INFORMATION SEARCH AND RECOMMENDATION SYSTEMS PROJECT: MOVIE RECOMMENDER

Evaluation report

Authors:
Weronika Zawadzka 12244068
Andrea Paletto 12243866
Gregoire Ville 12241475
Fatih Bugra Durmus 12246568

Klagenfurt University

15.07.2023

Contents

1 Introduction

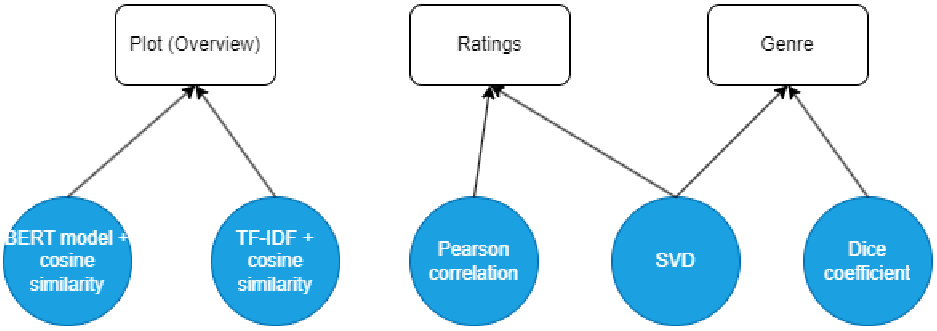
This document is an evaluation report for the final project realized in class Information Search and Recommendation Systems (623.610, 23S). The task was to create a movie recommender with front-end design, such that the client can search for the movie given its title and get in result 5 lists of movies recommended to the query movie. Each of 5 lists of 5 movies needed to be created using different technique. In this report, we will shortly describe our methods and later focus on the main topic, which is evaluation of our solution. It is important to note that we worked on restricted dataset (we removed movies having less than 300 ratings) as to not recommend too niche or obscure movies.

2 Project description

Short reminder of our approach before describing the evaluation.

2.1 Functions

Summary of our functions and what type of information they access is visible on the diagram below.



2.1.1 SVD

SVD, that is Singular Value Decomposition method, has been used. The whole approach can be described as hybrid one, combining collaborative filtering and content based methods. SVD is a matrix factorization technique, in which by representing the data in a lower-dimensional space we try to capture latent information and so measure similarity more effectively.

We computed the similarity scores between movies using a combination of ratings and genre information.

2.1.2 Bert model

Here we used BERT - powerful language model. We performed preprocessing on movies' overviews and tokenized them. Tokens were fed to the pretrained model and cosine similarity was calculated on the outputs.

2.1.3 Pearson

The Pearson similarity function measures the linear correlation between two sets of data. Here, Pearson correlation was computed using ratings of movies.

2.1.4 Dice

Here we used simple Dice coefficient, similarity measure commonly used in natural language processing tasks. We used the Dice coefficient to compare movie genres and find similar movies based on genre similarities.

2.1.5 TF-IDF

We used the technique of term frequency - inverse document frequency using preprocessed overviews data.

2.1.6 Add-ons on our similarity functions

We created "getNFirst" function, which given a list of movieID, returns the n first which have different normalized title (so that it doesn't return several times the same movie with different episodes). Of course, sometimes such behaviour of recommender could be desirable, but we decided to focus on diversity of our results.

