MPEI - Métodos Probabilísticos para Engenharia Informática (2015/2016) **PL 06**

Palavras-chave: Bloom Filters, Hash Functions.

O presente trabalho prático tem por objectivo criar um módulo que suporte a criação de Bloom filters (ex: [2]) e testar esse módulo.

Para tal, execute os seguintes passos:

1) Crie, em Matlab, um conjunto de funções que implementem as funcionalidades de um Bloom Filter básico. As funções devem ter os parâmetros necessários para que seja possível criar Bloom filters de diferentes dimensões e usando números diferentes de funções de hash (k).

   Sugestão 1: Deve criar pelo menos 3 funções [1, sec. 3.2]: uma para inicializar a estruturas de dados; outra para inserir um elemento (ou elementos) no filtro; uma terceira para verificar se um elemento pertence ao conjunto.

   Sugestão 2: Deve procurar, seleccionar e implementar uma função de hash que tenha bom desempenho.

   Nota 1: É obrigatório manter a informação original sobre autores e afins em todas as funções que utilizarem e que não sejam da vossa autoria. Adaptações de código existente apenas podem ser feitas se as condições de utilização definidas pelos autores o permitirem, mantendo sempre informação sobre o autor original e adicionando informação sobre quem fez a alteração/evolução. Neste trabalho sugere-se criação de código original para todas as funções com a excepção de funções de hash.

2) Teste as funções criadas na criação de um pequeno Bloom Filter para guardar uma lista de cidades. Insira alguns nomes de cidades no filtro e teste a pertença desses e de algumas cidades adicionais que não pertença a essa lista inicial.

3) Para um teste mais exaustivo:

   (a) Gere m=1000 strings aleatórias com 40 caracteres (considere como caracteres possíveis o conjunto de caracteres minúsculos, maiúsculos e algarismos) e preencha um bloom filter, de tamanho n=8000. Este bloom filter deve ter k= 3 hash functions.

   (b) Gere um novo conjunto de 10000 strings aleatórias com 40 caracteres e teste a pertença das mesmas ao bloom filter que preencheu.

4) Repita o teste da questão anterior para um número diferente de hash functions ($k = 1, \ldots, 15$), obtendo o número de falsos positivos para cada k. Represente num gráfico os valores obtidos, em função de k e sobreponha nesse gráfico os valores teóricos (Assuma a independência de hash functions e que cada uma seleciona cada posição do bloom filter com igual probabilidade).

   Nota: Assume-se que as 10000 strings de teste são todas diferentes das 1000 inseridas no Bloom filter. No entanto pode haver strings iguais.

Referências

[1] James Blustein and Amal El-Maazawi. Bloom filters - a tutorial, analysis, and survey. Tecnhical Report CS-2002-10, Dalhousie University, Dec 2002.

[2] Jure Leskovec, Anand Rajaraman, and Jeff Ullman. *Mining of Massive Datasets*, chapter Mining Data Streams. Cambridge University Press, 2014.