

**TECNOLOGIA EM SISTEMAS PARA INTERNET**

**Daniel Evangelista Pereira**

**Ribson Coelho Cardoch Valdés**

**Douglas Seidi Shibata**

**RELATÓRIO DE PRÁTICA INTEGRADA**

**DE**

**CIÊNCIA DE DADOS E INTELIGÊNCIA ARTIFICIAL**

19/08/2020

**Brasília - DF**

**2020**

**Sumário**

[**1. Objetivos**](#_r578hjw9oxco) **3**

[**2. Descrição do problema**](#_jsmvxwf8smme) **4**

[**3. Desenvolvimento**](#_ybi6lms96bcw) **5**

[3.1 Código implementado](#_tcrscypicfq5) 5

[**4. Considerações Finais**](#_sp9srgapjf6s) **6**

[**Referências**](#_5hyd2pq8xv2g) **7**

# 1. Objetivos

Está etapa do projeto tem como finalidade descorrer sobre a coleta de dados por meio de web scraping, com o script feito na linguagem python e gerando um arquivo Comma-separated values. Que será de suma importância para o restante do projeto. E armazenar os dados no repositório do Github com o versionamento de código do git.

# 2. Descrição do problema

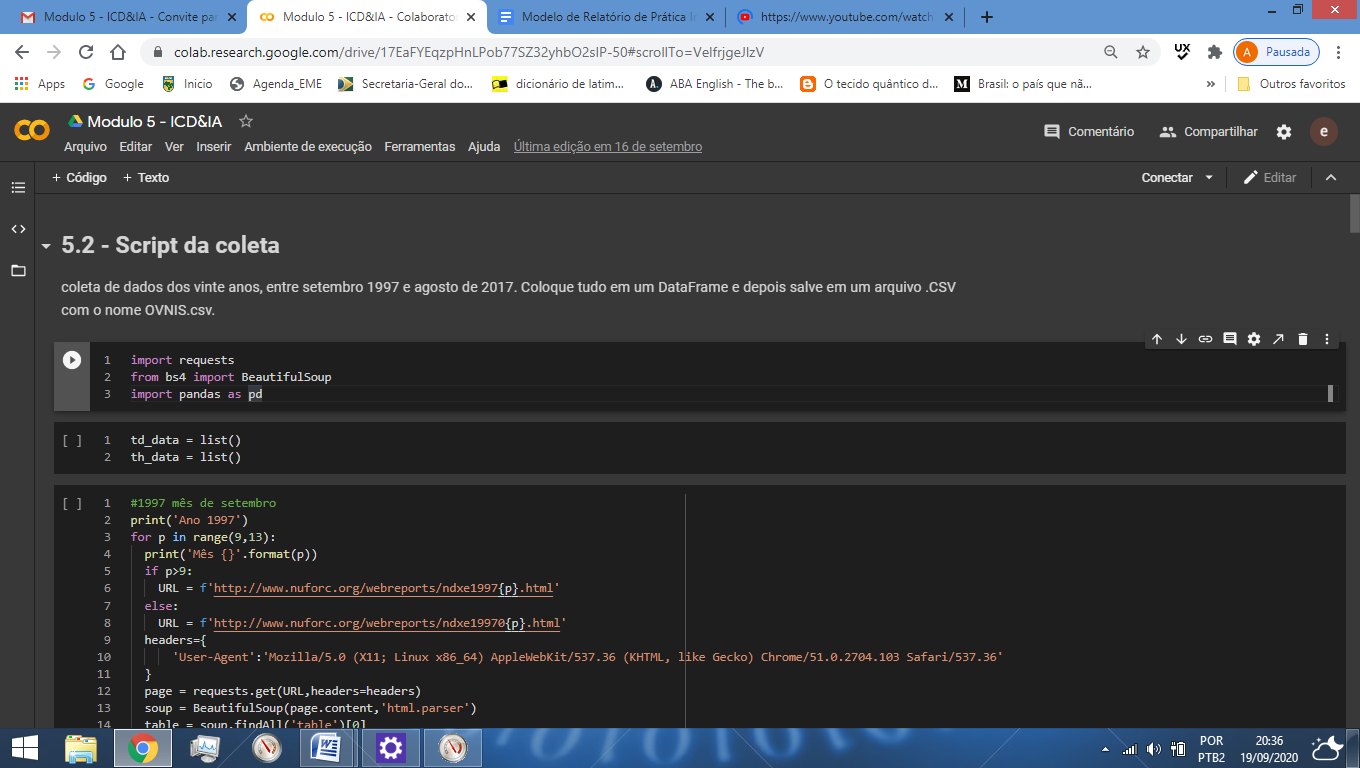
Ao iniciar o desenvolvimento do projeto podemos destacar a forma para se coletar os dados e como estão distribuídos no site e a partir desta análise inicial, foi visto que havia uma grande quantidade de dados em diversas páginas que precisar ser extraídas do site para realizar a coleta de dados e armazenar em DataFrames para que se possa gerar um arquivo csv para começar o estudo dos dados na primeira etapa do projeto.

