MPEI – Trabalho Práctico

# Informações

Este projecto está a ser realizado por Mário Alexandre Lopes Liberato (NMEC: 84917) e Jorge Miguel Aires de Matos Oliveira (NMEC: 84983). O projecto pretende utilizar alguns módulos desenvolvidos no âmbito dos guiões prácticos da unidade curricular de Métodos Probabilísticos para Engenharia Informática, convertidos para Java quando conveniente. Quando não o for, pretende-se utilizar a biblioteca JavaOctave ou outra que sirva o mesmo propósito, a fim de manter o projecto maioritariamente escrito em Java. Em último caso, utilizar-se-á apenas um script Octave.

# Conceito Inicial

Programa que sirva para gerir aplicações de pessoas a um determinado emprego. O lado do potencial trabalhador fornece um questionário que depois de completo adiciona os dados a um ficheiro ou base de dados simples (Exemplo: USER ID, SKILLSET ID, EDUCATION LEVEL ID, TIMESTAMP), guardado em CSV. Este ficheiro é processado pelo programa do empregador, que cria, para cada emprego diferente , um modelo de funcionário com o skillset ideal, com o nível de educação ideal, etc. O programa depois filtrará todos os empregados pelo seu índice de semelhança e opcionalmente eliminar aqueles que não possuem semelhança alguma(ou baixa semelhança).

# Descrição Breve do Programa

Três programas são utilizados. Um deles será o programa utilizado pela pessoa que submete a aplicação à empresa. Este é denominado por de programa B no documento. O segundo, o programa A, é utilizado pelo empregador para gerir todas as aplicações aos empregos disponíveis. O último programa, dito programa C, serve para gerar dados simulados para utilização pelo programa A.

## Descrição Breve do Programa B

Este programa contém uma interface simples, com um questionário com escolhas múltiplas. Quando o questionário é submetido, este gera dados em CSV, os quais são adicionados ao ficheiro principal, no formato seguinte:

* USER\_ID, SKILL\_ID, EDUCATION\_LEVEL\_ID

O programa gera ainda dados CSV para um segundo ficheiro no formato:

* USER\_NAME, USER\_BIRTHDATE, TIMESTAMP

Estes dados serão armazenados em dois ficheiros CSV. Idealmente, em vez disso, seriam armazenados numa única base de dados que use duas tabelas. Dependendo do progresso do projecto até à data de entrega, esta funcionalidade poderá ou não ser implementada.

### Descrição dos campos das tabelas

* USER\_ID – ID do utilizador, atribuído por ordem crescente. Na segunda tabela, os dados da linha X correspondem ao utilizador com ID X+1.
* SKILL\_ID – ID da capacidade.
* EDUCATION\_LEVEL\_ID – Nível de educação (0 – Nenhum, 1 – Primário, 2 – Secundário, 3 – Licenciatura, 4 – Mestrado, 5 – Doutoramento).
* USER\_NAME – Nome completo do utilizador.
* USER\_BIRTHDAY – Data de nascimento do utilizador.
* TIMESTAMP – A timestamp de criação dos dados.

## Descrição Breve do Programa A

Este programa trata do processamento dos dados obtidos. Essencialmente, lê os dados da primeira tabela, constrói uma lista dos skillsets de cada pessoa, insere os “ideais” como novas pessoas no fim da lista, calcula a distância de Jaccard entre todos os utilizadores e filtra todos os que estiverem abaixo de um determinado nível de semelhança. Depois apresentará os utilizadores mais relevantes para cada emprego, por ordem de nível de escolaridade. Esta parte do programa utilizará predominantemente MatLab.

A interface é simples, e consiste apenas numa opção que serve para iniciar o processamento. Eventualmente, esta interface poderá ser expandida e aumentar em complexidade, mas não faz parte do plano inicial.

## Descrição Breve do Programa C

Este programa gera dados aleatórios para os ficheiros utilizados pelo programa A. Este programa pede apenas o número de utilizadores que se deve criar e gera esse número de utilizadores nas duas tabelas. É gerado usando um programa escrito em Java.

# Funcionamento do programa A

## Classe BloomFilter

A classe BloomFilter é uma adaptação da versão escrita em Octave para os guiões das aulas prácticas da unidade curricular. Esta classe implementa um “Bloom Filter”. Utiliza a função de dispersão “Dbj2”, tal como a versão na qual se baseia.

