Educação Profissional Paulista

Técnico em Ciência de Dados

Secretaria da Educação

SÃO PAULO
GOVERNO DO ESTADO

Jargões e inglês da ciência de dados

Vocabulário, em inglês, na ciência de dados (parte II)

Aula 1

Código da aula: [DADOS]ANO1C1B2S10A1

Secretaria da Educação
SÃO PAULO
GOVERNO DO ESTADO

Exposição

Objetivos da aula:

- Conhecer os termos técnicos, em inglês, utilizados na área de ciência de dados e inteligência artificial.

Competências da unidade (técnicas e socioemocionais):

- Aprender a pensar de forma crítica e analítica;
- Trabalhar em equipe em busca de um objetivo;
- Desenvolver *networking*, curiosidade e autonomia;
- Expandir o conhecimento e vocabulário técnico.

Recursos didáticos:

- Recurso audiovisual para exibição de textos, vídeos e imagens;
- Acesso ao laboratório de informática e/ou internet.

Duração da aula:

50 minutos.

Exposição

Cronograma da aula de hoje

Vocabulário, em inglês, na ciência de dados (parte II)

- Introdução de vocabulário técnico em inglês
- Exemplos e seus significados
- Atividade

Exposição

Introdução

A ciência de dados é um campo de estudos multidisciplinar que usa técnicas científicas, métodos, algoritmos e sistemas para extrair conhecimento e *insights* de dados estruturados e não estruturados.

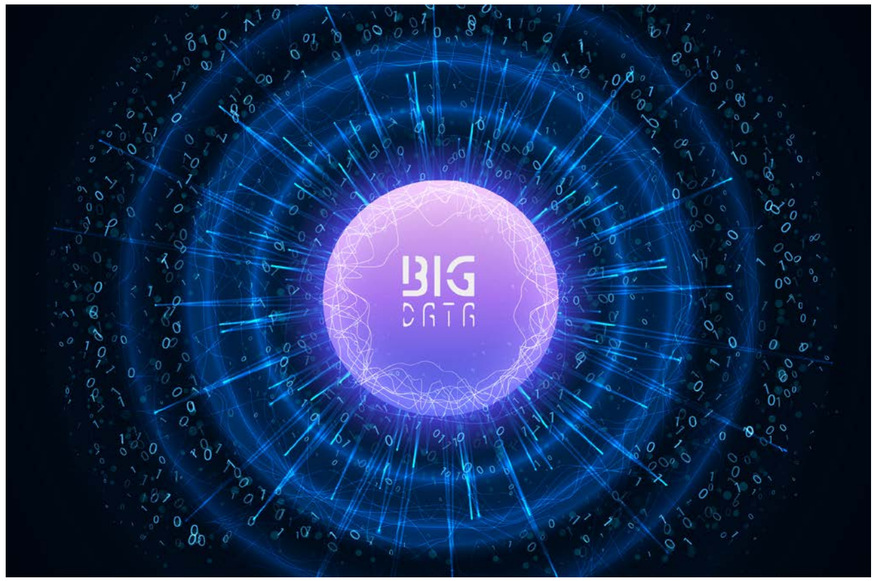
A capacidade de entender e comunicar, eficazmente, essas técnicas e métodos é essencial para qualquer cientista de dados. E uma grande parte disso é o **domínio do vocabulário técnico em inglês**.

Importante

Em geral, para as carreiras técnicas, o **inglês** é o idioma mais importante. Isso ocorre pela facilidade de escrita, leitura, contato (estamos, diariamente, em contato com o idioma!), pela produção acadêmica e diversos outros aspectos.

Exposição

Big data (grandes volumes de dados)



© Getty Images

Big data é um termo que se refere a grandes volumes de dados (tanto estruturados quanto não estruturados), que são tão extensos que os métodos tradicionais de processamento são inadequados para lidar com eles. Podem incluir imagens, textos, vídeos, logs de transações e muito mais

Esses dados são gerados por várias fontes, como:

- ✓ redes sociais;
- ✓ dispositivos móveis;
- ✓ sensores;
- ✓ máquinas.

