

Sistemas de Recomendação

Comparação de técnicas e como eles moldam as plataformas de conteúdo

Nome: Raissa Cavalcante Correia
Orientadora: Esther Luna Colombini

O que são os sistemas de recomendação?

Esses sistemas são os que dado um histórico de usuário de avaliação dos itens de uma plataforma de conteúdo, uma rede social, ou um e-commerce, seja essa avaliação explícita ou implícita, fornece um conjunto de novos itens para compor a página inicial deste usuário. Tentando manter assim, a plataforma adaptada, dinâmica e interessante para o usuário continuar consumindo os produtos ou o conteúdo.

Motivação

Com o advento das redes sociais, do e-commerce, e de plataformas de conteúdo como Netflix, Spotify e Youtube, surgiu a necessidade de algoritmos capazes de recomendar o conteúdo ao usuário para facilitar e dinamizar seu consumo. Além de manter o usuário interessado.

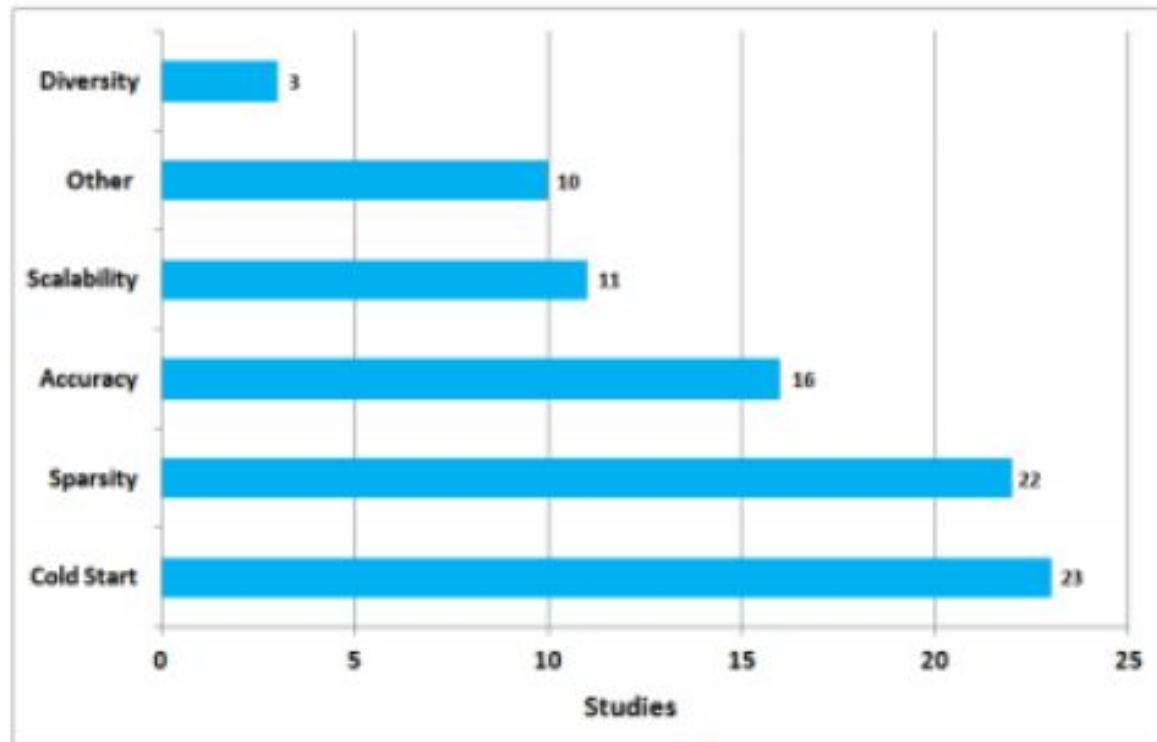
Nesse mesmo período ocorreram grandes avanços no aprendizado de máquina, e na ciência de dados aplicada, graças ao aumento de oferta de poder computacional e a disponibilidade de dados abundante.

Tais sistemas além de terem um imenso valor comercial, moldam o que assistimos, que notícias ficamos sabendo, quais produtos cujas propagandas ficamos sabendo.

É extremamente importante um conhecimento maior tanto de seu funcionamento, para entendermos como tais plataformas funcionam e como podem ser aperfeiçoados.

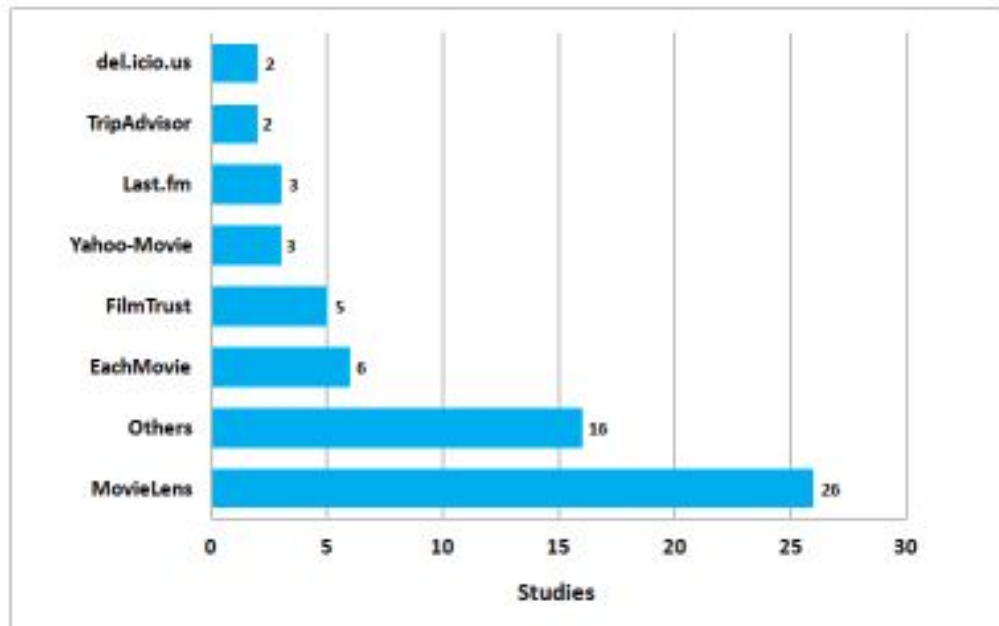
Principais Problemas Abordados

- Dados Esparsos
- Escalabilidade
- Itens Sinônimos
- Diversidade
- Ovelha Cinza
- Cold Start



