

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 Kaggle¹ e a partir deles foram gerados os conjuntos de dados disponíveis para este trabalho. Os dados dos jogadores foram extraídos do site <https://sofifa.com> e contém dados extraídos do modo carreira do FIFA 15 ao FIFA 21. 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 sofifa id, nome (curto e longo), lista de posições e nacionalidade. A Figura 1 ilustra o conteúdo deste arquivo.

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 (acima)). Também disponibilizamos um arquivo com 10,000 avaliações 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 bibliotecas 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 362,700 anotações de texto livre (tags) (ex.: Brazil, FK Specialist, Speedster, Playmaker, Paris Saint-Germain) para 18,944 jogadores (Figura 2 (abaixo)).

¹<https://www.kaggle.com/stefanoleone992/fifa-21-complete-player-dataset>

	sofifa_id	short_name	long_name	player_positions	nationality
0	158023	L. Messi	Lionel Andres Messi Cuccittini	RW, ST, CF	Argentina
1	20801	Cristiano Ronaldo	Cristiano Ronaldo dos Santos Aveiro	ST, LW	Portugal
2	200389	J. Oblak	Jan Oblak	GK	Slovenia
3	188545	R. Lewandowski	Robert Lewandowski	ST	Poland
4	190871	Neymar Jr	Neymar da Silva Santos Junior	LW, CAM	Brazil
...
18939	256679	K. Angulo	Kevin Angulo	CM	Colombia
18940	257710	Zhang Mengxuan	Mengxuan Zhang	CB	China PR
18941	250989	Wang Zhenghao	Wang Zheng Hao	CB	China PR
18942	257697	Chen Zitong	Zitong Chen	CM	China PR
18943	257936	Song Yue	Yue Song	CM	China PR

Figura 1: Arquivo `players.csv`: diversos campos descrevendo 18.944 jogadores.

ratings.csv			
	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

tags.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 2: Arquivo ratings.csv (acima): 24,188,078 de avaliações (notas entre 1 e 5) de usuários para jogadores. Arquivo tags.csv (abaixo): 364,950 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 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 players.csv descrito anteriormente somadas às informações de revisões de usuários 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.

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 curtos 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 **short_name** 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 deve ser impresso compacto e organizado em colunas tabuladas.

A sintaxe dessa consulta é *prefixo < stringprefixo >*. Um exemplo da consulta *prefixo Pedro* é dado na Figura 3.

sofifa_id	short_name	long_name	player_positions	nationality	rating	count
189505	Pedro	Pedro Eliezer Rodriguez Ledesma	RW, LW	Spain	3.639815	8454
175379	Pedro Leon	Pedro Leon Sanchez Gil	RM, LM	Spain	3.259415	2018
200054	Pedro Obiang	Pedro Mba Obiang Avomo	CDM, CM	Equatorial Guinea	3.207776	2469
243576	Pedro Porro	Pedro Antonio Porro Saucedo	RB, RM	Spain	3.030612	49
240950	Pedro Goncalves	Pedro Antonio Pereira Goncalves	CM	Portugal	3.014358	1393
230824	Pedro Pereira	Pedro Miguel Almeida Lopes Pereira	RB, RWB, RM	Portugal	2.918112	519
200677	Pedro Mendes	Pedro Filipe Teodosio Mendes	CB	Portugal	2.903026	1289
219258	Pedro Henrique	Pedro Henrique Pereira da Silva	CB	Brazil	2.901639	305
238856	Pedro Sa	Pedro Miguel Cunha Sa	CDM, CM	Portugal	2.882920	363
254824	Pedro Mendes	Pedro Manuel Lobo Peixoto Mineiro Mendes	ST	Portugal	2.881075	967
238616	Pedro Neto	Pedro Lomba Neto	LW, CF, RW	Portugal	2.852298	457
224538	Pedro Nuno	Pedro Nuno Fernandes Ferreira	LW, CAM, CM	Portugal	2.820621	1770
157479	Pedro Lopez	Pedro Lopez Munoz	RB	Spain	2.807500	400
251469	Pedro Amaral	Pedro Miguel Gaspar Amaral	LB	Portugal	2.797060	1769
211119	Pedro Santos	Pedro Miguel Martins Santos	CAM, RM, LM	Portugal	2.780439	1002
234930	Pedro Rebocho	Pedro Miguel Braga Rebocho	LB	Portugal	2.754962	655
248572	Pedro Brazao	Brazao Teixeira Pedro David	RW, CAM	Portugal	2.727139	678
236492	Pedro Diaz	Pedro Diaz Fanjul	CDM, CM	Spain	2.716883	385
207929	Pedro Henrique	Pedro Henrique Konzen Medina da Silva	LM, RM	Brazil	2.710774	1383
219181	Pedro Tiba	Pedro Miguel Amorim Pereira Silva	CM, CDM	Portugal	2.697205	1932
246866	Pedro Pelagio	Pedro Henrique Rocha Pelagio	CM, CDM	Portugal	2.677117	1547
257703	Pedro Henrique	Pedro Henrique Alves de Almeida	ST	Brazil	2.654539	1333
193870	Pedro Trigueira	Pedro Jose da Silva Trigueira	GK	Portugal	2.649930	1424
239915	Pedro Martelo	Pedro Alves Correia	ST	Portugal	2.625000	1108
256044	Pedro Amador	Pedro Miguel Santos Amador	LB	Portugal	2.579491	1258
224019	Pedro Chirivella	Pedro Chirivella Burgos	CM, CDM	Spain	2.578261	115
258186	Pedro Simoes	Pedro Miguel Goncalves Simoes	CM	Portugal	2.551957	1482
233132	Pedro Marques	Pedro Pinho Marques	CB	Portugal	2.543191	1285
244619	Pedro Mateus	Pedro Leonardo Goncalves Mateus	GK	Portugal	2.533203	256
227913	Pedro Lopez	Pedro Lopez Rodriguez	CB	Spain	2.508306	301
258365	Pedro Ferreira	Pedro Miguel Dinis Ferreira	CDM, CM	Portugal	2.503236	927
230419	Pedro Sousenha	Pedro Marcio Sousenha Botelho	RM	Brazil	2.482027	1391
252291	Pedro Augusto	Pedro Augusto Borges da Costa	CM, CDM	Brazil	2.468687	495
257575	Pedro Capo	Pedro Luis Capo Payeras	CAM, CM	Spain	2.436709	711

