Sistemas de Recomendação para Filmes “Zephyrus”

SCC0530 - Inteligência Artificial
Grupo 10

Objetivo

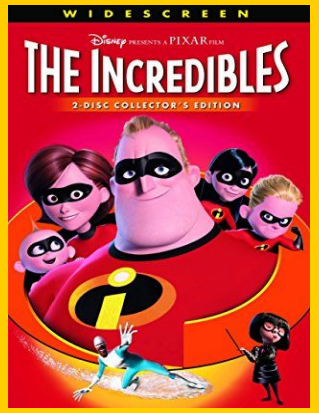
Recomendar **filmes relevantes** a usuários de um serviço de streaming de filmes

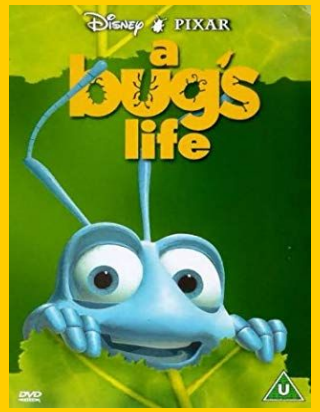
Olá, *Solange*!

Você avaliou:

★★★★★

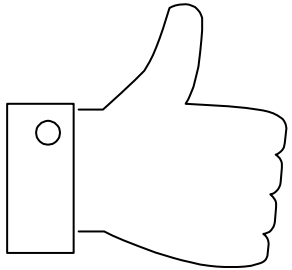
Recomendados para **você**!

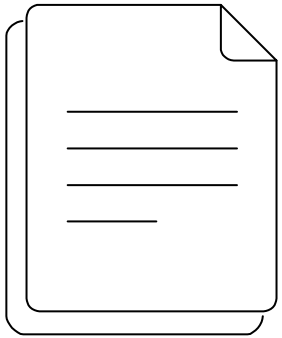


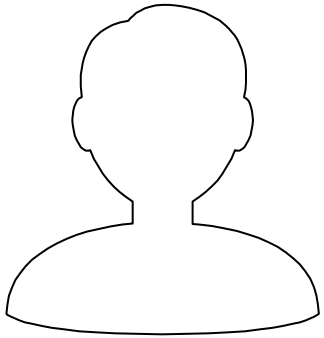


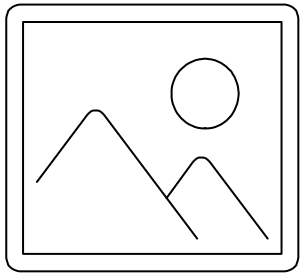

Dataset &
Mineração

movielens









Notas

train_data.csv

Críticas

movie_reviews.csv

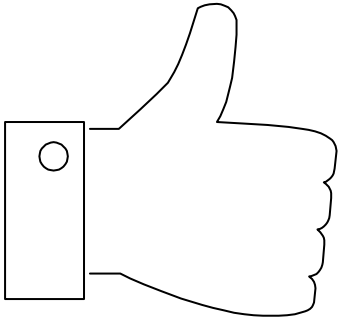
Usuários

users_data.csv

Filmes

movies_data.csv

movielens



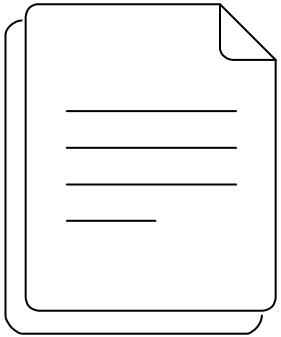
Notas

3974 usuários
3564 filmes
535785 avaliações

Dados Estruturados
Notas de **1-5**

user_id	movie_id	rating
1	1160	5
2	980	3
3	2079	3
4	842	5

movielens



Críticas

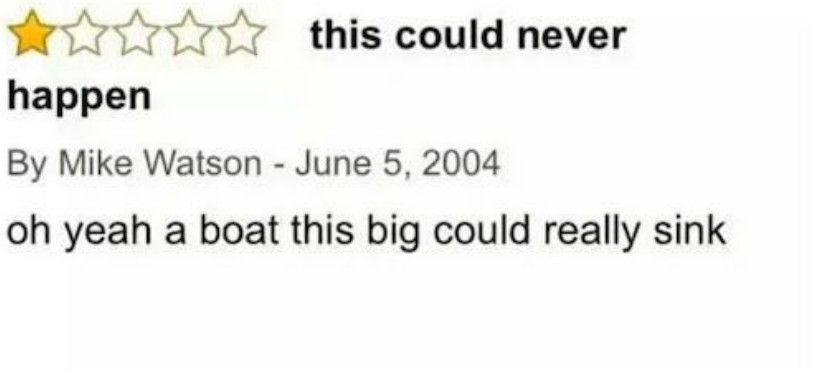
24882 avaliações
Não revela o usuário

Dados Não-Estruturados
Avaliações **textuais**

movie_id	rating
1	Andy's **toys** live a reasonable life of **fun** and peace, their only worries are birthdays and Christmases (...)
1	I am a big fan of the **animated movies** coming from the **Pixar** Studios. (...)
1	Children play with **toys**. It is a known fact. (...)

