INSTITUTO FEDERAL
DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA
Brasília
Campus Brasília

TECNOLOGIA EM SISTEMAS PARA INTERNET

Daniel Evangelista Pereira
Leonardo Campos Muniz

RELATÓRIO DE PRÁTICA INTEGRADA
DE
CIÊNCIA DE DADOS E APRENDIZADO DE MÁQUINA

Brasília - DF

21/03/2021

Sumário

1. Objetivos

Este é o documento é o relatório referente à terceira e penúltima sprint de atividades da PRÁTICA INTEGRADA DE CIÊNCIA DE DADOS E APRENDIZADO DE MÁQUINA, sendo o trabalho apresentado aqui relativo ao primeiro dos dois desafios presentes na semana.

A equipe é composta pelos alunos Daniel Evangelista Pereira e Leonardo Campos Muniz, com a mentoria de Thiago Marinha e utilizará a metodologia de desenvolvimento ágil SCRUM, com um total de 4 sprints de uma semana, estando atualmente na terceira semana.

O presente relatório trata do desafio de **Limpeza dos dados**, coletando uma base de dados que tem como o foco relato de ocorrências de OVNIs no território estadunidense. O intuito desta atividade é fazer o processo de ETL na mesma e no final entregar uma tabela onde serão apenas mostradas as ocorrências que aconteceram em determinados estados dos EUA e as que têm os formatos de OVNI mais populares que possuem mais de mil ocorrências.

2. Descrição do problema

O problema do desafio da semana se dá por alguns fatores, coletar a lista de ocorrência dos OVNIs nos EUA, após conseguir a lista de ocorrências coletar a lista de estados para a filtragem da tabela e por fim descobrir quais são os formatos de OVNIs mais populares, depois que todas essas informações foram obtidas basta apenas construir a nova tabela limpa.

3. Desenvolvimento

Para facilitar o desenvolvimento do código e da criação dos gráficos, foram utilizadas as seguintes bibliotecas:

- Pandas: Biblioteca utilizada para fazer a carga de dados e a transformação da mesma em forma de tabela;
- Numpy: Biblioteca utilizada para realização e elaboração de operações matemáticas complexas e manipulação de array;

3.1 Código implementado

Segue abaixo o código utilizado para realizar o desafio 5.7

```
import pandas as pd
import numpy as np

#coleta de dados dos ovnis
dados_ovni = pd.read_csv('OVNIS.csv')

#transformar os dados em um data frame e limpeza dos dados
tabela_ovni = pd.DataFrame(dados_ovni)
tabela_ovni.fillna(method='ffill').dropna()

#filtragem das colunas principais para a tabela
filtro_tabela = ['row','Date / Time','State','Shape']
tabela_filtrada = tabela_ovni.filter( items = filtro_tabela)

#coleta das formas de shape mais populares que possuem mais de
1000 reigstros
shape_filtrado = tabela_filtrada.groupby('Shape').count()
shape_filtrado =shape_filtrado['row'].sort_values(ascending=False)
shape_filtrado = shape_filtrado[shape_filtrado >1000]
shape_filtrado = shape_filtrado.to_frame()
lista_shapes = shape_filtrado.index.tolist()

# coleta dos estados
estados_permitidos = pd.read_excel('states.xlsx')
```

```
      #transformação da tabela de estados
      transformacao_tabela_estados=estados_permitidos["State"].str.split
(',')

      #transformação dos dados coletados e inserção das novas colunas no
data frame
      estado = transformacao_tabela_estados.str.get(0)
      abreviacao = transformacao_tabela_estados.str.get(1)
      estados_permitidos["Estado"] = estado
      estados_permitidos["Abreviação"] = abreviacao.str.replace('"','')
      estados = estados_permitidos[['Estado','Abreviação']]
      lista_estados = estados['Abreviação'].tolist()


      #transformação das tabelas tratadas em csv
      data_frame =
tabela_filtrada[(tabela_filtrada.State.isin(lista_estados)) &
(tabela_filtrada.Shape.isin(lista_shapes))]
      data_frame.to_csv('df_OVNI_limpo.csv', mode='w', header=True)
```

Com a execução do código após ser implementado os filtros é gerado em um arquivo csv a tabela que segue abaixo.

	row	Date / Time	State	Shape	City
1	1	9/22/97 20:00	MD	Disk	Solomons Island
2	2	9/19/97	CA	Rectangle	Garden Grove
3	3	9/18/97 20:15	FL	Unknown	Panama City
4	4	9/15/97 00:00	TX	Disk	Houston
5	5	9/15/97 20:00	NM	Light	Santa Fe
...	...	...	...	...	...
71969	412	8/1/17 06:15	GA	Fireball	Columbus (North)
71970	413	8/1/17 02:45	MN	Light	Corcoran
71971	414	8/1/17 02:00	CA	Other	Moreno Valley
71972	415	8/1/17 01:00	FL	Other	Bradenton
71974	417	8/1/17	MD	Other	Laurel

61672 rows × 5 columns

4. Considerações Finais

Em relação ao trabalho proposto e o que foi feito a ideia é realmente muitíssimo boa pois trabalha toda a ideia do etl de extração, transformação e carga dos dados, fazendo a coleta de dados de dois bancos e depois fazer toda a parte de transformação dos dados coletados, para aí sim criar todos os filtros que iriam resultar numa tabela limpa.

As considerações a se dar é que o pandas é uma biblioteca extremamente completa e faz juz a sua popularidade, basicamente não foi preciso nenhuma outra biblioteca para realizar o serviço.Da primeira versão do código até a última as mudanças foram gigantescas, refatoramento do código, mudança de linhas para utilizar uma única que realizava todo o trabalho sozinha, realmente muito boa, perdendo somente para ferramentas como power bi.

Referências

PANDAS DOCUMENTATION. Pandas, 2021. Disponível em: < https://pandas.pydata.org/docs/ >. Acesso em 21/03/2021.