2.2 Backend and Frontend

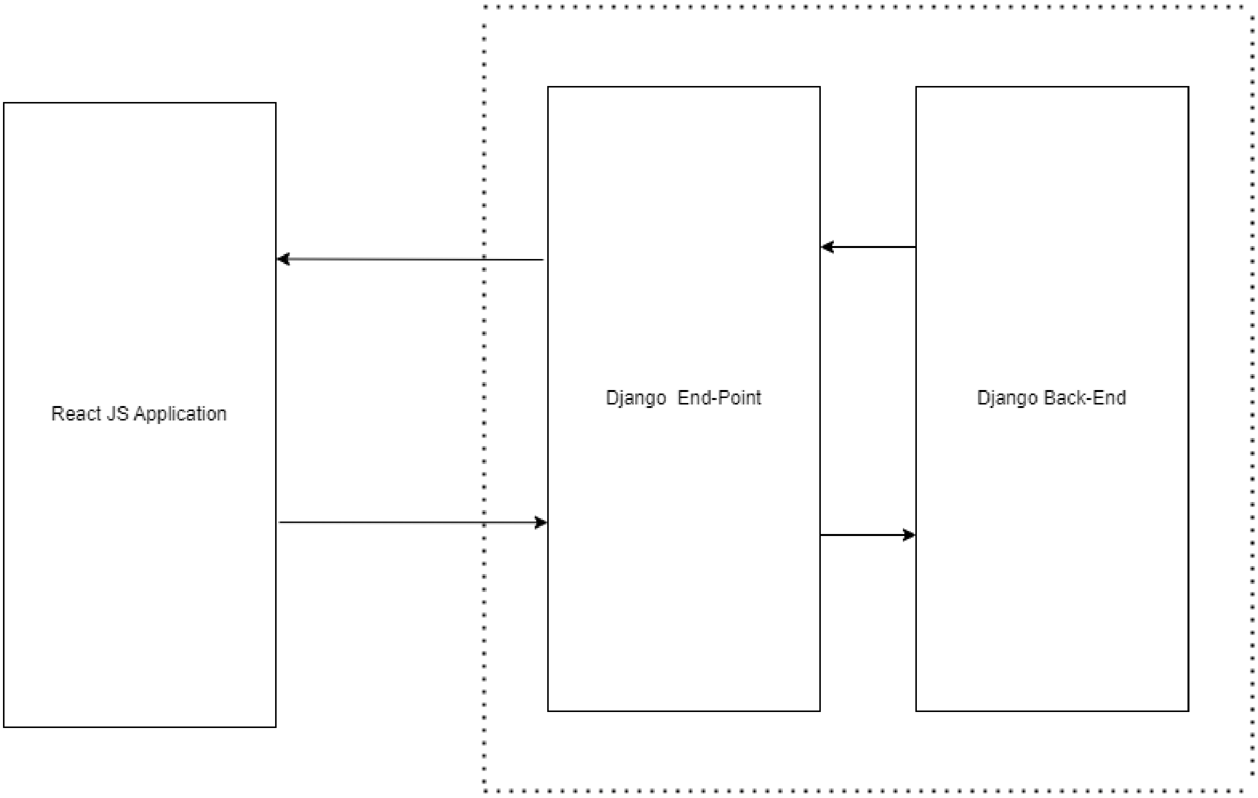


Figure 1: Diagram of our web app architecture

The architecture we chose for similar movies recommendations consists in two components. First, we made a backend with the framework Django, to handle data processing and implement similarity functions. This framework consists in splitting backend into applications, which notably contain a model file (in our case we used it to define our five similarity functions), and a view file to interact with the frontend. Then, we built an interactive frontend with ReactJS, to provide a user-friendly interface for movie similarity analysis. The final product looks like this:

First, we can write a movie in the search bar, and click on the correct movie in the dropdown select.

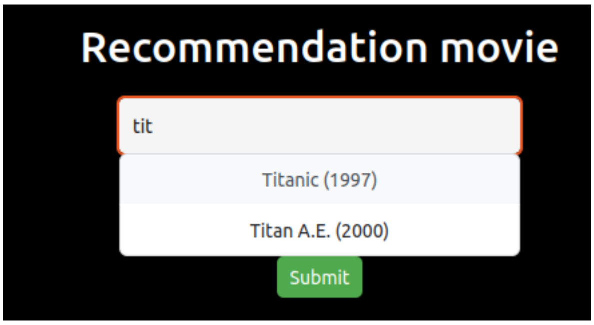
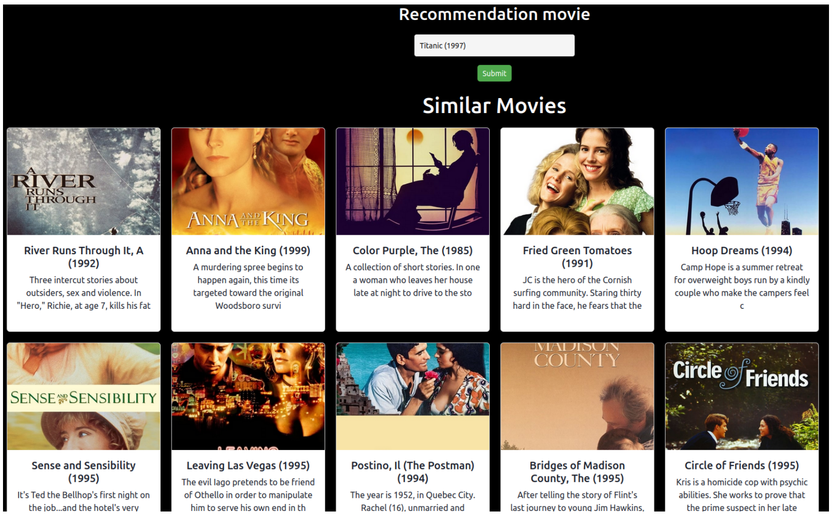