Exposição

Big data (grandes volumes de dados)

Vamos conhecer os termos em inglês associados ao *big data*:

	Significado
Hadoop	Um *software* de código aberto projetado para armazenamento e processamento de grandes conjuntos de dados distribuídos.
Spark	Uma plataforma de computação em *cluster* rápida, projetada para processamento de *big data*.
Data mining (mineração de dados)	Processo de descoberta de padrões em grandes conjuntos de dados.
Data warehouse (depósito de dados)	Sistema de armazenamento de dados em grande escala que atua como um repositório central para integrar dados de uma ou mais fontes distintas.
ETL (extract, transform, load - extração, transformação, carga)	É o processo que permite que as empresas reformatem e limpem dados, assim como movam-nos de várias fontes e os carreguem em outro banco de dados, data mart ou data warehouse, para análise, ou em outro sistema operacional, para apoiar um processo de negócios.
Data wrangling (preparo de dados)	Processo de transformação de dados similar a ETL, porém com foco maior na limpeza e preparação de dados para análise.

Elaborado especialmente para o curso.

Exposição

Machine learning (aprendizado de máquina)

	Significado
NoSQL (not only SQL)	Um tipo de banco de dados capaz de lidar com grandes volumes de dados estruturados e não estruturados.
Data lake (lago de dados)	Um repositório de armazenamento que contém uma grande quantidade de dados brutos em seu formato nativo.
Data analytics (análise de dados)	O processo de examinar conjuntos de dados para tirar conclusões sobre as informações que contêm.
MapReduce	Modelo de programação para processar e gerar grandes conjuntos de dados com um método paralelo e distribuído.
Hive	Uma infraestrutura de *data warehouse* que fornece consulta e análise de dados armazenados no *hadoop*.
Pig	Uma plataforma de alto nível para criar programas *mapreduce* com o *hadoop*.

Elaborado especialmente para o curso.

Exposição

Machine learning (aprendizado de máquina)

	Significado
HBase	Um banco de dados distribuído e não relacional construído no *hadoop*.
Zookeeper	Serviço de coordenação de alta *performance* para sistemas distribuídos.
Kafka	Plataforma de *streaming* de eventos distribuída, capaz de lidar com trilhões de eventos por dia.
Data scientist (cientista de dados)	Profissional responsável por extrair *insights* e conhecimentos de grandes volumes de dados.

Elaborado especialmente para o curso.

Exposição

Machine learning (aprendizado de máquina)

	Significado
Distributed systems (sistemas distribuídos)	Sistemas de *software* e *hardware* de rede que funcionam em conjunto para um objetivo comum.
Cluster computing (computação em cluster)	Uso de vários computadores conectados para formar um único e poderoso sistema de computação.
Cloud computing (computação em nuvem)	Entrega de recursos de computação, como servidores, armazenamento, bancos de dados, rede, *software*, análise e inteligência, pela Internet (“a nuvem").
Batch processing (processamento em lote)	Um método de execução de tarefas de computação (*jobs*), em que os dados são processados em grandes quantidades em um período de tempo específico.

Elaborado especialmente para o curso.

Exposição

Machine learning (aprendizado de máquina)

	Significado
Real-time processing (processamento em tempo real)	Um método de processamento de dados em que as informações recebidas são processadas e transmitidas quase instantaneamente.
Structured data (dados estruturados)	Dados que são organizados de forma pré-definida, ou que seguem um modelo específico, como dados armazenados em um banco de dados relacional.
Unstructured data (dados não estruturados)	Dados que não possuem uma estrutura ou modelo predefinido, como *e-mails*, *posts* de redes sociais, vídeos, fotos etc.
Semi-structured data (dados semi-estruturados)	Dados que não são totalmente estruturados, mas contêm marcadores para separar elementos de dados e impor hierarquias de registros e campos.

Elaborado especialmente para o curso.

Exposição

Machine learning (aprendizado de máquina)

	Significado
Data cleansing (limpeza de dados)	Processo de detecção e correção (ou remoção) de erros e inconsistências em um conjunto de dados ou base de dados.
Predictive analytics (análise preditiva)	A prática de extrair informações de conjuntos de dados existentes para prever futuras tendências ou comportamentos.
Data privacy (privacidade de dados)	Envolve o tratamento de dados com respeito à sua relevância e seu propósito, garantindo que os dados estejam protegidos contra acesso ou uso não autorizado.
In-memory computing (computação em memória)	Armazenamento de informações na memória principal de um servidor dedicado, em vez de um disco rígido, para produzir tempos de resposta mais curtos.

Elaborado especialmente para o curso.

