ATAD 2019/20

Algoritmos e Tipos Abstratos de Dados

Enunciado do Projeto – Epoca de Recurso

Importação, Normalização e Classificação de Dados

1. Objetivo do Projeto

O IMDB é a mais popular plataforma agregadora de informação relativa à indústria cinematográfica, com informação relativa a filmes, classificações, avaliações, entre outros.

Pretende-se desenvolver um programa em C para extrair informação útil de um ficheiro com dados relativos aos lançamentos de filmes dos últimos anos. Este programa, consiste num interpretador de comandos que o utilizador usa para obter diversos tipos de informação, principalmente informação estatística.

1.1 Representação dos dados em Memória

Cada filme é representado, obrigatoriamente, pelo tipo de dados `Movie` apresentada na Figura 1.

Cada pontuação é representada, obrigatoriamente, pelo tipo de dados `Rating` apresentada no quadro abaixo.

```
typedef struct movie {
    char id[10];        // Id format: CCNNNNNNN
    char title[100];    // Movie title
    int year;           // Year production
    char genre[10];     // Main genre
    int duration;       // Duration in minutes
    char country[40];   // Country production
    char director[50];  // Movie director
 } Movie;

typedef struct rating {
    char movieId[9];    // Movie ID
    int votes[10];      // Votes by scale
    int maleVotes;      // Total of male votes
    int femaleVotes;    // Total of female votes
    double score;       // Weighted average
} Rating;
```

Figura 1 – Tipos de dados

Considerações importantes:

- É comum determinado filme pertencer a várias categorias (géneros). No entanto, qualquer filme tem uma categoria predominante, sendo que as mesmas aparecem por ordem de significância. Deve por isso, ser considerada a primeira categoria que surgir no respetivo campo.
- O mesmo princípio descrito acima, aplica-se ao atributo *country*.

- Um filme pode ser realizado por uma equipa de realização composta por um realizador e diversos co-realizadores. O realizador, é sempre o primeiro nome que consta na listagem do respetivo campo, sendo este o que deve ser tido em consideração.
- Dentro de diversas categorias (géneros) existentes na indústria cinematográfica, consideram-se como primárias as seguintes:

```
DRAMA, HORROR, COMEDY, ANIMATION, THRILLER, ACTION, ROMANCE, MUSICAL
```

Sendo estas as categorias a considerar sempre que determinado comando se refira a filtragens por categoria.

Nota:

- Na implementação dos comandos descritos neste enunciado podem definir/utilizar outros tipos de dados auxiliares que achem úteis para a resolução dos problemas.
- **Todas as questões relativas à implementação, cujas especificidades não sejam determinadas neste enunciado, deverão ficar ao critério do programador.**

1.2 Dados de entrada

Existem dois tipos de ficheiro com dados:

- Ficheiro de dados sobre os filmes, com dados relativos aos anos de 2009 a 2018 (inclusive);
- Ficheiros de dados sobre as pontuações, dos respetivos filmes.

Ambos os ficheiros se encontram em formato CSV. E a primeira linha dos ficheiros é uma linha com os cabeçalhos e não contem dados.

Ficheiro dos filmes (cada linha corresponde a informação sobre um filme)

```
<id>;<title>;<year>;<genre>;<duration>;<country>;<director>
...
```

Ficheiro com os dados das pontuações (cada linha corresponde a informação sobre uma pontuação)

```
<movie_id>;<males_votes>;<females_votes>;<votes_10>;<votes_9>;<votes_8>;<votes_7>;<votes_6>;<votes_5>;<votes_4>;
<votes_3>;<votes_2>;<votes_1>
...
```

Deve-se assumir que não existem ficheiros "mal-formados" e que todos os campos se encontram preenchidos.

Juntamente com este enunciado são disponibilizados 2 ficheiros de entrada para testes:

- movies.csv
- ratings.csv

Após a divulgação do enunciado será disponibilizado no Moodle um exemplo com **alguns** dos resultados esperados na execução da aplicação para estes ficheiros.

1.3 Utilização de TADs

É obrigatória a manutenção em memória da informação importada:

- dos filmes exclusivamente numa instância do ADT List, sendo ListElem o tipo Movie (definido em 1.1)
- das pontuações exclusivamente numa instância de ADT Map, sendo ValueElem do tipo Rating (definido em 1.1) e o KeyElem de um tipo apropriado que permita guardar uma *string* (*movieID*);

Não é permitido alterar as interfaces lecionadas dos TAD, nomeadamente os ficheiros list.h e map.h. Estas instâncias serão designadas doravante por "coleções".

