PL 4

Algoritmos Probabilísticos

Secção para avaliação

Considere uma aplicação, a desenvolver em Matlab, com algumas funcionalidades de um sistema de informação para pesquisa de filmes.

Dados de entrada:

Considere o ficheiro movies.csv com uma lista de cerca de 58000 filmes existentes. A primeira coluna do ficheiro contém o nome do filme, a segunda coluna contém o ano do filme e as colunas a partir da terceira identificam os géneros a que o filme pertence (um filme pode pertencer a mais do que um género). NOTA: executando no Matlab a instrução:

movies readcell ('movies.csv', 'Delimiter', ',');

é criado o cell array movies em que a célula movies {i, j} contém a informação da linha i e da coluna j do ficheiro movies.csv.

Descrição da aplicação a desenvolver:

A aplicação deve começar por permitir ao utilizador selecionar uma de 6 opções:

1 - Display available genres

2 - Number of movies of a genre

3 – Number of movies of a genre on a given year

4 – Search movie titles

5 - Search movies based on genres

6 - Exit

Select an option:

Opção 1: A aplicação apresenta a lista de géneros existentes.

Opção 2: A aplicação começa por pedir ao utilizador a seleção de um género:

Select a genre:

garantindo que a palavra introduzida é um género existente (se não for, a aplicação indica erro na introdução dos dados e volta ao menu inicial). Em seguida, a aplicação indica o número (estimado) de filmes classificados no género selecionado.

A execução desta secção será objeto de avaliação. Assim, deverá fazer um relatório em PDF com todos os códigos Matlab desenvolvidos devidamente explicados e as opções de desenvolvimento devidamente justificadas. O relatório deverá começar por identificar o ano letivo, a disciplina, a turma prática e os elementos do grupo (nome e No. Mec.) que realizou o

trabalho. Deverá submeter um ficheiro comprimido com o relatório e todos os ficheiros necessários à execução da aplicação desenvolvida. Tenha em atenção os prazos estipulados

Para a introdução de dados pelo teclado, investigue a utilidade da função Matlab input

Opção 3: A aplicação começa por pedir ao utilizador a seleção de um género e de um ano (separados por uma vírgula) 3:

Select a genre and a year (separated by ','):

garantindo que a palavra e o número introduzidos sejam, respetivamente, um género existente e um número válido (se não forem, a aplicação indica erro na introdução dos dados e volta ao menu inicial). Em seguida, a aplicação indica o número (estimado) de filmes do ano selecionado e classificados no género selecionado.

Opção 4: A aplicação começa por pedir para o utilizador inserir uma string:

Insert a string:

e depois apresenta os nomes dos filmes (um filme por cada linha) com os nomes mais similares à string introduzida. Para cada filme, deve ser também apresentada a (estimativa da) similaridade de Jaccard entre a string introduzida e o nome do filme. A lista apresentada deve conter 5 filmes ordenados por ordem decrescente de similaridade de Jaccard e deve apresentar também em cada linha os géneros do respetivo filme.

Opção 5: A aplicação começa por pedir ao utilizador a seleção de um conjunto de géneros diferentes separados por vírgulas:

Select one or more genres (separated by ','):

garantindo que as palavras introduzidas sejam géneros válidos (se não forem, a aplicação repete o pedido até serem) e depois apresenta os nomes dos filmes (um filme por cada linha) cujos géneros sejam mais similares ao conjunto dos géneros introduzido. Para cada filme, deve ser também apresentada a (estimativa da) similaridade de Jaccard entre o conjunto de géneros introduzido e os géneros do filme. A lista apresentada deve conter 5 filmes ordenados por ordem decrescente de similaridade de Jaccard e, para o mesmo valor de similaridade, por ordem decrescente de ano (dando assim preferência à visualização dos filmes mais recentes) e deve apresentar também em cada linha o ano do respetivo filme.

Opção 6: A aplicação termina.

Notas sobre a implementação das funcionalidades da aplicação a desenvolver:

Nas Opções 2 e 3, a estimativa do número de filmes de um género (Opção 2) e do número de filmes de um género e de um ano (Opção 3) terão de ser implementadas usando um filtro de Bloom com contagem. Os parâmetros do filtro devem ser adequados ao problema e a sua escolha devidamente fundamentada no relatório.

Na Opção 4, a similaridade de Jaccard entre a string introduzida e o nome de cada filme terá de ser implementada por um método MinHash adequado à similaridade entre vetores de

caracteres escolhendo de forma fundamentada tanto o tamanho dos shingles como o número de funções de dispersão a usar.

Na Opção 5, a similaridade de Jaccard entre o conjunto de géneros introduzido e os géneros de cada filme terá de ser implementada por um método MinHash adequado a estimar a similaridade entre conjuntos de vetores de caracteres escolhendo de forma fundamentada o número de funções de dispersão a usar.

Requisitos para a implementação em Matlab

É obrigatório desenvolver 2 scripts Matlab.

O primeiro corre uma única vez para ler o ficheiro de entrada e guardar num ficheiro em disco todos os dados necessários à implementação da aplicação, incluindo:

- o(s) filtro(s) de Bloom com contagem (com o número de filmes por género e o número de filmes por género e por ano) de suporte às Opções 2 e 3;
- a matriz de assinaturas de suporte à Opção 4;
- a matriz de assinaturas de suporte à Opção 5.

O segundo script começa por ler do ficheiro em disco todos os dados previamente guardadas pelo primeiro script e depois implementa todas as interações com o utilizador descritas anteriormente.

Para a separação de um vetor de caracteres, investigue a utilidade da função Matlab strsplit

Avaliação do trabalho:

1. Opção 1 a funcionar corretamente (máximo 2 valores)

2. Opção 2 a funcionar corretamente (máximo 2 valores)

3. Opção 3 a funcionar corretamente (máximo 3 valores)

4. Opção 4 a funcionar corretamente (máximo 5 valores)

5. Opção 5 a funcionar corretamente (máximo 5 valores)

6. Fundamentação/avaliação das opções tomadas na implementação dos métodos probabilísticos (exem- plos: número de funções de dispersão, tamanho de shingles, dimensionamento dos filtros de Bloom com contagem) (máximo 2 valores)

7. Qualidade do relatório (máximo 1 valor)