INF01124 - Classificação e Pesquisa de Dados - Trabalho Final
Professor João Comba

Neste trabalho aplicamos diversas técnicas vistas em aula para explorar o dataset MovieLens 20M. Estes dados foram disponibilizados publicamente[1] e a partir deles foram gerados os conjuntos de dados disponíveis para este trabalho. O enunciado do trabalho inicia com a descrição dos dados, seguido das tarefas solicitadas.

1 Dados

Os dados são compostos de três arquivos, `movies.csv`, `ratings.csv` e `tags.csv` contendo respectivamente informações sobre filmes, avaliações de usuários e anotações em texto-livre (tags). O arquivo movies.csv contém informações de 27.256 filmes, composto de um movieId, título, gêneros e ano. A Figura 1 ilustra este arquivo.

O arquivo `ratings.csv` contém 20,000,263 avaliações (notas entre 1 e 5) de usuários para filmes. Também disponibilizamos um arquivo com 10,000 avaliações para ajudar nos testes (`miniratings.csv`). A leitura dos dados a partir do CSV pode demorar, em especial para o arquivo `ratings.csv` que possui mais de 520MB de dados. É permitido usar código externo para leitura eficiente de arquivos CSV. Exemplos de bibilotecas para a leitura rápida de arquivos CSV são disponibilizados no Moodle para as linguagens C e C++. Em Python pode-se usar Pandas para ler o arquivo CSV.

O arquivo `tags.csv` contém 465,548 anotações de texto livre (tags) (ex.: dark hero, noir thriller, sad, happy, feel-good, etc) para os 26,744 filmes (Figura 2 (abaixo)).

[1] `https://grouplens.org/datasets/movielens/`

	movieId	title	genres	year
0	1	Toy Story	Adventure|Animation|Children|Comedy|Fantasy	1995
1	2	Jumanji	Adventure|Children|Fantasy	1995
2	3	Grumpier Old Men	Comedy|Romance	1995
3	4	Waiting to Exhale	Comedy|Drama|Romance	1995
4	5	Father of the Bride Part II	Comedy	1995
...	...	...	...	...
27251	131254	Kein Bund für's Leben	Comedy	2007
27252	131256	Feuer, Eis & Dosenbier	Comedy	2002
27253	131258	The Pirates	Adventure	2014
27254	131260	Rentun Ruusu	(no genres listed)	2001
27255	131262	Innocence	Adventure|Fantasy|Horror	2014

Figura 1: Arquivo movies.csv: diversos campos descrevendo 18.256 filmes.

ratings.csv

	userId	movieId	rating	date
0	1	2	3.5	02-04-2005
1	1	29	3.5	02-04-2005
2	1	32	3.5	02-04-2005
3	1	47	3.5	02-04-2005
4	1	50	3.5	02-04-2005
...	...	...	...	...
20000258	138493	68954	4.5	13-11-2009
20000259	138493	69526	4.5	03-12-2009
20000260	138493	69644	3.0	07-12-2009
20000261	138493	70286	5.0	13-11-2009
20000262	138493	71619	2.5	17-10-2009

tags.csv

	userId	movieId	tag	timestamp
0	18	4141	Mark Waters	1240597180
1	65	208	dark hero	1368150078
2	65	353	dark hero	1368150079
3	65	521	noir thriller	1368149983
4	65	592	dark hero	1368150078
...	...	...	...	...
465543	138446	55999	dragged	1358983772
465544	138446	55999	Jason Bateman	1358983778
465545	138446	55999	quirky	1358983778
465546	138446	55999	sad	1358983772
465547	138472	923	rise to power	1194037967

Figura 2: Arquivo ratings.csv (acima): 20,000,263 avaliações (notas entre 1 e 5) de usuários para filmes. Arquivo tags.csv (abaixo): 465,548 anotações de texto livre (tags)

2 Criando Estruturas de Dados de Pesquisa

Esta seção descreve estruturas que devem ser construídas em pré-processamento, para suportar as consultas interativas.

2.1 Estrutura 1: Armazenando Dados Sobre filmes

Uma tabela Hash deve ser construída para armazenar as informações associadas aos filmes. A chave de acesso desta tabela Hash é o id do filme, e os dados satélites correspondem aos dados adicionais presentes no arquivo movies.csv descrito anteriormente somadas às informações de revisões de usuários sobre filmes do arquivo. Estas informações adicionais precisam ser calculadas. Por exemplo, o filme de id 3578 (Gladiator) recebeu 32878 revisões no arquivo ratings.csv. Para saber a média global das avaliações (de todos os usuários), é necessário ler e calcular a média de todos as avaliações para cada filme. Uma forma de fazer isso é adicionar nos dados satélites do filme um contador que armazena o número de revisões e um campo que contém a soma das notas de todas as revisões. Após processar o arquivo de revisões, basta dividir, para cada filme, esta soma pelo total de revisões atribuídas a esse filme. Por exemplo, a média global do Gladiator é 3.952248.