Dataset Movie Lens

O dataset mais usado por uma boa margem é o MovieLens por sua qualidade, simplicidade, apenas os csv necessários e por vir numa diversidade razoável de tamanhos. Foi feito o download do 100K e do 1M, porém apenas o 100K foi computacionalmente viável no notebook.



Redução de dimensionalidade

A redução de dimensionalidade, feita através do PCA, é uma etapa fundamental para um processo de clusterização de itens ser bem feito. No entanto o dataset MovieLens já é muito limpo e curado, especialmente a versão menor. Podemos ver a contribuição das 18 features nos 2 gráficos abaixo, e para os filmes é ainda pior que o usuário. Dessa forma a ideia de quebrar o dataset em sub grupos de itens e/ou de usuários estrategicamente não foi possível.

```
PCA FIT: PCA(copy=True, iterated_power='auto', n_components='mle', random_state=None,  
| svd_solver='auto', tol=0.0, whiten=False)  
Original shape: (2000, 19)
```

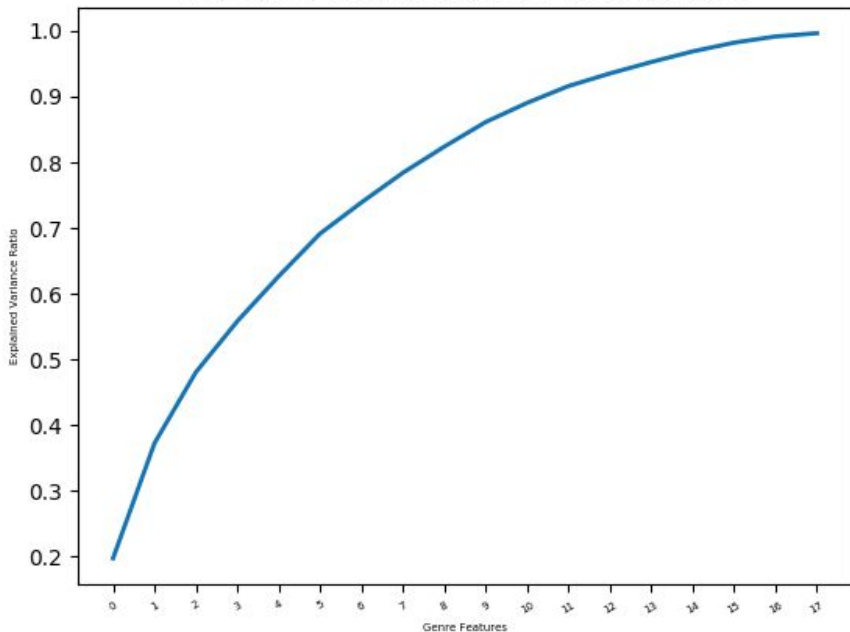
```
PCA FIT: PCA(copy=True, iterated_power='auto', n_components='mle', random_state=None,  
| svd_solver='auto', tol=0.0, whiten=False)  
Original shape: (589, 19)
```

```
Explained Variance Ratio: [0.1973232 0.1749443 0.10822461 0.07753266 0.06851518 0.06495379  
0.04747243 0.04511621 0.03979861 0.03746309 0.02913248 0.02565409  
0.01915558 0.01745262 0.01612411 0.01331549 0.00939311 0.00494625]
```

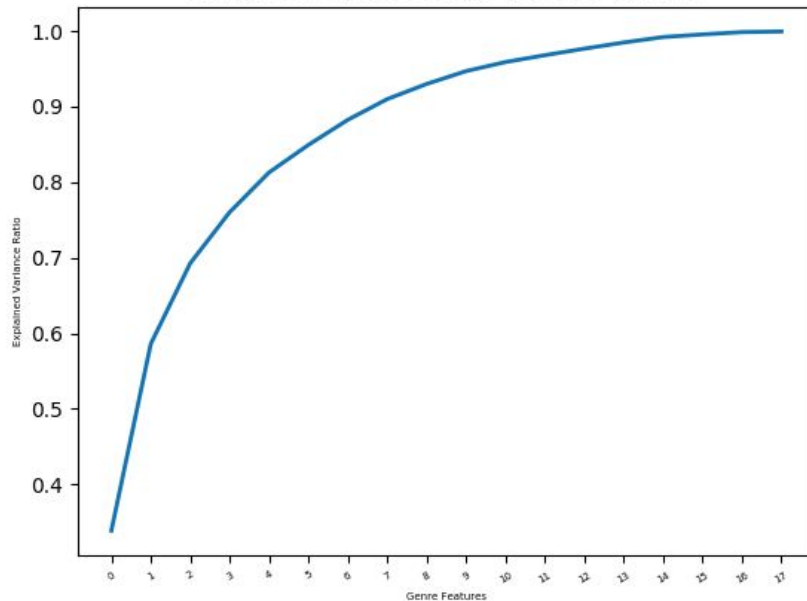
```
Explained Variance Ratio: [0.33865714 0.24721484 0.10668785 0.0676169 0.05263799 0.03652536  
0.03321709 0.02740602 0.0200626 0.01698555 0.0120138 0.00906401  
0.0086407 0.00818422 0.00719693 0.00348899 0.00306723 0.00078554]
```