movielens

Titanic (1997)



this could never happen

By Mike Watson - June 5, 2004

oh yeah a boat this big could really sink

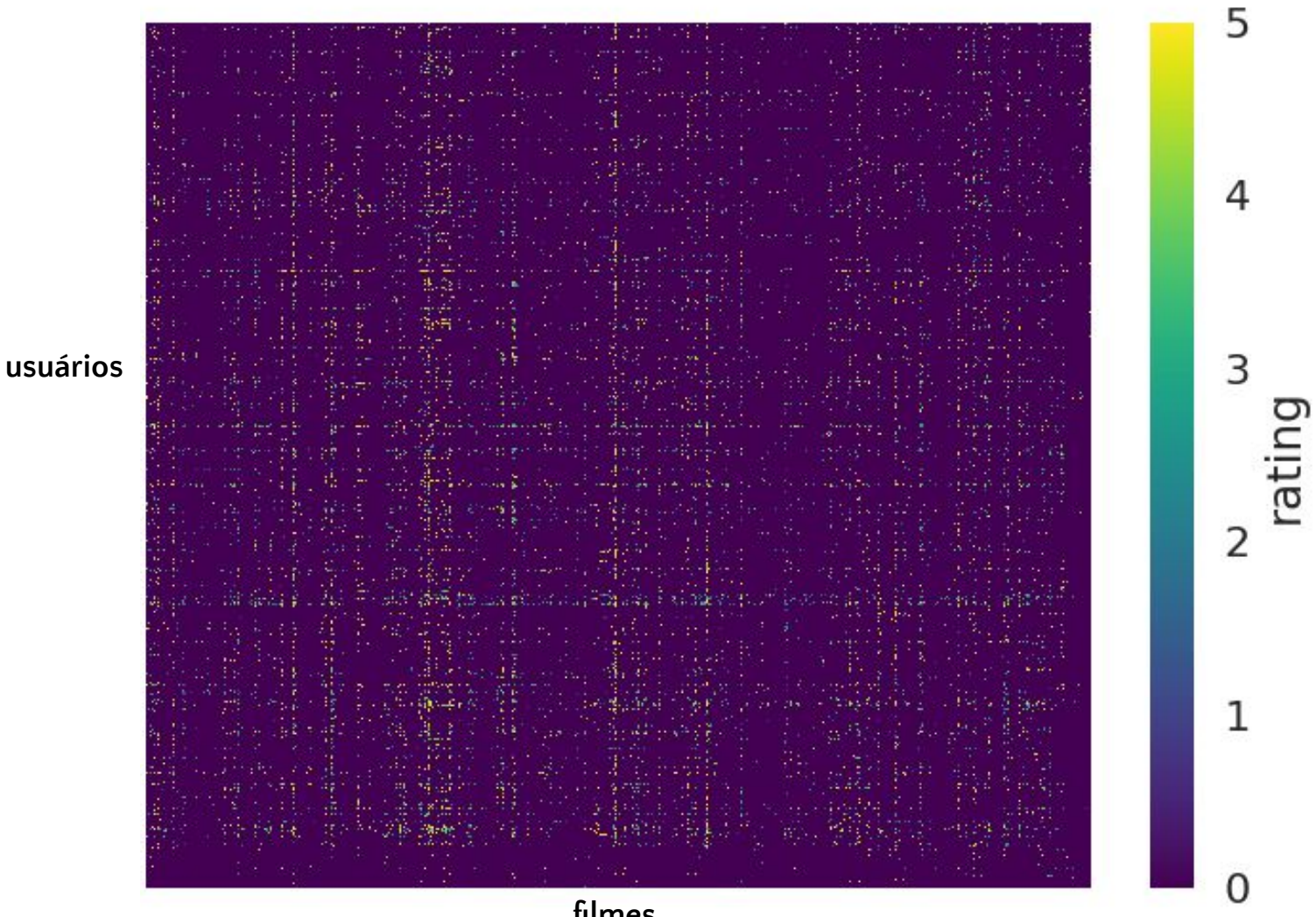
movie_id	rating
1	Andy's **toys** live a reasonable life of **fun** and peace, their only worries are birthdays and Christmases (...)
1	I am a big fan of the **animated movies** coming from the **Pixar** Studios. (...)
1	Children play with **toys**. It is a known fact. (...)

