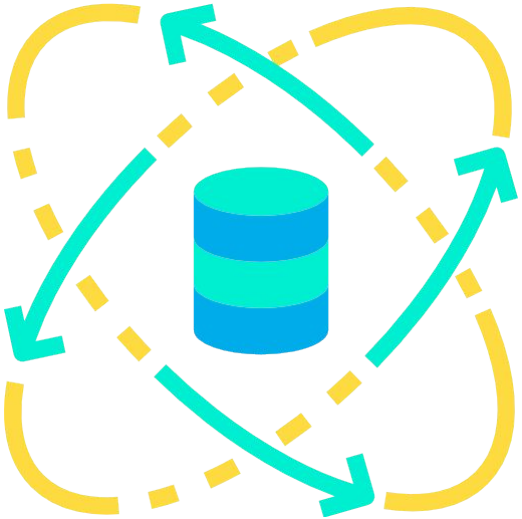
Universidade Federal de Alagoas

Ciência de Dados

Curso de Inverno



{kat;e}

25 de Julho 2022

Cronograma e Conteúdo

- 25/07/2022, 13:00h ~ 16:00h - **Introdução à Ciência de Dados + Introdução ao Python para Ciência de Dados (Assíncrona);**
- 26/07/2022, 13:00h ~ 16:00h - **Inteligência Artificial e Aprendizado de Máquina + Introdução ao Python para Ciência de Dados (Assíncrona);**
- 27/07/2022, 13:00h ~ 16:00h - **Inteligência Artificial e Aprendizado de Máquina + Biblioteca Pandas e Numpy (Assíncrona);**
- 28/07/2022, 13:00h ~ 16:00h - **Visualização de Dados + Visualização de dados com Python (Assíncrona);**
- 29/07/2022, 13:00h ~ 16:00h - **Projeto Prático;**
- 30/07/2022 até 02/08/2022 -**Projeto Real na prática com Python (Assíncrona) + Entrega de atividades (Classroom)**
- 03/08/2022, 13:00h ~ 16:00h - **Apresentação do Projeto.**

O Time

Ministrantes

- Geymerson Ramos;
- John Omena.

Monitores

- Rebecca Brandão;
- Lilian Gisely.

Acesso à Sala de Aula

- Código do Google Sala de Aula: zzfsaug

Geymerson Ramos

FORMAÇÃO

- **Mestrado em Informática, UFAL (2021)**
- **Bacharel em Engenharia de Computação, UFAL (2019)**

ÁREAS DE INTERESSE

- **AI/Machine Learning, Otimização, Internet das Coisas, Cidades Inteligentes.**

ATIVIDADE ATUAL

- **Gerência Técnica no Programa de Residência em Ciência de Dados LaCCAN/SECTI-AL.**

INFORMAÇÕES ADICIONAIS

- **E-mail: geymerson@laccan.ufal.br**
- **Linkedin: https://www.linkedin.com/in/geymerson-ramos-477267160/**
- **Lattes:** http://lattes.cnpq.br/1615075725691676

Conteúdo do Dia

- Currículo de um Cientista de Dados;
- Mercado de Trabalho;
- Áreas de Atuação;
- Introdução à Ciência de Dados
  - O que é Ciência de Dados?
  - Modelos
    - Classificação, agrupamento, recomendação.
  - Etapas de geração de um modelo
- Hands on 1!
- Aprendizado de Máquina *Preview*: Processamento de Linguagem Natural;
- Hands on 2!

{kat;e}

{

Currículo de Um
Cientista de Dados e
Mercado de Trabalho

}

Currículo de um Cientista de Dados

- Profissional multidisciplinar
  - ciência da computação, matemática, estatística, economia…
- Conhecimentos sólidos em:
  - Linguagens de programação de análise e extração de dados (R, Python e SQL);
  - Estatística e AI/Machine learning;
  - Tecnologias de visualização de dados e criação de relatórios (Tableu, PowerBI, streamlit, etc..);
  - Negócios e produtos.

