Tópicos Especiais SI

Fundamentos de Ciência de Dados

PROF SERGIO SERRA E JORGE ZAVALETA

{SERRA, JORGE.ZAVALETA} @PPGI.UFRJ.BR

2024.2

Datas Importantes

CALENDÁRIO DE ATIVIDADES ACADÊMICAS PARA 2024

Aprovado na Sessão do CEPG de **10/11/2023**
Aprovado na Sessão do CONSUNI de **14/12/2023 (Resolução CONSUNI/CET Nº 126/2023)**

Atos Acadêmicos no SIGA - Calendário Semestral	**1º período**	**2º período**
Início de atividades	11/03/2024	12/08/2024
Rematrícula de matrícula trancada (destrancamento de matrícula)	Até 08/03/2024	Até 09/08/2024
Previsão de turma	Até 23/02/2024	Até 26/07/2024
Trancamento de matrícula	Até 29/03/2024	Até 30/08/2024
Pedido de inscrição em disciplinas	**De 24/02/2024 a 05/03/2024**	**De 27/07/2024 a 06/08/2024**
Concordância do pedido de inscrição em disciplina	De 06/03/2024 a 07/03/2024	De 07/08/2024 a 08/08/2024
Efetivação do Pedido de Inscrição (Divisão de Ensino – PR2)	08/03/2024	09/08/2024
Pedido de alteração de inscrição em disciplina	De 09/03/2024 a 12/03/2024	De 10/08/2024 a 13/08/2024
Concordância do pedido de alteração de inscrição em disciplina	De 13/03/2024 a 14/03/2024	De 14/08/2024 a 15/08/2024
Efetivação de Alteração do Pedido de Inscrição (Divisão de Ensino – PR2)	15/03/2024	16/08/2024
Trancamento do pedido de inscrição (desistência de inscrição)	De 16/03/2024 a 19/03/2024	De 17/08/2024 a 20/08/2024
Concordância do pedido de trancamento de inscrição	De 20/03/2024 a 21/03/2024	De 21/08/2024 a 22/08/2024
Efetivação do Trancamento do Pedido de Inscrição (Divisão de Ensino – PR2)	22/03/2024	23/08/2024
Término de atividades	20/07/2024	14/12/2024
Notas – Pautas de graus e frequência	De 21/07/2024 a 20/08/2024	De 15/12/2024 a 14/01/2025

Programa

Terças das ~ 13:30 até ~17:00 Teórico–práticas
Lab NCE e Google Meet (excepcionalmente)

Módulo 1:

1. O que É data science?
2. Reprodutibilidade em Pesquisa Computacional
3. Introdução a Proveniência de Dados
4. Gestão de Grandes Volumes de Dados de Pesquisa
5. Ambiente de Programação: python 3, jupyter notebook, JupyterLab, Google Colab, DeepNot pacotes e github
6. Python I: tipos de dados, sequências e operações, estruturas de controle e repetição
7. Prática dos conteúdos estudados: construindo e operando listas e strings (básico)

Módulo 2:

1. Técnicas de coleta e preparação de dados
2. Numpy I: array, slicing, fancy index, copy and view
3. Pandas I: dataframes, series, index, Pandas I/O (csv, json, excel)
4. Prática dos conteúdos estudados: Processando e extraindo informações de arquivos csv, Jason, rdf

Módulo 3:

1. Técnicas de análise de dados
2. Numpy II e Matplotlib: operações com array, broadcasting, construção de gráficos usuais
3. Pandas II: estatísticas básicas
4. Prática dos conteúdos estudados: manipulando dados de saúde, ambiente, agricultura, cidades inteligentes

Módulo 4:

1. Introdução a técnicas de modelagem de fluxo de dados
2. Algoritmos e técnicas de extração inteligente de conhecimento
3. Scikit learn: introdução a mecanismos de regressão, classificação, clustering e PCA
4. Prática dos conteúdos estudados: clusterização e predição

Módulo 5:

1. *Seminários sobre Ciência de Dados aplicados domínio específicos (e.g. Saúde, Educação, Sustentabilidade, Agricultura, Cidades Inteligentes, COVID-19, entre outros)*
2. *Apresentação de trabalhos + artigos*

Avaliação e Atendimento

Critérios de aprovação são os do PPGI/UFRJ.

A avaliação da disciplina consiste em participação em sala de aula (P); protótipos de DS desenvolvidos com boas práticas (E); apresentações e elaboração de Dataset/Executable Paper (A).

$$MF = 0.1 * P + 0.4 * E + 0.5 * A$$

O aluno que desejar atendimento deverá requisitar o mesmo por e-mail e um horário será agendado pelos responsáveis para o atendimento.

serra@ppgi.ufrj.br jorge.zavaleta@ppgi.ufrj.br

Bibliografia

Materiais apresentados em sala de aula + extas

1. National Academies of Sciences, Engineering, andMedicine. Reproducibility and Replicability in Science. Washington, DC: The National Academies Press, 1st Edition, 2019.
2. Victoria Stodden, Friedrich Leisch, Roger D. Peng, Implementing Reproducible Research, CRC Press, 1st Edition, 2014.
3. Kleppmann, M., Designing Data-Intensive Applications: The Big Ideas Behind Reliable, Scalable, and Maintainable Systems, O'Reilly, 2017.
4. Taylor, E. Deelman, D.B. Gannon, M. Shields (Eds.), Workflows for e-Science: Scientific Workflows for Grids, Springer, 2006.
5. Wes McKinny, Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython, 2nd edition O'Reilly Media, 2017
6. Mark Lutz, Learning Python, 5th Edition, O'Reilly Media, 2013
7. Jonh Hearty, Advanced Machine Learning with Python. Packt Publishing, 2016.
8. Andreas C. Mueller and Sarah Guido, Machine Learning with Python. O'Reilly Media, 2016.
9. John D. Kelleher, Brian Mac Namee, and Aoife DArcy. Fundamentals of machine learning for predictive data analytics: algorithms, worked examples, and case studies. MIT, 2015.
10. Bibliografia completa no github da disciplina **https://github.com/zavaleta/Fundamentos_DS**
11. Artigos ou apresentações selecionados