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

Neste trabalho aplicamos diversas técnicas vistas em aula para explorar o dataset FIFA21 - Players. Estes dados foram disponibilizados no Kaggle[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, seguida das tarefas solicitadas.

1 Dados

Os dados são compostos de três arquivos, `players.csv`, `rating.csv` e `tags.csv` contendo respectivamente informações sobre jogadores, avaliações de usuários e anotações em texto-livre (tags). O arquivo players.csv contém informações de 18.944 jogadores, composto de um FIFA id, nome (curto e longo), lista de posições, nacionalidade, nome do clube e nome da liga. A Figura 1 ilustra o conteúdo deste arquivo.

	sofifa_id	short_name	long_name	player_positions	nationality	club_name	league_name
0	158023	L. Messi	Lionel Andres Messi Cuccittini	RW, ST, CF	Argentina	FC Barcelona	Spain Primera Division
1	20801	Cristiano Ronaldo	Cristiano Ronaldo dos Santos Aveiro	ST, LW	Portugal	Juventus	Italian Serie A
2	200389	J. Oblak	Jan Oblak	GK	Slovenia	Atlético Madrid	Spain Primera Division
3	188545	R. Lewandowski	Robert Lewandowski	ST	Poland	FC Bayern München	German 1. Bundesliga
4	190871	Neymar Jr	Neymar da Silva Santos Junior	LW, CAM	Brazil	Paris Saint-Germain	French Ligue 1
...	...	...	...	...	...	...	...
18939	256679	K. Angulo	Kevin Angulo	CM	Colombia	América de Cali	Colombian Liga Postobón
18940	257710	Zhang Mengxuan	Mengxuan Zhang	CB	China PR	Chongqing Dangdai Lifan FC SWM Team	Chinese Super League
18941	250989	Wang Zhenghao	Wang Zheng Hao	CB	China PR	Tianjin TEDA FC	Chinese Super League
18942	257697	Chen Zitong	Zitong Chen	CM	China PR	Shijiazhuang Ever Bright F.C.	Chinese Super League
18943	257936	Song Yue	Yue Song	CM	China PR	Tianjin TEDA FC	Chinese Super League

18944 rows × 7 columns

Figura 1: Arquivo players.csv

O arquivo `rating.csv` contém 24,188,078 de avaliações (notas entre 1 e 5) de usuários para jogadores. Esses dados foram simulados por avaliações de usuários para cada jogador (Figura 2). Também disponibilizamos um arquivo com 10,000 avaliações ao invés das 24 milhões para ajudar nos testes (`minirating.csv`). A leitura dos dados a partir do CSV pode demorar, em especial para o arquivo `rating.csv` que possui mais de 400MB 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++.

O arquivo `tags.csv` contém 362,700 anotações de texto livre (tags) (ex.: Brazil, FK Specialist, Speedster, Playmaker, Paris Saint-Germain) para 18,944 jogadores (Figura 3).

2 Criando Estruturas de Dados de Pesquisa

2.1 Estrutura 1: Armazenando Dados Sobre Jogadores

Uma tabela Hash deve ser construída para armazenar as informações associadas aos jogadores. A chave de acesso desta tabela Hash é o id do jogador, e os dados satélites correspondem aos dados adicionais presentes no arquivo

[1] `https://www.kaggle.com/stefanoleone992/fifa-21-complete-player-dataset`

	user_id	sofifa_id	rating
0	52505	158023	4.0
1	54989	158023	5.0
2	5409	158023	4.5
3	126061	158023	5.0
4	2782	158023	4.0
...	...	...	...
24188073	21795	257936	1.0
24188074	54766	257936	2.5
24188075	40824	257936	1.5
24188076	134921	257936	1.5
24188077	9005	257936	2.5