# 3. Desenvolvimento

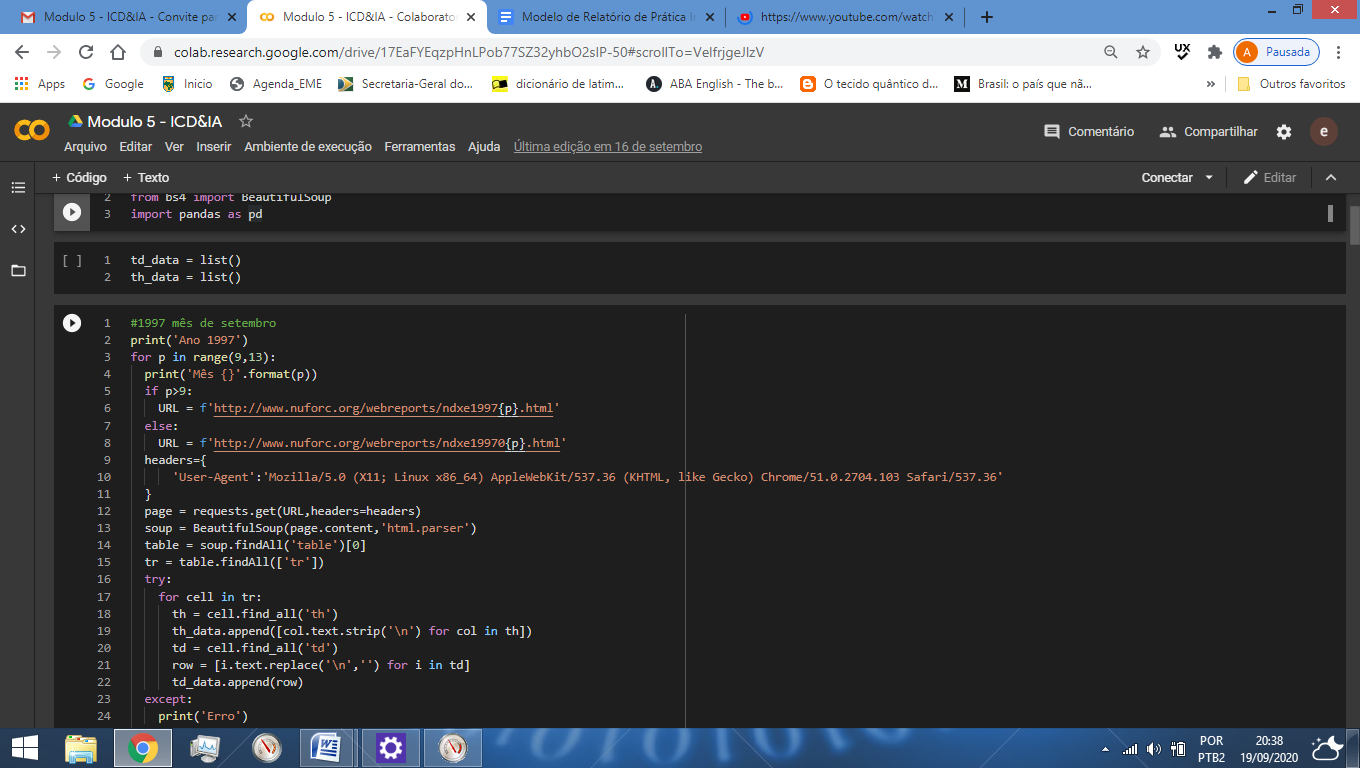
As tecnologias utilizadas para a elaboração da primeira parte do projeto na fase de coleta de dados consiste na linguagem de programação python com o auxílio de suas bibliotecas, como o requests para fazer a requisição do site e retornar o status code, a biblioteca BeautifulSoup para a leitura das páginas em Hyper Text Markup Language(HTML) e para o armazenamento de dados foi utilizada a biblioteca pandas, por meio do notebook e o ambiente de desenvolvimento Google Colab.