2.2 Estrutura 2: Estrutura para buscas por strings de nomes

Uma das consultas que iremos solicitar refere-se a uma busca por prefixos de nomes de filmes. Para suportar esta consulta, é solicitada a construção de uma árvore que suporta consultas de prefixos em strings (TRIE, RADIX TREE ou TST). A estrutura escolhida deve ser construída para armazenar os nomes dos filmes. Ao incluir um nome nessa estrutura, o identificador que sinaliza o final do string deve ser o id do filme. As consultas por prefixos devem portanto saber percorrer a estrutura implementada e retornar a lista de IDs de filmes que satisfazem a consulta. Todos os nomes presentes no arquivo movies.csv devem ser incluídos nessa estrutura.

2.3 Estrutura 3: Estrutura para guardar revisões de usuários

As avaliações descrevem as notas atribuídas para filmes por cada usuário. Para poder responder perguntas sobre quais filmes um usuário avaliou é preciso criar uma estrutura de dados que retorne, para um dado usuário, quais filmes foram avaliadas por este usuário e qual as notas que este atribui. A escolha de qual estrutura utilizar para guardar dados de usuários é livre.

2.4 Estrutura 4: Estrutura para guardar tags

Os usuários também atribuem comentários em texto livre sobre filmes no arquivo tags.csv. A estrutura que precisa ser construída deve suportar consultas por um string contendo uma tag, e retornar a lista de filmes que foram atribuídos esta tag. A escolha de qual estrutura utilizar para guardar dados de tags é livre.

3 Pesquisas

O objetivo do trabalho é implementar estruturas de dados e algoritmos que suportam pesquisas sobre os dados:

3.1 Pesquisa 1: prefixos de nomes de filmes

Esta pesquisa tem por objetivo retornar a lista de filmes cujo **nome** do filme começa com um string passado como parâmetro. Todos os filmes que satisfizerem o string de consulta devem ser retornados, um por linha, contendo o id do filme, o nome, a lista de gêneros dos filmes, avaliação média global e contagem de avaliações. **O resultado da consulta deve ser deve ser ordenado em ordem decrescente da nota global do filme, e a nota global de avaliação deve usar 6 casas decimais. Além disso, o resultado da consulta deve ser deve ser impresso compacto e organizado em colunas tabuladas**.

A sintaxe dessa consulta é $prefixo < string prefixo >$. Um exemplo da consulta *prefixo America* é dado na Figura 3.

movieId	title	genres	year	rating	count
102107	American Winter	Documentary|Drama	2013	4.250000	2
2858	American Beauty	Comedy|Drama	1999	4.155934	44987
2329	American History X	Crime|Drama	1998	4.151208	23309
108316	American Scary	Comedy|Documentary|Horror	2006	4.000000	1
105155	America the Beautiful	Documentary	2007	4.000000	1
66389	AmericanEast	Drama	2008	4.000000	1
26962	American Perfekt	Crime|Drama|Thriller	1997	4.000000	3
3007	American Movie	Documentary	1999	3.911550	2052
1169	American Dream	Documentary	1990	3.857372	312
55765	American Gangster	Crime|Drama|Thriller	2007	3.825420	5055
6620	American Splendor	Comedy|Drama	2003	3.809135	3678
900	American in Paris, An	Musical|Romance	1951	3.806792	3033
3363	American Graffiti	Comedy|Drama	1973	3.803874	7485
86290	American: The Bill Hicks Story	Comedy|Documentary	2009	3.763959	197
104827	American Vagabond	Documentary	2013	3.750000	2
6021	American Friend, The (Amerikanische Freund, Der)	Crime|Drama|Mystery|Thriller	1977	3.723140	242
59418	American Crime, An	Crime	2007	3.686528	193
106916	American Hustle	Crime|Drama	2013	3.670559	1199
11	American President, The	Comedy|Drama|Romance	1995	3.667713	18162
8190	Americanization of Emily, The	Comedy|Drama|War	1964	3.583942	137

Figura 3: Exemplo de resultado da consulta 1

A consulta deve ser feita diretamente pelo console (ou interface gráfica), e o resultado também deve ser impresso no console. Para responder esta pesquisa, deve-se consultar a árvore de pesquisa em strings para buscar todos os identificadores de filmes que correspondem ao string da consulta. Com essa lista de identificadores, pode-se buscar na tabela hash as informações complementares dos filmes.

