Aula 01 - Minicurso Python

Property	Ítalo Lima

Introdução

O que é Ciência de Dados ?

Antes de iniciarmos o minicurso de Python com ênfase em ciência de dados, é de fundamental importância conhecer um pouco sobre essa área de estudo e a importância da mesma. Em uma busca no Google, é possível obter uma breve definição: "A ciência de dados é um campo interdisciplinar que utiliza métodos, processos, algoritmos e sistemas científicos para extrair valor dos dados. Os cientistas de dados combinam uma série de habilidades, incluindo estatísticas, ciência da computação e conhecimento comercial, para analisar dados coletados da web, smartphones, clientes, sensores e outras fontes, de forma sintetizada a ciência de dados tem como objetivo transformar dados, utilizando matemática e estatística em insights e valores".

Fatores que influenciaram o crescimento da Área

- Era do Big Data.
- Aumento do poderio computacional, o que possibilita maior armazenamento e processamento dos dados.
- Otimização de recursos.
- Amadurecimento em técnicas de aprendizado de máquina.

Dados

Estamos em uma era denominada de Big Data, devido a imensa quantidade de dados que se encontram disponíveis e os valores que os mesmos podem apresentar, há inclusive uma comparação na qual os mesmos são tratados como o "novo petróleo". Os dados revelam padrões, e estes podem ser cruciais na melhora de um produto ou serviço, dessa forma há um aperfeiçoamento na tomada de decisão.

Estima-se que 90% dos dados do mundo tenham sido criados nos últimos dois anos.

Por exemplo, os usuários do Facebook carregam 10 milhões de fotos a cada hora.

Habilidades

Como já foi citado, matemática estatística e programação se apresentam como habilidades cruciais para a atuação de um cientista de dados, mas além disso existem alguns outros conhecimentos e habilidades também importantes, como:

- Criatividades e Atenção
- Capacidade de resolver problemas
- Conhecimento em Banco de Dados
- Conhecimento sobre Data Mining
- Visão de Negócios

Como aplicar os conhecimentos teóricos ?

A linguagem de programação Python se apresenta como uma das principais utilizadas por um cientista de dados, pois com a mesma é possível colocar em prática os conceitos e teorias da estatística e da matemática. Somado à linguagem, existe um conjunto de bibliotecas e ferramentas que auxiliam nas atividades envoltas pelo processo da ciência de dados.

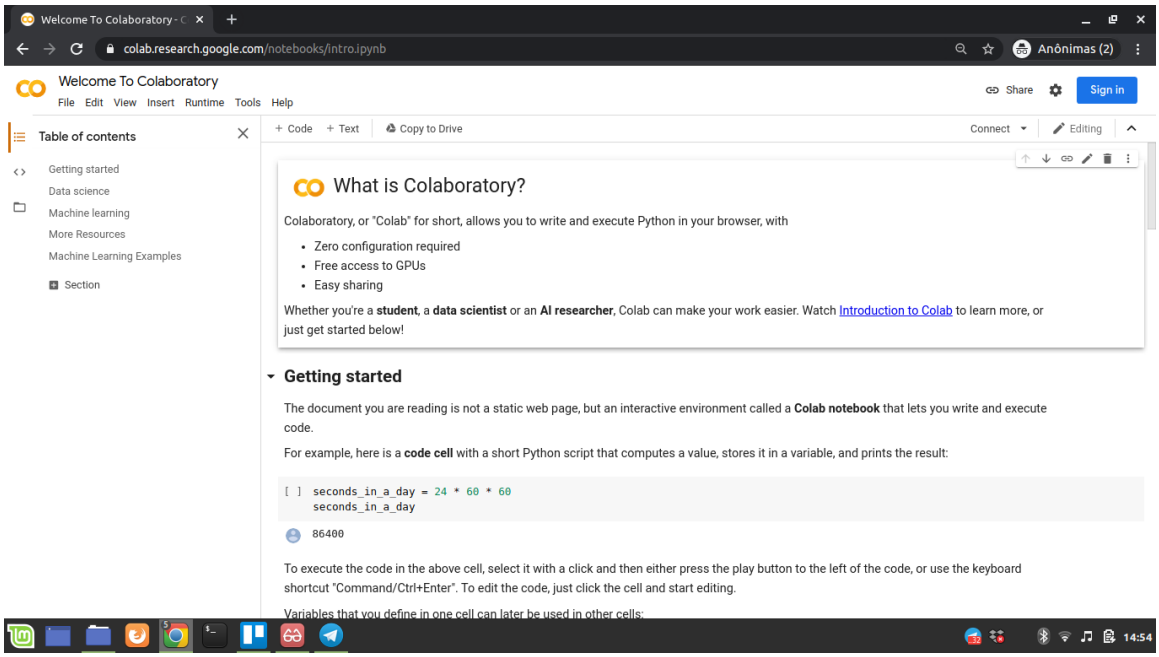
Ao longo do minicurso iremos aplicar conhecimentos da disciplina de probabilidade e estatística, utilizando Python e bibliotecas como Pandas, Numpy e Matplotlib.