Pré-Processamento

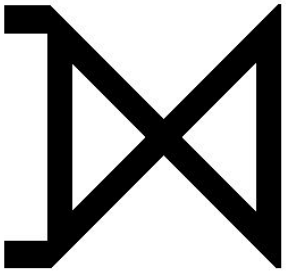
- Gerar **matriz** de notas
  - Usuários x Itens
  - Permite encontrar facilmente uma relação de um usuário para certo filme e todos os filmes que ele avaliou.

Content-based recommender systems

Movie	Alice (1)	Bob (2)	Carol (3)	Dave (4)
Love at last	5	5	0	0
Romance forever	5	?	?	0
Cute puppies of love	?	4	0	?
Nonstop car chases	0	0	5	4
Swords vs. karate	0	0	5	?


usuários
filmes
rating
5
4
3
2
1
0

Pré-Processamento



- **Unir** conjuntos de dados para conseguir exibir informações completas
  - Conjunto de notas possui apenas *movie_id*. Para obter o **nome e gênero desse filme**, unimos com *movies_data*.
  - Para fazer o mesmo com **usuários**, unimos com *users_data*.

Pré-Processamento

train_data

user_id = 10
movie_id = 1
rating = 5

movies_data

movie_id = 1
title = Toy Story (1995)
genre = Animation, Children's, Comedy

users_data

user_id = 10
gender = M
age = 25

Usuário 10

Pré-Processamento

Preparar textos

- Tokenização
- Remoção de stop words
- Normalização de termos
- Lematização
- Radicalização
- Desambiguação

Cada um que passa em nossa vida, passa sozinho, mas não vai só, nem nos deixa sós; leva um pouco de nós mesmos, deixa um pouco de si mesmo

(Antoine De Saint Exupery)

cada, pass, noss, vida, pass, sozinh, vai, só, deix, só, lev, pouc, nós, mesm, deix, pouc, si, mesm

passar→bn:00091458v
vida→bn:00002761n
sozinho→bn:00110682a
ir→bn:00088912v
deixar→bn:00087966v
levar→bn:00084554v
pouco→bn:00051575n
só→bn:00110682a

Pré-Processamento

Sampling

- Divisão simples do conjunto de dados.
- Conseguimos comparar as notas da predição com as notas esperadas.

