Trabalho final de CPD 2019 - MovieLens 20M

Neste trabalho aplicaremos diversas técnicas vistas em aula para explorar o dataset **MovieLens 20M**. O dataset contém avaliações e tags (anotações em texto-livre) do serviço de recomendações MovieLens. São 20,000,263 avaliações (notas entre 1 e 5) e 465,564 aplicações de tags (ex.: narrated, jazz, England, hero) em 27,278 filmes. Esse dataset foi criado por 138,493 usuários entre 09/01/1995 e 31/03/2015. Os usuários foram selecionados aleatoriamente e todos avaliaram no mínimo 20 filmes.

Para realizar as tarefas, utilizaremos 3 arquivos:

`movie.csv`, que contém informações sobre os filmes:

	movieId	title	genres
0	1	Toy Story (1995)	Adventure|Animation|Children|Comedy|Fantasy
1	2	Jumanji (1995)	Adventure|Children|Fantasy
2	3	Grumpier Old Men (1995)	Comedy|Romance
3	4	Waiting to Exhale (1995)	Comedy|Drama|Romance
4	5	Father of the Bride Part II (1995)	Comedy
5	6	Heat (1995)	Action|Crime|Thriller
6	7	Sabrina (1995)	Comedy|Romance
7	8	Tom and Huck (1995)	Adventure|Children
8	9	Sudden Death (1995)	Action
9	10	GoldenEye (1995)	Action|Adventure|Thriller
10	11	American President, The (1995)	Comedy|Drama|Romance
11	12	Dracula: Dead and Loving It (1995)	Comedy|Horror
12	13	Balto (1995)	Adventure|Animation|Children
13	14	Nixon (1995)	Drama
14	15	Cutthroat Island (1995)	Action|Adventure|Romance
15	16	Casino (1995)	Crime|Drama
16	17	Sense and Sensibility (1995)	Drama|Romance
17	18	Four Rooms (1995)	Comedy
18	19	Ace Ventura: When Nature Calls (1995)	Comedy
19	20	Money Train (1995)	Action|Comedy|Crime|Drama|Thriller

`tag.csv`, que contém as tags que os usuários aplicam aos filmes:

	userId	movieId	tag	timestamp
0	18	4141	Mark Waters	2009-04-24 18:19:40
1	65	208	dark hero	2013-05-10 01:41:18
2	65	353	dark hero	2013-05-10 01:41:19
3	65	521	noir thriller	2013-05-10 01:39:43
4	65	592	dark hero	2013-05-10 01:41:18

`rating.csv`, que contém as avaliações que os usuários fizeram a cada filme:

	userId	movieId	rating	timestamp
0	1	2	3.5	2005-04-02 23:53:47
1	1	29	3.5	2005-04-02 23:31:16
2	1	32	3.5	2005-04-02 23:33:39
3	1	47	3.5	2005-04-02 23:32:07
4	1	50	3.5	2005-04-02 23:29:40

Também disponibilizamos um arquivo com 10,000 avaliações ao invés das 20 milhões para ajudar nos testes (**`minirating.csv`** - disponível para download apenas no moodle**).**

Os dados podem ser baixados no moodle ou em:
https://www.kaggle.com/grouplens/movielens-20m-dataset

Referência: *F. Maxwell Harper and Joseph A. Konstan. 2015. The MovieLens Datasets: History and Context. ACM Transactions on Interactive Intelligent Systems (TiiS) 5, 4, Article 19 (December 2015), 19 pages.*

Tarefas

Os usuários devem construir uma aplicação que funciona em duas fases: Construção e inicialização das estruturas de dados necessárias, e modo console. Ao executar a fase de construção, esta não deve demorar mais de 3 minutos. Quem conseguir fazer esta etapa em menos de 1 minuto ganha um bônus de 10%.

A primeira fase engloba as seguintes tarefas:

1. Árvore trie
Construção de uma árvore trie cuja chave é o nome dos filmes. Cada folha deve conter o *movieId* associado ao filme. Essa estrutura será usada na busca por prefixo.

