

PL 4

Algoritmos Probabilísticos

Secção para avaliação¹

Considere uma aplicação, a desenvolver em Matlab, com algumas funcionalidades de um sistema online de disponibilização de filmes. A aplicação deve considerar um conjunto de utilizadores identificados por um ID e um conjunto de filmes também identificados por um ID (ambos os IDs definidos por um inteiro positivo).

Dados de entrada:

Dados de entrada: Considere o ficheiro u.data do conjunto de dados (release 4/1998) MovieLens 100k, disponível em <http://grouplens.org/datasets/movielens/> e utilize os dados das duas primeiras colunas deste ficheiro para identificar os utilizadores do sistema e os filmes que cada utilizador viu. A terceira coluna do ficheiro contém a avaliação atribuída por cada utilizador.

Do mesmo conjunto de dados, foi gerado o ficheiro `u_item.txt`, disponibilizado em separado, com o seguinte conteúdo:

Toy Story (1995)	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
GoldenEye (1995)	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	
Four Rooms (1995)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	

em que os dados de cada coluna estão separados por tabs. A linha número n contém a informação do filme com o ID n usado na segunda coluna do ficheiro u.data. A primeira coluna contém o nome do filme. As colunas da 2 à 20 contêm 1 ou 0 consoante o filme é classificado segundo um (ou mais) dos seguintes géneros:

Coluna: Género:	2 unknown	3 Action	4 Adventure	5 Animation	6 Children's	7 Comedy	
Coluna: Género:	8 Crime	9 Documentary	10 Drama	11 Fantasy	12 Film-Noir	13 Horror	14 Musical
Coluna: Género:	15 Mystery	16 Romance	17 Sci-Fi	18 Thriller	19 War	20 Western	

NOTA: executando no Matlab a instrução:

```
dic= readcell('u-item.txt', 'Delimiter', '\t');
```

é criado o cell array `dic` em que a célula `dic{i, j}` contém a informação da linha i e da coluna j do ficheiro `uitem.txt`.

¹A 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

Descrição da aplicação a desenvolver:

A aplicação deve começar por pedir o ID do utilizador que se torna o utilizador actual²:

Insert User ID (1 to 943):

certificando-se que o número introduzido é um ID válido (no ficheiro u.data, os IDs dos utilizadores são de 1 até 943). Depois, a aplicação deve permitir ao utilizador seleccionar uma de 5 opções:

- 1 - Your movies
 - 2 - Suggestion of movies based on other users
 - 3 - Suggestion of movies based on already evaluated movies
 - 4 - Search Title
 - 5 - Exit
- Select choice:

Opção 1: A aplicação lista os títulos dos filmes que o utilizador atual viu e, para cada nome, o número de vezes que o filme foi avaliado por todos os utilizadores (usando um Counting Bloom filter).

Opção 2: A aplicação determina os 3 utilizadores mais similares ao utilizador atual (em termos de conjuntos de filmes avaliados com nota superior ou igual a 3) e apresenta os títulos dos filmes que foram avaliados por pelo menos um dos 3 utilizadores e que ainda não foram avaliados pelo utilizador atual.

Opção 3: Para cada filme já avaliado pelo utilizador atual com nota superior ou igual a 3, a aplicação seleciona os filmes cuja distância de Jaccard estimada (em termos de géneros cinematográficos) seja menor que 0.9 e que ainda não tenham sido avaliados pelo utilizador atual. Isto resulta num conjunto de filmes por cada filme já avaliado pelo utilizador atual. No fim, a aplicação apresenta os títulos dos dois filmes que aparecem no maior número de conjuntos.

Opção 4: O utilizador insere uma string com o nome de um filme (ou parte do nome). A aplicação devolve os 5 nomes de filmes com os títulos mais similares à string introduzida.

Opção 5: A aplicação termina.

Notas sobre a implementação das funcionalidades da aplicação a desenvolver:

A estimativa da similaridade entre conjuntos (i.e., entre o conjunto de filmes vistos por 2 utilizadores na Opção 2, entre conjuntos de géneros cinematográficos de cada filme na opção 3 e entre 2 vectores de caracteres na Opção 4) tem de ser obrigatoriamente implementada por um método *MinHash*.

Na Opção 2, pode reutilizar a implementação que efectuou na secção 4.3 deste guião (PL04). O número adequado de funções de dispersão *k* pode ser escolhido de acordo com as conclusões que retirou nessa altura. A contagem do número de avaliações de cada filme tem de ser implementado com um Counting Bloom Filter.

Na Opção 3, deve desenvolver um método *MinHash* adequado à similaridade entre conjuntos de vectores de caracteres (géneros cinematográficos).

Na Opção 4, deve desenvolver um método *MinHash* adequado a estimar a similaridade entre vetores de caracteres escolhendo de forma fundamentada tanto o tamanho dos *shingles* como o número adequado de funções de dispersão *k* (sugere-se que experimente tamanhos de *shingle* entre 2 e 5 caracteres).

²Para introdução de dados pelo teclado, investigue a utilidade da função Matlab *input*

Requisitos para a implementação em Matlab

É obrigatório desenvolver 2 scripts Matlab.

O primeiro corre uma única vez para ler os dois ficheiros de entrada e guardar em ficheiro todas as estruturas de dados associadas aos utilizadores e aos filmes, incluindo:

- a matriz de assinaturas com os vectores *MinHash* correspondente ao conjunto de filmes avaliados por cada utilizador (suporte à Opção 2);
- a(s) estrutura(s) de dados do Counting Bloom filter para armazenamento do número de avaliações com nota superior ou igual a 3 (suporte à Opção 2);
- a matriz de assinaturas com os vectores *MinHash* correspondentes ao conjunto de géneros cinematográficos de cada filme (suporte à Opção 3);
- a matriz de assinaturas com os vectores *MinHash* associados aos títulos dos filmes (suporte à Opção 4);

O segundo script começa por ler do disco todas as estruturas previamente guardadas pelo primeiro script e depois implementa todas as interacções com o utilizador descritas anteriormente.

Avaliação do trabalho:

1. Opção 1 a funcionar corretamente (**máximo 4 valores**)
2. Opção 2 a funcionar corretamente (**máximo 4 valores**)
3. Opção 3 a funcionar corretamente (**máximo 5 valores**)
4. Opção 4 a funcionar corretamente (**máximo 4 valores**)
5. Fundamentação/avaliação das opções tomadas na implementação dos métodos probabilísticos (exemplos: número de funções de dispersão, tamanho de *shingles*, dimensionamento dos filtros de Bloom) (**máximo 2 valores**)
6. Qualidade do relatório (**máximo 1**)