Departamento de Electrónica, Telecomunicações e Informática da Universidade de Aveiro

Algoritmos Avançados

2023/2024 — 1º Semestre

3rd Project — Most Frequent Letters

Deadline: January 7, 2024

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 the most frequent letters in text files (literary works)** using different methods, and to **evaluate the quality of estimates** regarding the **exact counts**.

In order 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 has 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 letters** are identified, and in the **same relative order**.

And if the **most frequent letters** are **similar** in the text files of the same literary work in **different languages**.

For this you must:

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

b) Estimate the **n most frequent letters**, running your data stream algorithm for **n = 3, n = 5** and **n = 10**.

c) Perform a set of tests, **repeating the approximate counts a few times**.

b) **Compare the performance** of the approximate counters and the data stream algorithm (for the different k values), between themselves and regarding the exact counts.

c) Write a report (max. 8 pages).

Data for the computational experiments – Simulating data streams

Obtain **text files from literary works**, in **different languages** – e.g., from the Project Gutenberg.

Process the text files to:

- Remove the Project Gutenberg file headers.
- Remove all stop-words and punctuation marks.
- Convert all letters to uppercase.

J. Madeira, December 2023