FORMAÇÃO COMPLEMENTAR

Mentor: Felipe Moreira de Assunção
Conteúdo para formação complementar do Módulo Análise e Visualização de Dados

INFORMAÇÕES DO MÓDULO

Descrição

Conheça os conceitos básicos sobre Análise e Visualização de Dados e a partir de um projeto, análise e visualize dados em uma situação real do mercado de trabalho.

Objetivos do Ensino

Espera-se que o aluno, ao final do módulo, conheça:

- Principais bibliotecas de análise e visualização de dados
- Automatização de tarefas a partir dessas bibliotecas
- Trabalhar com arquivos de dados e fazer operações sobre eles
- Gerar visualização de dados / gráficos para análise e interpretação
- Desenvolver um projeto / Estudo de Caso real que contemple o processo de análise e visualização dos dados

Demonstração das Ferramentas

Utilização do Anaconda com Jupyter Notebook com as principais bibliotecas para Análise e Visualização de Dados: Pandas, NumPy e Matplotlib

BIBLIOTECAS PARA ANÁLISE DE DADOS

1. NumPy

O NumPy é um dos principais pacotes para área de análise de dados. Ele é destinado ao processamento de grandes matrizes e matrizes multidimensionais. Possui uma extensa coleção de funções matemáticas de alto nível e métodos implementados que possibilitam a execução de várias operações com esses objetos.

Ao longo do tempo, um grande número de melhorias vem sendo desenvolvida na biblioteca. Além de algumas correções de bugs e problemas de compatibilidade, algumas das principais mudanças dizem respeito as variedades de estilo, em outras palavras, o formato de impressão dos objetos NumPy. Além disso, algumas funções passaram a poder manipular arquivos de qualquer codificação disponível no Python.

2. SciPy

O SciPy é baseado no NumPy e por isso, estende seus recursos. A principal estrutura de dados SciPy é uma matriz multidimensional, implementada pelo Numpy. O pacote contém

ferramentas que ajudam a resolver álgebra linear, teoria da probabilidade, cálculo integral e muitas outras tarefas.

Ao longo do tempo, o SciPy recebeu grandes melhorias na construção, na forma de integração contínua em diferentes sistemas operacionais, novas funções e métodos e, algo bastante relevante, como os otimizadores atualizados.

3. Pandas

Pandas é uma biblioteca Python que fornece estruturas de dados de alto nível e uma grande variedade de ferramentas para análise. A grande característica deste pacote é a capacidade de traduzir operações bastante complexas com dados em um ou dois comandos. O Pandas contém muitos métodos internos para agrupar, filtrar e combinar dados, bem como a funcionalidade de séries temporais.

4. StatsModels

Statsmodels é um módulo Python que oferece muitas oportunidades para análise de dados estatísticos, como a estimação de modelos estatísticos, a realização de testes estatísticos, etc. Com este pacote, você pode implementar muitos métodos de aprendizado de máquina e explorar diferentes possibilidades de plotagem.

BIBLIOTECAS PARA VISUALIZAÇÃO DE DADOS

1. Matplotlib

O Matplotlib é uma biblioteca de baixo nível para criar diagramas e gráficos bidimensionais. Com este pacote, você pode construir gráficos diversos, desde histogramas e gráficos de dispersão a gráficos de coordenadas não cartesianas. Além disso, muitas bibliotecas de plotagem populares são projetadas para trabalhar em conjunto com o matplotlib.

Houve recentemente mudanças de estilo em cores, tamanhos, fontes, legendas, etc. Recebeu ainda melhorias no alinhamento automático de legendas nos eixos e, entre melhorias significativas de cores, há um novo ciclo de cores compatível com daltônicos.

2. Seaborn

O Seaborn é essencialmente uma API de alto nível baseada na biblioteca Matplotlib. Ele contém configurações padrão mais adequadas para o processamento de gráficos. Além disso,

há uma rica galeria de visualizações, incluindo alguns tipos complexos, como séries temporais, diagramas conjuntos e diagramas de violino.

3. Plotly

Plotly é uma biblioteca popular que permite construir facilmente gráficos sofisticados. O pacote é adaptado para trabalhar em aplicativos da web interativos. Entre suas visualizações notáveis estão gráficos de contorno, gráficos ternários e gráficos 3D.

4. Bokeh

A biblioteca Bokeh cria visualizações interativas e escalonáveis em um navegador usando widgets JavaScript. A biblioteca oferece uma coleção versátil de gráficos, possibilidades de estilo, habilidades de interação na forma de vincular gráficos, adicionar widgets e definir retornos de chamada, além de muitos outros recursos úteis.

O Bokeh possui habilidades interativas aprimoradas, como uma rotação de rótulos categóricos, bem como aprimoramentos de pequenos campos de ferramenta de zoom e de dicas de ferramentas personalizadas.

5. Pydot

Pydot é uma biblioteca para gerar grafos complexos orientados e não orientados. É uma interface para o Graphviz, escrita em Python puro. Com a sua ajuda, é possível mostrar a estrutura dos grafos, que muitas vezes são necessários ao construir algoritmos baseados em redes neurais e árvores de decisão.

LINKS INTERESSANTES

Um tutorial completo para aprender Data Science com Python do Zero: https://www.vooo.pro/insights/um-tutorial-completo-para-aprender-data-science-com-python-do-zero/

BIBLIOGRAFIA

- Data Science Para Negócios
- Estatística: O que é, para que serve, como funciona

- Data Science do Zero: Primeiras Regras com o Python
- Python Para Análise de Dados: Tratamento de Dados com Pandas, NumPy e IPython