Redução de Dimensionalidade

Explained Variance Ratio for Movie Profiles



Explained Variance Ratio for User Profiles



Clusterização

Foi escolhido o HDBSCAN pela sua capacidade por criar clusters através de densidade de distribuição, pois os pontos no espaço de 17 dimensões não tinha uma forma definida, ou uma forma perceptível pela leitura da distribuição dos pontos. A conclusão é que um percentual muito significativo dos filmes, e quase todos os usuários não são possíveis de serem postos em clusters.

```
You, 17 days ago | 1 author (you)
HDBSCAN(algorithm='best', allow_single_cluster=False, alpha=1.0,
        approx_min_span_tree=True, cluster_selection_method='leaf',
        core_dist_n_jobs=4, gen_min_span_tree=False, leaf_size=40,
        match_reference_implementation=False, memory=Memory(location=None),
        metric='euclidean', min_cluster_size=5, min_samples=5, p=None,
        prediction_data=False)

Outline Rate in Movies: 0.2645

HDBSCAN(algorithm='best', allow_single_cluster=False, alpha=1.0,
        approx_min_span_tree=True, cluster_selection_method='leaf',
        core_dist_n_jobs=4, gen_min_span_tree=False, leaf_size=40,
        match_reference_implementation=False, memory=Memory(location=None),
        metric='correlation', min_cluster_size=2, min_samples=2, p=None,
        prediction_data=False)

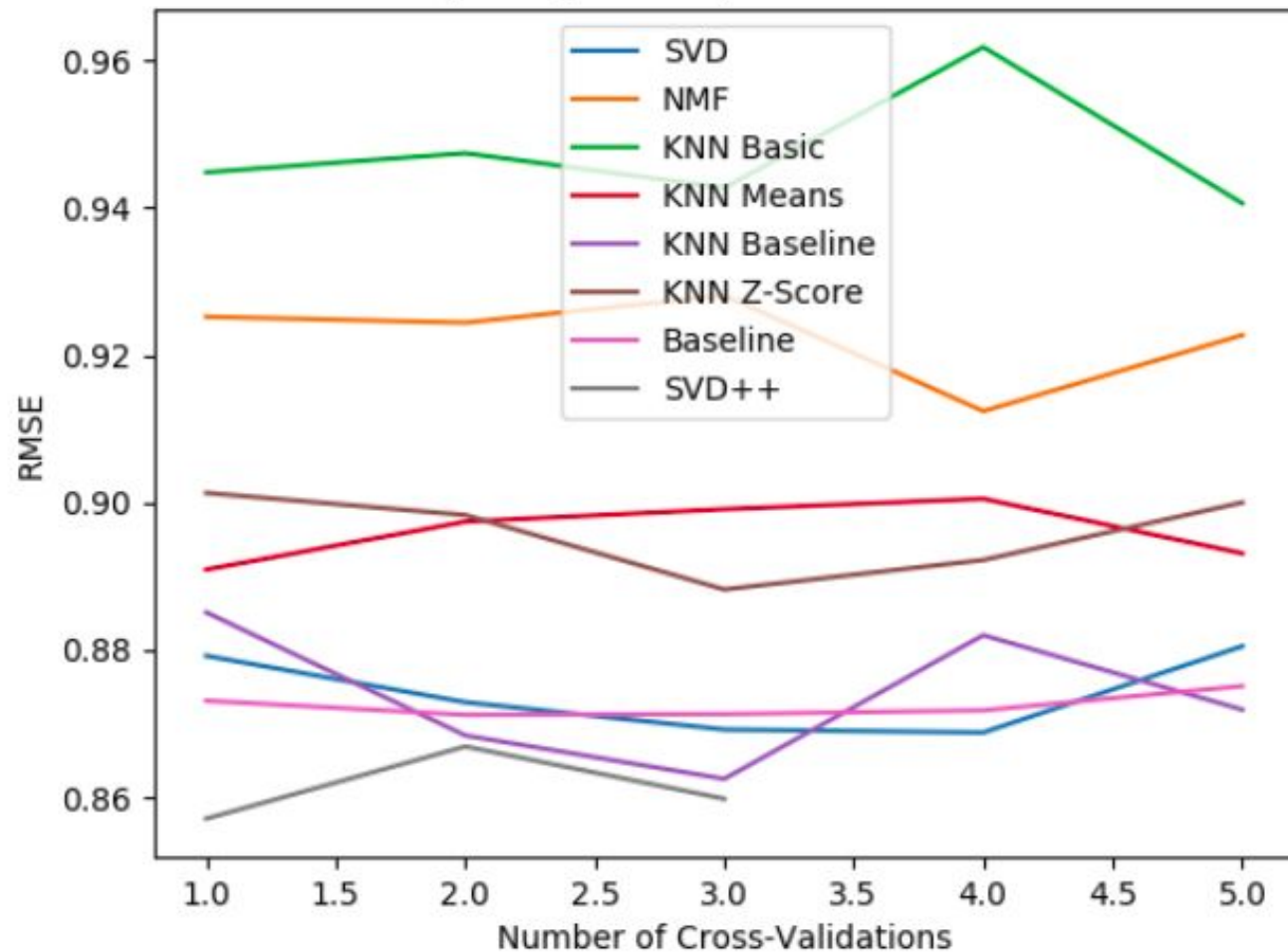
Outline Rate in Users: 0.733446519524618 | You, 17 days ago • HDBSCAN execu
```


