Samuel Regis Nascimento Barbosa

Daniella Rodrigues Vieira

Recomendação de filmes, através da plataforma de streaming Netflix.

São Paulo

2022

Sumário

Membros do grupo:

- Daniella Rodrigues Vieira
- Samuel Regis do Nascimento Barbosa

Sobre a empresa:

A Netflix foi fundada em 1997, iniciando seus negócios por aluguel de filmes por correio, até evoluir seu produto para uma plataforma de streaming com um catálogo de entretenimento para todas as idades.

Ela é uma empresa de entretenimento com a missão de entreter o mundo com séries, documentários, filmes e jogos para celulares e dispositivos móveis. Nessa plataforma os assinantes são capazes de controlar o que assistem no melhor horário e sem anúncios, a partir de uma única assinatura. O serviço de streaming da empresa está disponível em mais de 30 idiomas e 190 países.

Atualmente a rede de streaming é líder de mercado, com 31% do share no Brasil (Teleco,2022). Com cerca de 11 mil funcionários, a plataforma tem no seu DNA a tecnologia. Devido a seu tempo de mercado a marca tem um projeto maduro de Data Science para definições de UX, escolha de catálogo, e também o foco de nosso trabalho que é a recomendação de filmes e séries para os usuários, pois demonstrar para o usuário a relevância dos títulos conforme seu perfil diminui o churn da ferramenta (cancelamento da assinatura).

Objetivo:

O objetivo deste projeto é oferecer um serviço de recomendação de filmes baseando-se na atividade do usuário.

Contexto do estudo:

A Netflix é uma plataforma streaming, onde disponibiliza, aos seus usuários, filmes on demand. Além dos filmes para livre escolha, a plataforma também possui um sistema de recomendação, que é uma das aplicações interessantes em Ciência de Dados. Nosso estudo visa criar esse sistema de recomendação baseado em algumas características dos filmes baseado no dataset recolhido.O contexto deste trabalho é

identificar oportunidades de melhoria no algoritmo de recomendação, com finalidade de trazer melhor "match" entre os usuários com o catálogo que muda constantemente. Atualmente as informações relacionadas a filmes tem poucas informações numéricas e muitos textos para descrevê-las, então o desafio será classificar e entender como esses textos e informações podem transformar-se em um modelo estatístico para que as máquinas de aprendizado classifiquem os filmes assistidos e encontrem a correlação entre eles.

Cronograma:

Etapa	Descrição da atividade	Prazo
Etapa 1	Montagem do grupo Escolha da temática Escolha do Dataset Definição de ferramenta para análise	15/08 a 10/09
Etapa 2	Elaboração da proposta analitica Apresentação dos Scripts da Análise Exploratória em Python	11/09 a 08/10
Etapa 3	Construção grafíca dos resultados Elaboração do datastorytelling	10/10 a 01/11
Etapa 4	Ajuste do relatório final Vídeo de apresenrtação do projeto	02/11 a 28/11

Referências de aquisição do dataset:

Os dados estão disponíveis na plataforma aberta Kaggle, publicada em 11 de outubro de 2021, por Migara De Mel. Os dados foram gerados com o objetivo de entender o perfil de filmes disponíveis na plataforma Netflix e recomendar novos filmes aos usuários, com base em um título específico.
https://www.kaggle.com/code/migdev/netflix-movie-recommendation/data

Descrição do dataset:

O dataset recolhido no Kaggle é de uma fonte pública e está no formato csv. Devido ao modelo de recomendação ser um dos principais modelos da empresa e por causa da LGPD, a Netflix não fornece informações de usuários. Portanto, a base utilizada não possui dados sensíveis de nenhum tipo. A data de validade desses dados depende da troca de catálogo da operadora de streaming, e essa informação não é pré-definida. Porém, a troca de catálogo não interfere na lógica do modelo que estamos propondo inicialmente. Os arquivos em python e informações sobre os algoritmos estão disponíveis neste

endereço:https://drive.google.com/drive/folders/1IkkqIbDz1XBM8T04O6B_ExKSgjwht0UX?usp=sharing