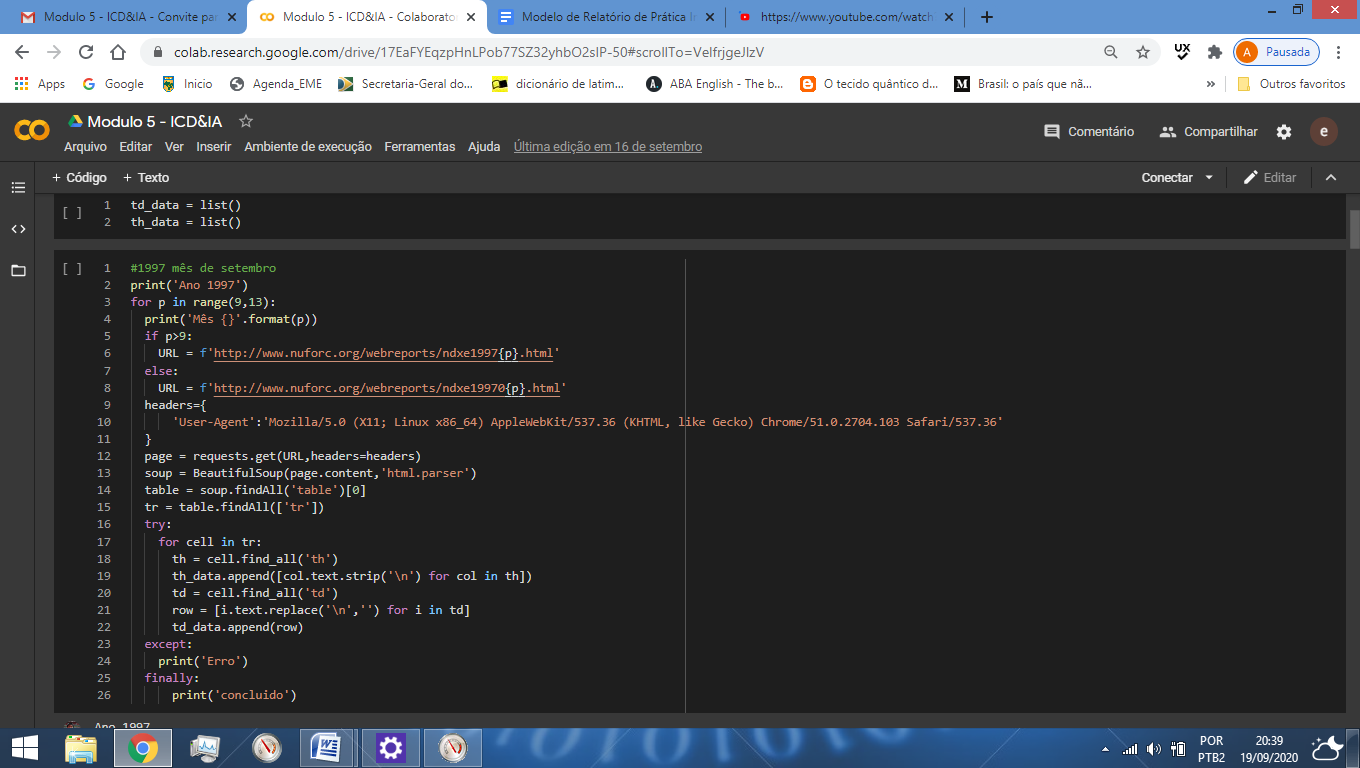
## 3.1 Código implementado

5.2 - Script