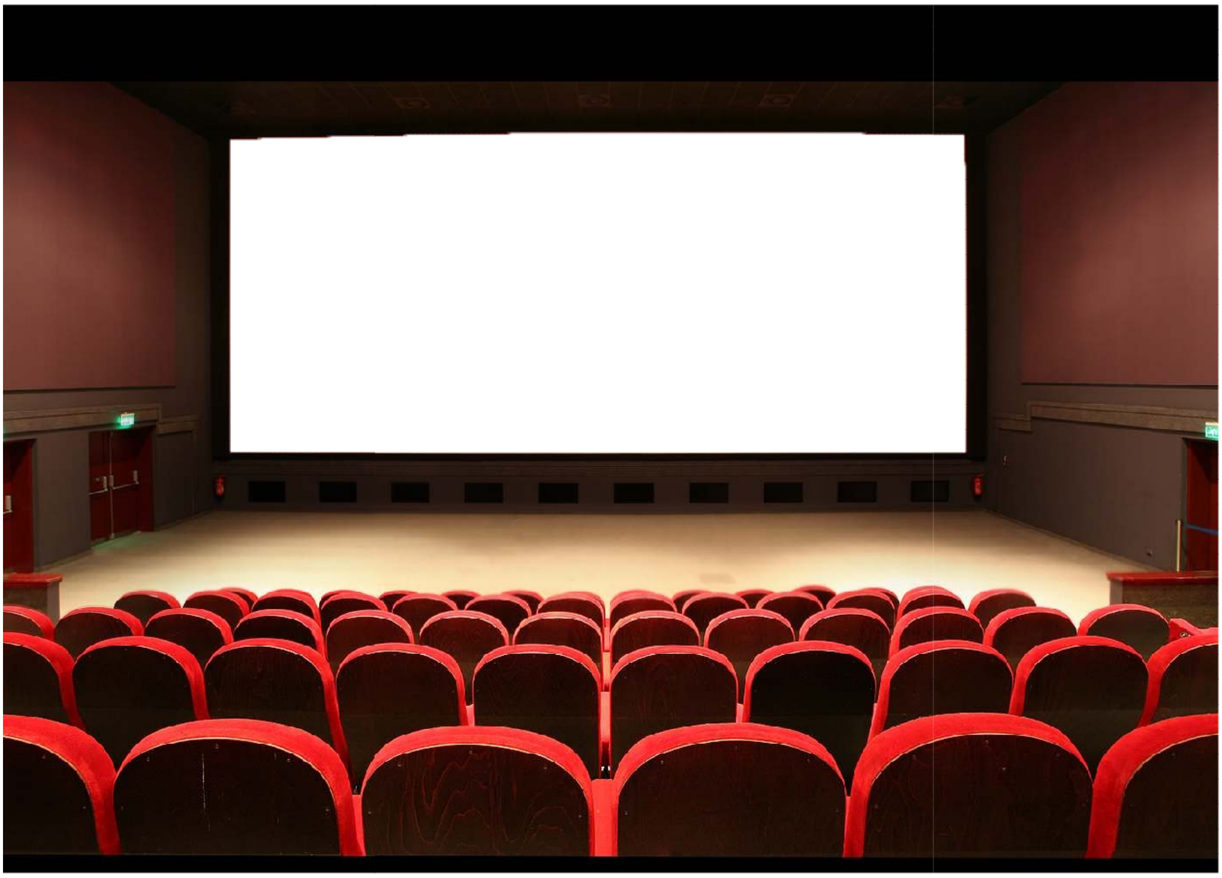


My movie recommendation system

Introduction

Founded in 1998, "Netflix is a streaming service that allows its members to watch a wide variety of TV series, movies, documentaries, etc. on thousands of Internet-connected devices." => **Wikipédia Netflix**

Today, Netflix is worth more than 20 billion dollars in revenue and consumes 12.6% of the world's Internet bandwidth.