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 jogadores que correspondem ao string da consulta. Com essa lista de identificadores, pode-se 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 30 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 54766* é dado na Figura 4.

sofifa_id	short_name	long_name	global_rating	count	rating
182521	T. Kroos	Toni Kroos	4.022486	10184	5.0
220814	L. Hernandez	Lucas Hernandez Pi	3.736523	8663	5.0
45186	Joaquin	Joaquin Sanchez Rodriguez	3.300786	5090	5.0
210035	Grimaldo	Alejandro Grimaldo Garcia	3.773125	9373	4.5
193348	X. Shaqiri	Xherdan Shaqiri	3.657221	8822	4.5
179846	S. Khedira	Sami Khedira	3.654610	6594	4.5
228941	Andre Silva	Andre Miguel Valente da Silva	3.376036	5066	4.5
173731	G. Bale	Gareth Frank Bale	3.923208	13211	4.0
212814	Joao Mario	Joao Mario Naval da Costa Eduardo	3.638770	5610	4.0
203775	L. Karius	Loris Karius	3.319157	4103	4.0
204277	Roque Mesa	Roque Mesa Quevedo	3.232939	4059	4.0
184789	A. Szalai	Adam Szalai	3.040927	4899	4.0
226380	Hwang Hee Chan	hwanghyican Huang Xi Can	2.965158	1966	4.0
229348	A. Robinson	Antonee Robinson	2.881773	1827	4.0
212715	S. Palacios	Sebastian Alberto Palacios	2.809633	872	4.0
235024	S. Diaz	Sergio Ismael Diaz Velazquez	2.799708	1715	4.0
224271	T. Goiginger	Thomas Goiginger	2.785931	988	4.0
258165	A. Musaba	Anthony Musaba	2.749218	959	4.0
238756	J. Larsen	Jorgen Strand Larsen	2.741071	952	4.0
191488	L. Orban	Lucas Alfonso Orban Alegre	2.736420	1620	4.0
222693	H. Fertoli	Hector Fertoli	2.690200	1551	4.0
239669	J. Levi	Jonathan Levi	2.681441	1568	4.0
234291	L. Vaisanen	Leo Vaisanen	2.642813	1635	4.0
240776	S. Laiton	Sonny Patrick Laiton	2.614407	826	4.0
189885	C. Noone	Craig Noone	2.572354	926	4.0
153244	A. Gignac	Andre-Pierre Gignac	3.661428	9131	3.5
150724	J. Hart	Joe Hart	3.486538	8468	3.5
194209	Y. El Arabi	Youssef El Arabi	3.272260	4571	3.5
228618	F. Mendy	Ferland Mendy	3.186525	1410	3.5
47201	S. Proto	Silvio Proto	3.144547	3677	3.5

Figura 4: 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 *top20 'RWB'* é dado na Figura 5.