Figure 2: Search bar of our web app

Then, the backend is doing computations, and after about 10 to 15 seconds, we get the following output:



River Runs Through It, A (1992)
Three intercut stories about outsiders, sex and violence. In "Hero," Richie, at age 7, kills his fat

Anna and the King (1999)
A murdering spree begins to happen again, this time its targeted toward the original Woodsboro survi

Color Purple, The (1985)
A collection of short stories. In one a woman who leaves her house late at night to drive to the sto

Fried Green Tomatoes (1991)
JC is the hero of the Cornish surfing community. Staring thirty hard in the face, he fears that the

Hoop Dreams (1994)
Camp Hope is a summer retreat for overweight boys run by a kindly couple who make the campers feel c

Sense and Sensibility (1995)
It's Ted the Bellhop's first night on the job...and the hotel's very

Leaving Las Vegas (1995)
The evil Iago pretends to be friend of Othello in order to manipulate him to serve his own end in th

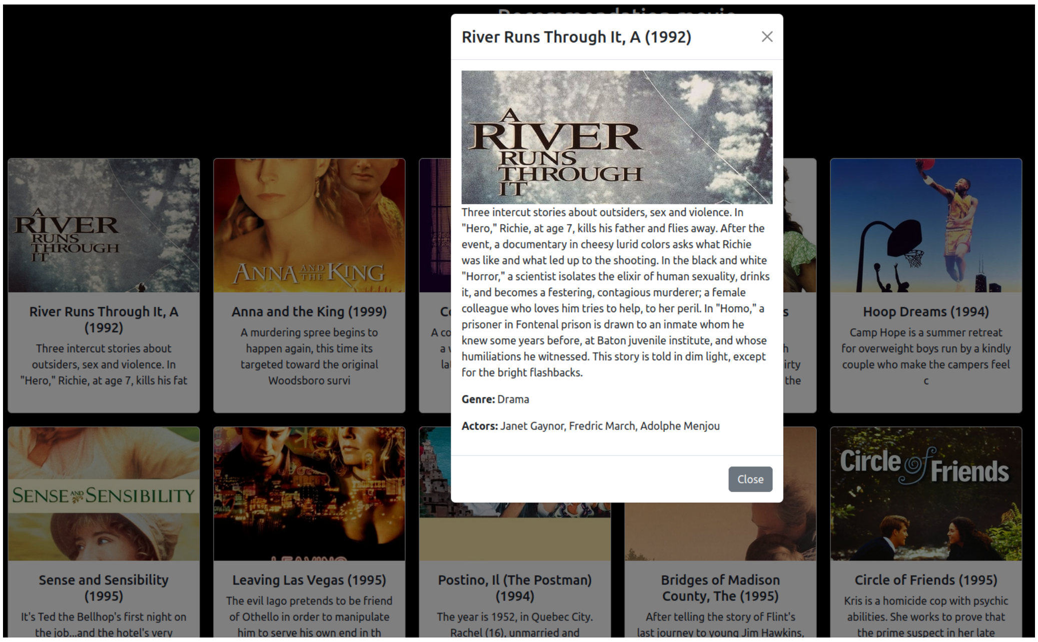
Postino, Il (The Postman) (1994)
The year is 1952, in Quebec City. Rachel (16), unmarried and