When accessing the Netflix service, the recommendation system helps the user to find as easily as possible the TV series or movies he might enjoy. Netflix calculates the probability that the user will watch a given title from the Netflix catalog, and can thus optimize these partnerships or more globally its marketing strategy. Netflix is the archetype of a data-driven company.

Your client is not Netflix, but he has big ambitions!

Goals

You are a freelance data analyst. A movie theater, located in the middle of nowhere, contacts you. It has decided to go digital by creating a website designed for local people.

To go even further, he asks you to create **a movie recommendation engine** that will eventually send notifications to customers via the Internet.

For the moment, no customer has entered his preferences, you are in a **cold start situation**. But fortunately, the customer gives you a database of movies based on the IMDb platform.

You will start by proposing a complete analysis of the database (which countries produce the most films? Which actors are the most present? At what period? Does the average length of the films get longer or shorter over the years? Are the actors in TV series the same as in movies? What is the average age of the actors? Which films have the highest ratings? Do they share common characteristics? Etc...) After a first analysis, you may decide to specialize your cinema, for example in the "90's period", or on "action and adventure films", in order to refine your analysis.

After this analytical step, at the end of the project, you will use machine learning algorithms to recommend movies based on movies that have been enjoyed by the viewer.

Data sources

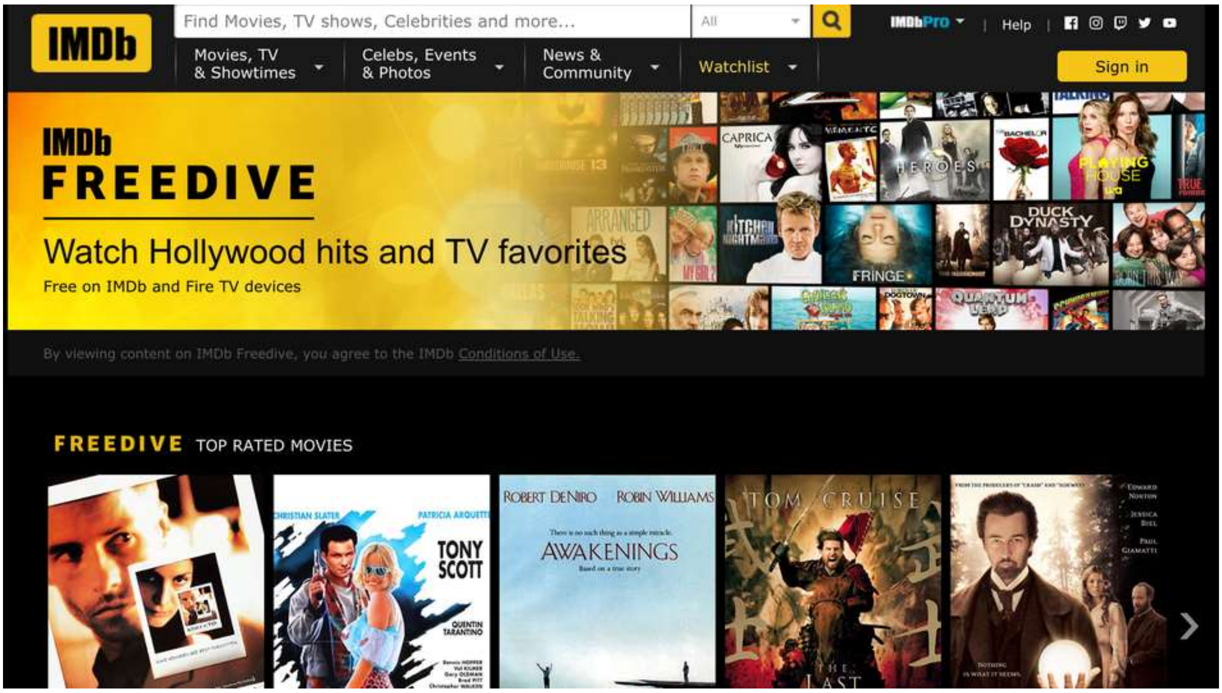
The data are available on the IMDB website. The data is divided into several tables (movies, actors, directors, notes, etc...)

