



Python for Data Engineer

Aula 1 - Apresentação da Disciplina e Introdução

Leandro Mendes Ferreira

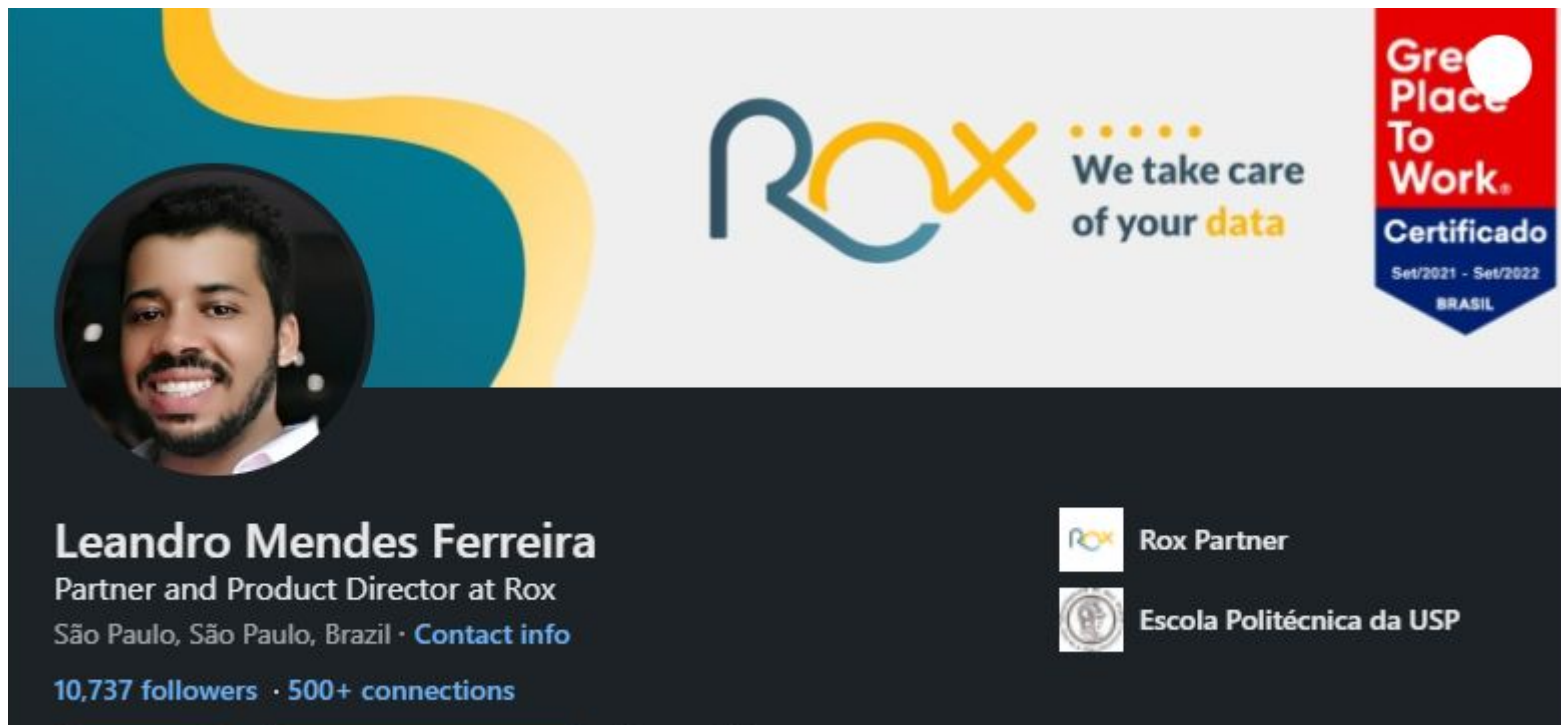
leandro.ferreira@faculdadeimpacta.com.br