Bridges of Madison County, The (1995)
After telling the story of Flint's last journey to young Jim Hawkins,

Circle of Friends (1995)
Kris is a homicide cop with psychic abilities. She works to prove that the prime suspect in her late

Figure 3: Output of our web app : similar movies

One line corresponds to the five most similar movies to the one written in search bar according to one strategy (leftmost movie is the most similar, rightmost is the fifth most similar). In the order, we have the recommendations of BERT, Dice, TF-IDF, SVD and Pearson correlation strategies. Here we can already see the picture, title and beginning of overview of each movie, but if we click on one, a pop-up opens itself and shows extra information (not only the picture and title, but also the complete overview, the genres and the main actors).



River Runs Through It, A (1992)

Three intercut stories about outsiders, sex and violence. In "Hero," Richie, at age 7, kills his father and flies away. After the event, a documentary in cheesy lurid colors asks what Richie was like and what led up to the shooting. In the black and white "Horror," a scientist isolates the elixir of human sexuality, drinks it, and becomes a festering, contagious murderer; a female colleague who loves him tries to help, to her peril. In "Homo," a prisoner in Fontenal prison is drawn to an inmate whom he knew some years before, at Baton juvenile institute, and whose humiliations he witnessed. This story is told in dim light, except for the bright flashbacks.

Genre: Drama

Actors: Janet Gaynor, Fredric March, Adolphe Menjou

River Runs Through It, A (1992)
Three intercut stories about outsiders, sex and violence. In "Hero," Richie, at age 7, kills his fat

Anna and the King (1999)
A murdering spree begins to happen again, this time its targeted toward the original Woodsboro survi

Hoop Dreams (1994)
Camp Hope is a summer retreat for overweight boys run by a kindly couple who make the campers feel c

Sense and Sensibility (1995)
It's Ted the Bellhop's first night on the job...and the hotel's very

Leaving Las Vegas (1995)
The evil Iago pretends to be friend of Othello in order to manipulate him to serve his own end in th

Postino, Il (The Postman) (1994)
The year is 1952, in Quebec City. Rachel (16), unmarried and

Bridges of Madison County, The (1995)
After telling the story of Flint's last journey to young Jim Hawkins,

Circle of Friends (1995)
Kris is a homicide cop with psychic abilities. She works to prove that the prime suspect in her late

Figure 4: Pop-up for extra movie information in our web app

3 Evaluation

In this section we will evaluate our implemented approaches of the movie recommendation. In the task description, we were asked to analyze at least 20 movies. We divided evaluation into two parts. Fist one was manual, where we, members of this group, looked at the recommendation lists and try to detect anomalies. And second, taking into consideration the view of potential "customers", a survey. In total we evaluated 20 movies.

3.1 Manual evaluation

For the manual evaluation we chose 10 known to us movies to analyse them in detail. The titles of the movies were: American Psycho, the Exorcist, Of Mice and Men, Peter Pan, Robin Hood, Karate Kid, Princess Mononoke, Midsummer Night's Dream and Sixth Sense. And so, we chose movies of different genres to test the quality of our recommender implementation.

3.2 Observations

Of course, answering the question of "is it a good recommendation?" is not easy, or even not fully possible - as it can mean different things depending on person, some individualistic subjective tastes, as well as on the purpose of recommendation, goal of it etc. And we do not have any measures like click rates, satisfaction or sales rates. So in this part we mostly focus on looking at (analysing one by one) the recommendations that strikes us as unfitting, inappropriate. That is to say, we do not focus on "good" recommendations, but on "bad" ones. First and positive observation is that vast majority of recommendations did not looked random or taken out of nowhere. That was truth for all of the 5 methods implemented. While reading the proposed lists, we thought that it works very well so that we would actually want to use it personally. However, some problems have been detected.