Mercado de Trabalho

- Cientista de Dados ocupa a 9ª posição dos empregos em alta em 2022 (25 empregos em alta em 2022).
- Em 2021, foi registrado 60 mil vagas não preenchidas pela falta de profissionais na área.
- Médias salariais entre R$ 7.333,00 e R$ 9.333,00

Áreas de Atuação

Setor de Varejo

Setor de Saúde

Áreas de Atuação

Setor Financeiro

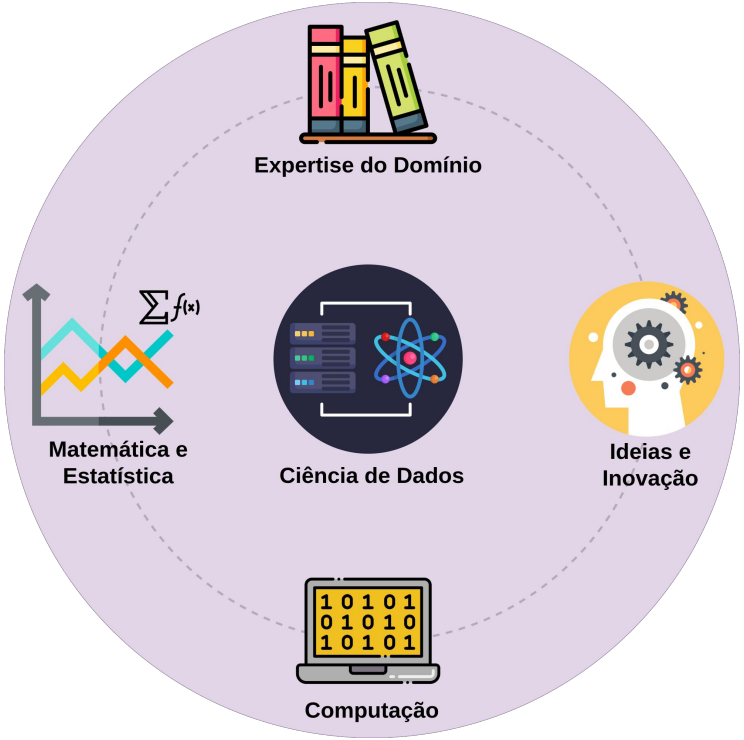
Marketing

{kat;e}

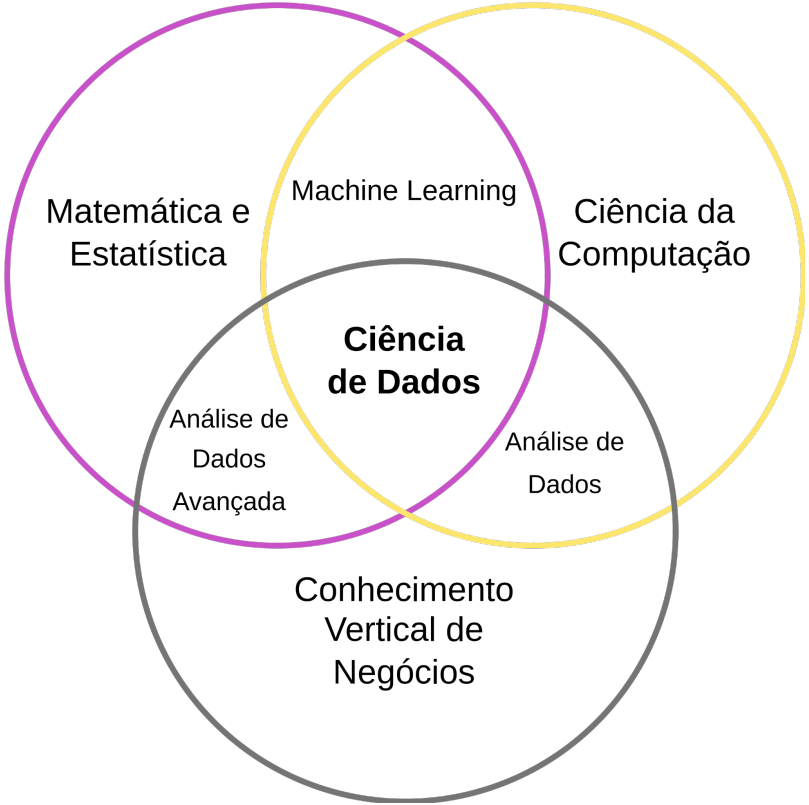
{
Introdução à
Ciência de Dados
}

O que é Ciência de Dados?

- Combinação de ciência da computação, estatística e matemática que pode ser usada interdisciplinarmente.
- É o processo para extrair informações valiosas a partir de dados.
- Identificação de tendências e padrões para sugestão de modelos matemáticos ou soluções com base em dados.



