Checkpoint #1 - K-Means em IMDB Top 250

Nome : Thiago Almança da Silva - RM 558108

Link do Repositório GitHub

https://github.com/ThiagoSilva15/CP-Front-End

Entrega Esperada

```
/
├── notebooks/
│  ├── 01scrapeandkmeanssynopsis.ipynb
│  └── 02kmeansallfeatures.ipynb
├── data/
│  ├── imdbtop250raw.csv
│  ├── imdbtop250k5synopsis.csv
│  └── imdbtop250k5_allfeatures.csv
├── README.md
```

Comparação dos Modelos

Modelo 1 — só sinopse (TF-IDF)
Pontos fortes: clusters temáticos muito interpretáveis (os “top termos” explicam bem cada grupo).
Limitações: filmes com sinopses parecidas, mas anos/ratings muito diferentes, acabam juntos; sensível ao texto curto/ruído.

Modelo 2 — todas as features (sinopse + gêneros + notas/votos/ano/runtime)
Pontos fortes: clusters mais coesos e balanceados; aproxima filmes por semântica + sinal objetivo (popularidade, época, duração).

Métricas: tipicamente Calinski-Harabasz ↑ e Davies-Bouldin ↓ em relação ao Modelo 1; a silhouette pode ficar similar ou até um pouco menor (efeito comum quando combinamos muitas variáveis), mas a separação prática melhora.

Leitura dos clusters 3D: nuvens menos sobrepostas e “faixas” por gênero/ano mais nítidas.

Qual modelo é melhor?
Escolha: Modelo 2 (todas as features).

Justificativa:
Combina conteúdo (sinopse) com contexto (gênero, ano, duração, rating e votos), formando grupos mais consistentes e úteis para análise.
Nas visualizações 3D, os agrupamentos ficam menos difusos, e as estatísticas por cluster (médias de ano/rating/runtime e distribuição de gêneros) são mais coerentes.
As métricas de validação interna tendem a favorecer o Modelo 2 (em especial CH ↑ e DB ↓), indicando clusters mais compactos e bem separados.

Análise dos Resultados

Foram gerados dois modelos de clusterização utilizando o algoritmo KMeans (k=5):

• Modelo 1 — baseado apenas na sinopse (TF-IDF + SVD):
- Silhouette: 0.1413
- Calinski-Harabasz: 5.64
- Davies-Bouldin: 1.0516

• Modelo 2 — utilizando todas as features (sinopse, gêneros, notas, votos, ano e duração):
- Silhouette: 0.2013
- Calinski-Harabasz: 16.55
- Davies-Bouldin: 1.2497

Conclusão:
O Modelo 2 apresentou melhor separação entre os clusters em métricas de qualidade (maior Calinski-Harabasz e menor Davies-Bouldin), mostrando que incluir múltiplas variáveis além do texto melhora a coerência e densidade dos grupos. O Modelo 1 obteve desempenho inferior e clusters mais difusos, por depender apenas do conteúdo textual das sinopses.