Conhecendo as Ferramentas

Antes de aplicar os conhecimentos em forma de código, é necessário conhecer ferramentas que facilitam o trabalho e disponibilizam um kit completo para programação, processamento e armazenamento de código Python.

Ambiente de Desenvolvimento

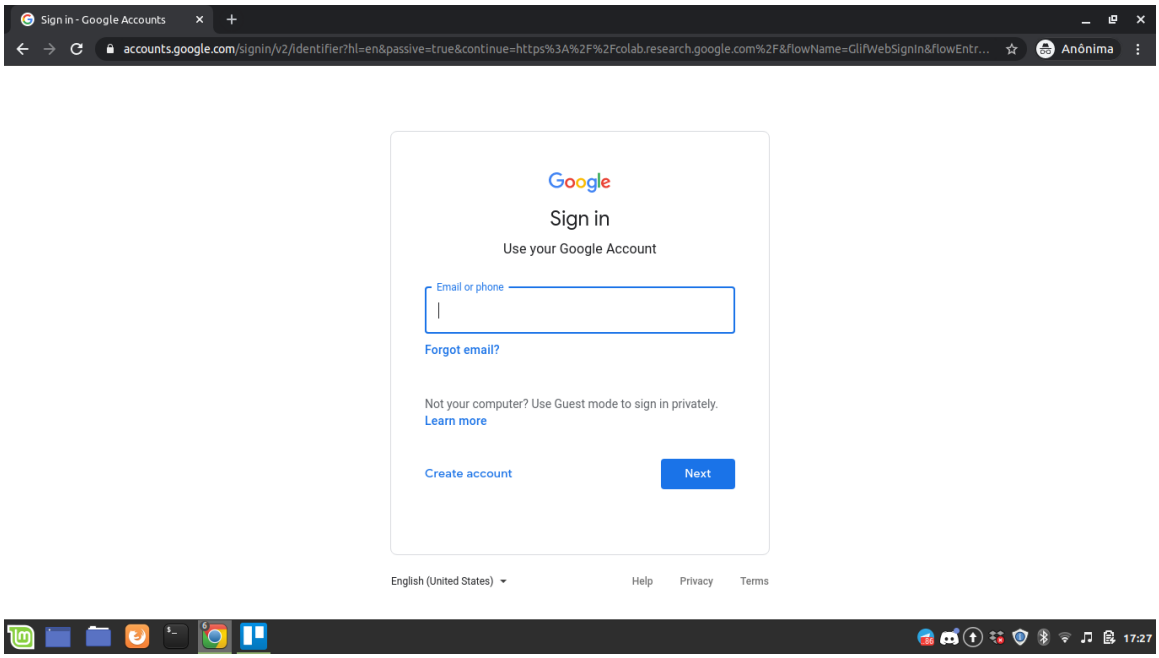
A ferramenta utilizada será o Colab do Google, devido a facilidade do uso e não dependência de instalações das principais bibliotecas da linguagem Python, além do suporte para a mesma em diferentes versões (2.7 e 3.6+).