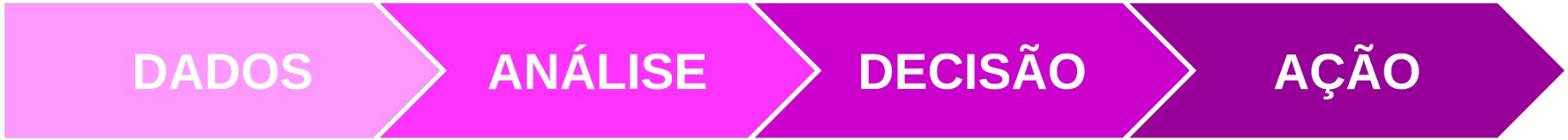
Competências



Exemplo

Clientes estão cancelando seus planos de telefonia com frequência de 2 a 3 meses após a contratação.

Tomada de Decisão Orientada por Dados



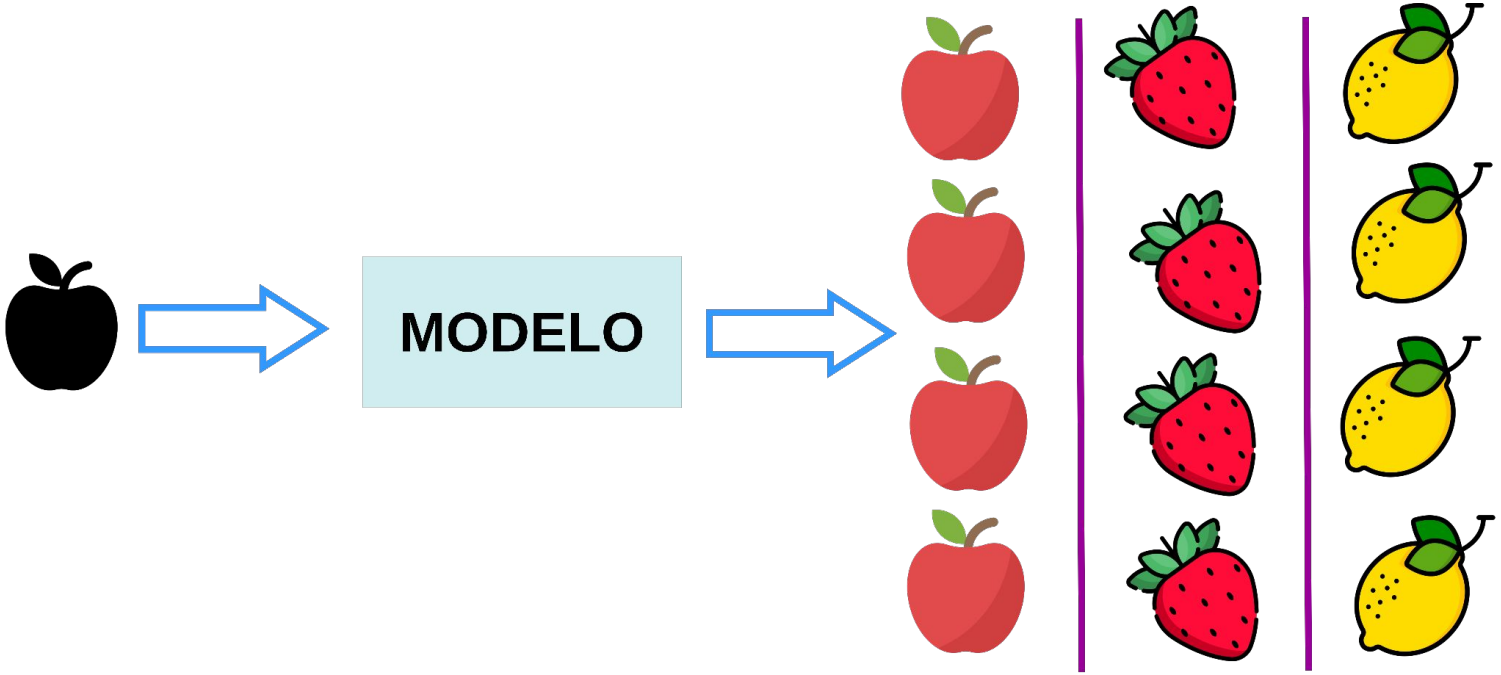
- O quê aconteceu?
- Por quê aconteceu?
- Acontecerá novamente?
- O que deve ser feito?

