IESP.UERJ
Instituto de Estudos Sociais e Políticos

Escola de Inverno em Ciências Sociais

Introdução à Ciência de Dados

Matheus Pestana
Docente

Ementa

O curso que aqui apresentamos tem por finalidade imergir o estudante no vasto domínio da Ciência de Dados, propondo uma abrangente compreensão dos conceitos fundamentais e das ferramentas mais utilizadas neste campo de conhecimento.

Os tópicos abordados no decorrer do curso envolvem a discussão de algoritmos, estruturação de modelos, análise de fluxos de dados, aprendizado de máquina (*Machine Learning*), aprendizado profundo (*Deep Learning*), processamento de grandes volumes de dados (*Big Data*), Inteligência Artificial e Mineração de Texto, todos essenciais para o entendimento profundo da Ciência de Dados.

O curso está estruturado em quatro módulos distintos. O primeiro módulo é uma introdução à Ciência de Dados e sua relevância no mundo moderno. O segundo módulo se dedicará ao *webscraping*, que é a raspagem e obtenção de dados da *Web*. O terceiro módulo se concentra em apresentar as nuances do *Machine Learning* e *Deep Learning* e suas aplicações em diversas áreas. Finalmente, o quarto módulo contempla aplicações de Ciência de Dados em áreas como processamento de imagens, vídeos e áudios, modelos generativos de texto, entre outras, bem como uma discussão sobre o futuro da Ciência de Dados.

Em todas as aulas, empregaremos uma abordagem didática bifocal, combinando aspectos teóricos e práticos. Os fundamentos teóricos serão detalhadamente discutidos, enquanto na parte prática utilizaremos a plataforma Google Colab. Faremos uso de modelos pré-treinados como ferramentas de ensino, facilitando assim, a compreensão dos conceitos e princípios discutidos. Esses modelos não apenas simplificam a assimilação dos conceitos, mas também constituem ferramentas valiosas para a pesquisa científica, abrangendo áreas como processamento de texto, transcrição de áudios, reconhecimento de imagens, entre outras, exigindo apenas um conhecimento básico de programação em Python.

Não há um pré-requisito para o curso, mas é recomendável, em termos de melhor aproveitamento, que o(a) aluno(a) tenha algum conhecimento prévio de programação em Python. Caso não tenha, noções gerais de Python serão apresentadas no decorrer do curso, o que permitirá que mesmo que o(a) aluno(a) sem conhecimento prévio de programação possa acompanhar as aulas e extrair o máximo de seu conteúdo.

Ao fim do curso, o(a) aluno(a) será capaz de:

- Compreender como são realizadas as pesquisas e trabalhos que utilizam métodos sofisticados de Ciência de Dados
- Compreender o fluxo utilizado em Ciência de Dados, da coleta à modelagem
- Coletar, tratar, visualizar e tomar decisões a partir de dados

Programa

Aula 1: Entendendo Ciência de Dados

1.1 O que é Ciência de Dados?

1.2 Dados e suas dimensões

1.3 Algoritmos, Modelos, Data Mining e Big Data

1.4 *Machine Learning*, *Deep Learning* e Inteligência Artificial: conceitos e aplicações

- Prática: Introdução ao Python e ao Google Colab

Aula 2: *Raspagem de dados da Web*

2.1 O que é *webscraping*?

2.2 *Webscraping* com Python: a biblioteca `requests` e `BeautifulSoup4`

2.3 Trabalhando com *APIs*

- Prática: Obtendo dados de redes sociais

Aula 3: Introdução ao *Machine Learning* e *Deep Learning*

3.1 Fluxo de criação de modelos de Aprendizagem de Máquina e Deep Learning

3.2 *Machine Learning* com Python: a biblioteca `scikit-learn`

3.3 *Deep Learning* com Python: a plataforma *HuggingFace*

- Prática: Análise de sentimento de textos

Aula 4: Aplicações de Ciência de Dados

4.1 Processamento de Imagens, Vídeos e Áudios

4.2 Modelos Generativos de Texto

4.3 O futuro da Ciência de Dados

- Transcrevendo áudios a partir do uso de Inteligência Artificial

Recursos

Google Colab

Para as abordagens práticas, será utilizada a plataforma Google Colab, que é uma plataforma de nuvem gratuita que oferece suporte ao Jupyter notebook e ao ambiente de execução em nuvem para máquinas virtuais, executando códigos em Python ou em R. Os notebooks do Colab são armazenados no Google Drive e podem ser compartilhados facilmente com colegas de trabalho ou amigos.

Além disso, através do Google Colab, é possível ter acesso à GPUs e TPUs gratuitamente, o que facilita a execução de tarefas complexas que envolvem inteligência artificial.

O acesso ao Colab se dá através do link: `https://colab.research.google.com/`

Outras opções

É possível também instalar o Python localmente, através do Anaconda, que é uma software de código aberto que visa facilitar o gerenciamento e a implantação do Python e suas bibliotecas. O download se dá através do link `https://www.anaconda.com/download`

Com o Anaconda instalado, é possível utilizar o JupyterLab, o PyCharm ou o VScode. Contudo, a preferência é utilizar o Google Colab dada a sua facilidade de uso.

Referências

AMARAL, F. Introdução à ciência de dados: mineração de dados e big data. Rio de Janeiro: Alta Books, 2016.

GRUS, J. Data Science do zero: primeiras regras com o Python. Rio de Janeiro: Alta Books, 2019.

HAN, J.; KAMBER, M.; PEI, J. Data mining: concepts and techniques. Burlington: Morgan Kaufmann, 2000.

JAMES, Gareth et al. An Introduction to Statistical Learning. New York: Springer, 2013

MCKINNEY, W. Python para análise de dados: tratamento de dados com Pandas, NumPy e IPython. Rio de Janeiro: Novatec, 2019

VASWANI, Ashish et al. Attention is all you need. Advances in neural information processing systems, v. 30, 2017.

ZHOU, Lina et al. Machine learning on big data: Opportunities and challenges. Neurocomputing, v. 237, p. 350-361, 2017.

IESP.UERJ
Instituto de Estudos Sociais e Políticos

Docente

Matheus Pestana

Professor na Escola de Comunicação, Mídia e Informação da Fundação Getulio Vargas (FGV-ECMI), Cientista de Dados no Instituto de Estudos da Religião, e Mestre e Doutorando em Ciência Política pelo IESP-UERJ. Trabalha nas áreas de Partidos e Eleições, Aprendizado de Máquina, Inteligência Artificial e Mineração de Texto.