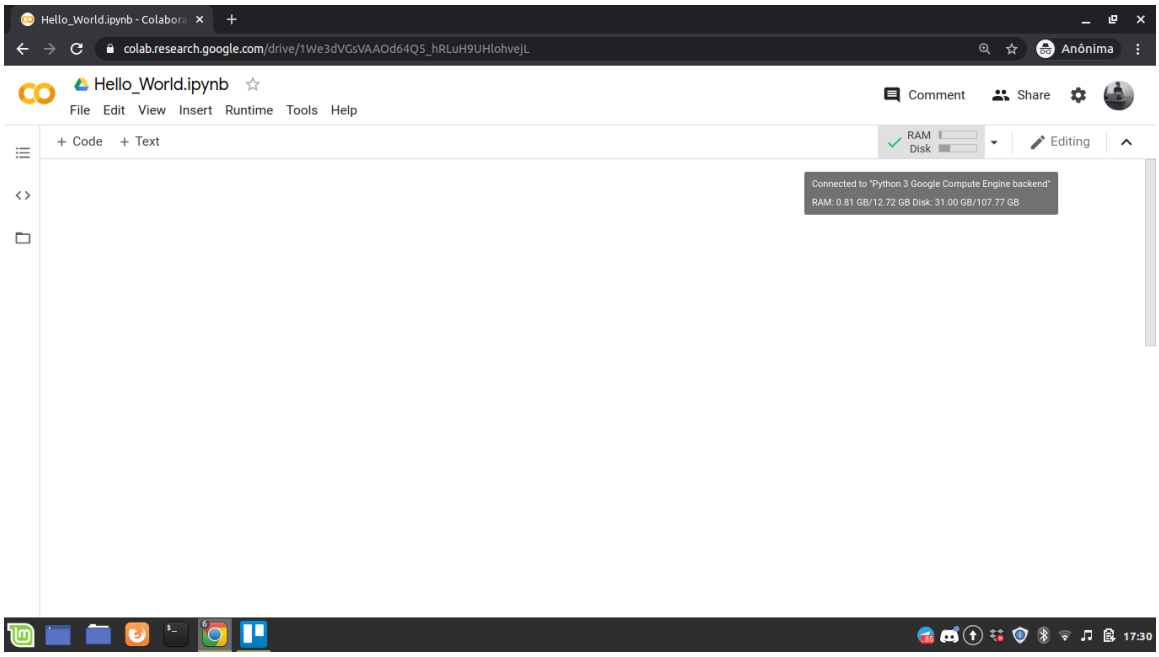
Além de código, o Google Colab fornece apoio a textos, links, imagens, comentários e sessões, assim o código além de criado e executado, pode ser também comentado, tornando completo o material de estudo. Um arquivo criado é chamado de Notebook.

Como Iniciar no Google Colab ?

- [ ] Após abrir a ferramenta, é necessário realizar a autenticação na Conta do Google → **Sign In.**