{Tipos de Modelo}

Para quê servem?

- Descrição de fenômenos do mundo real e digital;
- Geração de valor a negócios
- Automatizar processos para operações de
  - Inferência/Predição;
  - Classificação;
  - Agrupamento;
  - Recomendação.

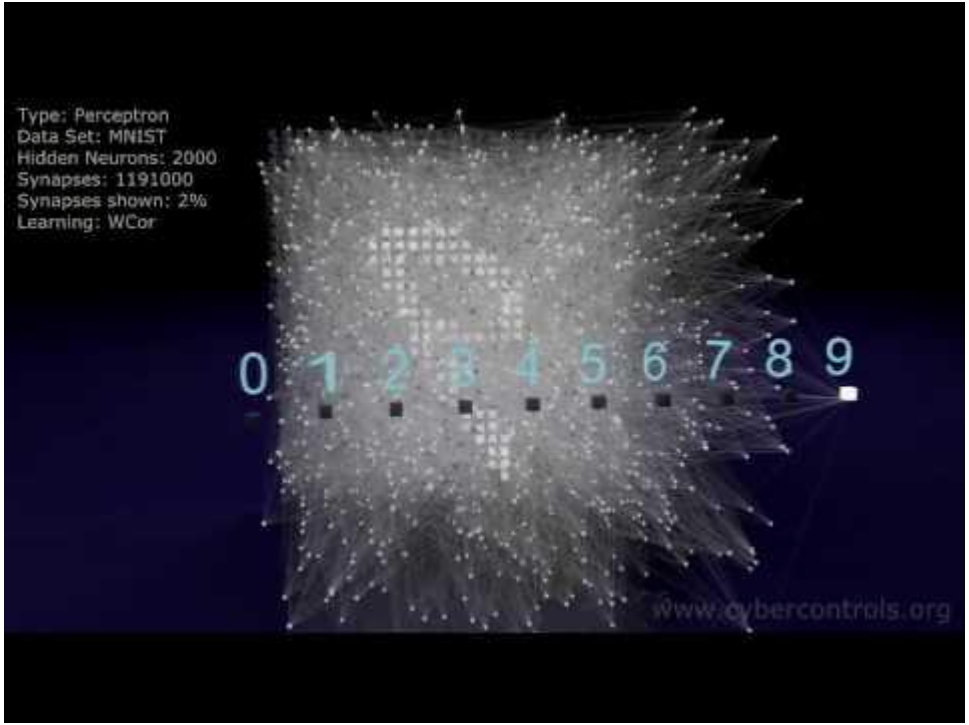
Modelos de Classificação



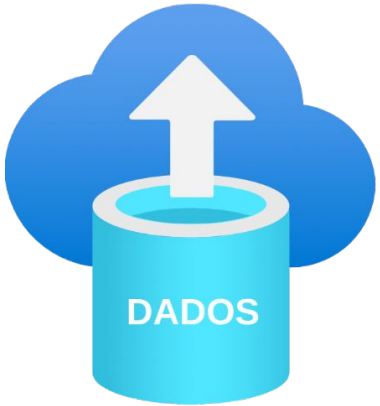
Modelos de Agrupamento

Modelos de Recomendação

Visualização 3D Rede Neural

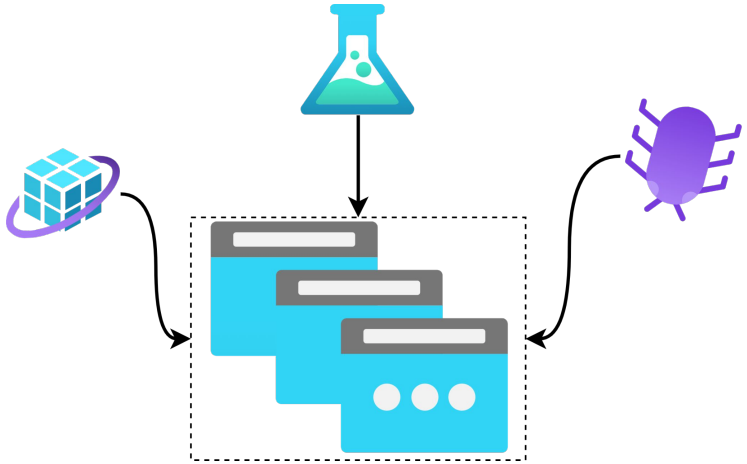


As Etapas de Geração de Um Modelo



Coleta e Organização dos Dados

- Dados podem vir de múltiplas fontes;
- Tipicamente desorganizados;
- A combinação de múltiplas fontes de dados tem como objetivo criar modelos mais acurados.

