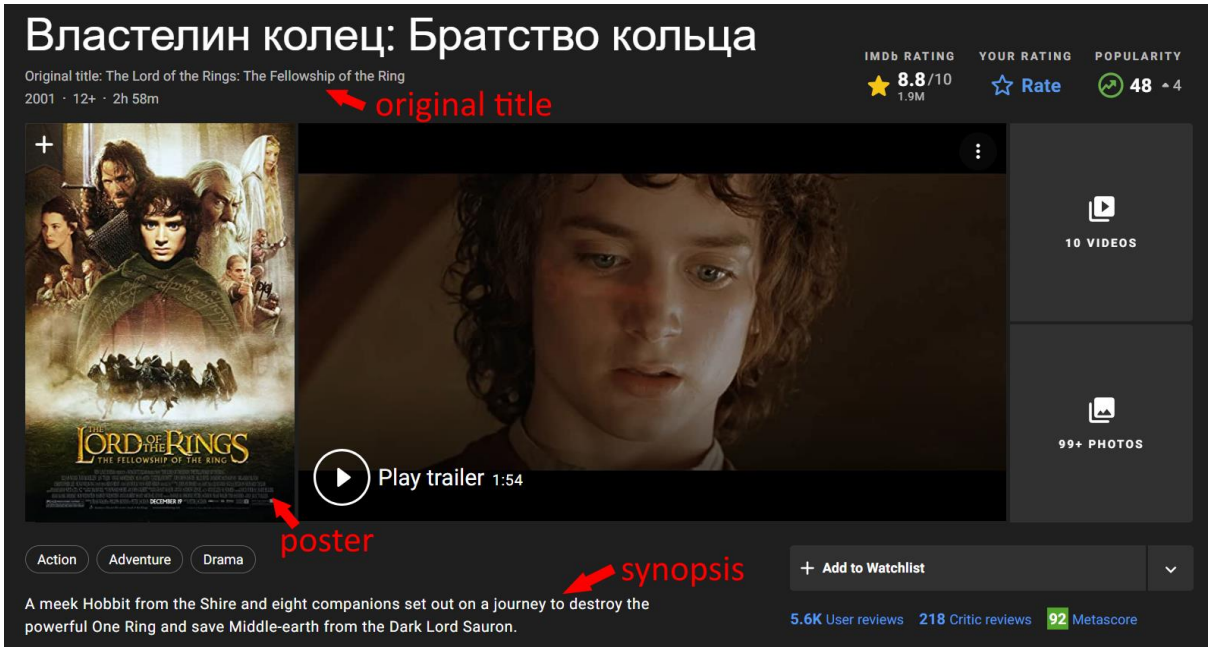
Laboratory Assignment №2

Abstract

In real-world industrial tasks it is often necessary to analyze and make predictions using data of different modalities. Modern models allow to combine text and image vector representations effectively, to select the most informative features and to solve a wide range of classification and regression problems.

Assignment

In this assignment you will be asked to predict the genre of a film from a pre-approved list of categories based on the title, poster (image), and a synopsis of the film.



To complete the assignment, you will need to:

1) Collect data for training and validation: film titles, posters, and short descriptions (synopsis) from **imdb.com**. For this purpose, we suggest using IMDb API: https://github.com/IMDb-API/IMDbApiLib.

Requirement for the dataset – at least 200 examples for each of the described 24 categories:

1. Action
2. Adventure
3. Animation
4. Biography
5. Comedy
6. Crime
7. Documentary
8. Drama
9. Family
10. Fantasy
11. Film Noir
12. History
13. Horror
14. Music
15. Musical
16. Mystery
17. Romance
18. Sci-Fi
19. Short Film
20. Sport
21. Superhero
22. Thriller
23. War
24. Western

2) Select algorithms and models to extract informative features from text and images.

3) Choose a model for predicting movie genre labels from extracted features.

4) Train the model to solve the problem of classifying movies, whose information is represented by title, poster, and synopsis, into genre categories.

5) Validate the model on the collected data.

6) Prepare a presentation describing the algorithms and models used, ways to account for multimodality of data, rationale for setting hyperparameters of models, results on the validation and test datasets.

7) Test dataset for quality assessment will be presented a week before the defense date.

Defense

The defense of the results of the assignment will be made in the format of a face-to-face presentation. The presentation should contain the information described in paragraph 6. In addition, the quality of the classification will be evaluated using examples from the test dataset.