Departamento de Electrónica, Telecomunicações e Informática da Universidade de Aveiro

Algoritmos Avançados

2020/2021 — 1º Semestre

3º Trabalho

Data limite de entrega: 5 de fevereiro de 2021

Hipótese A – Contagem dos Itens Mais Frequentes

Pretende-se determinar os itens mais frequentes de um conjunto de dados, explorando métodos que permitem processar conjuntos de dados de grande dimensão. Para isso deve implementar (**Python 3**) e analisar o comportamento de **um dos seguintes métodos**:

– O algoritmo de Misra & Gries – FREQUENT-COUNT

– O algoritmo de Manku & Motwani – LOSSY-COUNT

– O algoritmo de Metwally et al. – SPACE-SAVING-COUNT

– The Count-Min Sketch – use um **número fixo de funções de hashing**, por exemplo 5.

Tarefa: Analise o comportamento do método que desenvolveu quando altera algum dos seus parâmetros. Qual é a influência dessas alterações nos resultados dos testes computacionais?

Hipótese A-1 – Strings de letras minúsculas

Como aplicação, e para analisar o comportamento do método desenvolvido, deve utilizar conjuntos de dados contendo letras minúsculas, que lhe permitem "simular" *data streams* de um modo simples:

- gere ficheiros de texto com letras minúsculas aleatórias separadas por um espaço, com 100, 1000, 10000, 100000, 1000000, 10000000, etc. elementos. **Deve atribuir maior probabilidade a algumas das letras**.

- use o método desenvolvido para determinar as letras mais frequentes.

- há alguma letra que ocorra mais de 5% ou 10% das vezes?

- compare os resultados obtidos com as **contagens exatas**.

Hipótese A-2 – Ficheiros de texto de obras literárias

Como aplicação, e para analisar o comportamento do método desenvolvido, deve utilizar ficheiros de texto de obras literárias, que lhe permitem "simular" *data streams* de um modo simples:

- use o método desenvolvido para determinar as letras mais frequentes.

- compare os resultados obtidos com as contagens exatas.

Hipótese B – Estimativa do Número de Itens Distintos

Pretende-se estimar o número de itens distintos de um conjunto de dados, explorando métodos que permitem processar conjuntos de dados de grande dimensão. Para isso deve implementar (**Python 3**) e analisar o comportamento de **um dos seguintes métodos**:

– Tabela de Hashing Simplificada, sem resolução de colisões

– Filtro de Bloom – use um **número fixo de funções de hashing**: por exemplo 5.

Tarefa: Analise o comportamento do método que desenvolveu quando altera o tamanho da tabela/filtro. Qual é a influência dessa alteração nos resultados dos testes computacionais?

Como aplicação, e para analisar o comportamento do método desenvolvido, deve utilizar ficheiros de texto de obras literárias, que lhe permitem "simular" *data streams* de um modo simples:

- use o método desenvolvido para estimar o número de palavras distintas.

- compare os resultados obtidos com as **contagens exatas**.

J. Madeira, 11 de janeiro de 2021