sofifa_id		short_name	long_name	player_positions	nationality	rating	count
203747	Hector Bellerin		Hector Bellerin Moruno	RB, RWB	Spain	3.688981	6752
158626	M. Debuchy		Mathieu Debuchy	RB, CB, RWB	France	3.530062	5472
193525	Mario Fernandes		Mario Figueira Fernandes	RWB, RM	Russia	3.371619	5842
203605	P. Kaderabek		Pavel Kaderabek	RWB, RB, RM	Czech Republic	3.336950	3180
197853	S. Aurier		Serge Aurier	RB, RWB	Ivory Coast	3.327930	5675
172962	V. Moses		Victor Moses	RM, RWB	Nigeria	3.244079	5193
216150	D. Zappacosta		Davide Zappacosta	RB, RWB	Italy	3.232441	4257
201982	J. Schmid		Jonathan Schmid	RB, RM, RWB	France	3.225020	5044
194644	Montoya		Martin Montoya Torralbo	RB, RWB	Spain	3.220943	3734
182896	R. Rosales		Roberto Jose Rosales Altuve	RWB, RB	Venezuela	3.209732	4737
197083	D. Caligiuri		Daniel Caligiuri	RM, RWB, RB	Italy	3.206976	4544
201118	Cedric		Cedric Ricardo Alves Soares	RB, RWB	Portugal	3.165056	4656
193470	A. Souquet		Arnaud Souquet	RWB, RB	France	3.157910	5215
229880	A. Wan-Bissaka		Aaron Wan-Bissaka	RB, RWB	England	3.154173	1414
202316	T. Chandler		Timothy Chandler	RM, RWB, RB	United States	3.149570	4182
188155	D. Janmaat		Daryl Janmaat	RB, RWB	Netherlands	3.127154	3830
247204	Emerson	Emerson Aparecido Leite de Souza Junior		RB, RM, RWB	Brazil	3.098384	1733
188135	Juanfran	Juan Francisco Moreno Fuertes		RB, RWB	Spain	3.093109	3265
226166	N. Mukiele		Nordi Mukiele Mulere	RB, CB, RWB	France	3.076042	1440
197786	G. Donati		Giulio Donati	RB, RWB	Italy	3.065950	3533

Figura 5: 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’ ‘Team Player’* é dado na Figura 6. 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.

sofifa_id	short_name	long_name	player_positions	nationality	rating	count
176676	Marcelo	Marcelo Vieira da Silva Junior	LB	Brazil	3.929943	11562
200145	Casemiro	Carlos Henrique Venancio Casimiro	CDM	Brazil	3.828389	5157
212462	Alex Telles	Alex Nicolao Telles	LB	Brazil	3.744148	9142
158625	Dante	Dante Bonfim da Costa Santos	CB	Brazil	3.524409	7907
153260	Hilton	Vitorino Hilton da Silva	CB	Brazil	3.250181	4147
168530	Jardel	Jardel Nivaldo Vieira	CB	Brazil	3.116277	5388
203888	Eric Botteghin	Eric Fernando Botteghin	CB	Brazil	3.074889	4066
230168	Raphaelito Anjos	Raphael William Anjos Rochedo	GK	Brazil	3.025253	1683
230183	Gazzolisco	Gerson Adriano Gutierrez Serra	LB, LWB, LM	Brazil	2.937276	1116
219258	Pedro Henrique	Pedro Henrique Pereira da Silva	CB	Brazil	2.901639	305
195096	Fransergio	Fransergio Rodrigues Barbosa	CM, CAM	Brazil	2.860704	1023
230318	Nelsildo Reis	Nelson Arturo Reis Lopes	GK	Brazil	2.838221	1057
230252	Jorginhson	Jorginho Silas Ruiz Prestes	CDM, CM, CAM	Brazil	2.805867	1517
230336	Danisco Fachini	Daniel Clayton Fachini Lobato	CDM	Brazil	2.804677	727
230224	Juli Freitinho	Julio Leonardo Dourado de Freitas	CB	Brazil	2.790087	686
230230	Eltildo Correia	Eltildo Lucas Correia Pitta	CB, RB, LB	Brazil	2.782532	1414
207733	Filipe Augusto	Filipe Augusto Carvalho Souza	CM, CDM	Brazil	2.779261	1760
230215	Osvaldo Lodeiro	Osvaldo Murilo Lodeiro Ferreira	CB, RB, LB	Brazil	2.771960	1719
230461	Leordinho Paes	Manuel Leonardo Conceicao Paes	CAM, LM	Brazil	2.761261	111
230282	Guto Costinho	Guto Ramon Costinho Ribeiro	CB	Brazil	2.755102	784
230446	Marcos Paneira	Marcos Alam Paneira Almeida	CDM, CM, RM	Brazil	2.688679	53
224655	Rafael Defendi	Rafael Garcia Tonioli Defendi	GK	Brazil	2.639971	693
205340	Marcelo	Marcelo dos Santos Ferreira	CB	Brazil	2.603950	1924
230418	Henrique Jardinel	Henrique Alex Jardinel Zonta	CDM, CM	Brazil	2.513412	1193
230236	Leonardo Freijao	Leonardo Miguel Freijao Jasper	CDM, CM, CAM	Brazil	2.448905	1233

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