Esta classe contém um array que contém o bloom filter propriamente dito, denominado de bloomFilter. Esta variável é privada e só pode ser alterada pelos métodos implementados.

Para inicializar o bloom filter, deve-se criar um novo objecto do tipo BloomFilter, e fornecer-lhe como argumento o tamanho desejado. O construtor inicializa então o array com zeros.

Para obter o array, deve-se recorrer ao método getArray, que retorna o dito-cujo.

Para obter o tamanho, deve-se utilizar o método getSize, que retorna a length do array bloomFilter.

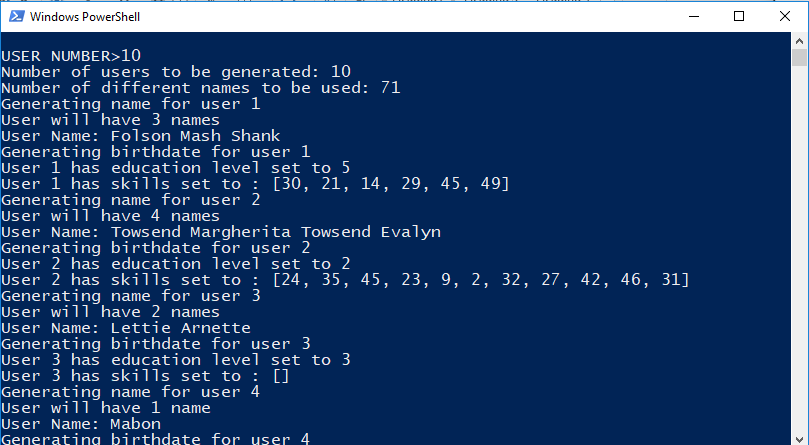
Para adicionar um membro deve-se utilizar o método addMember, que aceita como argumento o o número de funções de dispersão e a String que se pretende adicionar. Para remover um membro ou verificar se ele existe, os mesmos argumentos são fornecidos, e os métodos são removeMember e existsMember, respectivamente. O último retorna um valor de true ou false, dependendo da não existência ou possível existência do membro.

# Funcionamento do programa C

## Parte 1 – Gerar os valores

O programa C serve para gerar dados aleatórios. Para atingir este fim, precisa de um argumento, que é o número de utilizadores gerados. Se não for fornecido um argumento, o programa pedirá que um valor seja inserido. Depois, lê um ficheiro “names.txt”, que deverá conter **apenas** nomes, mais especificamente um nome por cada linha, os quais são depois utilizados para gerar nomes completos, os quais variam entre 1 e 4 nomes no total. Os nomes do ficheiro foram escritos pelos desenvolvedores do projecto ou obtidos a partir de um gerador de nomes aleatório. Como apenas existe para propósito de testes, não há qualquer problema no facto dos nomes não serem coerentes e misturarem várias línguas.

É utilizado um ciclo for para gerar cada utilizador, e este é adicionado ao ArrayList de objectos User denominado de userList no fim deste. Depois do nome ser gerado, é gerada a timestamp do nascimento, o qual varia entre 657170474 e 909631274. Não há propriamente uma razão para a escolha destes valores, uma delas é apenas a data de nascimento de ML, e a outra é a data 8 anos antes. Assim, as idades dos utilizadores aleatórios deve variar entre 19 e 27 anos.

O nível de educação é gerado de seguida, sendo um simples valor inteiro entre 0 e 5. Posteriormente, o utilizador recebe skills aleatórias, as quais variam entre 1 e 50 (havendo então 50 skills separadas). Neste programa, cada utilizador recebe um máximo de 13 skills, e pode ter 0. Esta limitação não é parte dos outros programas. No caso do programa B, mais skills podem ser escolhidas, e no caso do programa A, mais skills podem ser processadas. Tanto o limite de 13 como as 50 skills diferentes são valores arbitrários, e apenas este último se apresenta nos outros programas. Por fim, o utilizador recebe uma timestamp, que é a timestamp actual, e é adicionado à lista.

*Imagem 1 – Exemplo do output do programa para o terminal*

## Parte 2 – Armazenar os valores