

Tema para Dissertação do 2º ciclo em Engenharia Informática

Título: Processamento Paralelo de Dados em Spark R: 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 produção de conhecimento útil a partir da análise desses dados tem de extrema importância. 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 usar uma framework de programação paralela, o Apache SparkR para analisar os dados do Exame Nacional do Ensino Médio (ENEM) [1] do Brasil, com cerca de 8 milhões de registos. O método *bootstrap* [2] [3] e variantes serão usados para calcular várias estatísticas descritivas. 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. Num contexto de grande volume de dados (Big Data) foi introduzida a variante bag-of-little bootstraps [4] que permite reduzir consideravelmente o tempo de execução. Estudar-se-á o desempenho de implementações distribuídas de ambas as técnicas, analisando a validade dos resultados.

Tarefas

- T1 – Estudar a linguagem R e respectivos mecanismos de paralelismo.
- T2 – Estudar a framework Apache SparkR.
- 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 e de um artigo para divulgação de resultados.

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	T2						
			T3	T3	T3				
					T4	T4			
						T5	T5		
						T6	T6	T6	T6

Referências

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

[2] <https://fullstackml.com/how-to-check-hypotheses-with-bootstrap-and-apache-spark-cd750775286a>

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

[4] 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.