- The documentation explaining briefly the definition of each column and each table is available here.
- The datasets are available here.

Technical notes:

- You can download the datasets locally, on your drive. But you can also not download them, and directly import the datasets in pandas by putting the link of the dataset. You could also fill your database and use both SQL and Python to work on this project.
- The datasets are very large, there are more than 7M movies and 10M actors referenced. You probably won't need the whole database. Once you have cleaned up and filtered what you find relevant, consider exporting your "light" data. It will be faster to re-import.
- As a reminder, Google Colab offers "shared" servers. The performance depends on the number of people connected at the same time. Sometimes, you won't be able to load all these large datasets. Don't hesitate to process them locally with Anaconda / Jupyter.
- IMDB datasets are in TSV format, for "Tabulation Separated Values". This is similar to the CSV format but separated by tabs instead of commas. You can use the following function, which indicates that the separator is a tab:

```
pd.read_csv(“dataset_link”, sep = “\t”)
```


IMDb
Find Movies, TV shows, Celebrities and more...
All
IMDbPro
Help
Movies, TV & Showtimes
Celebs, Events & Photos
News & Community
Watchlist
Sign in
IMDb FREEDIVE
Watch Hollywood hits and TV favorites
Free on IMDb and Fire TV devices
FREEDIVE TOP RATED MOVIES

By viewing content on IMDb Freedive, you agree to the IMDb Conditions of Use.

Organization and planning

You will need to do joins (in Pandas or SQL) between datasets, graphs in Python, reprocessing with Pandas, machine learning. Of course, you won't be able to do everything the first week, because you will learn these notions in parallel to the project. In order to give you some visibility, here is an indicative schedule, but you are free to organize yourselves:

- Week 1: Appropriation and first exploration of the data
  - Main tools: Pandas, Matplotlib
- Week 2 and 3: Joins, filters, cleaning, correlation search
  - Main tools: Pandas, Seaborn
- Week 4: Machine learning, recommendations
  - Main tools: scikit-learn
- Week 5: Refinement, interface, presentation, and Demo Day
  - Main tools: Python Programming

Final rendering

The client would have liked to integrate your analysis and recommendations into his website to be able to test it, but the timing is too tight. You propose to make it testable with a **tool of your choice**.

The client has 2 needs, which can be in 2 separate tools:

- Get **some statistics on movies** (type, duration), actors (number of movies, type of movies), and others. You will do this in particular with the help of visualizations. You can use a business intelligence tool or charts via Python.
- Return **a list of recommended movies** based on IDs or movie names chosen by a user. You can integrate these recommendations into a dashboarding tool, or make these recommendations directly from the command line ("input", streamlit, or other).

The goal is that the system works and you identify areas for improvement. So, it’s a **POC: a Proof of Concept**.

Demo day

You will explain your approach and the difficulties encountered during a **10 to a 15-minute oral presentation (to be specified with the client)** and will give a demonstration in front of client representatives. Your client will give you some names of films during the presentation, and you will make film recommendations based on these films. Other audience members will also be able to suggest a film.

Missions

- Make a presentation to present your work, explain your approach, the tools used, the difficulties encountered, and avenues for improvement.
- Present the relevant statistical indicators and KPIs on films.
- Create a film recommendation system using machine learning algorithms and demonstrate these recommendations on films proposed by the client.

Expected deliverables

- A notebook containing the data exploration and cleaning and visualizations. You will explain your cleaning choices and exploration findings in a document of your choice.
- Present the exploratory analysis and the relevant KPIs
- A notebook for the Recommendation System step with the source code and your comments

Ressources

- https://blogs.gartner.com/martin-kihn/how-to-build-a-recommender-system-in-python/