2. Tabela hash
Construção de uma tabela hash cujas chaves são os ids do filmes (*movieId*) e os dados são a lista de gêneros, a média das avaliações e número de avaliações para esse filme.

3. Estrutura Livre (árvore binária de busca, hash ou array ordenado, etc.)
Construção de uma estrutura de dados que usa como chave o id do usuário (userID - números inteiros, e retorna os filmes avaliados e as respectivas notas.

4. Pesquisas
Depois dessas estruturas serem construídas, o programa deve entrar no modo console, onde será possível fazer pesquisas (queries) nos dados utilizando as estruturas mencionadas anteriormente. Essas são:

`4.1: movie <title or prefix>` - Retornar a lista de filmes cujo título começam com esse prefixo e para cada filme mostrar os gêneros, avaliação média e número de avaliações. A forma como isso deve ser feito é usando as estruturas citadas acima. Deve-se buscar na trie todos os *movieIds* que correspondem ao título ou prefixo dado e com essa lista de prefixos, e depois buscar na tabela hash o resto das informações.

```
$ movie Star Wa
```

movieId	title	genres	rating	count
260	Star Wars: Episode IV - A New Hope (1977)	Action|Adventure|Sci-Fi	4.190672	54502
1196	Star Wars: Episode V - The Empire Strikes Back...	Action|Adventure|Sci-Fi	4.188202	45313
1210	Star Wars: Episode VI - Return of the Jedi (1983)	Action|Adventure|Sci-Fi	4.004622	46839
2628	Star Wars: Episode I - The Phantom Menace (1999)	Action|Adventure|Sci-Fi	3.080983	29574
5378	Star Wars: Episode II - Attack of the Clones (...	Action|Adventure|Sci-Fi|IMAX	3.108584	16425
33493	Star Wars: Episode III - Revenge of the Sith (...	Action|Adventure|Sci-Fi	3.465821	12303
61160	Star Wars: The Clone Wars (2008)	Action|Adventure|Animation|Sci-Fi	3.013642	843
100089	Star Wars Uncut: Director's Cut (2012)	Action|Animation|Comedy|Sci-Fi|Western	3.833333	21
109713	Star Wars: Threads of Destiny (2014)	Action|Adventure|Sci-Fi	1.250000	6

Como no exemplo acima, tente deixar o texto a ser impresso compacto.

`4.2: user <userID>` - Retorna a lista de filmes revisados pelo usuário e para cada filme mostra a nota dada pelo usuário, a média global e a contagem de avaliações. Usar a estrutura de dados livre escolhida no item 3.

$ user 4

user_rating	title	global_rating	count
3.0	Heat (1995)	3.834930	23899
4.0	GoldenEye (1995)	3.430029	29005
3.0	Ace Ventura: When Nature Calls (1995)	2.607412	20938
1.0	Twelve Monkeys (a.k.a. 12 Monkeys) (1995)	3.898055	44980
3.0	Die Hard: With a Vengeance (1995)	3.489025	33940
3.0	Star Trek: Generations (1994)	3.329628	26404
4.0	Client, The (1994)	3.467410	15833
4.0	Forrest Gump (1994)	4.029000	66172
3.0	Mask, The (1994)	3.174718	34384
4.0	Maverick (1994)	3.410737	17062
4.0	Naked Gun 33 1/3: The Final Insult (1994)	2.955508	14261
4.0	Speed (1994)	3.493203	41562
3.0	True Lies (1994)	3.491149	43159
3.0	Beverly Hills Cop III (1994)	2.749866	14934
4.0	Carlito's Way (1993)	3.695059	8946
3.0	Dave (1993)	3.603980	19297
5.0	Firm, The (1993)	3.495387	25689
4.0	Jurassic Park (1993)	3.664741	59715
4.0	Made in America (1993)	2.768635	3032
3.0	RoboCop 3 (1993)	2.196904	4812
4.0	Robin Hood: Men in Tights (1993)	3.040169	14439
3.0	Secret Garden, The (1993)	3.539080	7459
3.0	Terminal Velocity (1994)	2.913999	2843
4.0	Home Alone (1990)	3.086655	28348
4.0	Terminator 2: Judgment Day (1991)	3.931954	52244
4.0	Snow White and the Seven Dwarfs (1937)	3.610492	17766
4.0	Pinocchio (1940)	3.499687	12782
5.0	Rock, The (1996)	3.679536	31353