Agenda

- Objetivo
- Proposta
- Definição do Projeto
- Cronograma
- Bibliografia
- Modelo de Avaliação
- Composição da Nota

Leandro Mendes Ferreira

- LinkedIn : <https://www.linkedin.com/in/leandroimail/>



The banner features a circular profile picture of Leandro Mendes Ferreira on the left. To the right of the photo is the Rox logo with the tagline "We take care of your data". Further right is a "Great Place To Work" certification badge for Brazil, valid from September 2021 to September 2022. Below the profile picture, the name "Leandro Mendes Ferreira" is displayed, followed by his title "Partner and Product Director at Rox" and location "São Paulo, São Paulo, Brazil". A "Contact info" link is also present. At the bottom left, it shows "10,737 followers" and "500+ connections". On the bottom right, there are two logos: "Rox Partner" and "Escola Politécnica da USP".

Objetivos

Capacitar o aluno na linguagem de programação Python para que ele consiga analisar, planejar e desenvolver programas, pipelines, scripts, APIs e outros recursos necessários no processo de Engenharia de Dados

Proposta

- Disciplina totalmente prática
- Método ativo de ensino e aprendizagem
 - Foco no aluno não no professor
 - Atividades são o centro da metodologia
 - Aprendizado por erro e acerto
- Resolver problemas “reais”
- Professor como facilitador, mentor e avaliador

Proposta

- Disciplina será dada a partir de projeto único
- Serão divididos os alunos em 4 grupos
- Cada contará com 5 alunos (obrigatoriamente)
- Cada grupo desenvolverá uma parte do projeto a cada aula da disciplina
- ~~• Cada aula serão sorteados dois grupos para apresentação das soluções desenvolvidas~~
 - ~~— Inicialmente cada grupo terá 8 possibilidades de ser sorteado~~
 - ~~— A cada vez que o grupo for sorteado será retirado uma possibilidade~~
 - ~~— Todos os grupos obrigatoriamente deveram apresentar pelo menos uma vez~~
- Cada grupo apresentará a cada aula

Projeto

- Motor para Auxilio de Busca Acadêmica
 - Desenvolvido 100%
 - Objetivo é tratar dados acadêmicos e facilitar aa busca
- Com esse projetos os seguintes pontos serão abordados
 - Tratamento de dados em formatos diversos e desconhecidos
 - Tratamento de arquivos de configuração (JSON, YAML)
 - Leitura e tratamento de dados em CSV
 - Salvamento de dados em diversos formatos (CSV, JSON, PARQUET)
 - Leitura de APIs
 - Criação de APIs
 - Utilização do módulo Pandas
 - Utilização de banco de dados (NoSQL MongoDB)

Cronograma

- 1ª Aula
 - Separação dos grupos
 - Aula teste
 - Desenvolvimento de um script em Python que:
 - Le um arquivo CSV
 - Realiza um sorteio aleatório ponderado de acordo com o descrito no CSV de dois nomes

Cronograma

- 2ª Aula
 - Apresentação da 1ª etapa do projeto que consiste:
 - Ler três arquivos do tipo .bibtex gerados manualmente e previamente
 - Os arquivos Bibtex devem ser exportados dos seguintes sistemas de busca acadêmicas
 - IEEE - <https://ieeexplore.ieee.org/Xplore/home.jsp>
 - Science Direct - <https://www.sciencedirect.com/>
 - ACM - <https://dl.acm.org/>
 - String de Busca deve ser a seguinte:
 - "data quality" AND "big data"
 - Juntar todos os arquivos em tempo de processamento
 - Padronizar os seguintes campos
 - author, title, keywords, abstract, year, type_publication, doi
 - De acordo com um arquivo YAML de configuração , exportar os dados nos seguintes tipos de formato:
 - Json, CSV, YAML, (opcional XML)