3.2 Pesquisa 2: filmes revisados por usuários

Esta pesquisa deve retornar a lista com no máximo 20 filmes revisados pelo usuário e para cada filme mostrar a nota dada pelo usuário, a média global e a contagem de avaliações. **O resultado da consulta deve ser ordenado em ordem decrescente da nota atribuído pelo usuário (ordenação primária) e pela nota global do filme (ordenação secundária). Além disso, o resultado da consulta deve ser impresso compacto e organizado em colunas tabuladas.**

A sintaxe dessa consulta é: *user* < *userID* >. Um exemplo da consulta *user* 54766 é dado na Figura 4.

movieId	title	genres	year	global_rating	count	rating
318	Shawshank Redemption, The	Crime|Drama	1994	4.446990	63366	5.0
2571	Matrix, The	Action|Sci-Fi|Thriller	1999	4.187186	51334	5.0
593	Silence of the Lambs, The	Crime|Horror|Thriller	1991	4.177057	63299	5.0
2028	Saving Private Ryan	Action|Drama|War	1998	4.064417	37110	5.0
47	Seven (a.k.a. Se7en)	Mystery|Thriller	1995	4.053493	43249	5.0
3996	Crouching Tiger, Hidden Dragon (Wo hu cang long)	Action|Drama|Romance	2000	3.903248	25090	5.0
4027	O Brother, Where Art Thou?	Adventure|Comedy|Crime	2000	3.891130	19459	5.0
1271	Fried Green Tomatoes	Comedy|Crime|Drama	1991	3.738160	8868	5.0
2291	Edward Scissorhands	Drama|Fantasy|Romance	1990	3.727096	21394	5.0
1527	Fifth Element, The	Action|Adventure|Comedy|Sci-Fi	1997	3.714479	27660	5.0
733	Rock, The	Action|Adventure|Thriller	1996	3.679536	31353	5.0
480	Jurassic Park	Action|Adventure|Sci-Fi|Thriller	1993	3.664741	59715	5.0
471	Hudsucker Proxy, The	Comedy	1994	3.664182	11268	5.0
2700	South Park: Bigger, Longer and Uncut	Animation|Comedy|Musical	1999	3.626389	17371	5.0
2355	Bug's Life, A	Adventure|Animation|Children|Comedy	1998	3.610603	20465	5.0
3070	Adventures of Buckaroo Banzai Across the 8th D...	Adventure|Comedy|Sci-Fi	1984	3.368302	3770	5.0
50	Usual Suspects, The	Crime|Mystery|Thriller	1995	4.334372	47006	4.0
527	Schindler's List	Drama|War	1993	4.310175	50054	4.0
1617	L.A. Confidential	Crime|Film-Noir|Mystery|Thriller	1997	4.083377	26836	4.0
1214	Alien	Horror|Sci-Fi	1979	4.041784	30933	4.0

Figura 4: Exemplo de resultado da consulta 2

3.3 Pesquisa 3: melhores filmes de uma determinado gênero

Esta pesquisa tem por objetivo retornar a lista de filmes com melhores notas de um dado gênero. O resultado da pesquisa identifica os flmes que contém no campo gênero o parâmetro de pesquisa informado. Para evitar que um filme seja retornado com uma boa média mas com poucas avaliações, esta consulta somente deve retornar os melhores filmes com no mínimo 1000 avaliações. Para gerenciar o número de filmes a serem retornados, a consulta deve receber como parâmetro um número N que corresponde ao número máximo de filmes a serem retornados. **O resultado da consulta deve ser ordenado em ordem decrescente da nota global do filme, e a nota global de avaliação deve usar 6 casas decimais. Além disso, o resultado da consulta deve ser impresso compacto e organizado em colunas tabuladas.**

A sintaxe dessa consulta é: $top < N >< genero >$. Um exemplo da consulta $top20\ 'Mystery'$ é dado na Figura 5.

movieId	title	genres	year	rating	count
50	Usual Suspects, The	Crime|Mystery|Thriller	1995	4.334372	47006
904	Rear Window	Mystery|Thriller	1954	4.271334	17449
1212	Third Man, The	Film-Noir|Mystery|Thriller	1949	4.246002	6565
908	North by Northwest	Action|Adventure|Mystery|Romance|Thriller	1959	4.233538	15627
1284	Big Sleep, The	Crime|Film-Noir|Mystery	1946	4.207361	5529
1252	Chinatown	Crime|Film-Noir|Mystery|Thriller	1974	4.199673	15310
913	Maltese Falcon, The	Film-Noir|Mystery	1941	4.187212	12144
4226	Memento	Mystery|Thriller	2000	4.178547	30443
5291	Rashomon (Rashômon)	Crime|Drama|Mystery	1950	4.176724	3712
79132	Inception	Action|Crime|Drama|Mystery|Sci-Fi|Thriller|IMAX	2010	4.156172	14023
903	Vertigo	Drama|Mystery|Romance|Thriller	1958	4.148006	14094
5008	Witness for the Prosecution	Drama|Mystery|Thriller	1957	4.146605	1620
1131	Jean de Florette	Drama|Mystery	1986	4.139051	3247
923	Citizen Kane	Drama|Mystery	1941	4.130443	17774
928	Rebecca	Drama|Mystery|Romance|Thriller	1940	4.115972	4652
942	Laura	Crime|Film-Noir|Mystery	1944	4.111321	2641
27773	Old Boy	Mystery|Thriller	2003	4.091026	6207
1089	Reservoir Dogs	Crime|Mystery|Thriller	1992	4.089361	27635
3730	Conversation, The	Drama|Mystery	1974	4.089212	3856
2208	Lady Vanishes, The	Drama|Mystery|Thriller	1938	4.086521	2196