Figura 2: Arquivo ratings.csv

	user_id	sofifa_id	tag
0	17800	158023	Clinical Finisher
1	17800	158023	Complete Forward
2	17800	158023	Dribbler
3	17800	158023	Distance Shooter
4	17800	158023	FK Specialist
...	...	...	...
364945	28151	257936	Tianjin TEDA FC
364946	28151	257936	Chinese Super League
364947	110052	257936	China PR
364948	110052	257936	Tianjin TEDA FC
364949	110052	257936	Chinese Super League

Figura 3: Arquivo tags.csv

players.csv descrito anteriormente somadas às informações de revisões de jogadores sobre jogadores do arquivo. Estas informações adicionais precisam ser calculadas. Por exemplo, o jogador 158023 (L. Messi) recebeu várias revisões no arquivo rating.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 jogador. Uma forma de fazer isso é adicionar nos dados satélites do jogador 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 jogador, esta soma pelo total de revisões atribuídas a esse jogador. Por exemplo, a média global do L. Messi é 4.256382. A construção desta tabela Hash deve ser feita em pré-processamento.

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 jogadores. 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 longos de todos os jogadores. Ao incluir um nome nessa estrutura, o identificador que sinaliza o final do string deve ser o id do jogador. As consultas por prefixos devem portanto saber percorrer a estrutura implementada e retornar a lista de IDs de jogadores que satisfazem a consulta. Todos os nomes longos presentes no arquivo players.csv devem ser incluídos nessa estrutura.

2.3 Estrutura 3: Estrutura para guardar revisões de usuários

As avaliações descrevem as notas atribuídas para jogadores por cada usuário. Para poder responder perguntas sobre quais jogadores um usuário avaliou é preciso criar uma estrutura de dados que retorne, para um dado usuário, quais jogadores 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 jogadores no arquivo tags.csv. A estrutura que precisa ser construída deve suportar consultas por um string contendo uma tag, e retornar a lista de jogadores 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 jogadores

Esta pesquisa tem por objetivo retornar a lista de jogadores cujo **nome longo** do jogador começa com um string passado como parâmetro. Todos os jogadores que satisfizerem o string de consulta devem ser retornados, um por linha, contendo o id do jogador, o nome curto, o nome longo, a lista de posições dos jogadores, avaliação média global e número de avaliações. **O resultado da consulta deve ser deve ser ordenado em ordem decrescente da nota global do jogador, 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: $player < nameorprefix >$. Um exemplo da consulta $player\ \ Fer$ é dado na Figura 4.

sofifa_id	short_name	long_name	player_positions	rating	count
135507	Fernandinho	Fernando Luiz Rosa	CB, CDM	3.720244	6402
162131	Llorente	Fernando Javier Llorente Torres	ST	3.501560	9935
165517	F. Gago	Fernando Ruben Gago	CDM, CM, CB	3.440452	4467
184134	Fernando	Fernando Francisco Reges	CDM, CM	3.421439	4977
207707	Marcal	Fernando Marcal de Oliveira	LB, CB	3.214713	4513
...	...	...	...	...	...
258882	Fernan Lopez	Fernan Ferreiroa Lopez	CAM, CM	2.347139	664
256977	F. Saglam	Ferhat Saglam	RW, ST	2.304565	1599
255803	F. Souza	Fernando Esekiel Souza Piriz	CB	2.289773	440
254765	F. Martinez	Fernando Martinez Rojas	CM	2.278153	444
255808	A. Alfaro	Fernando Agustin Alfaro Bares	LM, LW	2.207756	361

Figura 4: 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 jogadores que correspondem ao string da consulta. Com essa lista de identificadores, buscar na tabela hash as informações complementares dos jogadores.