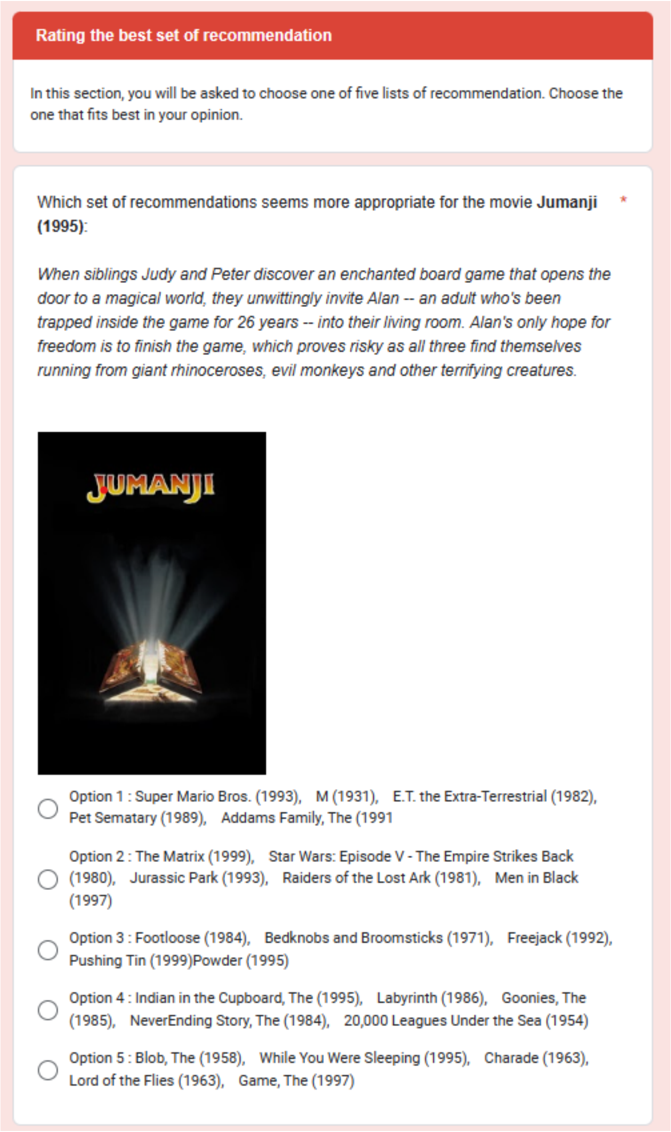
One thing that was prevalent is that the SVD method recommends Star Wars way too often. While it may be somehow fitting to movie like Robing Hood, it certainly was surprising to see Star Wars in the recommendation for Midsummer Night's Dream or the Exorcist. We asked ourselves why is that and looked into code. What the reason was, is that SVD highly depended on the ratings popularity, and as Star Wars is one of the most famous series, it would got high scores often, even if in terms of genre it was not a good fit. This can have some advantages, as recommending very popular movies usually works well, however it made the lists monotonic and unfitting, which we did not strive for.

Another problem we detected is that for Princess Mononoke we expected more Ghibli studio movies, japanese anime movies or at least some animation movies. We did not got that for the most part. The recommendations included movies like Madness of King George, Barbarella, Rebecca or American Beauty. None of the strategies gave us satisfactory results. We wondered why is that and looked manually at what movies are in our dataset and to our surprise, there were not really movies to recommend for Princess Mononoke. We worked on a smaller dataset, and it was the only Ghibli movie available there. That implies that not necessarily our methods were performing badly here, but were performing best to their abilities given that the input data was not enough. This is something we would change if we had to do this project again.

3.3 Survey

3.3.1 Survey Description

The drawback of the previous method is that only the four of us are evaluating the results. Therefore, for ten other movies not seen in 3.1, we asked the opinion of other people regarding the recommendations thanks to a survey. We decided to include only ten movies inside so as not to make it too long (otherwise, nobody would have taken the time to fill it in). For each movie, that we intentionally chose famous, we displayed its title, a picture to make the survey more attractive, an overview in case people wouldn't know the movie, and the results of our five recommendation strategies (the same as the ones we could see on the web application). The survey looks like this for each movie:



Rating the best set of recommendation

In this section, you will be asked to choose one of five lists of recommendation. Choose the one that fits best in your opinion.

Which set of recommendations seems more appropriate for the movie **Jumanji (1995)**: *

When siblings Judy and Peter discover an enchanted board game that opens the door to a magical world, they unwittingly invite Alan -- an adult who's been trapped inside the game for 26 years -- into their living room. Alan's only hope for freedom is to finish the game, which proves risky as all three find themselves running from giant rhinoceroses, evil monkeys and other terrifying creatures.