test_data.csv

train_data.csv

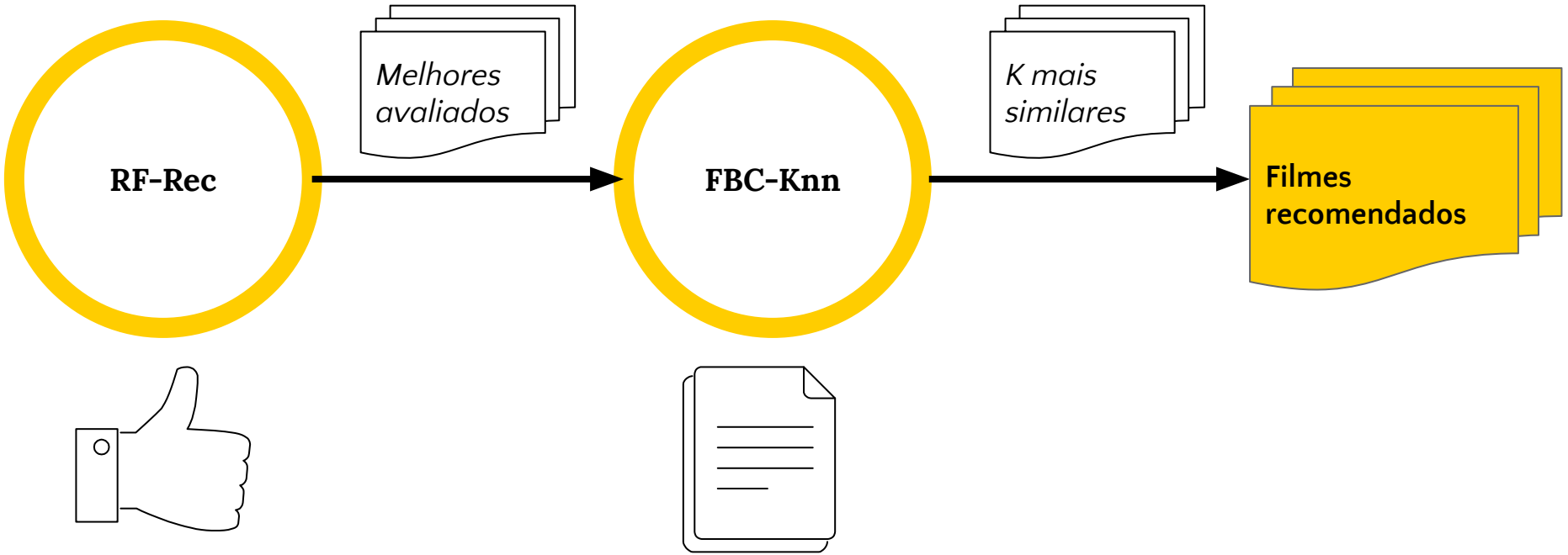
Sistema de Recomendação

Implementação em Python

- Rápida prototipagem
- Bibliotecas poderosas para pré-processamento

Recomendador híbrido canalizado utilizando paradigmas de filtragem colaborativa e baseada em conteúdo

Sistema de Recomendação em Cascata



Algoritmo RF-Rec

- Predição baseada nas frequências de avaliações de usuários e itens
- Onde:

$$\hat{r}_{ui} = \arg\max_{r \in R} f_{user}(u,i) \times f_{item}(i,r)$$

- R: conjunto de todas as avaliações. I.E: R = {1, 2, 3, 4, 5}
- $f_{user}(u,r)$ e $f_{item}(i,r)$ representam a (frequência+1) de uma avaliação r ter sido usada pelo usuário u ou atribuída ao item i

Algoritmo RF-Rec

- Consideramos a frequência de avaliações do usuário por item.
- Procedimento rápido, portanto pouco custoso – ideal para uma avaliação inicial.

	Item1	Item2	Item3	Item4	Item5
Alice	1	1	?	5	4
User1	2		5	5	5
User2			1	1	
User3		5	2		2
User4	3		1	1	
User5	1	2	3		4

Rating 1: (2 + 1) x (2 + 1) = 9
Rating 2: (0 + 1) x (1 + 1) = 2
Rating 3: (0 + 1) x (1 + 1) = 2
Rating 4: (1 + 1) x (0 + 1) = 2
Rating 5: (1 + 1) x (1 + 1) = 4

pred(Alice, Item3) = 1

Algoritmo FBC-Knn

Filtragem baseada em Conteúdo

$$pred(u,i) = \frac{\sum_{j \in I_{ui}} sim(i,j) * r_{uj}}{\sum_{j \in I_{ui}} sim(i,j)}$$

Procura os **vizinhos mais próximos** de itens avaliados positivamente pelo usuário

Vizinhança é dada pela **similaridade** entre itens, usando as **críticas textuais**