Exposição

Machine learning (aprendizado de máquina)

	Significado
Scalability (escalabilidade)	A capacidade de um sistema lidar com o aumento da carga de trabalho, adicionando recursos ao sistema.
Data governance (governança de dados)	É o gerenciamento geral da disponibilidade, usabilidade, integridade e segurança dos dados usados em uma empresa. É um conjunto de processos, funções, políticas, normas e métricas que garantem o uso efetivo e eficiente dos dados.
Business intelligence (inteligência de mercado)	Envolve a coleta, integração, análise e apresentação de dados empresariais para ajudar na tomada de decisões. A função de analista de *business intelligence* é a deriva deste conceito.

Elaborado especialmente para o curso.

Vídeo

Vídeo: assista ao vídeo indicado abaixo e saiba mais sobre *big data*

Exposição

SIMPLILEARN. *Big Data in 5 minutes* | What Is Big Data? | Big Data Analytics | Big Data Tutorial | Simplilearn. Disponível em: https://youtu.be/bAyrObI7TYE. Acesso em: 26 fev. 2024.

Vamos fazer uma **atividade**

Leitura de texto em inglês

1. Em duplas, façam a leitura do texto:

   SHAHZAN. *Big data explained in plain and simple english*. Medium, 2 maio 2019. Disponível em: https://medium.com/swlh/big-data-explained-38656c70d15d. Acesso em: 26 fev. 2024.

2. Extraia as principais ideias discutidas pelo autor.

3. Faça um resumo do que entendeu do texto e uma lista de tópicos e seus significados em inglês.

Registro

Vamos fazer um **quiz**

Qual é o foco do processo de *data wrangling*?

- Armazenamento seguro de dados
- Limpeza e preparação de dados
- Criptografia de dados sensíveis
- Distribuição de dados em rede

Registro

Vamos fazer um **quiz**

Qual é o foco do processo de *data wrangling*?

- ✗ **Armazenamento seguro de dados**
- ✓ **Limpeza e preparação de dados**
- ✗ **Criptografia de dados sensíveis**
- ✗ **Distribuição de dados em rede**

RESPOSTA CORRETA!

Data wrangling é o processo de transformação de dados com foco maior na limpeza e preparação de dados para análise.

Registro

Vamos fazer um **quiz**

O que é *hadoop*?

- **Um banco de dados NoSQL**
- **Uma linguagem de programação**
- ***Software* de código aberto para *big data***
- **Uma plataforma de análise de dados**

Registro

Vamos fazer um **quiz**

O que é *hadoop*?

- ✗ **Um banco de dados NoSQL**
- ✗ **Uma linguagem de programação**
- ✓ ***Software* de código aberto para *big data***
- ✗ **Uma plataforma de análise de dados**

RESPOSTA CORRETA!

Hadoop é um *software* de código aberto projetado para armazenamento e processamento de grandes conjuntos de dados distribuídos.

Registro

Vamos fazer um **quiz**

O que significa ETL?

- *Encode, transfer, load*
- *Extract, transform, load*
- *Execute, translate, link*
- *Encrypt, transmit, launch*

