Departamento de Electrónica, Telecomunicações e Informática da Universidade de Aveiro

Algoritmos Avançados

2021/2022 — 1º Semestre

2nd Project — Approximate Counting

Deadline: January 4, 2022

The goal is to count the **number of occurrences of letters in text files** and, for instance, identify the most common ones.

It should be possible to count letters using three types of counters: **exact counter, approximate counter with fixed probability** and **approximate counter with decreasing probability**.

An analysis of the computational efficiency and limitations of the developed counters has to be made. For this you must:

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

b) Compare the performance of the approximate counters.

c) Write a report (max. 6 pages).

Tasks

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

Remove the Project Gutenberg file headers and convert all letters to block (capital) letters.

Obtain the **number of occurrences** of all the **distinct letters** that appear in each of the text files.

The results of the exact counters should be compared with the **estimated counts** obtained from the values registered in the approximate counters.

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 approximate counts identify or not the same most frequent letters, and in the same relative order.

And if the most frequent letters are similar or not in the text files of the same literary work in different languages.

J. Madeira, December 2021