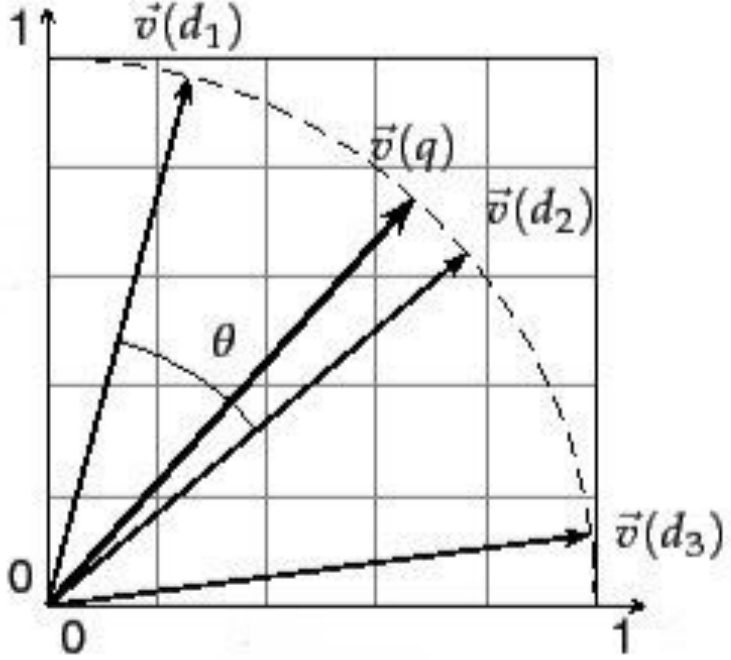
Algoritmo FBC-Knn

Cálculo de Similaridades

TF-IDF + Cosseno

term frequency–inverse document frequency

Ponderação de **frequência** dos termos

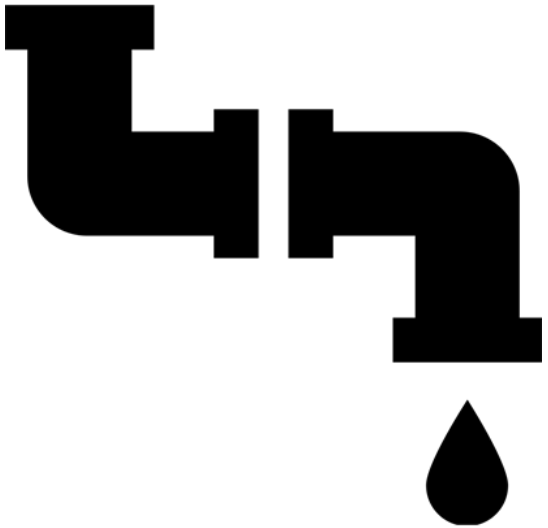


Arquitetura do Recomendador

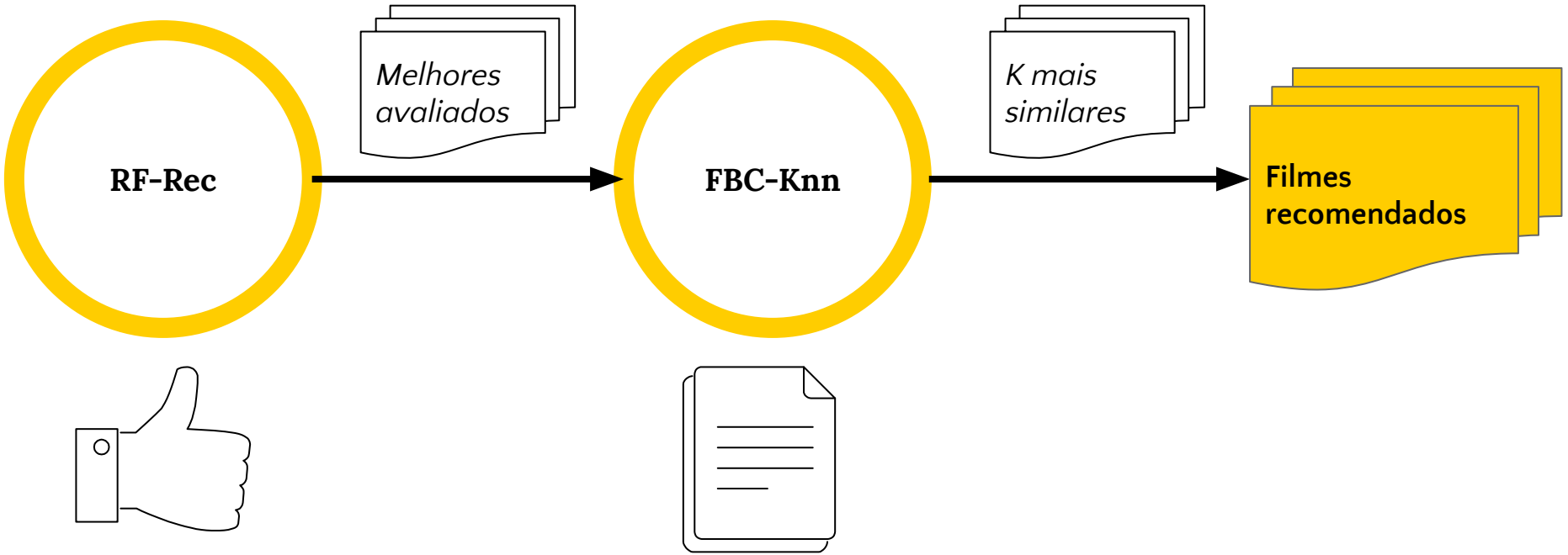
Recomendador híbrido: unir pontos positivos de paradigmas diferentes

Método Cascata: reduzir base de dados antes do algoritmo caro

RF-Rec (*Barato*) -> **FBC-Knn** (*Caro*)