Comparação Algoritmos

Tabela 1: Comparing the Best Algorithm Results

Algorithm	RMSE	MAE	MSE	FCP	Precision	Recall
KNN Basic	0.9406	0.7215	0.8848	0.6690	0.7875	0.2958
KNN Means	0.8909	0.6801	0.7937	0.6522	0.8432	0.2845
KNN Z-Score	0.8882	0.6734	0.7889	0.6511	0.8260	0.2771
KNN Baseline	0.8625	0.6602	0.7438	0.6698	0.8353	0.2721
SVD	0.8688	0.6685	0.7547	0.6592	0.8447	0.2522
NMF	0.9124	0.7021	0.8324	0.6452	0.8099	0.2738
SVD++	0.8571	0.6582	0.7347	0.6832	0.8642	0.2557
Co-Clustering	0.8961	0.6850	0.8031	0.6623	0.8400	0.2844
Slope One	0.8961	0.6850	0.8031	0.6623	0.8400	0.2844
Baseline	0.8712	0.6697	0.7591	0.6772	0.8773	0.2297
Normal	1.4217	1.1348	2.0212	0.4986	0.5933	0.2465

RMSE per Algorithm per Cross Validations



F1 per Algorithm per Cross Validations

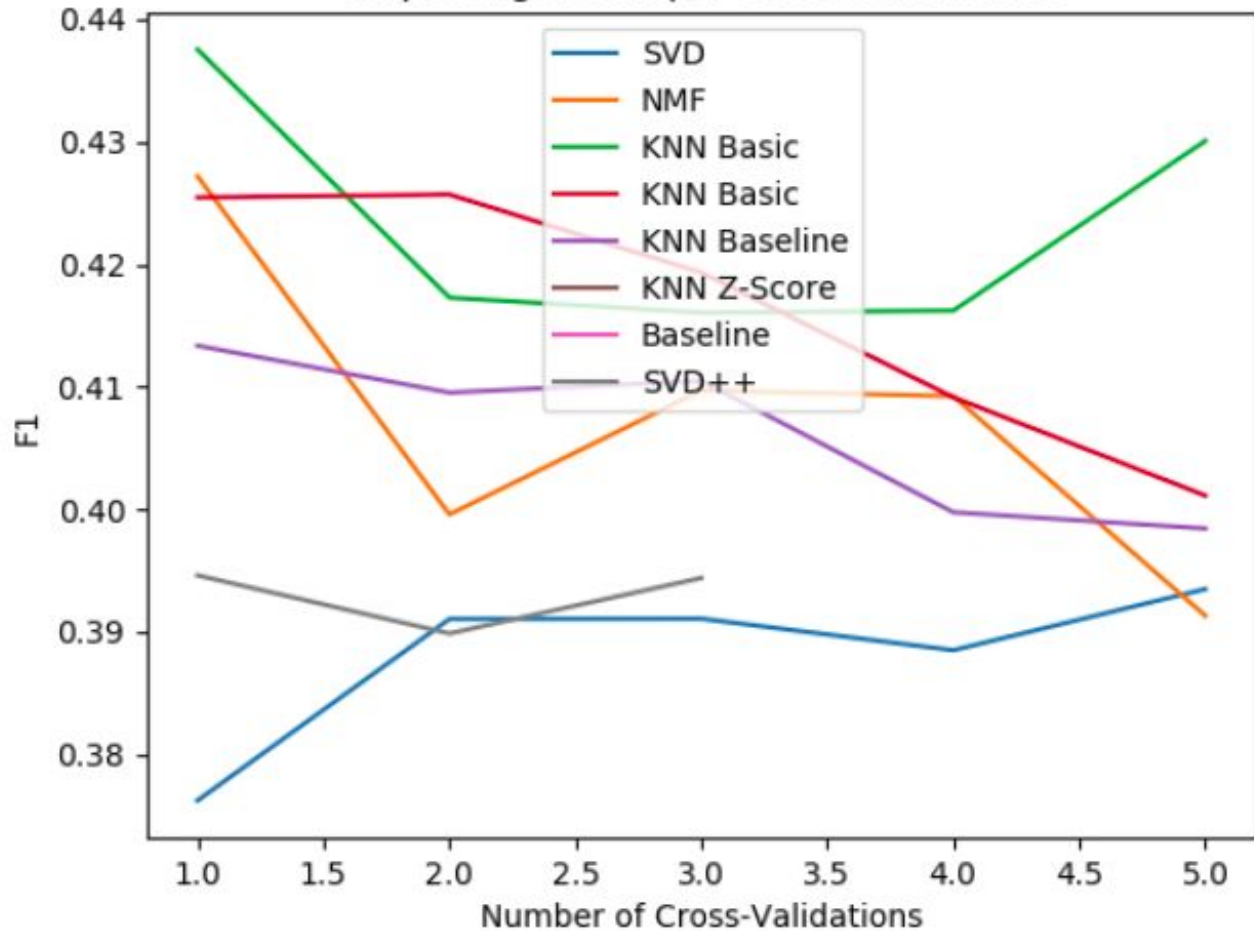


Figura 15: Distribuição de Erro nas Recomendações

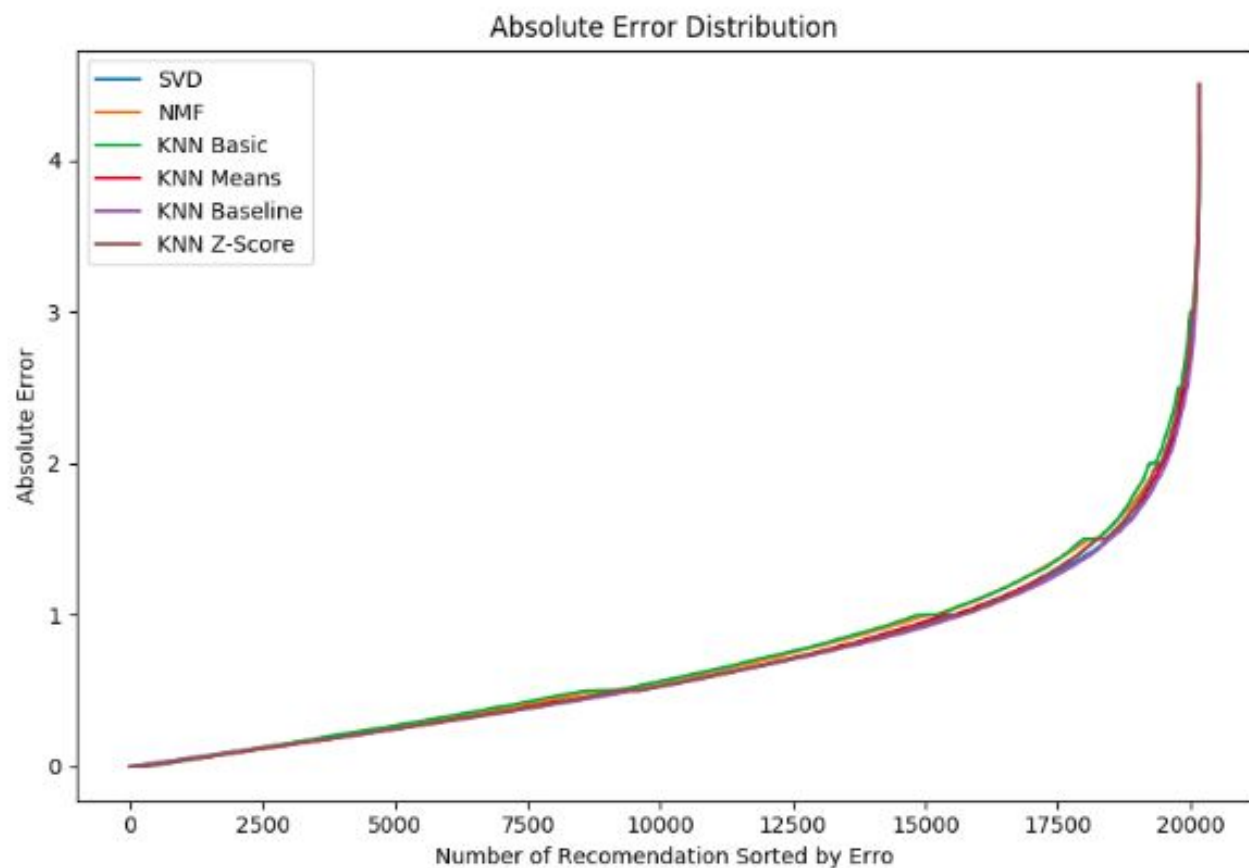
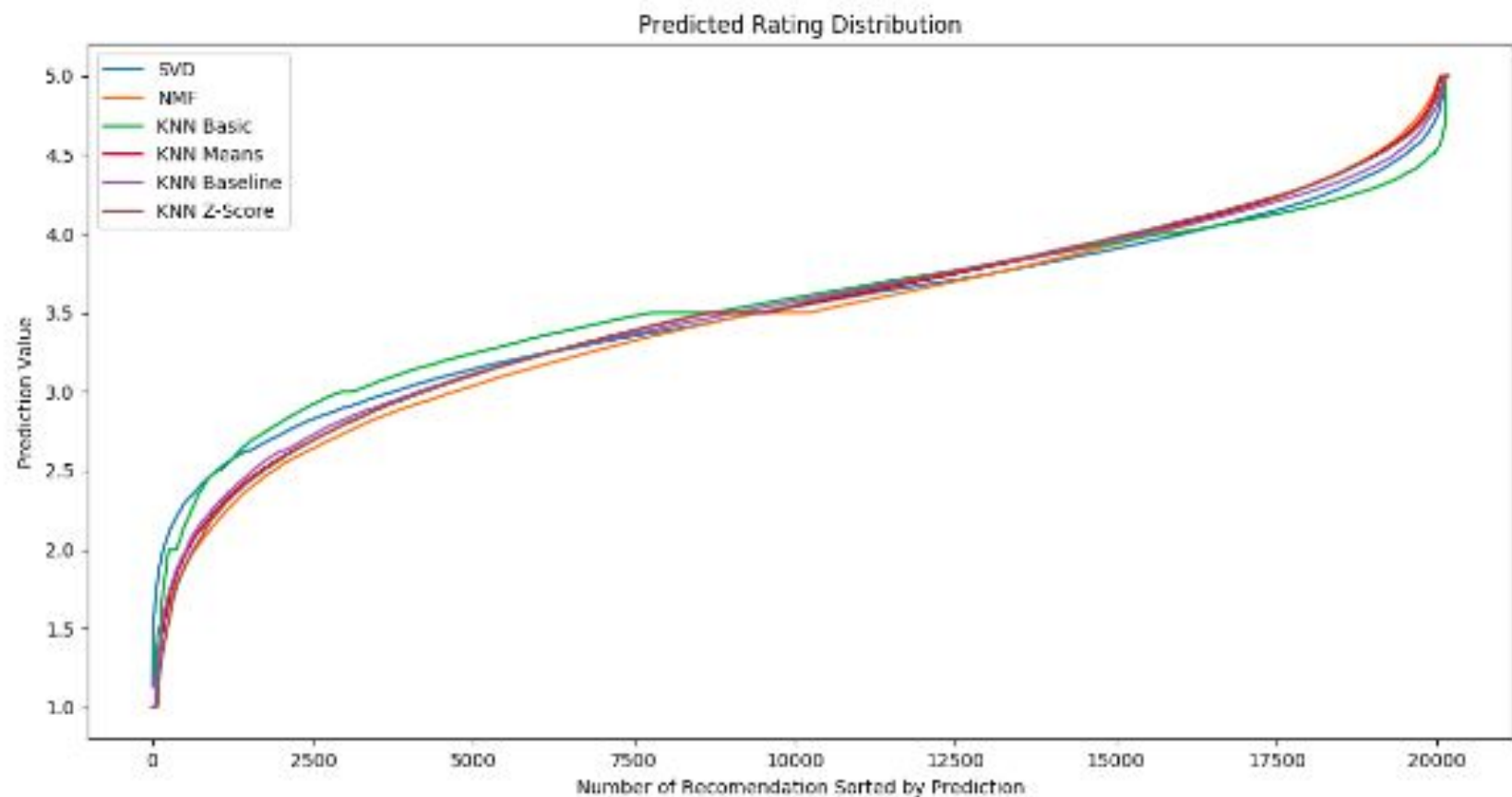


