Tema para Dissertação do 2º ciclo em Engenharia Informática

Título: Processamento Paralelo de Dados: Análise dos testes de seriação para acesso ao ensino superior no Brasil.

Orientador: Paula Prata
(e-mail: pprata@di.ubi.pt)

Co-orientador: Maria Eugénia Ferrão (Departamento de Matemática)

Contexto

Numa sociedade que produz cada vez maiores quantidades de dados, a análise desses dados de forma a deles retirar conhecimento útil é crucial. A análise de grandes quantidades de dados exige grande capacidade de processamento, sendo a programação paralela uma via óbvia para acelerar o processo.

Objetivos

Pretende-se explorar mecanismos de programação paralela numa linguagem tradicionalmente usada para análise de dados, como por exemplo *Python*, e explorar *frameworks* de programação distribuída como o *Apache Spark*.
Usando dados do Exame Nacional do Ensino Médio (ENEM) [1] do Brasil, com cerca de 8 milhões de registos, pretende-se calcular várias estatísticas usando o método *bootstrap* [2], [3]. O *bootstrap* consiste numa abordagem de simulação intensiva em que amostras – réplicas são extraídas de uma amostra original com a finalidade de obter a distribuição que permita a inferência populacional.

Tarefas

T1 – Estudar a linguagem Python e respectivos mecanismos de paralelismo.
T2 – Estudar a framework Apache Spark.
T3 – Estudar o método de bootstrap e variantes para grandes quantidades de dados.
T4 – Implementação de algoritmos paralelos de bootstrap para algumas estatísticas simples.
T5 – Testes com amostras de dados de grande dimensão (Dados do ENEM)
T6 – Escrever a dissertação.

Cronograma de Tarefas

Set 18	Out 18	Nov 18	Dez 18	Jan 19	Fev 19	Mar 19	Abr 19	Mai 19	Jun 19
T1	T1								
		T2	T2						

			T3	T3	T3				
					T4	T4			
						T5	T5		
						T6	T6	T6	T6

Referências

[1] ENEM, https://enem.inep.gov.br/#/crono?_k=kltpmt.

[2] B. Efron and R. Tibshirani, An Introduction to the Bootstrap. Boca Raton, FL: Chapman & Hall/CRC, 1993.

[3] A Scalable Bootstrap for Massive Data,
Ariel Kleiner, Ameet Talwalkar, Purnamrita Sarkar, Michael I. Jordan
Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2014.