1.4 Comandos

Há exatamente 13 comandos que o programa deve implementar, que serão apresentados de seguida; 2 comandos para carregamento de dados, 9 comandos para mostrar resultado de cálculos sobre os dados, 1 comando para sair da aplicação e 1 comando para limpeza dos dados em memória.

Notas:

- Cada comando é representado por uma palavra que pode ser escrita pelo utilizador em maiúsculas, minúsculas ou ambos, não importa.
- Sempre que um comando necessitar de algum input, e.g., Id de um filme, este deve ser solicitado ao utilizador.
- Sempre que um comando necessitar de informação que não está carregada, o comando deve indicar que informação está em falta, i.e., "No movie data available..." e/ou "No rating data available...".

A forma exata como os resultados devem ser mostrados no ecrã será descrita em seguida.

Os comandos são os seguintes:

✓ **LOADM**
Pede o nome dum ficheiro de filmes, abre o ficheiro e carrega-o em memória (ver Secção 1.2), mostrando o número de filmes importadas. Os restantes comandos passarão a atuar sobre o novo conteúdo da coleção. Se o ficheiro não puder ser aberto, escreve `File not found` e a coleção fica vazia.

✓ **LOADR**
Abre o ficheiro "ratings.csv" e carrega-o em memória (ver 1.2), mostrando o número de pontuações importadas. Se o ficheiro não puder ser aberto, escreve `File not found` e a coleção respetiva fica vazia. No momento de carregamento, calcule a média ponderada (considerando os pesos de cada voto), arredondando por excesso à primeira casa decimal, da classificação de todos os filmes. O valor calculado deve ser armazenado no atributo *score* do respetivo Rating.

- ✓ **CLEAR**
  Limpa a informação atualmente em memória. Deverá indicar o número de registos que foram descartados, e.g., `"<N> records deleted from <Movies | Ratings>"`

- ✓ **QUIT**
  Sai do programa, libertando toda a memória alocada para as coleções.

- ✓ **TOP5**
  Mostre de forma decrescente os 5 filmes com melhores classificação, por categoria. Cada categoria deve ser apresentada no seguinte formato:

```
<genre>:
1 - id: <id> | title: <title> | year: <year> | genre: <genre> | duration: <duration> | country: <country> | director: <director>
2 - id: <id> | title: <title> | year: <year> | genre: <genre> | duration: <duration> | country: <country> | director: <director>
3 - id: <id> | title: <title> | year: <year> | genre: <genre> | duration: <duration> | country: <country> | director: <director>
4 - id: <id> | title: <title> | year: <year> | genre: <genre> | duration: <duration> | country: <country> | director: <director>
5 - id: <id> | title: <title> | year: <year> | genre: <genre> | duration: <duration> | country: <country> | director: <director>
...
```

  Em caso de empate, os filmes devem ser apresentados por ordem cronológica decrescente.

- ✓ **RATING**
  Apresente uma lista com todos os filmes num intervalo de classificação. **Para tal, deve ser solicitado ao utilizador que introduza o intervalo mínimo e máximo de classificação, numa escala de 0 a 10**. A lista deve ser apresentada por ordem decrescente de classificação, com um número máximo de 25 resultados encontrados. Em caso de igualdade, os filmes devem ser apresentados por ordem alfabética.

```
Score <minimum>-<maximum>:
1 - id: <id> | title: <title> | year: <year> | genre: <genre> | duration: <duration> | country: <country> | director: <director>
2 - id: <id> | title: <title> | year: <year> | genre: <genre> | duration: <duration> | country: <country> | director: <director>
3 - id: <id> | title: <title> | year: <year> | genre: <genre> | duration: <duration> | country: <country> | director: <director>
4 - id: <id> | title: <title> | year: <year> | genre: <genre> | duration: <duration> | country: <country> | director: <director>
5 - id: <id> | title: <title> | year: <year> | genre: <genre> | duration: <duration> | country: <country> | director: <director>
...
```

- ✓ **SAMEDIR**
  Dado um id de um filme, apresente a informação referente ao mesmo. De seguida, liste por ordem cronológica os filmes do mesmo realizador, no seguinte formato:

```
Following Movie: Id: <id> | title: <title> | year: <year> | genre: <genre> | duration: <duration> | country: <country> |
director: <director>

Other movies directed by <director>:
- id: <id> | title: <title> | year: <year> | genre: <genre> | duration: <duration> | country: <country> | director: <director>
- id: <id> | title: <title> | year: <year> | genre: <genre> | duration: <duration> | country: <country> | director: <director>
- id: <id> | title: <title> | year: <year> | genre: <genre> | duration: <duration> | country: <country> | director: <director>
- id: <id> | title: <title> | year: <year> | genre: <genre> | duration: <duration> | country: <country> | director: <director>
...
```