Registro

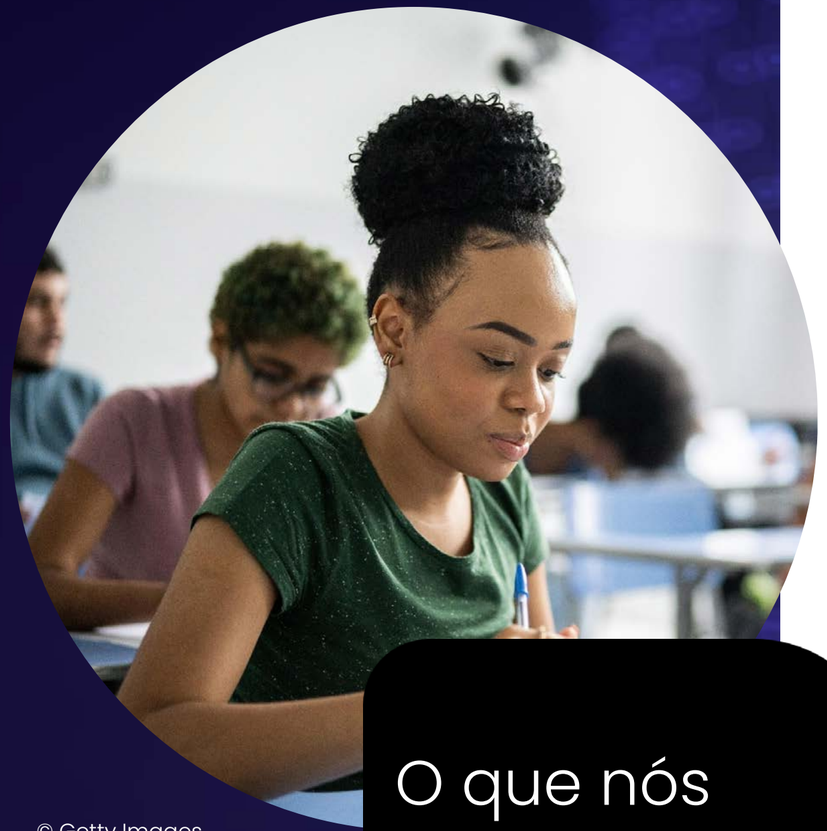
Vamos fazer um **quiz**

O que significa ETL?

- ✗ ***Encode, transfer, load***
- ✓ ***Extract, transform, load***
- ✗ ***Execute, translate, link***
- ✗ ***Encrypt, transmit, launch***

RESPOSTA CORRETA!

ETL é o processo que permite às empresas mover dados de várias fontes, reformatá-los e limpá-los, e carregá-los em outro banco de dados, *data mart*, ou *data warehouse* para análise, ou em outro sistema operacional para apoiar um processo de negócios.



© Getty Images

O que nós **aprendemos?**

Hoje, desenvolvemos:

1. Introdução ao vocabulário técnico em inglês com foco nos termos técnicos utilizados na área de ciência de dados e inteligência artificial.

2. Conhecimento sobre *big data* e *machine learning*, termos mais usados e o que significam.

3. Leitura de artigo em inglês e identificação das principais informações sobre *big data*.

Saiba mais

O *Dicionário do Programador* é o quadro semanal em que você aprende mais sobre termos, tecnologias ou palavras do nosso maravilhoso mundo da programação! Para saber mais sobre *big data*, assista:

CÓDIGO FONTE TV. *Big data* // Dicionário do Programador. Disponível em: https://youtu.be/IpfE8B9H9cI. Acesso em: 26 fev. 2024.

Referências da aula

AMARAL, F. *Introdução à ciência de dados*: mineração de dados e *big data*. Rio de Janeiro: Alta Books, 2018.

CHACON, S.; STRAUB, B. Pro Git. USA: Apress, 2022. Disponível em: https://github.com/progit/progit2-pt-br/releases/download/2.1.46/progit.pdf. Acesso em: 27 fev. 2024.

JUPYTER. *Jupyter project*, [s.d.]. Disponível em: https://jupyter.org/. Acesso em: 27 fev. 2024.

MATPLOTLIB. *Matplotlib documentation*, [s.d.]. Disponível em: https://matplotlib.org/. Acesso em: 26 fev. 2024.

NUMPY. *NumPy documentation*, [s.d.]. Disponível em: https://numpy.org/. Acesso em: 27 fev. 2024.

PYDATA. *Pandas documentation*, [s.d.]. Disponível em: https://pandas.pydata.org/. Acesso em: 27 fev. 2024.

PROVOST, F.; FAWCETT, T. Data science *para negócios*: o que você precisa saber sobre mineração de dados e pensamento analítico de dados. Rio de Janeiro: Alta Books, 2018.

RODRIGO, T. *Mapas mentais*: nosso jeito de raciocinar, desenhando. Medium, 27 jul. 2020. Disponível em: https://tiagorodrigos.medium.com/mapas-mentais-nosso-jeito-de-raciocinar-desenhando-c6cd4d125272. Acesso em: 26 fev. 2024.

SHAHZAN. *Big data explained in plain and simple english*. Medium, 2 maio 2019. Disponível em: https://medium.com/swlh/big-data-explained-38656c70d15d#:~:text=What%20is%20Big%20Data%3F,within%20a%20given%20time%20frame Acesso em: 26 fev. 2024.

SIMPLILEARN. *Big data in 5 minutes | What is big data? | Big data analytics | Big data tutorial.* Disponível em: https://youtu.be/bAyrObl7TYE. Acesso em: 26 fev. 2024.

Identidade visual: Imagens © Getty Images

Educação Profissional Paulista

Técnico em Ciência de Dados

Secretaria da Educação

SÃO PAULO
GOVERNO DO ESTADO