4.3: top<N> ‘<genre>’ - Retorna os N filmes com melhores notas de um dado gênero **com no mínimo 1000 avaliações**.

$ top10 ‘action’

title	genres	rating	count
Seven Samurai (Shichinin no samurai) (1954)	Action|Adventure|Drama	4.274180	11611
Band of Brothers (2001)	Action|Drama|War	4.263182	4305
City of God (Cidade de Deus) (2002)	Action|Adventure|Crime|Drama|Thriller	4.235410	12937
North by Northwest (1959)	Action|Adventure|Mystery|Romance|Thriller	4.233538	15627
Fight Club (1999)	Action|Crime|Drama|Thriller	4.227123	40106
Dark Knight, The (2008)	Action|Crime|Drama|IMAX	4.220129	20438
Raiders of the Lost Ark (Indiana Jones and the...	Action|Adventure	4.219009	43295
Yojimbo (1961)	Action|Adventure	4.211717	3559
Star Wars: Episode IV - A New Hope (1977)	Action|Adventure|Sci-Fi	4.190672	54502
Star Wars: Episode V - The Empire Strikes Back...	Action|Adventure|Sci-Fi	4.188202	45313

4.4: tags <list of tags> - Para uma lista de tags dada como entrada, a query deve retornar a lista de filmes que estão associados a esse conjunto de tags. Cabe ao aluno pensar em uma forma eficiente de resolver esse problema. Como as tags podem ser termos com espaço(ex.: dark hero, noir thriller), cada tag deve ser escrita entre apóstrofes.

$ tags ‘Brazil’ ‘drugs’

title	genres	rating	count
City of God (Cidade de Deus) (2002)	Action|Adventure|Crime|Drama|Thriller	4.235410	12937
Carandiru (2003)	Crime|Drama	3.705085	295
Cazuza - O Tempo Não Pára (2004)	Drama	3.156250	16
Elite Squad: The Enemy Within (Tropa de Elite ...	Action|Crime|Drama	3.946610	590

O que pode e o que não pode ser usado

É possível fazer o trabalho em C, C++, Python e Java. Não é permitido usar bibliotecas ou mecanismos da linguagem de alto nível, nem implementações prontas para lidar, buscar ou armazenar os dados (ex.: pandas, numpy, dicionários, maps, bancos de dados). Todas as estruturas citadas anteriormente, buscas e ordenações devem ser implementadas pelo aluno. Não é permitido reabrir os arquivos após a fase de construção e inicialização das estruturas. Qualquer dúvida, consulte um dos monitores da cadeira.

Extra

Interfaces gráficas e novas queries complexas serão recompensadas com até 20 por cento de nota extra.

Grupos

Os trabalhos podem ser feitos de grupos de até 2 pessoas. A definição dos componentes do grupo deve ser comunicada ao professor até a aula do dia 28/11.

Apresentação

Os trabalhos serão apresentados na aula do dia 12/12. Antes desta data será disponibilizado um link para os horários das apresentações dos respectivos grupos. Cada grupo terá aproximadamente 5 minutos para apresentar o trabalho. As seguintes instruções devem ser seguidas:

- grupo deve chegar 10 minutos antes do horário da apresentação

- acessar o computador e ficar pronto para demonstrar como resolveu cada tarefa (explicar decisões de implementação), e executar elas

- integrante não presente recebe nota 0

- enviar código fonte em um arquivo .zip no moodle