MPEI 2024-2025

Trabalho Prático

Objectivo Geral

- Desenvolver
- Testar e
- Demonstrar

uma aplicação que demonstre a utilização conjunta de:

- **Classificador Naïve Bayes**
- **Determinação pertença a um conjunto**
- **Deteção de itens similares** (finding similar items)

Como vão conseguir isto ?

- Dividindo em partes
  - Como em muitas outras situações

- Que partes ?
  1. Desenvolver os componentes/ módulos
  2. Testar os módulos
  3. Criar demonstração de uso conjunto
     - E testá-la
  4. Demonstrar aos Professores das Práticas

Desenvolver os componentes/ módulos

- Terão de desenvolver **três componentes**:
  - **Classificador Naïve Bayes**
  - **Um Bloom Filter**
  - Determinação de **itens similares** usando **Minhash**

- **Os guiões práticos servem de base para a compreensão dos algoritmos**

Testar os módulos

- Para além da criação dos módulos **terão de criar um conjunto de testes adequados para cada um dos módulos**

- Deverão ser adaptados os testes pedidos nos Guiões das aulas prática
- Podem e devem pensar e implementar outros testes que usem conhecimentos da cadeira …
- Será valorizado esse esforço e espírito criativo

Criar demonstração de uso conjunto (dos vários algoritmos)

- Esta parte apela particularmente à vossa imaginação

- Exemplo (que não devem replicar por deixar de ser uma ideia nova):
  - Sistema de Deteção de Spam em E-mails

Exemplo - Sistema de Deteção de Spam em E-mails

Para identificar e-mails de spam e verificar similaridades entre mensagens.

1. Classificador Naïve Bayes
   - O Naïve Bayes é usado para **classificar e-mails como spam ou não-spam**:
   - O modelo é treinado com um conjunto de e-mails marcados como spam e não-spam.

2. Filtro Bloom
   - O filtro Bloom ajuda a **detetar rapidamente e-mails duplicados ou com características conhecidas de spam**, economizando tempo e recursos:
   - Armazena hashes de e-mails já marcados como spam.

3. MinHash para Deteção de Similaridade
   - MinHash é utilizado para **identificar e-mails com alta similaridade a outros já classificados como spam**, detetando campanhas de spam que alteram levemente o conteúdo:
   - O conteúdo do e-mail é convertido em "shingles" (sequências de palavras), gerando uma assinatura MinHash.
   - As assinaturas são comparadas com e-mails anteriores, permitindo identificar grau de semelhança.

Exemplo - Fluxo do Sistema

1. Pré-processamento:
   - O sistema extrai características do e-mail (palavras, links).
2. Filtro Bloom:
   - Verifica se o e-mail já é conhecido como spam; se sim, marca como spam imediatamente.
3. Classificação de Spam:
   - Se não houver correspondência no filtro Bloom, o Naïve Bayes analisa o e-mail e calcula a probabilidade de ser spam.
4. Verificação de Similaridade:
   - MinHash compara o e-mail com mensagens anteriores para ver se é altamente similar a outros spams.
5. Decisão Final:
   - O sistema marca o e-mail como spam ou não com base nos resultados das análises.

Este sistema combina métodos eficientes para detetar rapidamente e-mails suspeitos e similares, tornando-se ideal para filtros de spam em tempo real.

Criar demonstração de uso conjunto

- Nas TP e nas práticas podem e devem "discutir" ideias

- Nas TPs haverá tempo no final das aulas para conversar sobre este assunto desde que o solicitem
  - Por vezes podem apenas ter resposta na aula seguinte …

Demonstrar aos Professores das Práticas

- No **final**, nas práticas, terão de mostrar **o que fizeram** e responder às questões do docente

- Terão, também, de entregar o código e uma explicação de como o usar (mini relatório)
  - em especial para correr os testes

Papel das aulas TP e Prática

- Nas TPs serão:
  - apresentados os conceitos base;
  - mostradas utilizações;
  - dadas dicas de como implementar partes dos módulos;
  - discutidas as vossas ideias, questões e problemas
- Resumindo: serão importantes

- Nas práticas irão ter ajuda na resolução de dois guiões que garantem uma parte importantes dos pontos 1 e 2 (criar os módulos e testá-los)
  - Resumindo: essenciais para que tenham trabalho para entregar no final

- No entanto será preciso mais do que ir às TPs e Práticas
- …. Trabalho de casa

Avaliação

Regras

- Trabalho de grupo
  - Máximo de 2 alunos
  - Da mesma turma prática
  - Terão de comunicar os grupos na primeira aula prática sobre Bloom Filters

- Solicitaremos submissão no Elearning do estado do trabalho ao longo das semanas

Regras (cont.)

- Obrigatório ter os 3 módulos a funcionar para ter >=7 (em 20 valores)

- Cuidado com as cópias:
  - se detetadas serão penalizados, dividindo a nota pelo número de cópias

Regras (continuação)

- Linguagem de programação:
  - Matlab

em **casos devidamente justificados** e **se aceites pelo Regente**, poderá ser utilizada outra linguagem de programação.

Cotações

Critério	Peso	Escala
Implementação dos módulos	50%	1 (Muito pouco) a 5 (Excelente)
Testes dos módulos	10%	1 a 5
Aplicação conjunta	25%	1 a 5
Apresentação	15%	1 a 5

Critério	Peso	Exemplo
Implementação do módulos		Bom → 4 [equivalente a 16 de 0 a 20]
Testes dos módulos		MB → 4.5 [equivalente a 18 de 0 a 20]
Aplicação conjunta		SUF → 3
Apresentação		SUF- → 2
		(1/2x4 + 1/10 x 4.5 + 1/4 x 3 + 15/100 x 2) x 4
		(2 + 0.45 + 3/4 + 0.3) x 5 = 3,5 x 4 = 14

Datas / Prazos

- Datas limite para o trabalho
  - Apresentação na última aula prática (dezembro)
  - Entrega até 2 dias antes da apresentação

Documentos a entregar

- Programas (separados e claramente identificados)
  - Código relativo aos módulos desenvolvidos (≥ 3 programas)
  - Testes dos módulos (3 programas)
  - Demontração conjunta dos vários algorítmos (1 programa)
  - Datasets
- Relatório
  - Máximo 5 páginas
  - Deve incluir:
    - Descrição de como correr os vários programas (testes, demonstração conjunta, etc)
    - Apresentação e análise dos resultados obtidos nos vários testes
    - Descrição da aplicação de uso conjunto bem como das vantajens e limitações das soluções propostas.
- Apresentação
  - se criarem slides para a apresentação devem submeter o respetivo documento
    - não é obrigatório fazer um conjunto de slide mas muito recomendado

Critérios de avaliação

- Complexidade e Ambição
  - Em particular da aplicação conjunta
- Quantidade de trabalho
- Quantidade e qualidade do código produzido

- Quantidade e qualidade dos testes realizados

- Qualidade da apresentação
  - Incluindo qualidade dos materiais de suporte à apresentação (ex.: apresentação PowerPoint)
- Resultados obtidos (e demonstrados)
- Qualidade do relatório

- Similaridade a outros trabalhos
  - [penalização para trabalhos similares]