  Caso não exista outros filmes do mesmo realizador:

```
Following Movie: Id: <id> | title: <title> | year: <year> | genre: <genre> | duration: <duration> | country: <country> |
director: <director>

No other films by the same director.
```

- ✓ **VOTERS**
  Mostra a percentagem de votantes, do sexo feminino, masculino ou desconhecido, para cada uma das categorias principais. Para contabilização de sujeitos de sexo desconhecido deve ser considerado o número total de votos existentes.

```
<genre>: Females-<value>%, Males-<value>%, unknown-<value>%, total votes-<value>
<genre>: Females-<value>%, Males-<value>%, unknown-<value>%, total votes-<value>
…
```

- ✓ **COUNTRY**
  Dado um determinado país, apresente o total de filmes produzidos, e a respetiva classificação média:

```
COUNTRY: <country>
TOTAL PRODUCED MOVIES: <total>
AVERAGE SCORE: <average_score>
```

- ✓ **YEAR**
  Para um determinado ano, apresentar a classificação média por género, bem como o número total de filmes desse género. Deve ser mostrado no seguinte formato:

```
<year>
<genre>:
      AVERAGE SCORE: <average_score>
      TOTAL PRODUCED MOVIES: <total>
<genre>:
      AVERAGE SCORE: <average_score>
      TOTAL PRODUCED MOVIES: <total>
<genre>:
      AVERAGE SCORE: <average_score>
      TOTAL PRODUCED MOVIES: <total>
…
```

- ✓ **MATRIX**
  Crie uma matriz 5x8 de inteiros com informação sobre o número total de filmes produzidos no intervalo de anos de 2014 a 2018, por categoria, tal como se ilustra na Figura 2.

	Drama	Horror	Comedy	Animation	Thriller	Action	Romance	Musical
2014	412	246	825	60	53	402	28	3
2015	438	242	769	72	54	409	26	2
2016	447	215	808	73	65	478	26	5
2017	428	249	758	77	72	466	28	5
2018	364	221	738	74	83	452	34	2

Figura 2 - Matrix

- ✓ **PREDICTION**
  Considere agora que se pretende fazer a previsão do número de filmes produzidos por cada uma das principais categorias, para o ano de 2019, tendo por base os dados

relativos aos 5 anos anteriores (2014 a 2018). Para esta previsão, utilize o método de Regressão Linear Simples baseado no método dos Mínimos Quadrados[1].
Seja x_i o valor do ano e y_i o valor do número de filmes produzidos. A fórmula que permite obter y_i em função de x_i é dada pela seguinte equação da reta:

$$y_i = a + bx_i$$

onde **a** e **b** podem ser calculados por:

$$a = \bar{y} - b\bar{x}$$

$$b = \frac{\sum_{i=1}^{n} x_i(y_i - \bar{y})}{\sum_{i=1}^{n} x_i(x_i - \bar{x})}$$

Para calcular o valor de y_i, correspondente a x_i igual a 2019, basta executar a equação da reta anterior, após ter calculado os valores de **a** e **b.**

As previsões devem ser apresentadas no seguinte formato:

```
2019 Predictions:
Drama         - <number_of_movies_predicted>
Horror        - <number_of_movies_predicted>
Comedy        - <number_of_movies_predicted>
Animation     - <number_of_movies_predicted>
Thriller      - <number_of_movies_predicted>
Action        - <number_of_movies_predicted>
Romance       - <number_of_movies_predicted>
Musical       - <number_of_movies_predicted>
```

2 Relatório e Documentação

2.1 Documentação

Todo o código deve ser documentado utilizando a **documentação Doxygen**.

A mesma deve ser gerada para formato HTML e entregue a respetiva pasta "html" junto com o projeto.

2.2 Relatório

No relatório deverão constar as seguintes secções (para além de capa com identificação dos alunos e índice):

a) Descrição breve dos ADTs utilizados, descrição da implementação e justificação da estrutura de dados escolhida para a implementação (comparação de eficiências para o problema de aplicação).