Cronograma

- 3ª Aula
 - Será disponibilizado dois arquivos de dados no formato CSV
 - JCR, SCIMAGO
 - Ler os arquivos, tratar os dados, as colunas, padronizar os dois arquivos em um único formato
 - Esse tratamento deve ser obrigatoriamente desenvolvido em Dataframes Pandas
 - Realizar uma operação de Join com os .bibtex já tratados da aula anterior
 - Uma função adicional de tratamento dos dados contidos no bibtex deve ser implementada:
 - Deduplicação dos dados
 - Disponibilizar a filtragem dos dados de acordo com um arquivo de configuração YAML ou um arquivo .py . Os campos que deverão ser possíveis a filtragem devem ser:
 - title, keywords, abstract, year, type_publication, doi, jcr_value, scimago_value
 - Ser possível exportar novamente os dados filtrados nos formatos abaixo:
 - Json, CSV, YAML, (opcional XML)

Cronograma

• 4ª Aula

- Os alunos deve desenvolver uma integração com as API das seguintes bases acadêmicas de buscas:
 - IEEE
 - Science Direct
- Deve ser possível passar a string de busca via arquivo de configuração YAML
- Desta forma será possível realizar o processamento dos dados tanto com arquivos .bibtex quanto através de integração via API
- Os dados agora devem ser armazenados em um banco de dados relacional ou NoSQL.

Cronograma

• 5ª Aula

- Os alunos disponibilizarão uma API para facilitar a realização de um projeto de busca científica
- A API deve receber/persistir um código para cada pesquisa científica realizada.
- A API deve suportar todas as ações desenvolvidas anteriormente
 - Receber um arquivo/texto .bibtex
 - Receber uma string de pesquisa para consulta integrada
 - Realizar filtragem na pesquisa
 - Retornar um arquivo no formato solicitado (JSON)*

– Opcional

- A API pode conter um end-point com os metadados da pesquisa
 - String consultadas
 - Quantidade de Artigos
 - » Total
 - » Por base de busca científica

Bibliografia

Livros On-Line

- Tutorial Python – Guido van Russom - <https://docs.python.org/3/tutorial/>
- Manual de Referência Python – Guido van Russom - <https://docs.python.org/pt-br/3/reference/index.html>
- Aprenda Computação com Python 3.0 - Allen Downey, Jeff Elkner e Chris Meyers - <https://mange.ifrn.edu.br/python/aprenda-com-py3/>
- Python para Desenvolvedores – Luis Eduardo Borges - <https://ricardoduarte.github.io/python-para-desenvolvedores/>

Bibliografia

Livros Impressos

- Introdução à Programação com Python (3ª Edição) - MENEZES, Nilo Ney Coutinho. Introdução à programação com Python—3ª edição: Algoritmos e lógica de programação para iniciantes. Novatec Editora, 2019.
- Python Cookbook - MARTELLI, Alex; RAVENSCROFT, Anna; ASCHER, David. Python cookbook. Novatec Editora, 2005.
- Python Fluente – RAMALHO, Luciano. Python Fluente: Programação clara, concisa e eficaz. Novatec Editora, 2015.
- Pense em Python - DOWNEY, Allen. Think python. Novatec Editora, 2016.
- Data Science do Zero. Primeiras Regras com o Python - GRUS, Joel. Data Science do zero: Primeiras regras com o Python. Alta books, 2019.

Bibliografia

- **Documentação Oficial**

- Python - <https://docs.python.org/pt-br/3/index.html>
- Pandas - <https://pandas.pydata.org/docs/>
- NumPy - <https://numpy.org/doc/stable/>