Sistema de Recomendação em Cascata

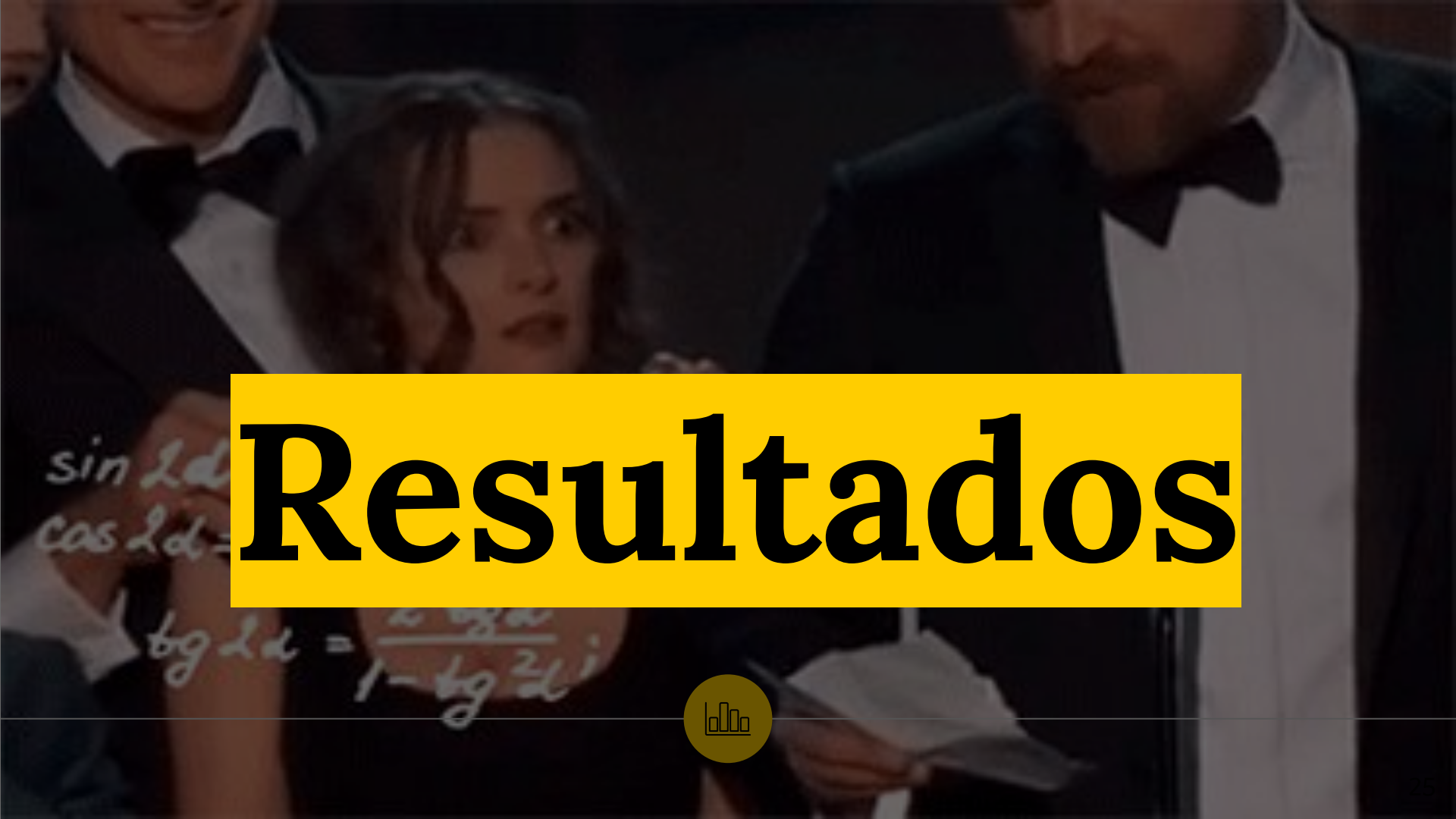


Z


IRON MAN
THE INCREDIBLE HULK
IRON MAN 2
THOR
CAPTAIN AMERICA
THE AVENGERS
IRON MAN 3
THOR: THE DARK WORLD
CAPTAIN AMERICA: THE WINTER SOLDIER
GUARDIANS OF THE GALAXY
AVENGERS: AGE OF ULTRON
ANT-MAN
CAPTAIN AMERICA: CIVIL WAR
DOCTOR STRANGE
GUARDIANS OF THE GALAXY VOL. 2
SPIDER-MAN: HOMECOMING
THOR: RAGNAROK
BLACK PANTHER