○ Option 1 : Super Mario Bros. (1993), M (1931), E.T. the Extra-Terrestrial (1982), Pet Sematary (1989), Addams Family, The (1991

○ Option 2 : The Matrix (1999), Star Wars: Episode V - The Empire Strikes Back (1980), Jurassic Park (1993), Raiders of the Lost Ark (1981), Men in Black (1997)

○ Option 3 : Footloose (1984), Bedknobs and Broomsticks (1971), Freejack (1992), Pushing Tin (1999)Powder (1995)

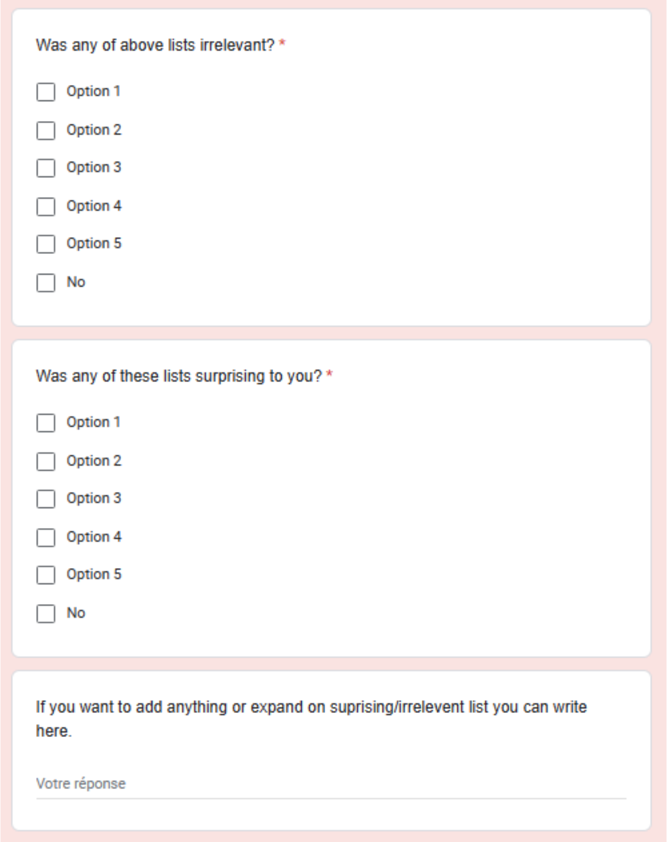
○ Option 4 : Indian in the Cupboard, The (1995), Labyrinth (1986), Goonies, The (1985), NeverEnding Story, The (1984), 20,000 Leagues Under the Sea (1954)

○ Option 5 : Blob, The (1958), While You Were Sleeping (1995), Charade (1963), Lord of the Flies (1963), Game, The (1997)

Figure 5: Extract of our survey : movie + recommendation lists

Each option corresponds to one recommendation strategy and contains the set of five most similar movies to the input one in descending order. We intentionally did not explain the strategy behind each option; we only showed “Option n”, so as to not bias participants.

For each movie, we asked the participants which set of movie recommendations was the most appropriate (namely the most similar) to the input one. Then we gave them the opportunity to select one or several set of movies which seemed irrelevant or surprising to them, and they could at the end give extra comments about these recommendations. We chose these questions, because besides our wish to know the best strategy (first question in the survey), we wanted to know if every strategy returned relevant results, or if some were inappropriate (second and third questions in the survey).



Was any of above lists irrelevant? *

□ Option 1

□ Option 2

□ Option 3

□ Option 4

□ Option 5

□ No

Was any of these lists surprising to you? *

□ Option 1

□ Option 2

□ Option 3

□ Option 4

□ Option 5

□ No

If you want to add anything or expand on suprising/irrelevent list you can write here.

Votre réponse

Figure 6: Extract of our survey : votes for surprising or irrelevant results

3.3.2 Results

We have got only fifteen answers for our survey. This number is disappointing, but still better than if only the four of us would have filled it in.

