Departamento de Electrónica, Telecomunicações e Informática da Universidade de Aveiro

Algoritmos Avançados

2021/2022 — 1º Semestre

3rd Project

Deadline: February 6, 2022

As in the previous project, each student will be assigned one of the following methods.

Check your assignment on the corresponding PDF file.

Data for the computational experiments – Simulating data streams

Obtain **text files from different editions of the same 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 lowercase.

A – The Most Frequent Words

Determine the most frequent words of each one of your text files. Compare the results obtained with the exact counts.

Use the method assigned to you. Implement it (Python 3) and analyze its behavior:

– Misra & Gries – **FREQUENT-COUNT**

– Manku & Motwani – **LOSSY-COUNT**

– Metwally et al. – **SPACE-SAVING-COUNT**

– **Count-Min Sketch** – at first, use a fixed number of hash functions, for example 5

Analyze the behavior your method when you change some of its parameters.

What is the influence of those changes on the results of the computational experiments?

B - Number of Distinct Words

Estimate the number of distinct words in each one of your text files. Compare the results obtained with the exact counts.

Use the method assigned to you. Implement it (Python 3) and analyze its behavior**:**

– **Simplified Hash Table**, no collision resolution

– **Bloom filter** – at first, use a fixed number of hash functions, for example 5

Analyze the behavior of your method when you change the size of the table/filter.

What is the influence of those changes on the results of the computational experiments?

J. Madeira, January 11, 2022