Aulas Resumo – Ciência de dados 1.1 05/05/22

Qual é a diferença entre Big Data e Data Science?

O BigData compõem as técnicas computacionais e armazenamento de dados

A Ciência de dados analisa dados brutos para gerar conhecimento útil

Qual é a definição de Small Data? Cite um exemplo.

Small Data está mais presente no dia, como uma planilha de excell que nos auxilia na administração de uma pequena empresa, a quantidade de dados é muito inferior se comparado ao Big Data.

Nesta unidade, falamos sobre dados tabulares, que são representados como tabelas, e também mencionamos alguns dados que não podem ser apresentados dessa forma. Um deles é o documento de texto. No entanto, após o processamento, um texto pode ser representado como um vetor de características.

Pesquise brevemente na internet algumas técnicas que permitem a realização dessa tarefa.

* [*Word2vec*](https://en.wikipedia.org/wiki/Word2vec);
* [TF-IDF](https://en.wikipedia.org/wiki/Tf%E2%80%93idf)e
* [*Bag-of-Words*](https://en.wikipedia.org/wiki/Bag-of-words_model).
* Coleta- formação da base de documentos ou Corpus (normalmente atraves de robôs de Crawling;
* Pré-processamento- Preparação dos dados atravéns de PNL;
* Indexação- que tem como objetivo fácil/ágil acesso à busca e recuperação da informação;
* Mineração- Cálculos, inferências e extração do conhecimentos, através da mineração de dados;
* Análise- Leitura e interpretação dos dados.

[Rich Morin](https://nc-www5.fgv.br/cursosgratuitos/cg/OCWICDEAD/base_rede/pag/1545_unidade-1-a-revolucao-dos-dados.html) descreve um problema existente em estudos com grandes amostras. Explique que problema é esse.

Quanto maior o tamanho da amostra, menores serão as diferenças para serem estatisticamente significativas, ou seja, é altamente provável que sejam verdadeiramente diferentes umas das outras.

Inicialmente, o Big Data foi definido pelos três Vs: velocidade, volume e variedade. No entanto, com o passar dos anos, outros Vs foram atribuídos a essa área do conhecimento. Cite alguns deles e descreva-os brevemente.

Veracidade - sobre como separar dados verdadeiros e falsos

Valor- quando o volume de dados agrega valor para a organização.

Aulas Resumo – Ciência de dados 1.2

**QUESTÃO 3**

Ainda sobre a seção *Data science, engineering, and data-driven decision making*, no exemplo mencionado sobre o WalMart e a Target, que problema ético eles veem nas suas ações?

Os habitos de consumo dos seus cliente e como influenciá-los, porém ao minerar os dados dos clientes pode haver uma invasão de sua privacidade(LGPD), como no caso da propaganda para uma mulher que o sistema enviava produtos referentes a gravidêz e o processo que o pai dela iniciou.

**QUESTÃO 4**

Na seção *Data science, engineering, and data-driven decision making*, os autores afirmam que, atualmente, estamos vendo uma revolução de propagandas e anúncios. Por que esse fenômeno está ocorrendo?

Porque os sistemas de decisão do merchandising está cada vez mais automatizado

Aulas Resumo – Ciência de dados 1.3 CONHECIMENTOS E HABILIDADES DE UM CIENTISTA DE DADOS

**QUESTÃO 1**

Uma das linguagens de programação amplamente utilizadas na ciência de dados é o *Python*, e alguns dos seus módulos são muito úteis, como *Pandas*, *Scikit-learn* e *Numpy*. Pesquise sobre cada um deles na internet e descreva-os brevemente.

O pandas pretende ser o bloco de construção fundamental de alto nível para fazer análises de dados práticas e do mundo real em Python.

O scikit-learn é uma biblioteca da linguagem Python desenvolvida especificamente para aplicação prática de machine learning.

Rápidos e versáteis, os conceitos de vetorização, indexação e transmissão do NumPy são os padrões de fato da computação de matriz hoje.

* Pandas: usado para manipulação e análise de dados – em particular, fornece estruturas e operações para manipular tabelas numéricas e séries temporais;
* *Scikit-learn*: biblioteca que contém muitos algoritmos de aprendizado de máquina – existem algoritmos para classificação, regressão, agrupamento, etc. e
* *Numpy*: uma das bibliotecas básicas de Python que nos permite manipular matrizes multidimensionais e também possui um grande número de operações matemáticas que nelas operam.

**QUESTÃO 2**

A Ciência de Dados está na interseção de três áreas. Qual(is) delas nos permite(m) movimentar entre diferentes domínios?

Ciência da computação, Matemática e Estatística, conhecimento do negócio

**QUESTÃO 3**

Um modelo de dados é uma relação organizada e formal de dados que geralmente finge simular um fenômeno do mundo real. No exemplo *spawer-recruit models*, descrito na leitura obrigatória (Ozdemir 2016), quais são os elementos ou as variáveis dos dados que estão sendo relacionados?

Na biologia para julgar a saúde biológica da espécie, analisando o numero de pares saudáveis em um grupo distinto, relaciona os grupos de salmons onde uns crescem mais do que outros.

O modelo mede a saúde biológica da espécie. É uma relação básica entre o número de unidades parentais saudáveis de uma espécie e o número de novas unidades no grupo de animais. Formalmente, são as variáveis *recruits* e *spawners*, encontradas por meio da equação a seguir:

*recruits* = 0,5 \* *spawner* + 60

**QUESTÃO 4**

Em muitos casos, não é possível ter um cientista de dados com todas as habilidades desejáveis da área. Se você fosse o dono de uma empresa com esse problema, o que faria?

Formaria uma equipe multidisciplinar onde cada colaborador seria focado em uma das áreas do conhecimento

Esse problema é muito frequente, pois dominar todo conhecimento necessário ou exigido é difícil. Por esse motivo, geralmente, são formados grupos de Ciência de Dados, nos quais cada pessoa tem uma habilidade diferente. Nesse caso, no entanto, é muito importante que todos saibam trabalhar em grupo. Podemos ter os melhores colaboradores, mas nada funcionará se eles não se comunicarem.

17:42 (pausa dos estudos do dia)

continuação 06/04/22

Aulas Resumo – Ciência de dados 1.4 CURIOSIDADES E INQUISIÇÕES (PERGUNTAS) SOBRE DADOS