de coleta

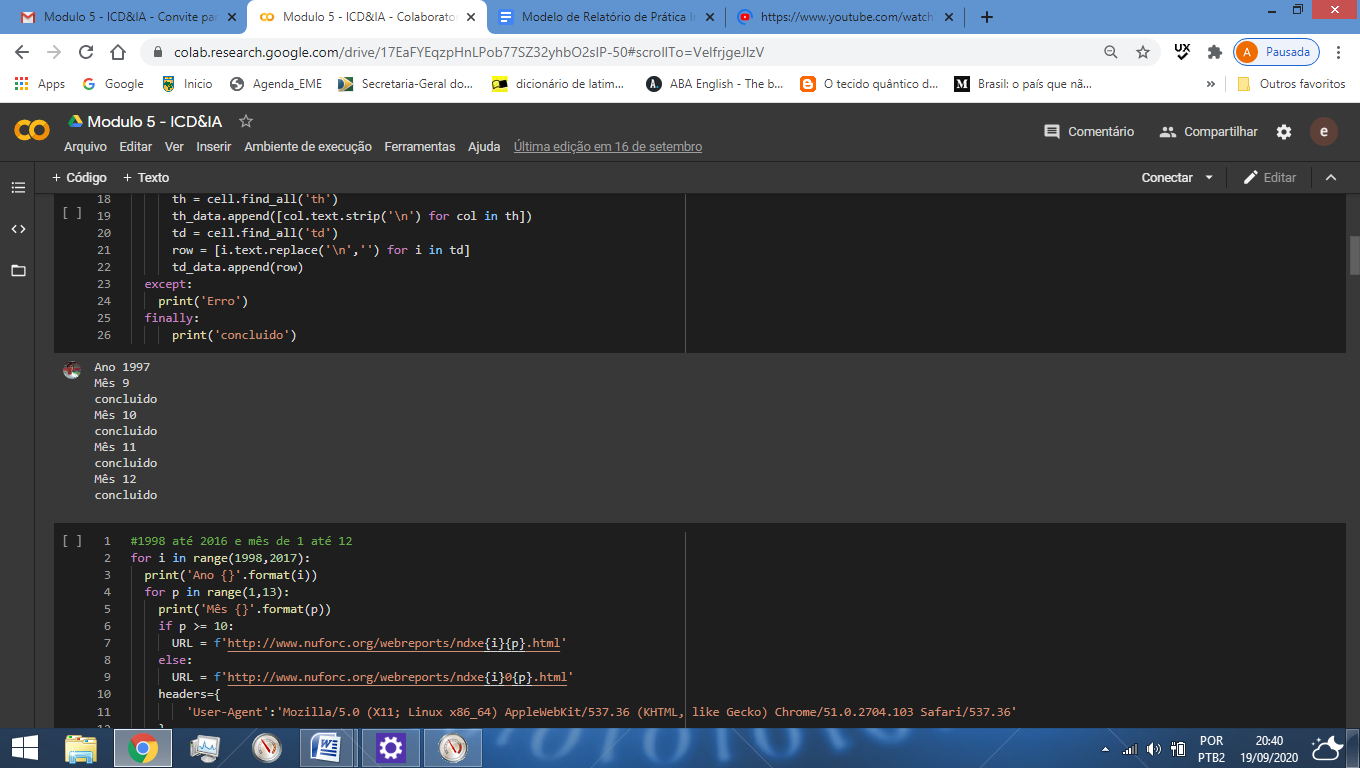


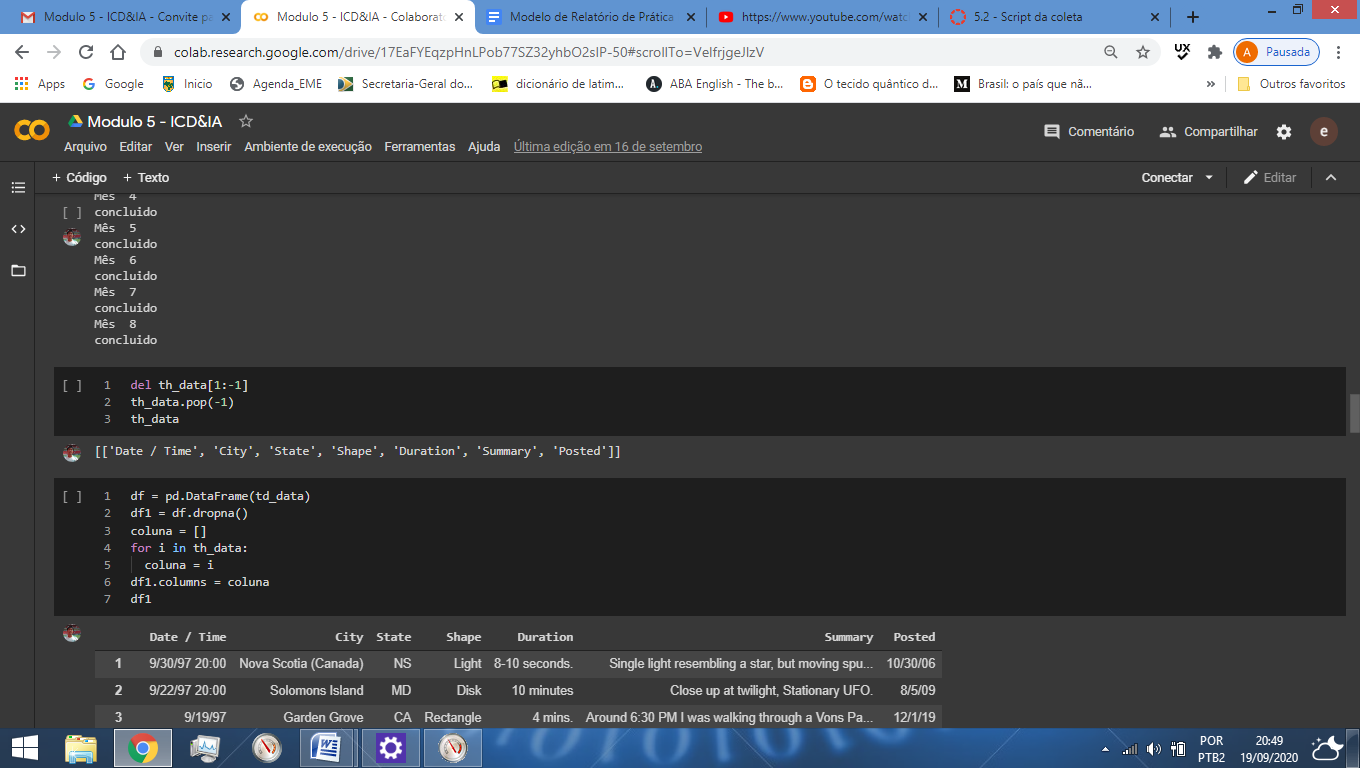
Importação das bibliotecas Requests, BeautifulSoup e Pandas.



