

# Métodos Probabilísticos para Engenharia Informática

## Projeto: Sistema de recomendação de notícias

Trabalho realizado por:

Tiago Costa (114629)

Tiago Almeida (113106)

## 1 Introdução

### 1.1 Contexto

Serve o presente relatório para expor o trabalho desenvolvido no projeto proposto na Unidade Curricular de *Métodos Probabilísticos para Engenharia Informática* (MPEI). O grupo responsável pela realização do projeto propôs-se a desenvolver um sistema capaz de recomendar notícias a um utilizador, baseando-se num *dataset* de treino. O modelo é posteriormente testado e posto em prática no programa final (apresentado ao utilizador), utilizando um *dataset* de teste. Todo o processo de desenvolvimento, bem como a descrição dos módulos apresentados e do *dataset* utilizado será explorado ao longo do relatório. Todo o projeto foi desenvolvido em MATLAB, colocando em prática três matérias lecionadas na Unidade Curricular:

- Classificadores Naive Bayes
- Filtros de Bloom
- Algoritmo MinHash

### 1.2 Objetivos

O objetivo final do projeto incide na criação de uma aplicação interativa em MATLAB, na qual um utilizador pode ver listas de artigos, procurar por artigos numa certa categoria, e

lê-los. É de notar que a categoria dos artigos não é explicitada no *dataset* de teste, pelo que a identificação dos artigos também fez parte do trabalho desenvolvido.

Com isto, ainda dentro de cada categoria, o sistema deve recomendar ao utilizador as notícias mais relevantes, tendo em conta o histórico de leitura do mesmo.

Assim, os três modelos utilizados têm utilidades únicas e fundamentais:

- Naive Bayes - Categoriza cada um dos artigos do *dataset* de teste, tendo por base o *dataset* de treino
- Bloom Filter - Assegura que nenhum artigo já lido é recomendado novamente ao utilizador
- MinHash - Dentro de cada categoria, recomenda ao utilizador apenas os artigos mais relevantes, tendo em conta o histórico de leitura

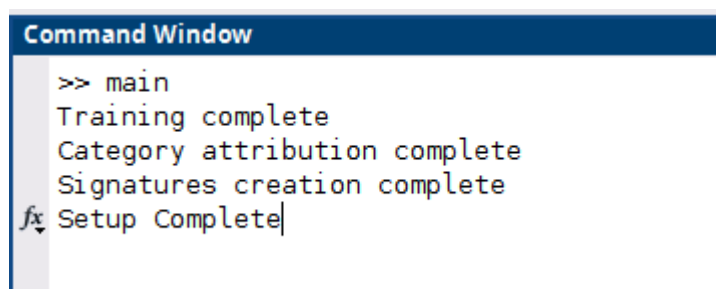
## 1.3 Dataset

O *dataset* utilizado é um *dataset* público que tem 3 ficheiros .csv. O ficheiro de treino, com o texto dos artigos e as suas categorias; o ficheiro de teste, que apenas tem o texto dos artigos; e o ficheiro das soluções, que contém a atribuição correta das categorias aos artigos do ficheiro de teste.

# 2 Módulos Utilizados

## 2.1 Naive Bayes

Ao correr o programa, o modelo Naive Bayes é logo treinado e posto em prática. Espera-se que no fim do treino, o modelo já tenha classificado corretamente todos os artigos e agrupado todos os artigos da mesma categoria dentro de listas respetivas.



```
Command Window
>> main
Training complete
Category attribution complete
Signatures creation complete
fx Setup Complete|
```

Este modelo começa por dividir o *dataset* de treino em *bags of words* para cada um dos artigos, removendo palavras duplicadas. De seguida, em cada artigo, calcula a quantidade de vezes que cada palavra surge.

Para cada um dos artigos é verificada a categoria à qual o artigo pertence e, com isto, calcula-se (de um modo muito resumido) a “probabilidade de cada palavra pertencer a uma dada categoria.

De seguida, o programa parte para a atribuição de categorias aos artigos desconhecidos (do *dataset* de teste). Para isto, utiliza-se o logaritmo das probabilidades, uma vez que a multiplicação de probabilidades teria como resultado números muito pequenos, o que traz consigo os seus problemas.

Com as categorias atribuídas, o programa agora já pode mostrar ao utilizador os artigos categorizados.

## 2.2 Filtro Bloom

A utilização do filtro Bloom é simples e direta. Sempre que o programa está prestes a mostrar artigos ao utilizador, todos esses artigos passam pelo filtro Bloom de modo a que não sejam apresentados artigos repetidos (isto é, que o utilizador já tenha lido).

Para isto, inicialmente é calculado o número ótimo  $k$  de *hash functions*, através do número de artigos e de alguns outros critérios.

Sempre que um artigo é lido, o seu ID passa pelas  $k$  *hash functions*, e os índices gerados são colocados a 1 no *bitmap*.

