Big Data Project
"Movie Recommendation System"

Apostolos Mikas

mikas.a@live.unic.ac.cy

Background and Motivation

During this project, we will create a movie recommendation system based on big data sourced from the popular streaming platform Netflix.

Netflix has revolutionized the way we watch movies. From its beginning, it has been motivated to give its users the best experience possible when watching movies. Today, it offers a massive selection of movies and recommendations, to help its viewers find the best fit for them and provide an optimum viewing experience.

Some interesting Netflix facts and statistics [1]:

- In Q2 2022, it counted 220 million subscribers worldwide.
- Accounts for 17% of all worldwide online video subscriptions.
- As of October 2022, it offers more than 17000 titles globally.
- In 2019, Netflix users spent 164 million hours per day watching content.

Project Objectives (Hypothesis)

The first question we will try to answer is how to personalize a streaming service as much as possible for each user who uses it.

Data Sources

We will use data sourced from Netflix. On the occasion of the Netflix prize, Netflix published a dataset with over 100 million ratings given to 17 thousand of movies by 480 thousand users. The total size of the Netflix dataset is a little above 2GB.[3]

Big Data Dimensions

1. Volume: With more than 220 million users and more than 17 thousand titles, Netflix is a "data-generating machine". In 2019, Netflix announced that it counted over 5 billion user reviews. With a rough calculation, this means more than 10,000GB of rating data. [2]
2. Velocity: Every time a user watches content in Netflix, it collects usage statistics such as viewing history, ratings, preferences on actors, genres, directors, years of release, and more. Additionally, it collects data related to the devices used for watching and the duration of watch. In 2020, Netflix users watched an average of 3.2 hours daily. [1]
3. Veracity: Netflix has thousands of content titles, some more popular and some less popular, with disproportionate positive or negative ratings. Also, some users like to upvote or downvote their favorite titles, while others don't. I expect the Netflix data to include many abnormalities, biases, and noise.

Tools/libraries

The tools that we used:

- Python / Jupyter Notebooks
- NoSQL database, i.e., MongoDB
- Apache Spark
- Laptop
  - OS: Windows 10
  - CPU: Intel(R) Core(TM) i7-7700HQ CPU @ 2.80GHz 2.81 GHz
  - RAM: 16,0 GB

References

1. Cook, S. (2022, October 6). *50+ Netflix statistics & facts that define the company's dominance in 2022*. Comparitech. https://www.comparitech.com/blog/vpn-privacy/netflix-statistics-facts-figures/
2. Kasula, C. P. (2021, December 15). *Netflix Recommender System — A Big Data Case Study - Towards Data Science*. Medium. https://towardsdatascience.com/netflix-recommender-system-a-big-data-case-study-19cfa6d56ff5
3. Wikipedia contributors. (2022, December 8). *Netflix Prize*. Wikipedia. https://en.wikipedia.org/wiki/Netflix_Prize