3.2 Pesquisa 2: jogadores revisados por usuários

Esta pesquisa deve retornar a lista com no máximo 20 jogadores revisados pelo usuário e para cada jogador 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 jogador (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* 106180 é dado na Figura 5.

sofifa_id	short_name	long_name	global_rating	count	rating
188545	R. Lewandowski	Robert Lewandowski	4.081675	7989	5.0
176676	Marcelo	Marcelo Vieira da Silva Junior	3.929943	11562	5.0
213565	T. Lemar	Thomas Lemar	3.704151	8768	5.0
207862	M. Ginter	Matthias Ginter	3.691176	8296	5.0
199482	A. Lopes	Anthony Lopes	3.438106	2270	5.0
158963	L. Biglia	Lucas Rodrigo Biglia	3.529665	6034	4.5
221491	N. Elvedi	Nico Elvedi	3.354515	4839	4.5
198946	D. D'Ambrosio	Danilo D'Ambrosio	3.227721	2334	4.5
201982	J. Schmid	Jonathan Schmid	3.225020	5044	4.5
194963	Andre	Andre Felipe Ribeiro de Souza	3.023390	5964	4.5
247106	A. Bernabei	Alexandro Ezequiel Bernabei	2.828586	1966	4.5
198710	J. Rodriguez	James David Rodriguez Rubio	3.893234	13347	4.0
197655	S. Coates	Sebastian Coates Nion	3.641980	8554	4.0
153260	Hilton	Vitorino Hilton da Silva	3.250181	4147	4.0
182896	R. Rosales	Roberto Jose Rosales Altuve	3.209732	4737	4.0
219536	I. Pussetto	Ignacio Pussetto	2.896732	1530	4.0
240753	A. Gouiri	Amine Gouiri	2.853595	765	4.0
226285	R. Gudino	Raul Manolo Gudino Vega	2.849102	835	4.0
241487	J. Ferreira	Jesus Ferreira	2.797297	1147	4.0
226401	K. Dowell	Kieran Dowell	2.761965	1588	4.0

Figura 5: Exemplo de resultado da consulta 2

3.3 Pesquisa 3: melhores jogadores de uma determinada posição

Esta pesquisa tem por objetivo retornar a lista de jogadores com melhores notas de uma dada posição. Para evitar que um jogador seja retornado com uma boa média mas com poucas avaliações, esta consulta somente deve retornar

os melhores jogadores com no mínimo 1000 avaliações. Para gerenciar o número de jogadores a serem retornados, a consulta deve receber como parâmetro um número N que corresponde ao número máximo de jogadores a serem retornados. **O resultado da consulta deve ser ordenado em ordem decrescente da nota global do jogador, 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: $top < N >< position >$. Um exemplo da consulta $top10\ 'ST'$ é dado na Figura 6.

sofifa_id	short_name	long_name	player_positions	nationality	club_name	league_name	rating	count
158023	L. Messi	Lionel Andres Messi Cuccittini	RW, ST, CF	Argentina	FC Barcelona	Spain Primera Division	4.256382	12887
20801	Cristiano Ronaldo	Cristiano Ronaldo dos Santos Aveiro	ST, LW	Portugal	Juventus	Italian Serie A	4.247577	11144
176580	L. Suarez	Luis Alberto Suarez Diaz	ST	Uruguay	FC Barcelona	Spain Primera Division	4.174944	9403
41236	Z. Ibrahimovic	Zlatan Ibrahimovic	ST	Sweden	Milan	Italian Serie A	4.105496	9389
188545	R. Lewandowski	Robert Lewandowski	ST	Poland	FC Bayern München	German 1. Bundesliga	4.081675	7989
153079	S. Aguero	Sergio Leonel Aguero del Castillo	ST	Argentina	Manchester City	English Premier League	4.052894	11249
165153	K. Benzema	Karim Benzema	CF, ST	France	Real Madrid	Spain Primera Division	4.036720	10471
183277	E. Hazard	Eden Hazard	LW, ST	Belgium	Real Madrid	Spain Primera Division	4.014172	9314
194765	A. Griezmann	Antoine Griezmann	ST, CF, LW	France	FC Barcelona	Spain Primera Division	4.006960	13937
167664	G. Higuain	Gonzalo Gerardo Higuain	ST	Argentina	Inter Miami	USA Major League Soccer	3.918668	7285

Figura 6: Exemplo de resultado da consulta 3

3.4 Pesquisa 4: prefixos de nomes de jogadores

Esta pesquisa tem por objetivo explorar a lista de tags adicionadas por cada usuário em cada revisão. Para uma lista de tags dada como entrada, a pesquisa deve retornar a lista de jogadores que estão associados a interseção de um conjunto de tags. **O resultado da consulta deve ser ordenado em ordem decrescente da nota global do jogador, 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 $tags\ 'Brazil'\ 'Dribbler'$ é dado na Figura 7. Como as tags podem ser termos com espaço (ex.: Solid Player, French Ligue 1, Manchester United), a tag passada na consulta deve ser escrita entre apóstrofes.