[1] Exemplo de regressão simples:
https://pt.wikipedia.org/wiki/M%C3%A9todo_dos_m%C3%ADnimos_quadrados#Exemplo_de_regress%C3%A3o_simples

b) Para cada comando (exceto CLEAR, SHOW e QUIT) fornecer:
   - A complexidade algorítmica da respetiva implementação, tendo em conta as complexidades algorítmicas das funções dos ADTs utilizadas (dependem da implementação escolhida).
c) Escolha de 2 funcionalidades do tipo B e C, onde apresentam o algoritmo implementado em pseudo-código;
d) Limitações: Quais os comandos que apresentam problemas ou não foram implementados;
e) Conclusões: Análise crítica do trabalho desenvolvido.

3 Tabela de Cotações e Penalizações

A avaliação do trabalho será feita de acordo com os seguintes princípios:

- **Estruturação:** o programa deve estar estruturado de uma forma modular e procedimental;
- **Correção:** o programa deve executar as funcionalidades, tal como pedido.
- **Legibilidade e documentação:** o código deve ser escrito, formatado e comentado de acordo com o standard de programação definido para a disciplina.
- **Desempenho:** Os algoritmos implementados devem ter em conta a complexidade do mesmo, valorizando-se a implementação de algoritmos com menor complexidade. A gestão da memória deverá ser feita corretamente, garantindo que a mesma é libertada quando não está a ser utilizada. Utilização da ferramenta Valgrind, para validar a correta gestão de memória.

A nota final obtida, cuja tabela de cotações se apresenta a seguir, será ponderada de acordo com os princípios acima descritos.

Descrição	Cotação (valores)
Leitura de comandos, tratamento de situação de ficheiro inexistente/vazio , limpeza de memória e saída do programa (QUIT)	2
Importação de dados (comandos LOAD)	2
Comando TOP5	1.5
Comando RATING	1.5
Comando SAMEDIR	2
Comando VOTERS	1
Comando COUNTRY	1
Comando YEAR	2
Comando MATRIX	2
Comando PREDICTION	2
Relatório e Documentação	3
TOTAL	20

A seguinte tabela contém penalizações a aplicar:

Descrição	Penalização
Uso de variáveis globais	até 2
Não separação de funcionalidades em funções/módulos	até 3
Não libertação de memória	até 3
Não comentar o programa	até 1
Não utilização dos ADTs obrigatórios	Anulado

4 Instruções e Regras Finais

O IDE a utilizar fica ao critério dos alunos, mas, caso não utilizem o IDE usado na disciplina (i.e., VS-Code), terão que, **antes de submeter, criar os respetivos projetos finais no IDE VS-Code.**
O não cumprimento das regras a seguir descritas implica uma penalização na nota do trabalho prático. Se ocorrer alguma situação não prevista nas regras a seguir expostas, essa ocorrência deverá ser comunicada ao respetivo docente de laboratório de ATAD.

Regras:

a) O Projeto deverá ser elaborado por no máximo **dois alunos do mesmo docente de laboratório**.

b) Em caso de melhoria de nota, o projeto tem de ser realizado individualmente.

c) A nota do Projeto será atribuída individualmente a cada um dos elementos do grupo após a discussão. As discussões poderão ser orais e/ou com perguntas escritas. As orais poderão ser feitas com todos os elementos do grupo presentes em simultâneo ou individualmente. E poderão ser feitas remotamente via plataforma zoom.

d) A apresentação de relatórios ou implementações plagiadas leva à imediata atribuição de nota zero a todos os trabalhos com semelhanças, quer tenham sido o original ou a cópia.

e) No rosto do relatório e nos ficheiros de implementação deverá constar o número, nome e turma dos autores e o nome do docente a que se destina.

f) O trabalho deverá ser submetido no moodle, no link do respetivo docente de laboratórios criado para o efeito, até às **11:00 do dia 21 de julho:**
   - **Todos os trabalhos submetidos para além do prazo estabelecido terão a penalização de 1 valor por cada hora de atraso**;
   - Para tal terão de criar uma pasta com o nome: **nomeAluno1_númeroAluno1-nomeAluno2_númeroAluno2**, onde colocarão o ficheiro do relatório em formato **pdf** e uma pasta com o projeto VS Code (pasta com os respetivos ficheiros) da implementação das aplicações a desenvolver.
   - Os alunos terão de submeter essa **pasta compactada em formato ZIP**. Apenas será permitido a submissão de um ficheiro.

g) Não serão aceites trabalhos entregues que não cumpram na íntegra os pontos anteriores.

h) As datas das discussões serão publicadas após a entrega dos trabalhos.

(fim de enunciado)