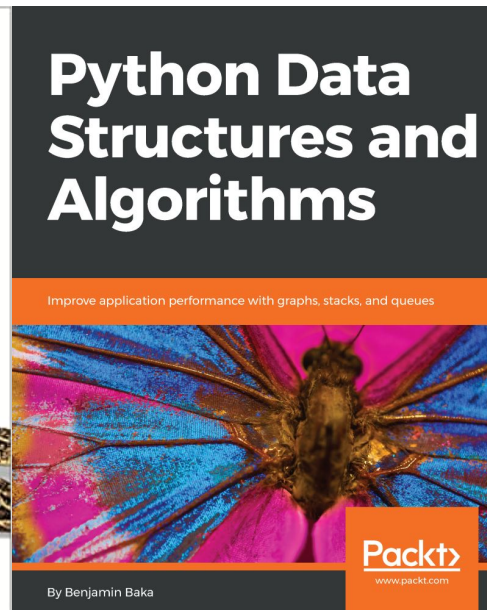
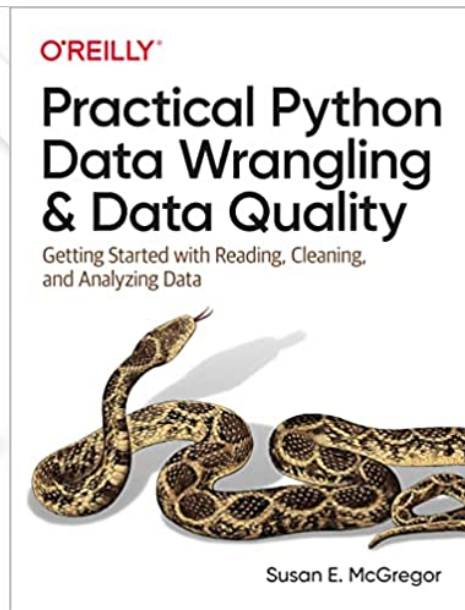
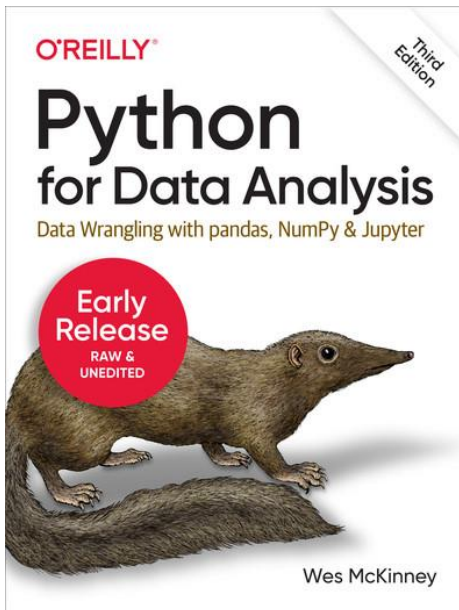
Bibliografia

McKinney, W., 2012. Python for data analysis: Data wrangling with Pandas, NumPy, and IPython. " O'Reilly Media, Inc."

Goodrich, M.T., Tamassia, R. and Goldwasser, M.H., 2013. Data structures and algorithms in Python. John Wiley & Sons Ltd

Kazil, J. and Jarmul, K., 2016. Data wrangling with python: tips and tools to make your life easier. " O'Reilly Media, Inc."

McGregor, S. and Jarmul, K., 2016. Practical Python Data Wrangling and Data Quality. " O'Reilly Media, Inc."



Metodologia de Ensino

- As aulas são divididas em 2 partes:
 - 19h00min às 20h45min
 - Exposição teórica
 - 21h00min às 23h00min
- Podem ser alteradas de acordo com o dinamismo do assunto e da turma
- **O horário de aula é até as 23:00.**

Critérios de Avaliação

- NF = Nota Final (Nota final da disciplina)
- FR = Frequência (Frequência em sala de aula)

Cenário	Resultado
<ul style="list-style-type: none"> • $NF > ou = 7$ e $FR > ou = 75\%$ 	Aprovado
<ul style="list-style-type: none"> • $NF < 7$ ou $FR < 75\%$ 	Reprovado

Composição da Nota

- ATG = Atividades em Grupo
- NF = Nota Final
- Composição da Nota Final:

$$NF = ((ATG1+ATG2+ATG3+ATG4)/4)$$