Assim, nos momentos em que o programa vai apresentar artigos ao utilizador, os IDs desses artigos passam todos pelo filtro Bloom. Isto é, todos os artigos passam pelas mesmas  $k$  *hash functions* para confirmar que o artigo não foi lido.

## 2.3 MinHash

O MinHash é usado na opção “Pesquisar por categoria” para mostrar opções mais similares ao que nós já lemos. As assinaturas são criadas no setup para por categoria para cada artigo. Com essas assinaturas, ao pesquisar por categoria ele vai ver dentro dos artigos lidos dessa categoria, e devolver os 10 mais similares.

# 3 Funcionamento da aplicação

## 3.1 Setup

Ao correr a aplicação (ficheiro main.m), existe um tempo de setup de cerca de 6 minutos, onde são feitas várias operações como treinamento do Naive Bayes, atribuição de categorias aos diversos artigos usando o modelo Naive Bayes treinado, criação de bitmaps para os bloom filters e até criação de assinaturas para uso de minhash.

É possível ver o progresso com o aparecimento de mensagens na “Command Window”.

```
Command Window
Training complete
Category attribution complete
Signatures creation complete
Setup Complete

Precione enter para continuar.
fx
```

Quando tiver terminado, é pedido para pressionar a tecla “enter” para continuar para o menu de utilizador.

## 3.2 Menu de utilizador

Ao terminar o setup, podemos finalmente usar de facto a aplicação. Inicialmente existe um menu de operações, com 4 operações diferentes. Para escolher uma operação basta escrever o número correspondente. Caso escolha um número que não existe, nada acontece.

```
Command Window
===== Menu de operações =====

1- Mostrar artigos aleatórios
2- Pesquisar por categoria
3- Mostrar artigos lidos
4- Sair

=====
fx Escolha uma operação: |
```

## 3.3 Mostrar artigos aleatórios

Ao escolher a primeira opção, serão escolhidos aleatoriamente 10 artigos aleatórios e serão apresentados junto com o seu ID, texto inicial (primeiras 15 palavras) e a categoria determinada pelo Naive Bayes calculada na hora.

```
Command Window
=====
ArticleID: 1712 | Category: sport
Text: souness backs smith for scotland graeme souness believes walter smith would be the perfect choice to succeed berti vogts as

ArticleID: 1644 | Category: business
Text: us economy still growing says fed most areas of the us saw their economy continue to expand in december and

ArticleID: 2136 | Category: business
Text: boeing secures giant japan order boeing is to supply japan airlines with up to 50 of its forthcoming 7e7 planes

ArticleID: 118 | Category: politics
Text: chancellor rallies labour voters gordon brown has issued a rallying cry telling supporters the stakes are too high to stay

ArticleID: 1898 | Category: sport
Text: henry tipped for fifa award fifa president sepp blatter hopes arsenal s thierry henry will be named world player of

ArticleID: 2138 | Category: sport
Text: captains lining up for aid match ireland s brian o driscoll is one of four six nations captains included in

ArticleID: 954 | Category: entertainment
Text: hollywood hunts hits at sundance the sundance film festival the movie industry s top destination for uncovering the next independent

ArticleID: 558 | Category: business
Text: us bank boss hails genius smith us federal reserve chairman alan greenspan has given a speech at a scottish church

ArticleID: 1138 | Category: sport
Text: d arcy injury adds to ireland woe gordon d arcy has been ruled out of the ireland team for saturday

ArticleID: 369 | Category: business
Text: emi shares hit by profit warning shares in music giant emi have sunk by more than 16% after the firm

=====
===== Menu de operações =====
1- Ler um artigo
2- Voltar ao menu inicial
=====
fx Escolha uma operação: |
```

No fim da página, aparece também um submenu que funciona como o menu inicial, mas desta vez com duas opções: voltar ao menu inicial ou ler um artigo. Se escolhermos voltar ao menu inicial, voltamos ao início, e nada de especial acontece. Se escolhermos a opção “Ler um artigo”, é nos pedido o “ArticleID” do artigo que pretendemos ler.

```
Command Window
ArticleID: 369 | Category: business
Text: emi shares hit by profit warning shares in music giant emi have sunk by more than 16% after the firm issued a profit warning following disappointing sales and delays to two albu

Precione enter para voltar.
fx
```

Após inserir, é nos apresentado o artigo na sua extensão total, e é acrescentado no bitmap dos artigos lidos usando bloom filter.

## 3.4 Pesquisar por categoria

Ao escolher “Pesquisar por categoria” é nos pedido para indicar a categoria. Ao indicar a categoria existem dois casos possíveis: Ou ainda não foram lidos nenhuns artigos e são apresentados 10 artigos aleatórios dessa categoria, ou já foram lidos artigos da categoria escolhida e são apresentados artigos similares aos já lidos usando MinHash.

## 3.5 Mostrar artigos lidos

Ao escolher a opção “Mostrar artigos lidos”, o programa vai buscar à matriz de artigos lidos todos os artigos que o utilizador já leu (isto é, todos os artigos cujo ID foi inserido na opção “Ler um artigo”).