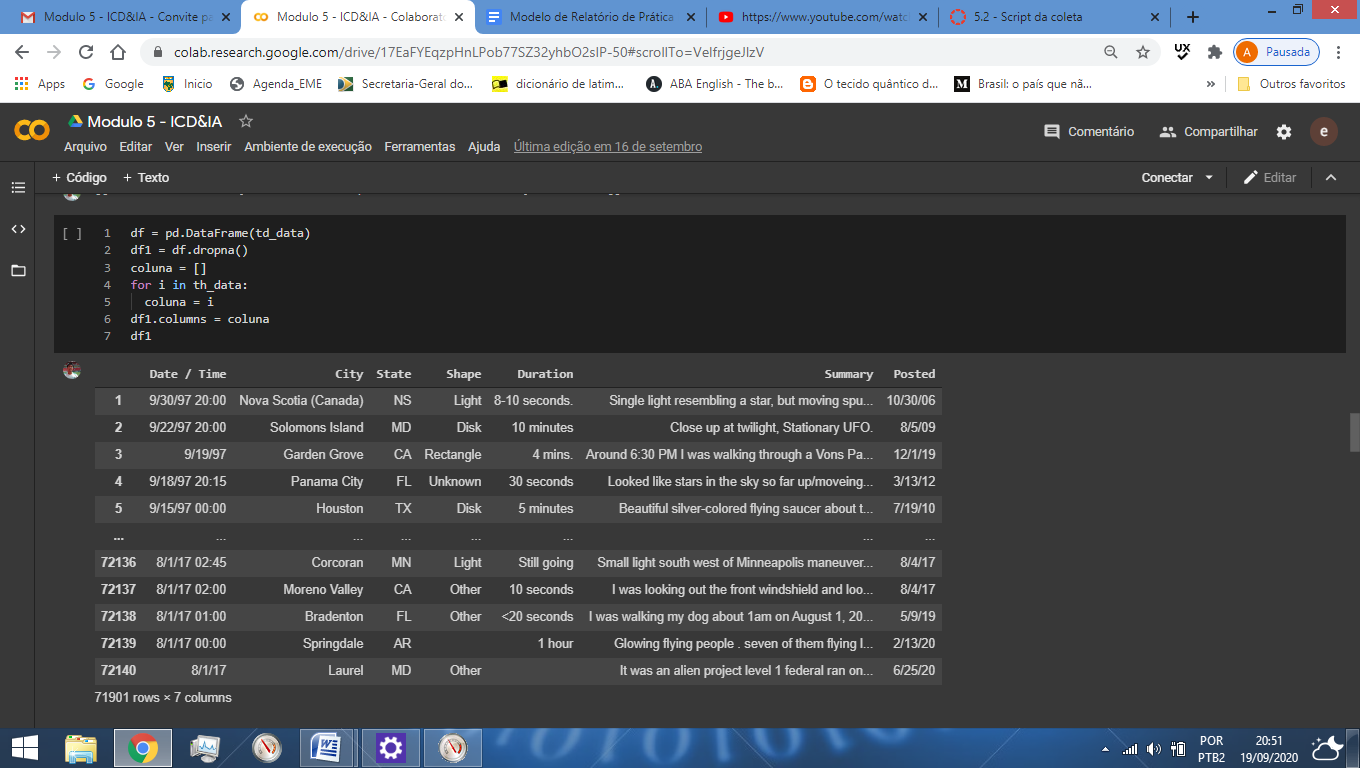
Declarando duas listas, uma para armazenar os dados da tabela que está no html da página. E segundo o cabeçalho da tabela.

Neste pedaço de código foi feito a coleta de dados de setembro de 1997 até dezembro do mesmo ano. Na terceira linha foi utilizado o laço de repetição for para que possa se percorrer o site a partir dos parâmetros passados na url do site. E para que se possa ter o controle dos parâmetros que são passados na url, foi printado os meses correspondentes. A variável Header pega os metadados para a coleta de dados do site, e a variável page recebe a url e o header dos metadados e faz a requisição da página html, e a variável soup irá utilizar a função BeatifulSoup para que se possa ler o conteúdo do site que foi requisitado. E a variável table receberá a variável soup que possui a função findall e irá buscar a primeira tag table que aparece no código html. E em seguida é declarada uma variável tr que receberá a variável table e será utilizado a função findall para que se possa encontrar a tag tr no código da página. E depois é feito um bloco de try para que se possa passar a requisição, se houver êxito, o código continuará. E foi implementado um laço de repetição for para percorrer as colunas da tabela para saber as informações, e a partir disso é adicionado ao array declarado inicialmente.