FROM THE STUDIO THAT BROUGHT
CAPTAIN AMERICA
THE WINTER SOLDIER
MARCH 26
IN REAL 3D AND IMAX 3D
MARVEL
AVENGERS
MAY 4


MARVEL STUDIOS
AVENGERS
INFINITY WAR
APRIL 27
IN REAL 3D AND IMAX



Resultados

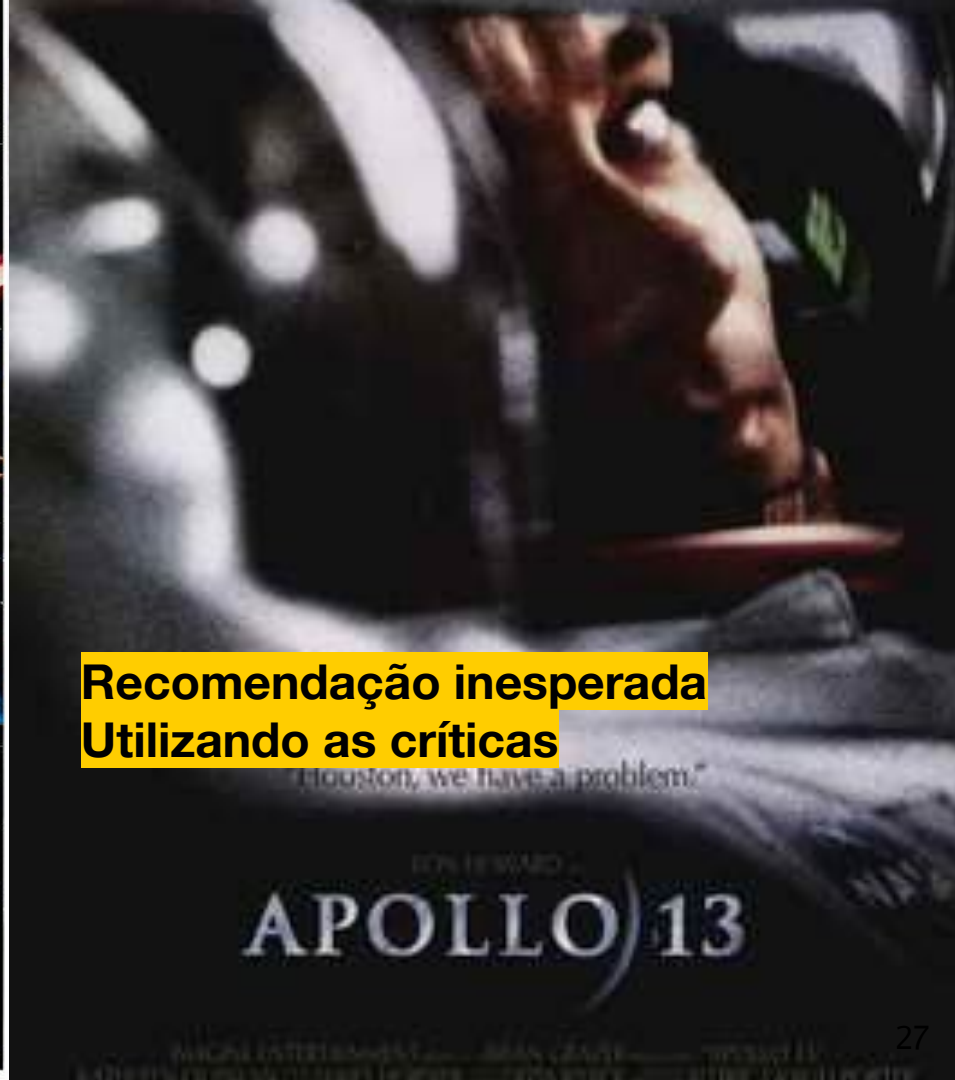
RF-Rec

RF_Rec results:
* Usual Suspects, The (1995)
* Lamerica (1994)
* Braveheart (1995)
* Jupiter's Wife (1994)
* Star Wars: Episode IV - A New Hope (1977)
* Pulp Fiction (1994)
* Shawshank Redemption, The (1994)
* Schindler's List (1993)
* Paris, France (1993)
* Silence of the Lambs, The (1991)

FBC-1NN

* Watcher, The (2000)
* Senseless (1998)
* Apollo 13 (1995)
* Beat the Devil (1954)
* Two Family House (2000)
* Winter Guest, The (1997)
* Deuce Bigalow: Male Gigolo (1999)
* Lonely Are the Brave (1962)
* Two Family House (2000)
* Replacements, The (2000)