Figura 16: Avaliação Prevista nas Recomendações

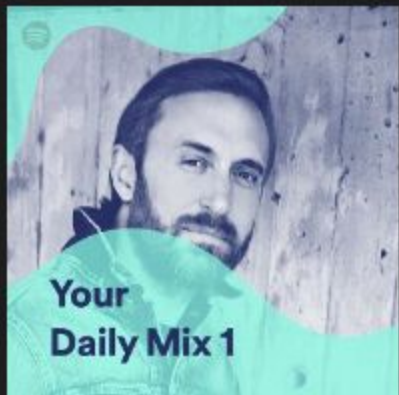


Conclusões

- Habilidade e Capacidade de clusterizar os itens por categorias
- Lidar com itens pouco avaliados, e como recomendar para usuários com muito histórico
- Lidar com usuários que tem pouco histórico
- Híbrido de algoritmos e as medidas já citadas para lidar com dados esparsos
- O problema dos itens sinônimos é necessário um sistema de constante exclusão do que foi consumido ou comprado e recalculado para aquele usuário, além da identificação do que é idêntico e do que é similar.
- Mais importante do que pedir para o usuário dar uma nota de 1 a 5, é apenas o like/dislike é suficiente, em conjunto com os dados de interação implícita.
- Evoluir os algoritmos para além do que temos hoje é muito difícil. O mais importante nos últimos anos tem sido aplicar os algoritmos e a preparação dos dados corretamente de acordo com o contexto, junto de híbridizações adequadas.

Exemplo de segmentação por fatores latentes do usuário, e recomendar itens por aproximação a partir disso. Interface do Spotify, Novembro/2019

Your Daily Mixes



Daily Mix 1

David Guetta, Don Diablo, Marshmello and more



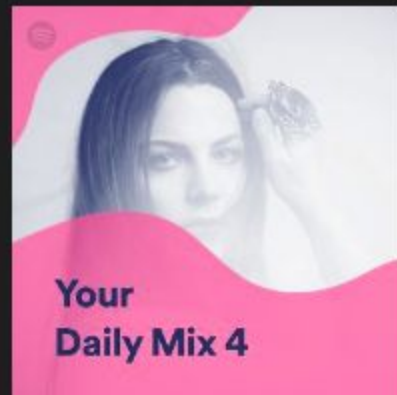
Daily Mix 2

Sia, Dua Lipa, Carly Rae Jepsen and more



Daily Mix 3


Lara Fabian, Zaz, Jackie Evancho and more



Daily Mix 4

Amy Lee, Nightwish, Within Temptation and more

Exemplo de avaliação por Like, ao invés de rating. Além disso, o like também adiciona na biblioteca. Interface do Spotify, Novembro/2019



PLAYLIST

This Is Alan Walker

The essential Alan Walker tracks and remixes.

Created by Spotify • 43 songs, 2 hr 13 min

PLAY

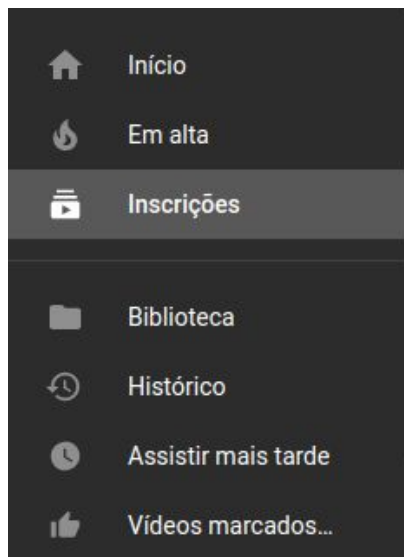
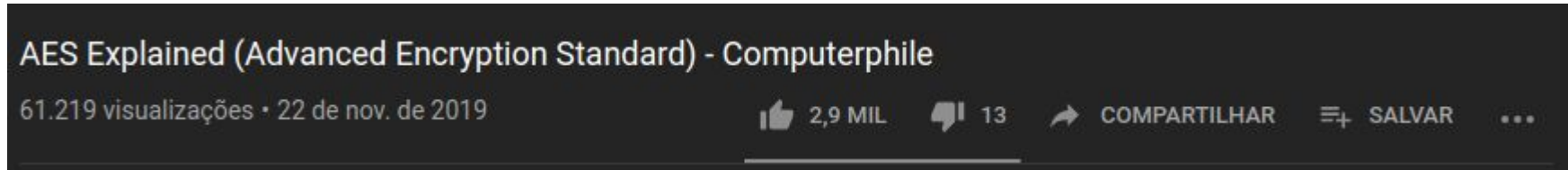
FOLLOWERS
904,920

Filter

Download

	TITLE	ARTIST	ALBUM	
	Avem (The Aviation Theme)	Alan Walker	Avem (The Aviati...	2019-10-30
	Play	K-391, Alan Walk...	Play	2019-10-30