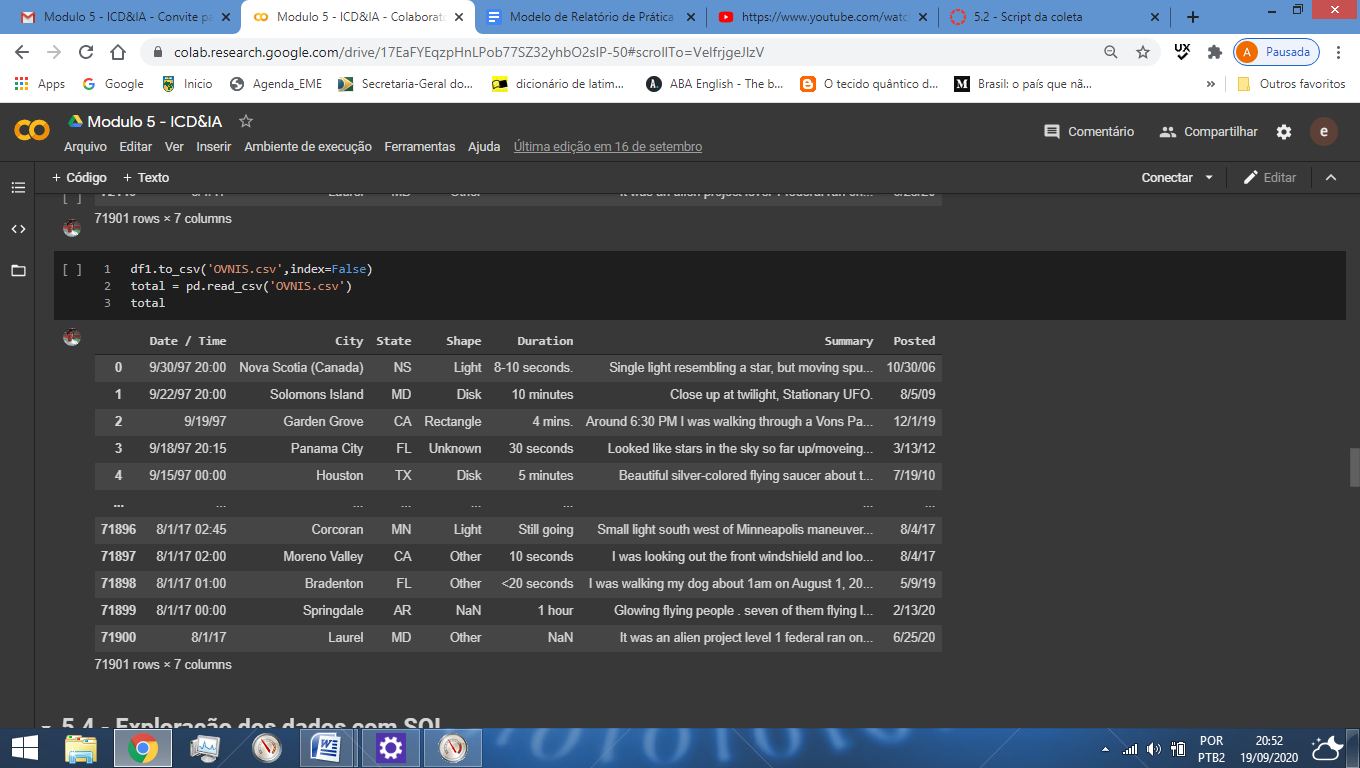




Nesta parte do código é feita a remoção de duplicadas no array do cabeçalho. Deixando apenas um lista com os cabeçalhos.



Nesta parte do código é feita o armazenamento dos dados coletados no DataFrame. e depois renomeando o nome das colunas.



Depois é a partir do Dataframe, foi utilizado a função to\_csv para gerar o arquivo csv.

# 4. Considerações Finais

Nesta primeira do projeto integrado de ciência de dados e inteligência artificial. Houve o estudo das bibliotecas sugeridas pelos professores para que se possa auxiliar na coleta de dados. E a partir do estudo dessas bibliotecas, a coleta de dados foi feita, e acabou gerando diversas versões de códigos de coletas, com melhorias em sua leitura e entendimento, tornando o código mais limpo.

# Referências

TAGLIAFERRI, Lisa. Como trabalhar com dados da web usando Requests e Beautiful soup com Python 3. **Community**. Estados Unidos, 09 de Jul de 2018. Disponível em: <https://www.digitalocean.com/community/tutorials/como-trabalhar-com-dados-da-web-usando-requests-e-beautiful-soup-com-python-3-ptt>. Acesso em: 08 de Set.2020

RICHARDSON, Leonard. Beautiful soup documentation**. Beautiful Soup 4.9.0 Documentation**. Estados Unidos, c2004-2020. Disponível em: <https://www.crummy.com/software/BeautifulSoup/bs4/doc/>. Acesso em: 12 Set. de 2020

FIGUEIREDO, Vinicius. Seus Primeiros Passos com Data Scientist: Introdução ao Pandas. **Data Hackers**. São Paulo, 30 de maio de 2018. Disponível em: <https://medium.com/data-hackers/uma-introdu%C3%A7%C3%A3o-simples-ao-pandas-1e15eea37fa1>. Acesso em: 08 de Set. de 2020

GUIA Rápido. **Requests.** Estados Unidos**,** c2013. Disponível em: <https://requests.readthedocs.io/pt\_BR/latest/user/quickstart.html>. Acesso em: 11 Set. de 2020