Recomendação esperada
Utilizando as notas



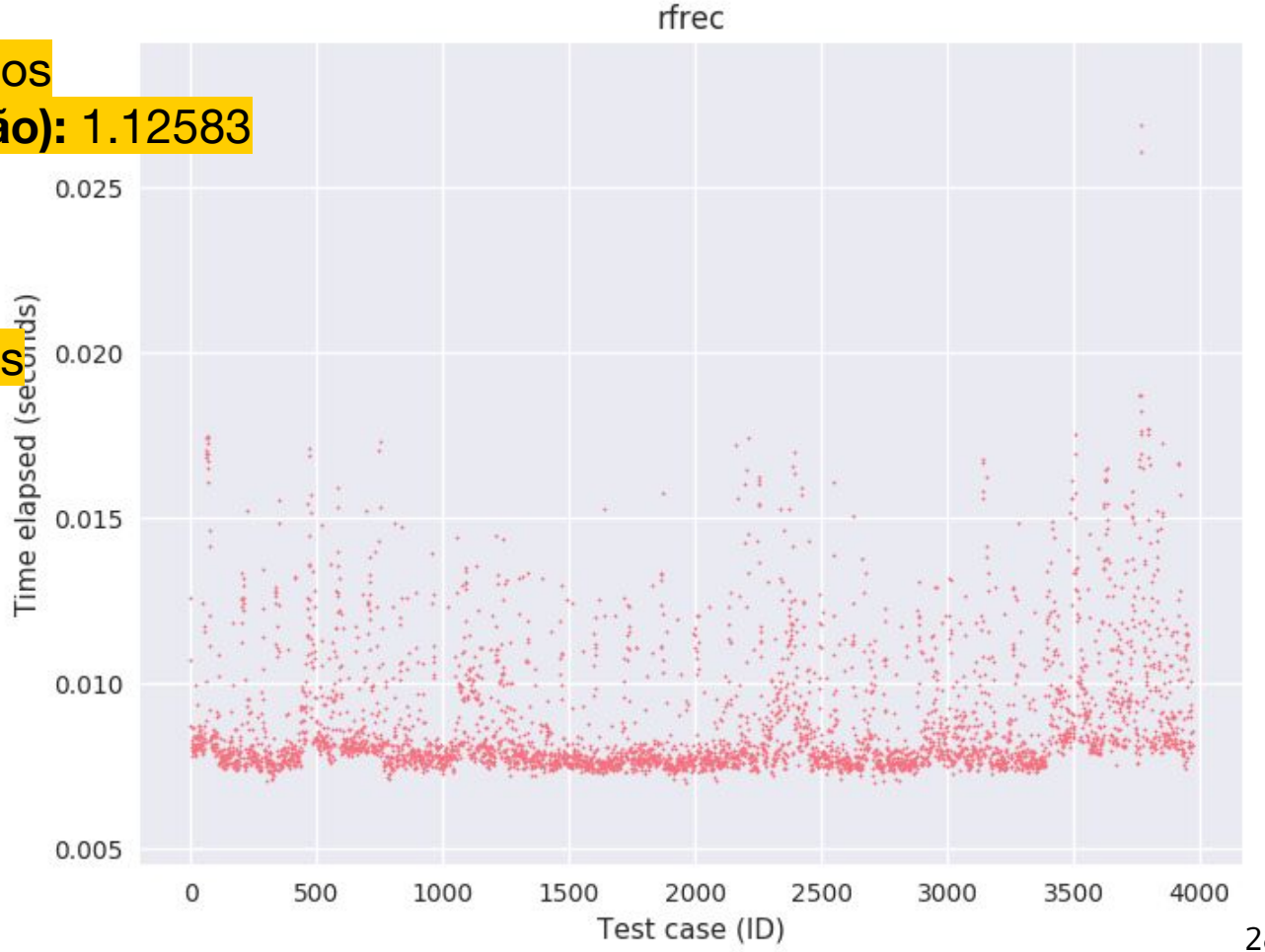
Recomendação inesperada
Utilizando as críticas

RF-Rec

Execução: ~ 30 segundos
RMSE (Notas x Predição): 1.12583

FBC-1NN

Execução: ~ 2 minutos



Obrigado!

***Perguntas*?**

Caroline Jesuíno Nunes da Silva – 9293925
Danilo Zecchin Nery – 8602430
Felipe Scrochio Custódio – 9442688
Henrique Fernandes M. Freitas – 8937225
Henrique Martins Loschiavo – 8936972
Isadora Maria Mendes – 8479318