First of all, we calculated for each strategy the global proportion of votes as best one. We then noticed that the proportions were close to each other (the difference between maximal and minimal value is only 10%), so according to this measure, each technique has a quite similar quality perception. But still, BERT and SVD strategies have slightly better results, while Dice Coefficient and TF-IDF have slightly worse results (see Figure 7).

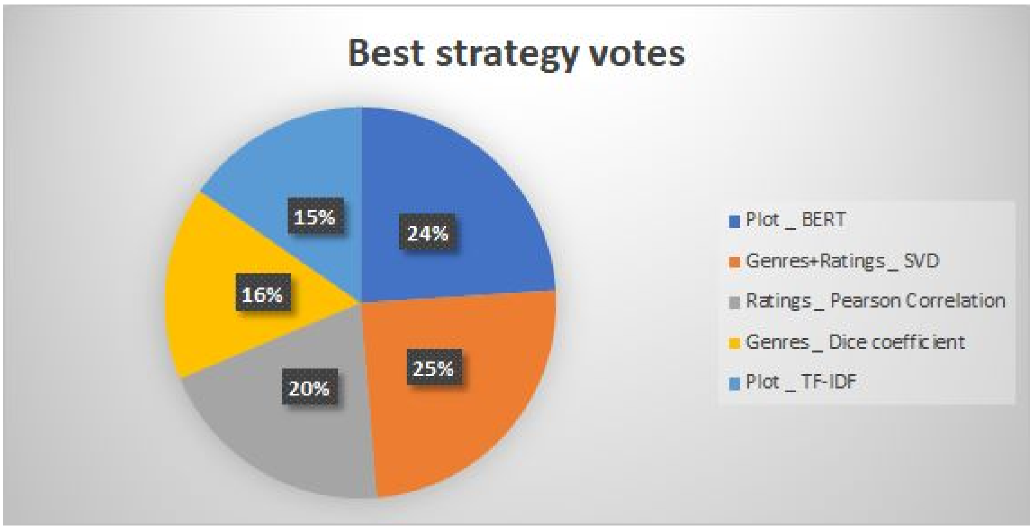


Figure 7: Extract of our survey : votes for surprising or irrelevant results

Furthermore, we computed the rate of times each recommendation has been considered as irrelevant. To do that, for a given option, we divided the total number of votes as irrelevant by the maximum number of times this vote could occur (namely 15 participants x 10 movies = 150).

As it can be seen in Figure 8, SVD and TF-IDF strategies are the worst with this measure, and BERT is the best one (for example, it has been considered as an irrelevant strategy more than three times less frequently than SVD). We did the same calculations and analysis for results considered as surprising (instead of irrelevant), and the results were quite the same; the only difference is that the proportions were a bit lower every time, since participants more often considered an option as irrelevant than surprising.

Examples of surprising results we noticed are that TF-IDF strategy considered Sleeping Beauty as similar to Star Wars, and that SVD strategy considered Star Wars as similar to Alice in Wonderland. More generally, SVD tended to recommend only famous movies, and especially Star Wars (as seen in 3.1); that can be explained by the fact that this strategy uses collaborative filtering methods.