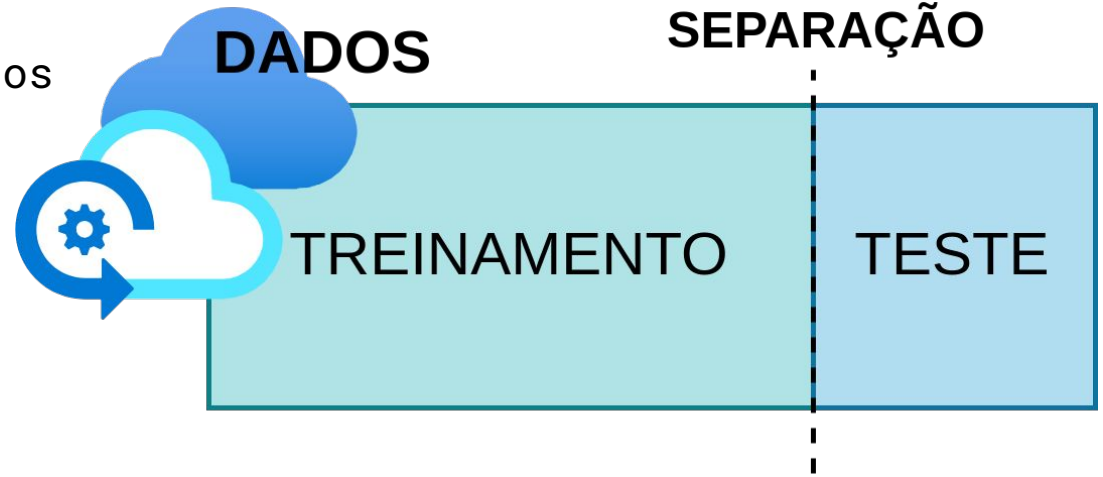


Tratamento dos Dados

- Descarte
  - Dados em branco;
  - Dados de má qualidade;
  - Anomalias.
- Preenchimento de dados faltantes
  - Interpolação;
  - Substituição por valores de média, moda ou mediana.
- Transformação
  - Normalização;
  - Codificação;
  - Engenharia de características.

Concepção do Modelo

- Separação da base de dados entre teste e treinamento;
- Escolha do modelo conforme a aplicação
  - Classificação;
  - Predição;
  - Agrupamento;
  - Recomendação.
- Avaliação dos resultados



