Departamento de Electrónica, Telecomunicações e Informática da Universidade de Aveiro

Algoritmos Avançados

2025/2026 — 1º Semestre

3rd Project — Most Frequent and Less Frequent Items

Deadline: December 22, 2025

In addition to the exact counters, each student will be assigned two additional methods. Check your assignment on the corresponding PDF file.

Objectives

The goal is to **identify frequent items in datasets** using different methods, and to **evaluate the quality of estimates** regarding the **exact counts**.

To accomplish that, develop and test **three different approaches**:

- **exact counters,**
- **approximate counters,**
- **one algorithm to identify frequent items in data streams**.

An analysis of the computational efficiency and limitations of the developed approaches is also to be carried out.

For example, in terms of **absolute and relative errors** (lowest value, highest value, average value, etc.), **average values**, etc.

It can also be verified whether the **same most frequent / less frequent items** are identified, and in the **same relative order**.

For this you must:

a) Compute the **exact number of occurrences** of each item.

b) **Estimate the number of occurrences** of each item using **approximate counters**. Perform a set of tests, **repeating the approximate counts a few times**.

c) Estimate the **n most frequent** items, running your **data stream algorithm** for **some values of n** (e.g., 5, 10, 15, 20, …). Also, **experiment** with **different** algorithm **parameters**, if applicable.

d) **Compare the performance** of the approximate counters and the data stream algorithm, between themselves and regarding the exact counts.

e) Write a report (max. 10 pages).

Data for the computational experiments

Each student is assigned a **dataset** from Kaggle – check the assignment on the corresponding PDF file.

For each dataset, the **attribute to be counted / analyzed** is also indicated.

J. Madeira, December 1, 2025