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

O texto cuja leitura sugerimos a seguir está referenciado na bibliografia sugerida desta disciplina.

The hardest thing In data science. Disponível em: https://buckwoody.wordpress.com. Acesso em: abril 22.

Todo mundo pensa que a Matemática é a parte mais difícil da Ciência de Dados, certo? No entanto, segundo Buck Woody, autor desse artigo, isso não é verdade. O autor afirma que a parte mais difícil é fazer a pergunta certa. Você concorda com ele? Leia o artigo criticamente para responder a essa pergunta.

Sistemas como Azure ML e HDInsight tornam a obtenção de respostas de dados mais rápida do que nunca

### QUESTÃO 1

Ter todos os dados seria o ideal, mas isso nunca é possível. Que alternativa nós temos?

Dividir o problema em questões menores para obter o máximo de resposta e fazer boas perguntas (boa preparação do questionário)

### QUESTÃO 2

No texto, vimos que fazer uma pergunta muito ampla é um problema muito comum. Analise a pergunta a seguir e informe que modificações você faria para torná-la mais precisa.

Por que as nossas vendas caíram?

Levando a analise de pareto, quais os produtos que tiveram queda?

Qual o nicho de produtos que tiveram desempenho menor?

Fizemos campanha de marketing no período analisado?

### QUESTÃO 3

De acordo com o autor, por que fazer a pergunta certa é a parte mais difícil da Ciência de Dados?

Pois perguntas amplas não trazem resultados específicos, e como o público analisado muitas vezes estão impacientes devemos coletar o máximo de informações verídicas para dar valor à pesquisa.

## CONCLUSÃO

Neste módulo, enfocamos os fundamentos da Ciência de Dados a partir da revolução que gerou o movimento *Big Data*. Também abordamos como o uso adequado de dados ajuda-nos a tomar decisões com mais fundamento, pois passamos a baseá-las em fatos e dados, e não apenas em suposições.

No final deste módulo, também conhecemos as habilidades necessárias para ser considerado um cientista de dados.

Aulas Resumo – Ciência de dados 2.0 PROBLEMAS E SOLUÇÕES EM CIÊNCIA DE DADOS

OBJETIVOS

Ao final deste módulo, esperamos que você seja capaz de:

apresentar casos em que a Ciência de Dados foi usada com sucesso e

explicar que fatores são importantes para o sucesso na Ciência de Dados bem como discutir as suas limitações e as estratégias para enfrentar possíveis problemas.

Aulas Resumo – Ciência de dados 2.1 CIÊNCIA DE DADOS NA VIDA REAL

### QUESTÃO 1

De acordo com o primeiro estudo de caso automating government paper pushing (Ozdemir, 2016, p. 16), o computador só conseguiu processar 20% das reclamações. Mencione duas razões pelas quais o computador pode ter falhado nos outros 80%.

Porque o computador era bom em formulários simples enquanto os mais complexos ele deixava a desejar, e quanto ao uso gramatical diferenciado por cada indivíduo o computador não conseguia realizar uma melhor distinção de características.

No segundo estudo de caso marketing dollars, qual dos três meios de comunicação parece ter um comportamento linear melhor e pior?

A meu ver a TV foi o melhor meio pois os pontos no gráfico estão todos próximos à linha de investimento, sendo mais constantes os resultados, o Rádio vem em segundo pois a dispersão dos dados estão muito espalhados e abaixo da linha de investimento e por último o jornal, que parece ter mais pontos abaixo da linha de investimento do que em seu topo.

### QUESTÃO 3

O código descrito no terceiro estudo de caso, que está em uma descrição de trabalho, usa uma classe em Python chamada CountVectorizer. Pesquise sobre essa função na internet e descreva-a brevemente.

É uma ferramenta do Python utilizado para transformar um texto em um vetor de frequência (contagem) de palavras que ocorrem em todo o texto

Aulas Resumo – Ciência de dados 2.2 FATORES DE SUCESSO NA CIÊNCIA DE DADOS

### QUESTÃO 1

Liste, no mínimo, quatro fatores que podem levar a falhas em um projeto de Ciência de Dados.

Decisões que desconsideram evidências claras dos dados, resultados equívocos que não lançam luz em uma direção ou outra, incerteza que impede a criação de novos conhecimentos e quando as medições tem muito ruidos e não geram resultado para um novo conhecimento.

### QUESTÃO 2

Algumas vezes, os dados são mencionados como fatores que podem afetar o sucesso de um projeto. Por que razões os nossos dados podem ser de baixa qualidade?

Pois os questionários foram mal formulados, não compreendem de forma correta o objeto a ser estudado, ou não produziram informação que resultasse em um grande impacto positivo.

### QUESTÃO 3

Indique que tipos de pessoa podem, normalmente, fazer parte de uma equipe de Ciência de Dados, marcando **V** para as que podem e **F** para as que não podem.

| **tipos de pessoas** | **V** | **F** |
| --- | --- | --- |
| gerente de projetos |  |  |
| analista de negócios |  |  |
| gerente de marketing |  |  |
| pessoas que integram dados |  |  |

exceto o gerente de marketing, talvez por ele ser parte interessada?!

### QUESTÃO 4

Uma das técnicas que está sendo usada com ênfase suficiente nos últimos cinco a oito anos, especificamente no contexto da aprendizagem profunda, envolve o conceito de redes neurais. Um cientista de dados deve, portanto, dominar esse conceito. Pesquise na internet e explique, brevemente, o que é uma rede neural.

Redes neurais são sistemas de computação com nós interconectados que funcionam como os neurônios do nosso cérebro. Utiliza-se de algoritmos, para reconhecer padrões e correlacioná-los em dados brutos.

17:23 (pausa dos estudos do dia)