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8807 entries, 0 to 8806
Data columns (total 12 columns):
 #   Column        Non-Null Count  Dtype 
---  ------        --------------  ----- 
 0   show_id       8807 non-null   object
 1   type          8807 non-null   object
 2   title         8807 non-null   object
 3   director      6173 non-null   object
 4   cast          7982 non-null   object
 5   country       7976 non-null   object
 6   date_added    8797 non-null   object
 7   release_year  8807 non-null   int64 
 8   rating        8803 non-null   object
 9   duration      8804 non-null   object
 10  listed_in     8807 non-null   object
 11  description   8807 non-null   object
dtypes: int64(1), object(11)
memory usage: 825.8+ KB
```

Figura 1 - Autoria Própria

Proposta análitica:

Análise exploratória de dados:

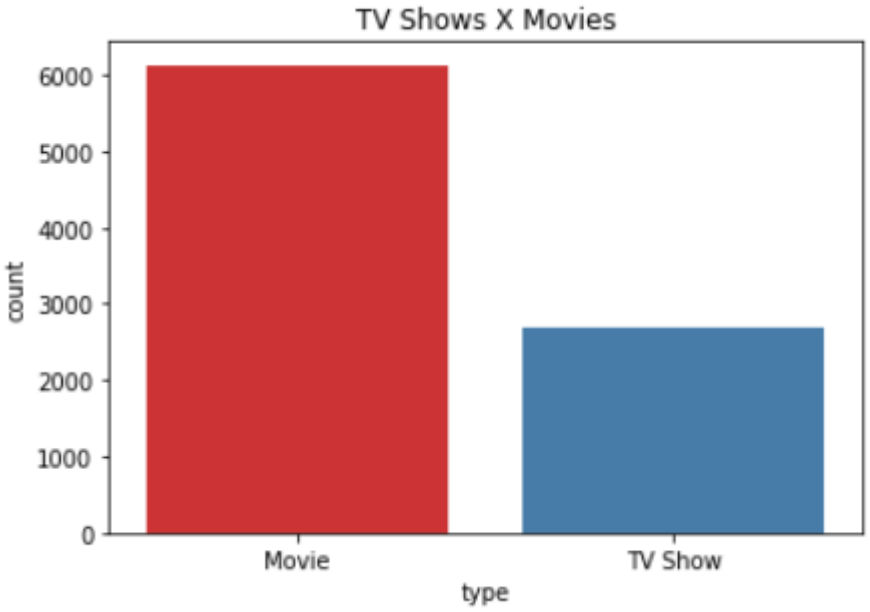
Temos uma base com 12 colunas, onde somente uma coluna é numérica. Nesse momento vamos começar as análises exploratórias dos dados e colunas para extrair mais informações e entender quais tipos de limpeza de dados devem ser feitas nas bases, quais são as informações relevantes com base na estatística e como se comportam as informações nulas.

Nossa proposta analítica baseia-se em correlacionar os dados estatísticos entre os filmes para recomendar um novo filme ao usuário. Para isso, nosso código em python extrai a seguintes informações:

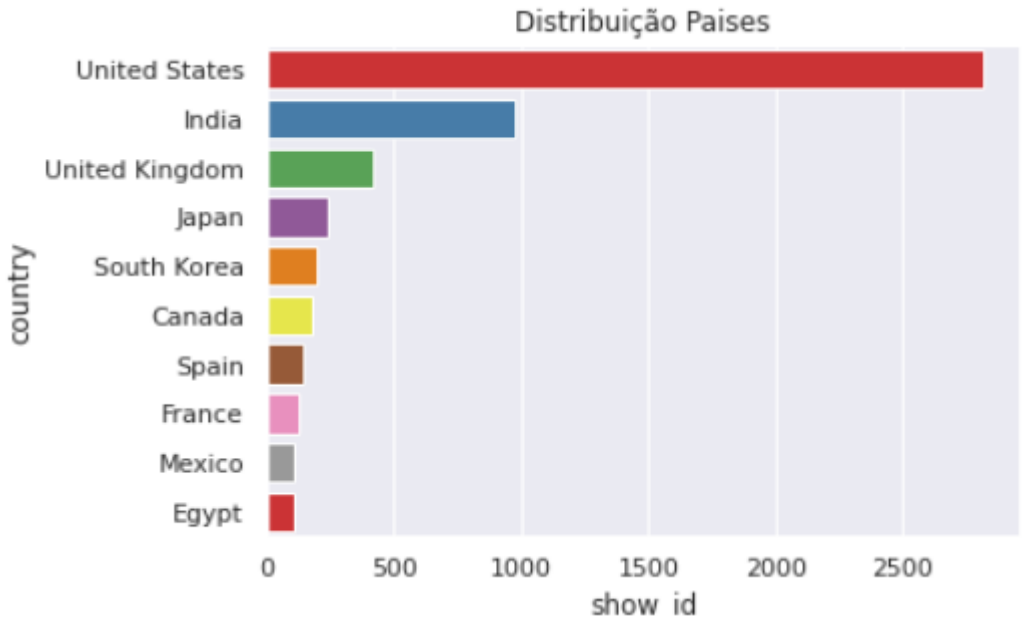
- Quantidade de linhas: 8807 com a seguinte composição de vazios nas colunas

	Colunas	Vazios
0	show_id	0
1	type	0
2	title	0
3	director	2634
4	cast	825
5	country	831
6	date_added	10
7	release_year	0
8	rating	4
9	duration	0
10	listed_in	0
11	description	0
12	duration_films	0