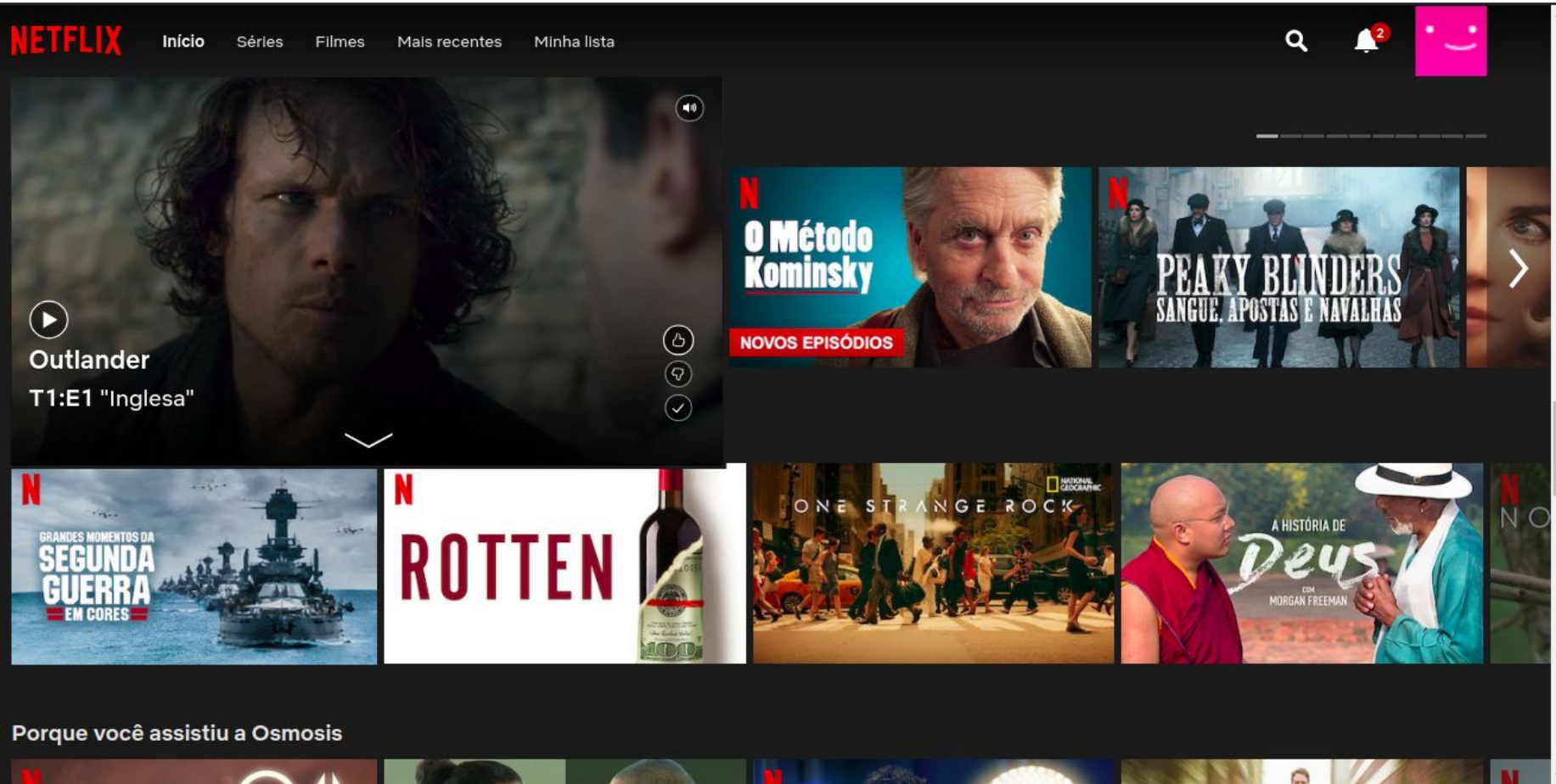
Exemplo de avaliação Binária, ao invés de Rating. Além de critérios implícitos, como compartilhar e salvar em playlist. Interface do Youtube, Novembro/2019



Outra questão interessante no Youtube, que demonstra esse sintoma: A página inicial é apenas de itens recomendados, se quiser ver os canais que voce se inscreveu postou, tem que ir na sub página de inscrições.

Provando novamente que a avaliação implícita de tempo assistido, juntamente da explícita de gostei ou não, é mais relevante para manter o interesse do usuário que o ato de se inscrever.

Outro exemplo de aplicação binária, com “adicionar a minha lista” e ter assistido como avaliação implícita. Interface Netflix, Novembro/2019



O e-commerce mantém a avaliação por estrelas de 1 a 5. Amplo uso do critério de “visualizado/pesquisado, e tem várias seções de promoção e indicações genérica na home. Interface da Amazon, 2019

Promoção em Destaque

[Veja mais](#)



Celular Xiaomi Note 8
64GB Rom 4GB Ram Du...

Redmi

R\$ 1.048⁰⁹ ~~R\$ 1.999,90~~

★★★★★ (139)



Relógio Inteligente Mi
Band 4 Original Xiaomi S...

Xiaomi

R\$ 164⁹⁰ ~~R\$ 299,90~~

★★★★★ (492)



Smartphone Xiaomi Mi A3
64GB 4GB RAM Preto -...

Xiaomi

R\$ 999⁰⁰ ~~R\$ 1.099,99~~

★★★★★ (776)



Smartphone Xiaomi Redmi
Note 7 64GB 4GB RAM P...

Xiaomi

R\$ 993³⁹ ~~R\$ 1.356,74~~

★★★★★ (4,872)



Smartphone Xiaomi Mi 9
Lite 128GB 6GB RAM On...


