TECHNOLOGICAL FUNDAMENTALS IN THE BIG DATA WORLD

PROTEIN MATCHING IN PYTHON

Lab 1

Description

- To create a program to match a pattern introduced using the keyboard against all the proteins in the file provides using Python.
  - File: proteins.csv

- You are asked to:
  - Write the serial version of the program
  - Write the parallel version of the program using multiprocessing and threads
  - Write a memory explaining your results (maximum 12 pages)

Serial program

- Make a program in Python that:

1. Reads the pattern to search from the keyboard.
2. Changes the pattern to UPPERCASE
3. Start metering exec time
4. Reads the protein patterns from the file in pair (id, sequence)
5. Look for occurrences of the pattern string inside each protein sequence
6. If there are occurrence, register the id of the protein and the number of occurrences in the sequence
7. Print a histogram using protein id as X and number of occurrences as Y (you can use matplotlib.pyplot). Show 10 proteins with more matches
8. Print the protein id with max occurrences.

Data set generation

- Run the command:

$> python protein-generator.py row_no

For development:

row_no 50000

For testing and delivery:

row_no 500000

Parallel program

- Multiprocessing version
  1. Write a parallel version of you program using multiprocessing
  2. Measure the time and optimize the program to get the fastest version you can.

- Threaded version
  1. Write a parallel version of you program using threads
  2. Measure the time and optimize the program to get the fastest version you can.

Rules

- Make the lab in groups maximum 3

- To deliver:
  1. Python programs with all versions
  2. Written report

- Program names:
  - serial-proteins.py
  - mp-proteins.py
  - th-proteins.py".

- Deadline: **October 18th 2020. 23:30 hours.**