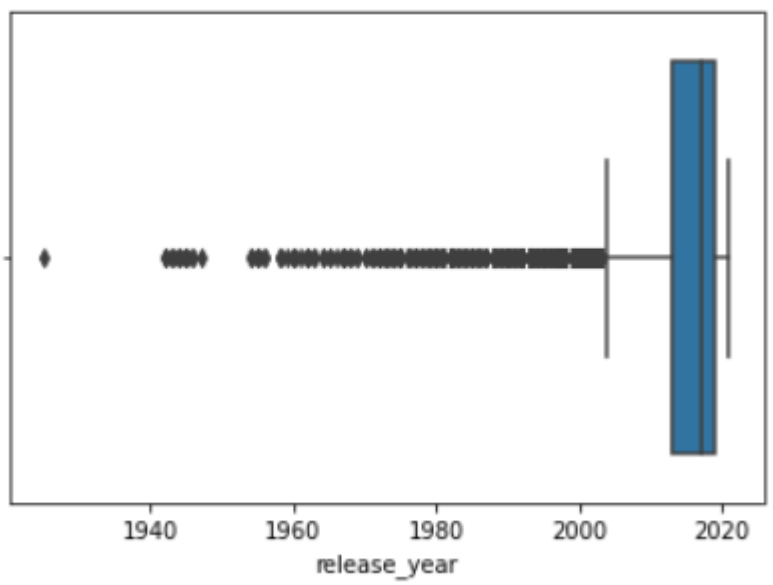
- Quantas linhas são filmes e quantas são séries de TV? 6131 filmes e 2676 séries de TV



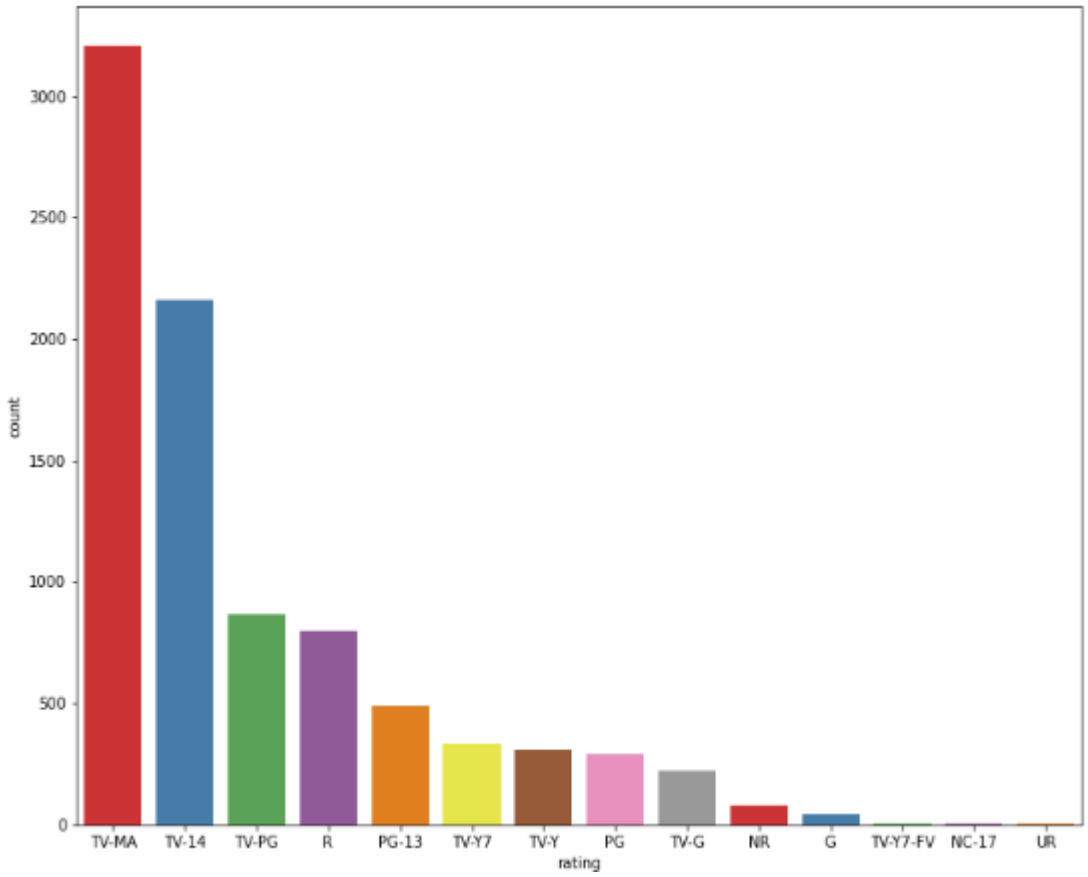
- Quais os países com maior índice de filmes? EUA, Índia, UK e Japão. (Lista completa no Colab).



- Quais os anos com maior índice de lançamento? 2018, 2017, 2019, 2020. (Lista completa no Colab). 50% da amostra são filmes lançados a partir de 2017, com desvio padrão de 9 anos.



- Qual a faixa etária dos filmes? Uma importante quantidade de filmes é para uma audiência mais madura(18+).O segundo maior números de filmes é para uma audiência 14+ (Gráfico disponível no Colab)



- Qual a duração média dos filmes? A mediana dos filmes está em 98 min, enquanto a média está em 99,5 min. Até metade dos filmes tem 98 min e 75% deles tem 114 min, o que significa que a maior parte dos filmes do catálogo tem menos de 2 horas de duração.

	release_year	duration_films
count	6131.000000	6131.000000
mean	2013.121514	99.528462
std	9.678169	28.369284
min	1942.000000	0.000000
25%	2012.000000	87.000000
50%	2016.000000	98.000000
75%	2018.000000	114.000000
max	2021.000000	312.000000

- As séries em sua maior parte tem 1 temporada, sendo o mínimo 1 temporada e o máximo 17. Isso mostra que há uma diferença importante entre a quantidade de temporadas já que o desvio padrão é 1,58, ou seja, 3 desvios padrões são cerca de 6,74 temporadas, mostrando que há outliers na amostra quando a questão são quantidades de temporadas.

	release_year	duration_films
count	2676.000000	2676.000000
mean	2016.605755	1.764948
std	5.740138	1.582752
min	1925.000000	1.000000
25%	2016.000000	1.000000
50%	2018.000000	1.000000
75%	2020.000000	2.000000
max	2021.000000	17.000000

