PL 4

Algoritmos Probabilísticos

Secção para avaliação [1]

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:

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

Toy Story (1995)	Adventure	Animation	Thriller
GoldenEye (1995)	Action	Adventure	Thriller
Four Rooms (1995)	Thriller		

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 e respetivo ano de estreia. As restantes colunas contém um número variável de géneros cinematográficos associados ao filme.

NOTA: executando no Matlab a instrução:

```
dic= readcell('films.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 `films.txt`.

Descrição da aplicação a desenvolver:

A aplicação deve começar por pedir o ID do utilizador que se torna o utilizador actual [2]:

```
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 6 opções:

```
1 - Your Movies
2 - Films from most similar user
3 - Search Title
4 - Find most similar films
5 - Estimate the number of films from a year
6 - Exit
Select choice:
```

[1] 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**

[2] Para introdução de dados pelo teclado, investigue a utilidade da função Matlab `input`

Opção 1: A aplicação lista os títulos dos filmes que o utilizador actual viu. Cada linha deve mostrar o ID e o título de um filme.

Opção 2: A aplicação lista os filmes avaliados pelo utilizador mais "similar"ao utilizador actual. A aplicação começa por determinar qual de todos os outros utilizadores é mais similar ao utilizador actual (em termos do conjunto de filmes vistos por cada um) e, finalmente, a aplicação lista todos os títulos dos filmes que o utilizador mais similar viu.

Opção 3: A aplicação pede primeiro para o utilizador inserir uma string:

```
Write a string:
```

e depois a aplicação apresenta os títulos dos filmes (um título por linha) que sejam mais similares à string introduzida indicando a (estimativa da) distância de Jaccard entre a string e cada título. Esta opção é independente do ID do utilizador actual. A lista apresentada deve ter no máximo 5 títulos, ordenados por ordem crescente de distância de Jaccard e deve apresentar apenas filmes cuja distância de Jaccard seja menor ou igual a 0.99. Se não houver nenhum título nestas condições, a aplicação deve indicar que não encontrou nenhum título.

Opção 4: A aplicação começa por listar os filmes vistos pelo utilizador atual (ID e título) e pede para escolher um desses filmes. Em seguida a aplicação deve apresentar os 3 filmes mais similares ao filme escolhido. Comece por implementar um primeiro nível de comparação baseado na similaridade do conjunto de géneros cinematográficos de cada filme (a implementar obrigatoriamente por MinHash) e numa segunda fase, para os casos de empate, implemente um segundo nível de comparação utilizando o valor médio de avaliação dos filmes (Nota: a coluna 3 do ficheiro `u.data` contém a avaliação de cada utilizador para cada filme).

Opção 5: A aplicação pede primeiro para o utilizador inserir um ano. Em seguida, a aplicação deve indicar uma estimativa do número de filmes desse ano que fazem parte da lista. A implementar obrigatoriamente com um filtro de Bloom com contagem.

Opção 6: A aplicação termina.

Notas sobre a implementação das funcionalidades da aplicação a desenvolver:

A estimativa da similaridade entre conjuntos (i.e., entre filmes vistos por 2 utilizadores na Opção 2, entre 2 vectores de caracteres na Opção 3 e entre conjuntos de géneros cinematográficos de cada filme 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.

Na Opção 3, deve desenvolver um método *MinHash* adequado à similaridade entre vectores 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).

Na Opção 4, deve desenvolver um método *MinHash* adequado a estimar a similaridade entre conjuntos de vetores de caracteres.

Na Opção 5 A estimativa do número de filmes de um determinado ano terá de ser implementada usando um filtro de Bloom com contagem. Os parâmetros do filtro devem ser adequados ao problema e a sua escolha devidamente fundamentada no relatório.

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* de cada utilizador (suporte à Opção 2);
- a matriz asssinaturas com os vectores *MinHash* de cada título (suporte à Opção 3);
- a matriz com os vectores *MinHash* associados ao conjunto de géneros cinematográficos de cada título (de suporte à Opção 4);
- a(s) estrutura(s) de dados do filtro de Bloom com contagem para armazenamento dos anos dos filmes (suporte à Opção 5).

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 2 valores)**
2. Opção 2 a funcionar corretamente **(máximo 4 valores)**
3. Opção 3 a funcionar corretamente **(máximo 4 valores)**
4. Opção 4 a funcionar corretamente **(máximo 5 valores)**
5. Opção 5 a funcionar corretamente **(máximo 2 valores)**
6. 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)**
7. Qualidade do relatório **(máximo 1)**