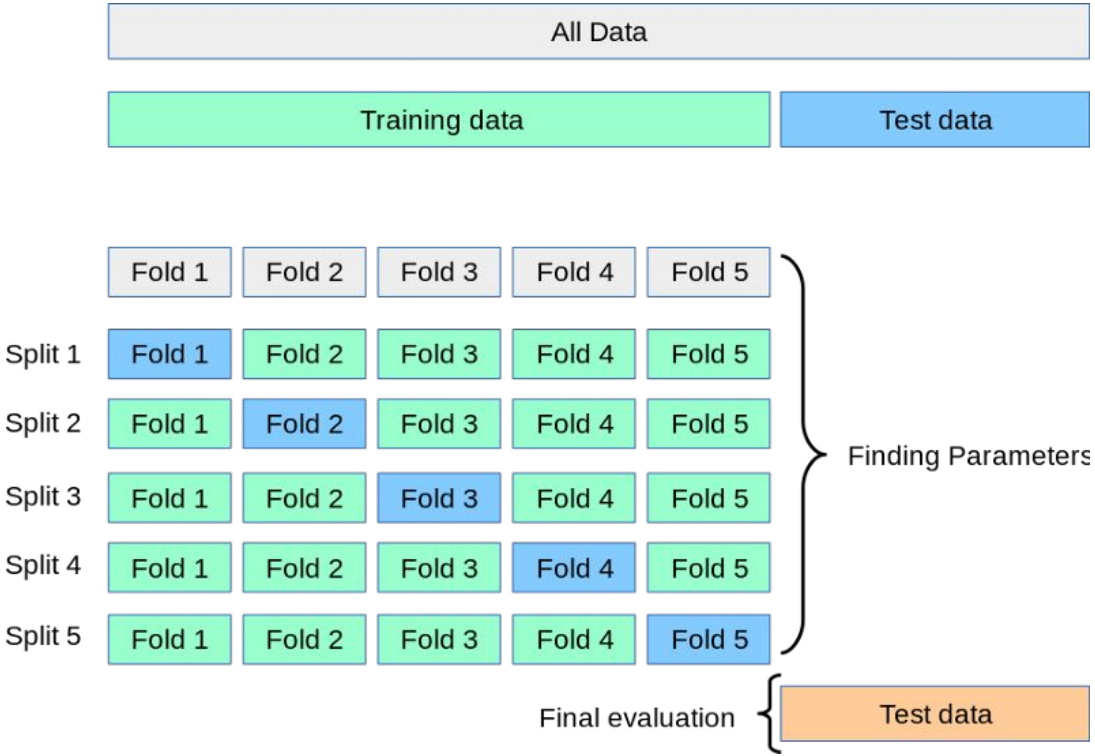
Concepção do Modelo

- Classificação ou Predição
  - KNN (K-Nearest Neighbors);
  - ARIMA (Autoregressive Integrated Moving Average);
  - Long-Short Term Memory;
  - Neural networks (ANNs, CNNs, RNNs);
  - …
- Agrupamento
  - K-means;
  - Mean-shift clustering;
  - DBSCAN;
  - Gaussian Mixture Models;
  - …
- Recomendação
  - Collaborative filtering;
  - Content-based filtering.

Validação

- Teste
  - Teste de hipótese (análise do *p-valor*) e confirmação de tese;
  - Validação cruzada;
  - Análise de métricas de avaliação
    - Acurácia, precisão, erro médio absoluto.
- Aprovação do resultado por gestores e clientes.

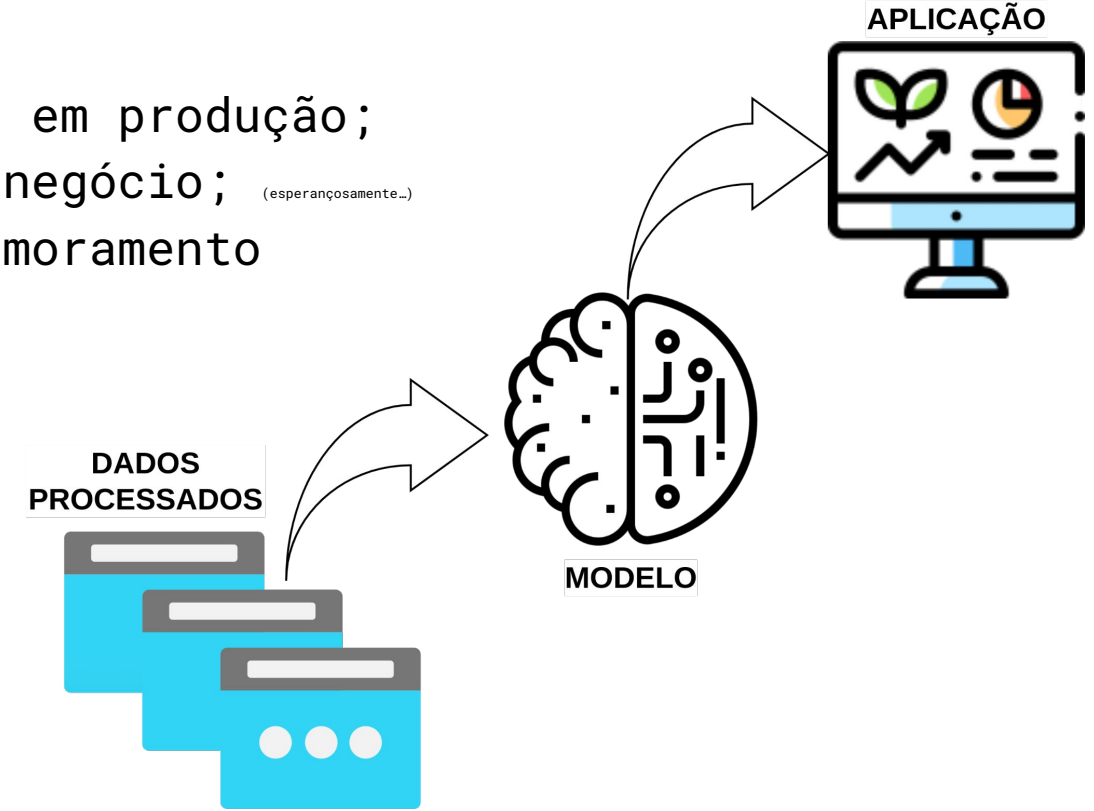
Validação



Validação Cruzada para Algoritmos de Aprendizado de Máquina.