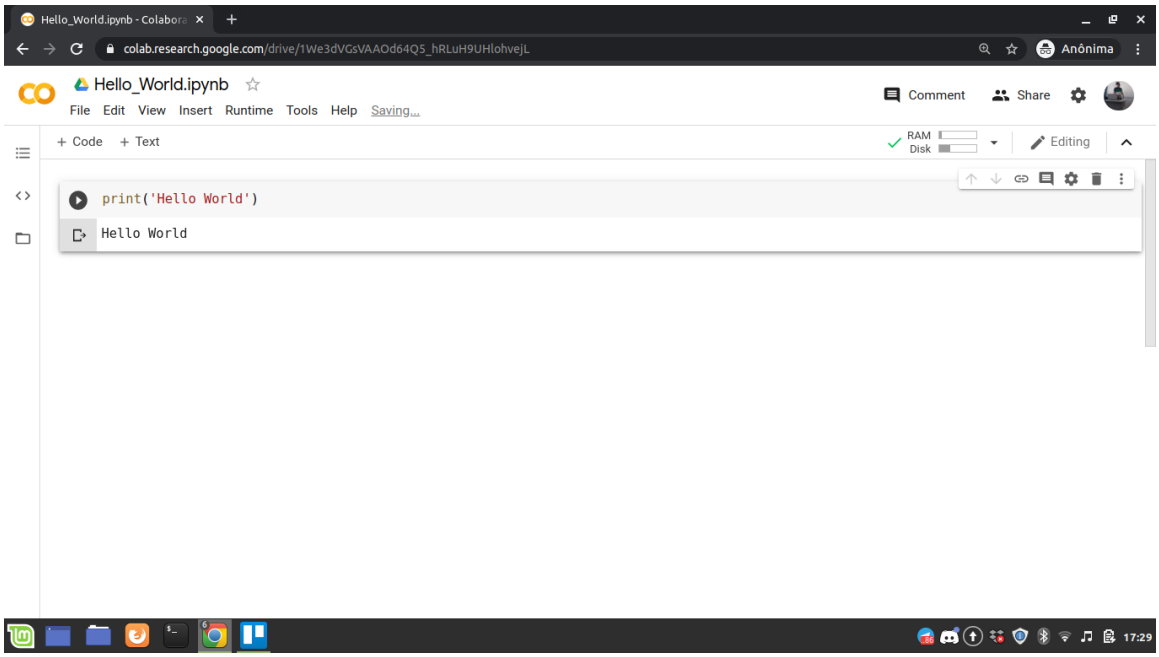
- [ ] No canto superior esquerdo da tela, acessar **File** → **New Notebook.**
- [ ] O notebook irá carregar em uma nova aba, feito isso é possível renomeá-lo.
- [ ] Conferir se o notebook está conectado, na máquina virtual. Se não, clicar em **Connect.**
- [ ] Verificar a quantidade de **Memória** em **Disco** e **RAM** disponibilizadas.



Além de mais de 12GB de RAM e mais de 100GB em disco, é possível configurar e ter acesso à uma GPU e/ou TPU.

Primeiros Passos

O primeiro Hello World no Colab, utilizando Python.



Um notebook é dividido em células, nestas estão os códigos, é necessário seguir também um fluxo de execução correto dessas células, de acordo com o fluxo do projeto.

Atalhos importantes do Google Colab

- Ctrl + M B → Inserir uma nova célula
- Ctrl + Enter → Executar célula selecionada
- Ctrl + F9 → Executar todas as células
- Ctrl + S → Salvar Notebook

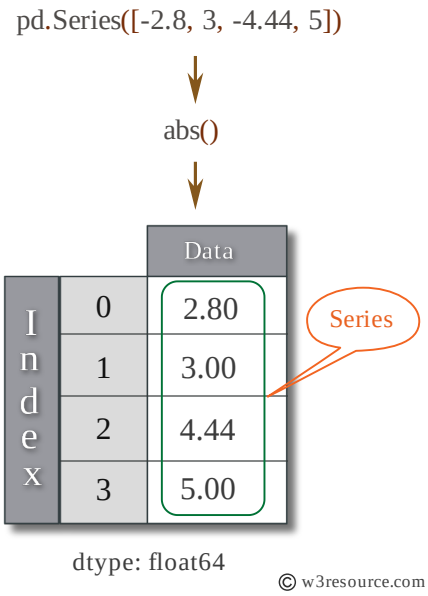
Conhecendo as Bibliotecas

Pandas

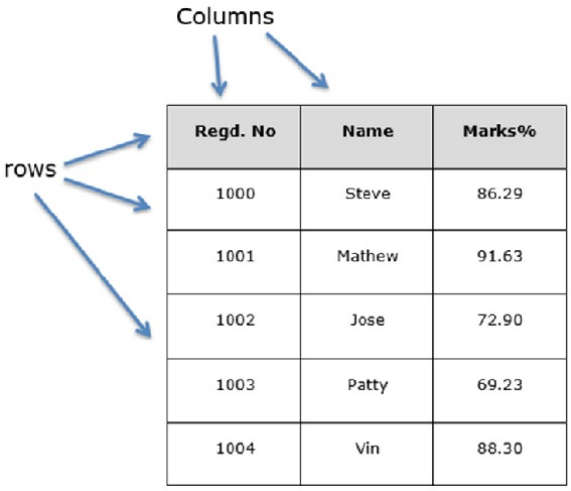
https://pandas.pydata.org/

Pandas é uma das bibliotecas mais utilizadas pelos cientistas de dados, esta fornece ferramentas para análise de dados e estruturas de dados de alta performance e fáceis de usar, são essas:

> Series é um array unidimensional, uma lista de valores, na qual cada um destes possui um índice associado, que é o index.