- Para a limpeza dos dados, foi feita a obtenção do índice de valores vazios em cada coluna.
- Retiramos a coluna "director" pois aproximadamente 1⁄4 do catálogo está com ausência desta informação.
- Limpar as linhas onde possuímos ausência de informação classificando como " não informado" as colunas country, cast, rating e date_added. Obs. Decidimos manter as colunas devido a menos de 10% da amostra estar em branco.
- Apresentação do catálogo limpo.

	show_id	type	title	cast	country	release_year	rating	duration	listed_in
0	s1	Movie	Dick Johnson Is Dead	nao_informado	United States	2020	PG-13	90 min	Documentaries
1	s2	TV Show	Blood & Water	Ama Qamata	South Africa	2021	TV-MA	2 Seasons	International TV Shows
2	s3	TV Show	Ganglands	Sami Bouajila	nao_informado	2021	TV-MA	1 Season	Crime TV Shows
3	s4	TV Show	Jailbirds New Orleans	nao_informado	nao_informado	2021	TV-MA	1 Season	Docuseries
4	s5	TV Show	Kota Factory	Mayur More	India	2021	TV-MA	2 Seasons	International TV Shows
...	...	...	...	...	...	...	...	...	...
8802	s8803	Movie	Zodiac	Mark Ruffalo	United States	2007	R	158 min	Cult Movies
8803	s8804	TV Show	Zombie Dumb	nao_informado	nao_informado	2018	TV-Y7	2 Seasons	Kids' TV
8804	s8805	Movie	Zombieland	Jesse Eisenberg	United States	2009	R	88 min	Comedies
8805	s8806	Movie	Zoom	Tim Allen	United States	2006	PG	88 min	Children & Family Movies
8806	s8807	Movie	Zubaan	Vicky Kaushal	India	2015	TV-14	111 min	Dramas

Proposta análitica:

A partir das estatísticas obtidas e com os dados modelados. Analisar a correlação entre os filmes pesquisados pelo usuário para recomendar um novo título.

- Considerando as características do dataframe, criamos um código para unificar os textos das colunas para uso do algoritmo.