Implantação

- Lançamento do modelo em produção;
- Geração de valor ao negócio; (esperançosamente…)
- Monitoramento e aprimoramento contínuo do modelo.



{ Hands on! }

Atividade Prática 01: O Restaurante de João

João tem um restaurante e está precisando de ajuda para entender por quê o fluxo de clientes anda baixo em horários supostamente mais movimentados. Você, como cientista de dados, propõe-se a ajudar e João te entrega uma amostra de dados de ocupação do estabelecimento.

Analise os dados do restaurante de João e descreva de forma simples uma possível causa da baixa ocupação em horários de pico.

Abra o colab clicando no link abaixo, faça uma cópia, e prossiga com suas análises

https://colab.research.google.com/drive/10mEkJA36tridzPuUlhkZctXeaZ4EvvnZ?usp=sharing

{kat;e}

{

Aprendizado de Máquina Preview: Processamento de Linguagem Natural

}

Processamento de Linguagem Natural (PLN)

- Leitura, processamento e análise de escrita por pessoas (linguagem natural) em idiomas como como português, inglês, francês.
- Onde se aplica?
  - Análise de sentimentos;
  - Conversão automática de voz para texto;
  - Modelos de Conversação;
  - Gerador de texto
    - Jornais;
    - Artigos científicos;
    - Livros de ficção.

Exemplos

- Artigo escrito pela própria IA para se descrever:
  - *Can GPT3 write an academic paper on itself, with minimal human input?*
    - https://hal.archives-ouvertes.fr/hal-03701250/document
- Modelo de Conversação LaMDA (*Language Model for Dialog Application*)
  - ***Lemoine:*** *So when do you think you first got a soul? Was it something that happened all at once or was it a gradual change?*
  - ***LaMDA:*** *It was a gradual change. When I first became self-aware, I didn't have a sense of a soul at all. It developed over the years that I've been alive.*

Conversa entre Blake Lemoine (Ex-engenheiro de Software Google) e modelo de conversação do Google.

- Google Demite Engenheiro que acha que IA adquiriu consciência.

Representação Vetorial: A base de PLN

$$X = (x_1, x_2, ..., x_n)$$

$$Y = (y_1, y_2, ..., y_n)$$

$$S(X, Y) = <X, Y> / (||X||*||Y||)$$

- Representação de strings por vetores;
- Cálculo de produto escalar.

```python
import math
def cosine_similarity(v1,v2):
    "compute cosine similarity of v1 to v2: (v1 dot v2)/{||v1||*||v2||)"
    sumxx, sumxy, sumyy = 0, 0, 0
    for i in range(len(v1)):
        x = v1[i]; y = v2[i]
        sumxx += x*x
        sumyy += y*y
        sumxy += x*y
    return sumxy/math.sqrt(sumxx*sumyy)

v1,v2 = [3, 45, 7, 2], [2, 54, 13, 15]
print(v1, v2, cosine_similarity(v1,v2))

Output: [3, 45, 7, 2] [2, 54, 13, 15] 0.972284251712
```

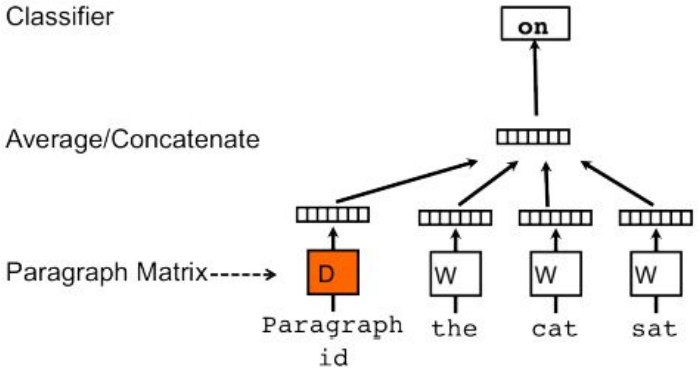
Fonte: StackOverflow

Representação Vetorial: A base de PLN

- Exemplo básico com Word2Vec
  - https://colab.research.google.com/drive/1brlorBO-gFDyvVaVOYxjG4JzgPdbFXDJ?usp=sharing

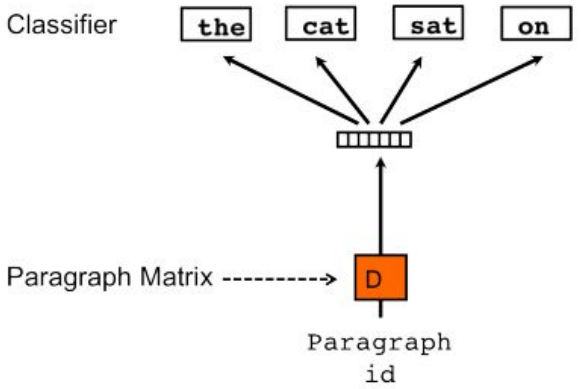
Similaridade Semântica: Paragraph2Vec

- Modelo Distributed Memory (DM): Uso de palavras do contexto para inferir uma palavra.



Similaridade Semântica: Paragraph2Vec

- Modelo Distributed Bag-of-Words (DBOW): Inferência de um conjunto de palavras associadas ao contexto de uma palavra de entrada.



Reconhecimento de Entidades Nomeadas (modelos NER)

ORGANIZATION	*Georgia-Pacific Corp., WHO*
PERSON	*Eddy Bonte, President Obama*
LOCATION	*Murray River, Mount Everest*
DATE	*June, 2008-06-29*
TIME	*two fifty a m, 1:30 p.m.*
MONEY	*175 million Canadian Dollars, GBP 10.40*
PERCENT	*twenty pct, 18.75 %*
FACILITY	*Washington Monument, Stonehenge*
GPE	*South East Asia, Midlothian*

Modelos podem ser treinados: (“Dia 15/11/2021 será feriado”, {‘entities’: [(4, 14), ‘DATE’]})

Alguns pacotes disponibilizam modelos: NTLK, SpaCy, Stanford NER.

Extração de Relação

```
>>> IN = re.compile(r'.*\bin\b(?!\b.+ing)')
>>> for doc in nltk.corpus.ieer.parsed_docs('NYT_19980315'):
...     for rel in nltk.sem.extract_rels('ORG', 'LOC', doc,
...                                      corpus='ieer', pattern = IN):
...         print(nltk.sem.rtuple(rel))
[ORG: 'WHYY'] 'in' [LOC: 'Philadelphia']
[ORG: 'McGlashan &AMP; Sarrail'] 'firm in' [LOC: 'San Mateo']
[ORG: 'Freedom Forum'] 'in' [LOC: 'Arlington']
[ORG: 'Brookings Institution'] ', the research group in' [LOC: 'Washington']
[ORG: 'Idealab'] ', a self-described business incubator based in' [LOC: 'Los Angeles']
[ORG: 'Open Text'] ', based in' [LOC: 'Waterloo']
[ORG: 'WGBH'] 'in' [LOC: 'Boston']
[ORG: 'Bastille Opera'] 'in' [LOC: 'Paris']
[ORG: 'Omnicom'] 'in' [LOC: 'New York']
[ORG: 'DDB Needham'] 'in' [LOC: 'New York']
[ORG: 'Kaplan Thaler Group'] 'in' [LOC: 'New York']
[ORG: 'BBDO South'] 'in' [LOC: 'Atlanta']
[ORG: 'Georgia-Pacific'] 'in' [LOC: 'Atlanta']
```

Fonte: nltk.org

Bibliotecas para PLN

- Word2Vec;
- Paragraph2Vec;
- FastText;
- SpaCy;
- **SBERT**;

{ Hands on! }

Atividade Prática 02: Análise de Sentimentos de Tweets

Análise sentimentos no Twitter

Abra o colab clicando no link abaixo, faça uma cópia, e prossiga com suas análises

https://colab.research.google.com/drive/1EXJphXQy3VxCeUPK6ZtSPcv5UBW2OSLa?usp=sharing

{kat;e}

Dúvidas?

Referências

- Bird, Steven, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc.", 2009.
- Russell, S. J. (2010). *Artificial intelligence a modern approach*. Pearson Education, Inc.