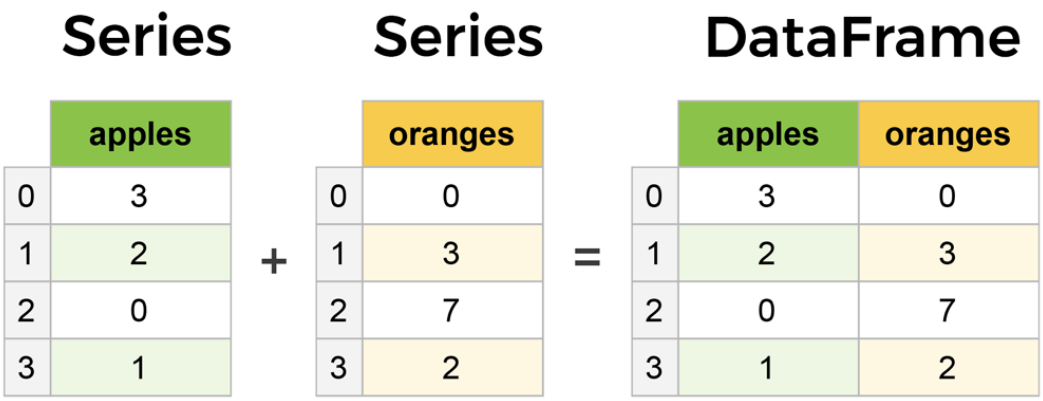


> Data Frame é uma estrutura Bidimensional de dados, possui linhas e colunas, como um tipo de tabela.

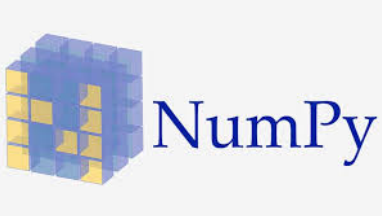


Regd. No	Name	Marks%
1000	Steve	86.29
1001	Mathew	91.63
1002	Jose	72.90
1003	Patty	69.23
1004	Vin	88.30

Exemplo



Numpy



https://numpy.org/

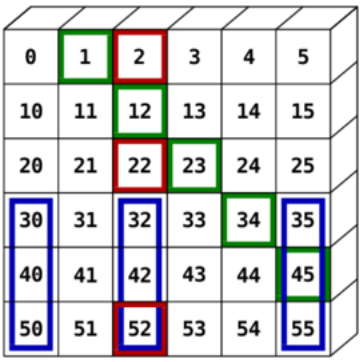
> Numpy é uma importante e conhecida biblioteca do Python para manipulação de arrays e matrizes multidimensionais, além de processamento e cálculos matemáticos.

💡 Exemplo

```
>>> a[(0,1,2,3,4), (1,2,3,4,5)]
array([1, 12, 23, 34, 45])

>>> a[3:, [0,2,5]]
array([[30, 32, 35],
       [40, 42, 45],
       [50, 52, 55]])

>>> mask = np.array([1,0,1,0,0,1], dtype=bool)
>>> a[mask, 2]
array([2, 22, 52])
```



Importando Pandas e Numpy

Essas bibliotecas não precisam ser baixadas quando se utiliza o Google Colab, mas é necessário importa-las para ter acesso às suas ferramentas.

Como importar ?

```
import pandas as pd
import numpy as np
```

💡 A importação padrão é somente: import pandas e import numpy, no entanto uma boa prática é utilizar aliases, que são como "apelidos" para comandos, assim, após esse processo é possível utilizar métodos das duas bibliotecas somente com np e pd,.