4 Implementação

Os usuários devem construir uma aplicação que funciona em duas fases. A fase 1 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**.

Depois dessas estruturas serem construídas, a aplicação entra na fase 2, que corresponde ao modo console. Nesta fase será possível fazer as pesquisas listadas na seção anterior, bem como a exibição dos resultados.

O arquivo players_21.csv está incluso entre os conjuntos de dados e trata-se do conjunto original. Neste temos diversas informações que podem ser utilizadas para melhorar as listagens como, por exemplo, clube, idade e nacionalidade dos jogadores.

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

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 abrir os arquivos após a fase de construção e inicialização das estruturas.

sofifa_id	short_name	long_name	player_positions	nationality	club_name	league_name	rating	count
190871	Neymar Jr	Neymar da Silva Santos Junior	LW, CAM	Brazil	Paris Saint-Germain	French Ligue 1	4.242804	17370
176676	Marcelo	Marcelo Vieira da Silva Junior	LB	Brazil	Real Madrid	Spain Primera Division	3.929943	11562
189242	Coutinho	Philippe Coutinho Correia	CAM, LW, CM	Brazil	FC Barcelona	Spain Primera Division	3.928441	7106
201942	Roberto Firmino	Roberto Firmino Barbosa de Oliveira	CF	Brazil	Liverpool	English Premier League	3.802751	5488
200949	Lucas Moura	Lucas Rodrigues Moura da Silva	RM, CF	Brazil	Tottenham Hotspur	English Premier League	3.721487	4761
236632	David Neres	David Neres Campos	LW, RW, CAM	Brazil	Ajax	Holland Eredivisie	3.706804	7246
190483	Douglas Costa	Douglas Costa de Souza	LM, RW, LW	Brazil	Juventus	Italian Serie A	3.704957	9663
180403	Willian	Willian Borges da Silva	RW, LW, RM	Brazil	Arsenal	English Premier League	3.678571	4004
201995	Felipe Anderson	Felipe Anderson Pereira Gomes	LM, CAM	Brazil	West Ham United	English Premier League	3.665974	5290
238794	Vinicius Jr.	Vinicius Jose Paixao de Oliveira Junior	LW	Brazil	Real Madrid	Spain Primera Division	3.547206	4385
230658	Arthur	Arthur Henrique Ramos de Oliveira Melo	CM	Brazil	Juventus	Italian Serie A	3.536732	2491
230666	Gabriel Jesus	Gabriel Fernando de Jesus	ST	Brazil	Manchester City	English Premier League	3.476959	4340
188803	Taison	Taison Barcellos Freda	LM, CAM	Brazil	Shakhtar Donetsk	Ukrainian Premier League	3.334329	4518
201400	Rafinha	Rafael Alcantara do Nascimento	CM, CAM, RM	Brazil	FC Barcelona	Spain Primera Division	3.292032	3652
222716	Everton	Everton Sousa Soares	LM	Brazil	SL Benfica	Portuguese Liga ZON SAGRES	3.114625	253
210411	Otavio	Otavio Edmilson da Silva Monteiro	RM, CAM, CM	Brazil	FC Porto	Portuguese Liga ZON SAGRES	2.922907	227
230475	Adilson Edrada	Adilson Patrick Edrada Pereira	CAM	Brazil	Santos	Campeonato Brasileiro Série A	2.888048	661

Figura 7: Exemplo de resultado da consulta 4

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 até a a data especificada no Moodle, 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;
- a aplicação deve já 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 paracada 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