

## Sobre o Ranking

Aqui descrevo melhor como posso abordar algumas features. Leve em conta que também estou aprendendo.

Nosso ranking base <sup>baseline</sup> vai ser apenas a feature de popularidade e nota geral do filme. O motivo dessa escolha é que estas features são mais genéricas, fazendo com que o ranking já tenha uma boa classificação logo de cara.

O principal objetivo é personalizar o ranking para cada usuário. A forma de saber se a feature construída por cada um surtiu efeito será comparando os testes do modelo sem personalização e do modelo desenvolvido (com a feature personalizada).

## Feature: Gênero

(I)

filme	Gênero
x	romance, ação
y	aventura, guerra
z	comédia
w	guerra, ação, comédia

Explode  
Gênero →

(II)

filme	Gênero
x	romance
x	ação
y	aventura
y	guerra
z	comédia
w	guerra
w	ação
w	comédia

Podemos gerar duas features e testar qual desempenha melhor.  
Treinar com os gêneros individualmente (II) e Treinar considerando todos  
os gêneros do filme como uma combinação (I). Só consegui tirar essa conclusão porque olhei  
como estava o dado no df, observando que um filme pode ter mais que um gênero. Ao lado  
vemos um exemplo de usuários e notas para os filmes.

usuário	filme	nota
a	x	1
a	y	5
a	z	10
b	z	8
b	w	4
c	x	4
c	y	6
c	w	8

Meu desejo era utilizar a técnica de bag of words, o problema é que não acho suficiente apenas contar o número de vezes que o usuário viu determinado gênero como uma feature suficientemente boa, portanto vou fazer outra. Para complementar que faz a média da nota dos filmes que tinham determinado gênero.

Como vamos mexer com média de notas do usuário, preste atenção para não gerar vazamento de dados. É importante usar o timestamp para que a média do filme seja feita com notas anteriores ao momento que o usuário viu o filme, caso contrário você estará prevendo o futuro.

Id-film	genero
1	musica
2	Ação
3	romance
1	romance
3	Ação

join

User	Id-film	nota
1	1	4
1	2	9
1	3	5
2	2	7
2	3	3

Window

Id-user  
genero

Count  
bag-of-genero

Id-User	Id-film	nota	genero
1	1	4	musica
1	2	9	Ação
1	3	5	romance
1	1	4	romance
1	3	5	Ação
2	3	7	Ação
2	3	7	romance
2	2	3	Ação

Id-User	Id-film	nota	genero	Count_genero
1	1	4	Música	1
1	2	9	Ação	2
1	3	5	Romance	2
1	1	4	Romance	2
1	3	5	Ação	2
2	3	7	Ação	2
2	3	7	Romance	1
2	2	3	Ação	2

Preferi usar window pois posso fazer a feature de nota por genero no mesmo df, economizando join. Adicionando coluna de timestamp para essa outra feature. Ela é necessária para evitar o problema de vazamento de dados, em resumo seria evitar de prever o que você quer assistir hoje usando os dados da semana que vem.

Ps: É importante fazer isso na feature de bag of words também. No código isso será considerado!

Ps<sub>2</sub>: A gente está usando a ideia de bag of words, mas não é exatamente um. Esse método é mais comum em textos longos, com + de 2 mil palavras.

Faremos o window usando estes 3 atributos. A nota será composta pela nota presente + notas passado dividido pelo total de notas Presente + passado

Presente = linha atual

(segundo) timestamp	Id-User	Id-film	nota	genero	Count_genero
20	1	2	9	Música	1
80	1	3	5	Ação	2
1	1	1	4	Romance	2
80	1	3	5	Ação	2
50	3	2	7	Ação	2
50	3	2	7	Romance	1
100	2	2	3	Ação	2

(segundo)	timestamp	Id-User	Id-film	nota	genero	count_genero	media_genero
1	1	1	1	4	Música	1	4
20	1	1	2	9	Ação	2	9
80	1	1	3	5	Romance	2	4,5
1	1	1	1	4	Romance	2	4
80	1	1	3	5	Ação	2	7
50	2	2	3	7	Ação	2	7
50	2	2	3	7	Romance	1	7
100	2	2	2	3	Ação	2	5

Temos a feature, agora precisamos colocar no dataframe de treino.

Como cada linha é uma nota do usuário, vamos agrupar as linhas dessa maneira no dataset com as features.

Como seguimos a abordagem II, temos 20 gêneros ao todo. Vamos pivotar de forma a termos 40 features, 20 de count e 20 de média. Nem todos os usuários tem notas em todos os filmes, nesse caso vamos deixar um valor neutro (talvez 5\*). Depois disso juntamos tudo em duas colunas como estava antes, porém em formato de array.

\* Gerei um notebook chamado análise de dispersão para saber qual valor faz mais sentido → Alguém pode ver isso

Posição:	0	1	2
Array =	Valores do gênero música	Valores do gênero Ação	Valores do gênero Romance

Saber a posição de cada gênero no array é importante para não nos perdemos na hora de jogar as features no modelo. No código vou ordenar por ordem alfabética.  
Resultado Pra média

Falta colocar os gêneros dos filmes

↳ Pra feature ser útil precisamos saber a quais gêneros o filme pertence. então criei um dataframe e coloque os identificadores do gênero.

Posteriormente ordenamos a lista de gêneros, onde música era o valor 0, ação valor 1, romance 2, e assim vai. O valor virou a informação sobre qual gênero o filme pertence. Se o filme é de ação e romance, ele terá [1,2] na coluna genero\_movie.

No final teremos esse dataset: