Departamento de Electrónica, Telecomunicações e Informática da Universidade de Aveiro

Algoritmos Avançados

2024/2025 — 1º Semestre

3rd Project — Most Frequent and Less Frequent Words

Deadline: January 3, 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 words in text files (books)** 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 words** are identified, and in the **same relative order**.

And if those **most frequent / less frequent words** are **similar** in the text files of the same book in **different languages**.

For this you must:

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

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

c) Estimate the **n most frequent words**, running your **data stream algorithm** for **some values of n, e.g., 5, 10, 15, 20, …**

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 – Simulating data streams

Obtain **text files from different editions of the same books**, and in **different languages** – e.g., from Project Gutenberg.

Process the text files to:

- Remove the Project Gutenberg **file header** and **file tail**, to only process the text of each book.
- Remove all **stop-words** and **punctuation marks**.
- Convert all letters to **lowercase**.

J. Madeira, December 9, 2024