Xiaomi

R\$ 1.639⁰⁰


★★★★★ (49)

Página inicial Amazon Novembro/2019 não logada


Navegue por loja




Ferramentas



Fire TV Stick



Games




Joias

Confira milhares de produtos e novas ofertas todos os dias.

[Veja mais](#)


Ofertas do Dia



Todos os dias, novas ofertas com até 80% de desconto.

[Ver todas](#)


Loja de R\$50



Aproveite nossos pequenos preços no Esquenta Black Friday.

[Confira os destaques](#)

Kindle Unlimited




Mais de um milhão de eBooks para você ler onde e quando quiser.

[Saiba mais](#)


Ofertas até 80% off no Esquenta Black Friday

[Ver mais](#)




Esquenta Black Friday


[Ver todas as ofertas](#)




Até 80% off em Livros




Até 50% off em Esportes




Até 60% off em eBooks



Até 40% off em Informática




Até 40% off em Moda




Até 40% off em marcas para Bebê

Oferas em alta


[Veja todas](#)




R\$59,90
~~R\$69,90~~
Termina em 10:42:2




R\$74,90
Termina em 00:27:54



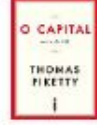
R\$16,40
~~R\$19,90~~
Termina em 10:42:2



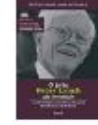
R\$129,00
~~R\$149,00~~
Termina em 10:42:53




R\$48,99 - R\$118,91
Termina em 10:42:5



R\$9,99
~~R\$14,99~~
Termina em 10:42:2









R\$14,00
~~R\$19,90~~
Termina em 10:42:1




R\$2,99
~~R\$3,99~~
Termina em 10:42:1

Oferas em casa e cozinha

[Veja mais](#)




Eletrônicos



Confira os eletrônicos em oferta no Esquenta Black Friday.

[Veja mais](#)


Relógios até 40% off



Encontre Casio, Technos e mais.

[Confira](#)


Oferas em Eletrônicos




Até 40% off em TVs, caixas de som, câmeras e muito mais.

[Confira](#)


Loja de Eletrodomésticos




Geladeiras



Fogões e Fornos



Lava-Louças



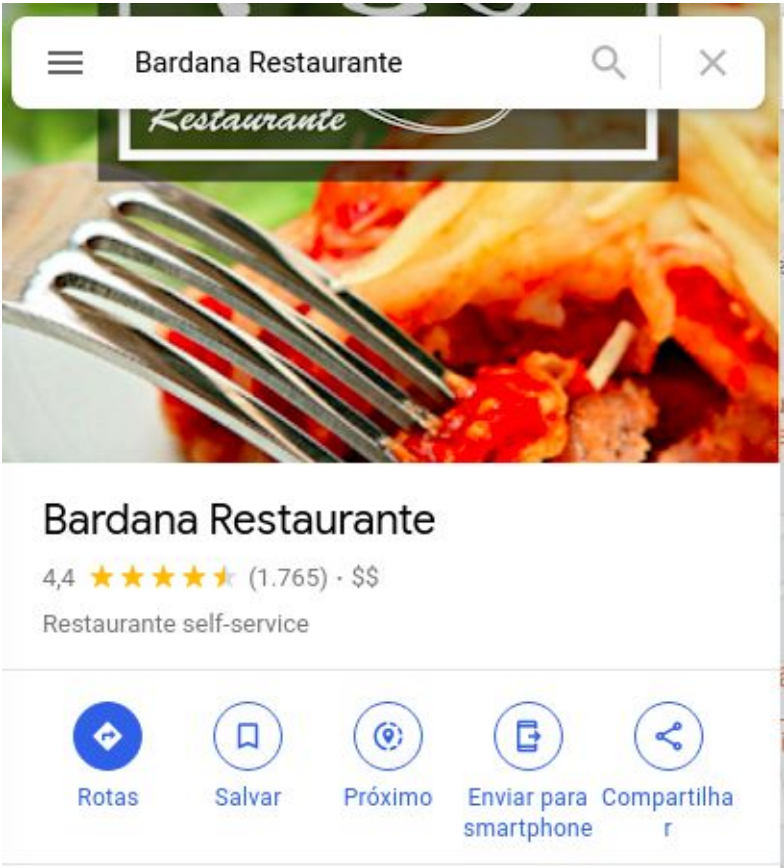
Lavadoras e Secadoras

Produtos essenciais para sua cozinha e lavanderia.

[Veja mais](#)

Exemplo das recomendações genérica oferecidas como base, as categorias mais populares, os itens mais populares no momento, e proximidade de itens entre si para calcular o ranking de itens a ser mostrado em cada sub página.

Avaliação por Estrelas de 1 a 5 se mantém em itens que envolvem uma avaliação maior da experiência do usuário. Google Maps, interface de Novembro/2019



Avaliação por Estrelas de 1 a 5 se mantém em itens que envolvem uma avaliação melhor da subjetividade do usuário, e da relação entre o número de estrelas. AirBNB, interface de Novembro/2019

Lugares para estadia em todo o mundo



SUPERHOST Stanardsville ★4,95
*YURT*Goats*MTNS*3rd NITE 25%...
R\$733/noite

Trullo del 1800 in Valle d'Itria



SUPERHOST Bali · Balian Beach, Bali ★4,84
BALIAN TREEHOUSE w beautiful pool
R\$348/noite



Pioneertown ★4,93
Off-grid itHouse
R\$1,886/noite



PLUS Old Town · Edinburgh ★4,98
Classical Apartment on the Royal Mile
R\$641/noite



LB of Hammersmith & Fulham ... ★4,87
Stylish house close to river thames
R\$352/noite



Bali · Ubud ★4,64
Bamboo eco cottage in rice fields
R\$277/noite



SUPERHOST Mount Washington ... ★4,96
LA Pool, Privacy and Amazing Views!
R\$629/noite