```
catalogo['titulo_limpo'] = catalogo['title'].apply(limpeza_titulo)
catalogo['release_year'] = catalogo['release_year'].astype(str)

catalogo['titulo_limpo'] = catalogo['title'].apply(limpeza_titulo) + ' '+ catalogo['country'] + ' ' + catalogo['type'] + ' '+catalogo['rating'] + " " + catalogo['cast'] + " " + catalogo['release_year']
```

- Com o código abaixo vetorizamos as informações do texto para análise do algoritmo.

```
vectorizer = TfidfVectorizer(ngram_range=(1,3))

tfidf = vectorizer.fit_transform(catalogo['titulo_limpo'])
```

- Utilizando a "Similaridade Cos", criamos uma função de busca de similaridade baseando-se nos valores da coluna limpeza de título.

```
def busca(titulo):

  titulo = limpeza_titulo(titulo)
  query_vec = vectorizer.transform([titulo])
  similaridade = cosine_similarity(query_vec,tfidf).flatten()
  indices = np.argpartition(similaridade,-5)[-5:]
  resultado = catalogo.iloc[indices][::-1]
  return resultado
```

- No código abaixo criamos uma tela para o usuário. Onde ele deve digitar um título e baseado neste título terá recomendações de um título similar.

```
movie_input = widgets.Text(
    value = 'Zodiac',
    description = 'Titulo do Filme:',
    disabled = False
)

lista_filmes= widgets.Output()

def digitacao(data):
  with lista_filmes:
    lista_filmes.clear_output()
    titulo = data["new"]
    if len(titulo)>4:
      display(busca(titulo))
```

- Para finalizar, o código abaixo foi criado para interação com o usuário. Onde ele digita o título dentro de um display e recebe quatro títulos de recomendação..

```
movie_input.observe(digitacao, names = 'value')

display(movie_input, lista_filmes)
```

Titulo do Fil... Zodiac

	show_id	type	title	cast	country	date_added	release_year	rating	duration	listed_in	description	titulo_limpo
8802	s8803	Movie	Zodiac	Mark Ruffalo	United States	November 20, 2019	2007	R	158 min	Cult Movies	A political cartoonist, a crime reporter and a...	Zodiac United States Movie R Mark Ruffalo 2007
2998	s2999	TV Show	SAINT SEIYA: Knights of the Zodiac	Bryson Baugus	Japan	January 23, 2020	2020	TV-14	2 Seasons	Anime Series	Seiya and the Knights of the Zodiac rise again...	SAINT SEIYA Knights of the Zodiac Japan TV Sho...
2934	s2935	Movie	The Coldest Game	Bill Pullman	Poland	February 8, 2020	2019	TV-MA	103 min	Dramas	During 1962's Cuban missile crisis, a troubled...	The Coldest Game Poland Movie TV-MA Bill Pullm...
2933	s2934	Movie	Polaroid	Kathryn Prescott	United States	February 9, 2020	2019	PG-13	88 min	Horror Movies	A teen's discovery of a vintage Polaroid camer...	Polaroid United States Movie PG-13 Kathryn Pre...
2932	s2933	Movie	Fronteras	Steve Oropeza	United States	February 9, 2020	2018	TV-MA	106 min	Dramas	When a new, deadly narcotic sweeps the America...	Fronteras United States Movie TV-MA Steve Orop...

Bibliografia:

https://about.netflix.com/pt_br

https://www.teleco.com.br/streaming_mshare.asp