Tópicos especiais em Estatística II

Computação em Estatística para dados e cálculos massivos

1. Identificação

Disciplina: Tópicos especiais em Estatística II
Código: PPGEST0024
Carga horária: 60h
Período: 2/2022
Local: Lab. 03 (Prédio CIC/EST)
Horário: segundas e quartas-feiras (14h - 16h)
Professor: Guilherme Souza Rodrigues
Sala: Dep. de Estatística - CIC/EST - A1-45/28
Email: guilhermerodrigues@unb.br

2. Objetivos

Introduzir conceitos, métodos e ferramentas para o processamento de grandes conjuntos de dados e para a realização de operações computacionalmente intensivas. Espera-se que, ao final do curso, o aluno esteja apto a manipular e analisar dados (de forma paralela e distribuída) que estejam armazenados em um servidor.

3. Conteúdo programático da disciplina

Pacotes de extração e manipulação de dados projetados com foco em eficiência computacional e escalabilidade (por exemplo, *data.table, vroom, future, furr, dtplyr e dbplyr*); *Apache Spark*; Plataformas de computação em nuvem (*Google Compute Engine, RStudio cloud, Google Colab*); Integração de *R* e *Python*; Tópicos adicionais (*H2O, Databricks*, entre outros).

4. Critério de avaliação

Os alunos realizarão trabalhos, em formato de listas de exercícios, durante o semestre (peso de 80%). Haverá ainda uma apresentação oral no final do curso (peso de 20%). A menção será atribuída de acordo com os padrões da UnB. Trabalhos entregues fora do prazo estabelecido não serão corrigidos.

5. Bibliografia

Básica:

- Luraschi, J., Kuo, K. & Ruiz E.; Mastering Spark with R. O'Reilly Media, 2019. (Disponível gratuitamente online em https://therinspark.com/)
- Wickham, H. & Grolemund, G.; R for Data Science: Import, Tidy, Transform, Visualize and Model Data. O'Reilly Media, 2016. (Disponível gratuitamente online em https://r4ds.had.co.nz)
- Kim, J. & Bengfort, B.: Interactive Spark using PySpark. O'Reilly Media, 2016.

Complementar:

- Rizzo, M. L; Statistical Computing with R. CRC Press, 2007.

6. Informações adicionais:

Como parte da disciplina, analistas que utilizam *Spark* rotineiramente em sua atuação profissional (na iniciativa privada ou em instituições públicas) serão convidados para apresentar palestras sobre a infraestrutura computacional em que operam. Isso trará uma grande oportunidade aos alunos de melhor entender como o trabalho do Estatístico é realizado fora da academia.

Os exemplos em sala serão apresentados usando o pacote *R*. Entretanto, alunos com maior familiaridade em *Python* poderão, a seu critério, usar esta linguagem para a realização das listas de exercício.

Os alunos deverão se matricular na equipe da disciplina na plataforma Microsoft Teams.

Bom semestre a todos!