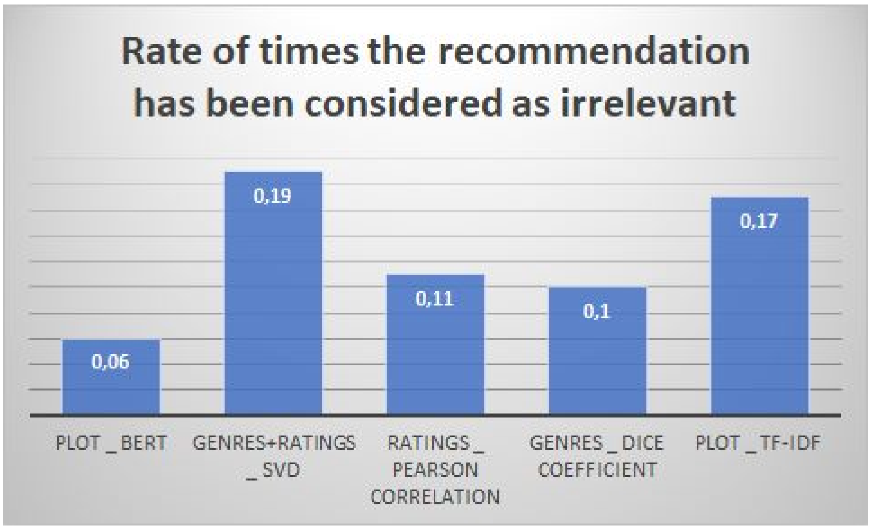


Figure 8: Extract of our survey : votes for surprising or irrelevant results

On the one hand, BERT strategy seems to be the best one, because it has one of the two biggest proportion of votes as most appropriate, and it is the strategy least often considered as irrelevant. But it doesn't mean that the strategy is perfect. It is satisfactory since it has been considered as irrelevant in 6% of cases and as surprising in 5% of cases, but these results could still be improved. On the other hand, TF-IDF strategy seems to be the worst one, because it has one of the poorest results in both criteria (best strategy votes and proportion of irrelevant or surprising votes). This time, the strategy is disappointing and unreliable, since it has been considered as surprising in 9% of cases and as irrelevant almost 1 time out of 5.

Besides, this difference between TF-IDF and BERT is surprising, because both strategies use the overviews to do their calculations; we could thus expect similar results. The only difference between these strategies is the technical method used to implement it; that seems to show that this factor has an influence which is far from being negligible.

What we can also notice is that except for BERT which is the best and most elaborate method at the same time, the technically simplest methods (Pearson Correlation and Dice Coefficient) have more satisfactory results than the most sophisticated ones (SVD and TF-IDF). We can mainly see it in Figure 8: Pearson Correlation and Dice Coefficient have been considered less often as irrelevant than SVD and TF-IDF. Therefore, the best strategies are not necessarily the most sophisticated ones.

However, as we have only 15 participants, we should take a critical look at our results. Indeed, the latter are not very reliable because of this low number of answers, which moreover only come from people with about the same age.

3.3.3 Limits of our survey

In addition, the survey itself also has limits which prevent us from having reliable results and bring bias.

First of all, the question asked to participants is not clear. We asked "which set of recommendations seems the most appropriate for the movie [movie-title]?”, but what does “appropriate” mean? We should have rather asked “which set of movies contains the most similar movies to [movie-title]?” (similar instead of appropriate). Better still, we could have asked “Which set of movies do you think is the most convenient to recommend to someone who liked [movie-title]?” because a set of similar movies is not necessarily a good set of recommendation movies. For example, given Star Wars IV, the set Star Wars I, Star Wars II, Star Wars III, Star Wars V, Star Wars VI is a set of similar movies but is not a convenient recommendation (since the same movie is always recommended).

Moreover, we didn’t randomize the strategies’ order. Option1 always corresponds to the same strategy, as well as Option 2, 3, 4 and 5. That is a source of bias, since if for example a user notices in the first questions that Option 1 has good results, he will tend to choose this option every time (and opposite tendency if he notices in the first questions that the results are irrelevant). We should have shuffled the strategy order and told it to participants to remove this bias.

Finally, most of the movies of the database (and so the movies in the recommendation set) are unknown to the participants. So, the latter can’t evaluate the relevance of the recommendation. We could tell participants to take into account only the movies they know, but we still have a problem if they don’t know any of the five movies of an option’s set. That’s why we should have given access to a short description of each movie. But even with this, an issue remains, since the description we add can bias participants. For example, if we show an overview, they will unconsciously favour BERT and TF-IDF strategies, based on overview.

4 Conclusion

To conclude, the evaluation showed us that the vast majority of our movie recommendations were relevant, even if some inconsistencies remained. Other potential customers' quality perception is about the same for each strategy, but BERT, for its low proportion of movie recommendation considered as irrelevant or surprising, has been taken on as the best strategy. According to our survey, even if its results can still be improved, BERT method is a reliable and good recommendation strategy (though "good" is subjective and depends on the customer and objectives).

However, some limits remain in our evaluation, because of our lack of possible ways to evaluate (no click rate, sale rate...), and because of sources of bias from our survey, as well as its low number of participants.

If we had to do this project again, we would have paid attention to work on a bigger dataset, to correct the drawbacks of our survey, and to improve the performance or our web application, as well as its user interface.