Figura 5: Exemplo de resultado da consulta 3

3.4 Pesquisa 4: tags atribuídas por usuários para filmes

Esta pesquisa tem por objetivo explorar a lista de tags adicionadas por cada usuário em cada revisão. Para dois strings contendo tags informados na entrada, a pesquisa deve retornar a lista de filmes que estão associados a interseção de um conjunto de tags. **O resultado da consulta deve ser ordenado em ordem decrescente da nota global do filme, e a nota global de avaliação deve usar 6 casas decimais. Além disso, o resultado da consulta deve ser impresso compacto e organizado em colunas tabuladas.**

A sintaxe dessa consulta é: $tags < list\ of\ tags >$. Um exemplo da consulta $tags$ '$feelgood$' '$predicatable$ é dado na Figura 6. Como as tags podem ser termos com espaço (ex.: 'feel good', 'dark hero'), a tag passada na consulta deve ser escrita entre apóstrofes.

movieId	title	genres	year	rating	count
81845	King's Speech, The	Drama	2010	4.022206	5629
63082	Slumdog Millionaire	Crime|Drama|Romance	2008	3.950478	9208
72641	Blind Side, The	Drama	2009	3.865756	2123
88129	Drive	Crime|Drama|Film-Noir|Thriller	2011	3.844993	2916
5377	About a Boy	Comedy|Drama|Romance	2002	3.697639	9401
44709	Akeelah and the Bee	Drama	2006	3.676505	864
6218	Bend It Like Beckham	Comedy|Drama|Romance	2002	3.572797	7308
339	While You Were Sleeping	Comedy|Romance	1995	3.500345	23214
64969	Yes Man	Comedy	2008	3.456522	1863
91653	We Bought a Zoo	Comedy|Drama	2011	3.419444	360
90249	Real Steel	Action|Drama|Sci-Fi|IMAX	2011	3.310160	561
91873	Joyful Noise	Comedy|Musical	2012	1.888889	18

Figura 6: Exemplo de resultado da consulta 4

4 Implementação

Os usuários devem construir uma aplicação que funciona em duas fases. A primeira fase corresponde a construção e inicialização das estruturas de dados necessárias para suportar as consultas. 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 5% na nota final**. Após as estruturas serem construídas, a aplicação entra na segunda fase, que corresponde ao modo console. Nesta fase será possível fazer as pesquisas listadas na seção anterior.

É possível fazer o trabalho em C, C++, Python e Java, ou outras linguagens. Não é permitido usar bibliotecas ou mecanismos da linguagem de alto nível, nem implementações prontas para lidar, buscar ou armazenar os dados (dicionários, maps, bancos de dados). Todas as estruturas citadas anteriormente, buscas e ordenações devem ser implementadas pelo aluno. Não é permitido abrir os arquivos após a fase de construção e inicialização das estruturas.

Bônus: Interfaces gráficas e consultas novas serão recompensadas com até 20% na nota final.

5 Apresentação do Trabalho Final

Os trabalhos podem ser feitos de grupos de até 2 pessoas. A definição dos componentes do grupo deve ser comunicada ao professor, bem como o horário da apresentação. Cada grupo terá aproximadamente 5 minutos para apresentar o trabalho. As seguintes instruções devem ser seguidas:

- cada grupo deve estar disponível 10 minutos antes do horário da apresentação;
- antes da apresentação iniciar, a aplicação deve ter construído as estruturas de dados de suporte e estar pronta para responder as pesquisas;
- o grupo deve relatar brevemente as seguintes informações no começo da apresentação: tempo de construção das estruturas de dados, e explicação das estruturas de dados usadas para cada uma das quatro consultas acima;
- cada integrante deve estar apto para demonstrar como resolveu cada tarefa (explicar decisões de implementação), integrante não presente recebe nota 0.

6 Entrega

A solução deve ser enviada pelo Moodle dentro de um arquivo .zip, contendo os seguintes arquivos:

- integrantes.txt: coloque o nome dos integrantes do grupo (até 2 pessoas) , com um nome por linha
- código fonte correspondente a solução