Projeto de prática Integrada de ciência de dados, inteligência artificial e machine learning

Relatório de avistamento de Objetos Voadores Não Identificados.

Sprint → Limpeza de dados

Curso: Tecnologia em sistemas para internet

Estudantes:

Brenda Lopes Miranda Teixeira
Mateus Gomes da Silva Fonteles
Rickson Queiroz Marques de Souza
Samuel Araújo Lopes

Professores

Fábio Henrique
Diego Queiroz
Ana Régia

Brasília, agosto de 2021

INSTITUTO FEDERAL
Brasília

Sumário

1. Objetivos

Nesta etapa do trabalho será executada a seguinte sequência de tarefas:

1. Carregar o seu arquivo OVNIS.csv em um dataframe;
2. Remover registros que tenham valores vazios (None, Unknown, ...) para City, State e Shape;
3. Manter somente os registros referentes aos 51 Estados dos Estados Unidos.
4. Remover variáveis irrelevantes para a análise (Duration, Summary e Posted);
5. Manter somente os registros de Shapes mais populares (com mais de 1000 ocorrências);
6. Salvar o *dataframe* final em um arquivo CSV com o nome "df_OVNI_limpo".

2. Descrição do problema

A limpeza e preparação dos dados é uma tarefa central no trabalho de qualquer analista de dados. Falhas nos dados podem levar a resultados falsos refletidos nos nossos relatórios, de pouco vale ter gráficos avançados se a ideia que estes transmitem não é válida e um relatório com informação falsa pode levar a parte interessada a tomar decisões equivocadas.

Os dados “sujos” podem ser caracteres que não foram digitados corretamente, campos preenchidos com informações de outros campos, campos vazios ou desatualizados, dados que contém “pontos fora da curva” entre outros problemas e cabe ao analista não somente detectar tais problemas, mas tomar a decisão de como resolvê-los de acordo com com a demanda que deve cumprir.

Por vezes um ponto fora da curva ou campo vazio pode atrapalhar o cálculo de uma média de valores e deve ser retirado do *dataset,* por outras podem ser justamente estas as informações buscadas.

No presente caso, faremos o processo de limpeza por eliminação, como o objetivo deste relatório está mais ligado à médias do que a casos particulares.

3. Desenvolvimento

Este trabalho está sendo desenvolvido usando um Script Python por ser uma linguagem orientada a objetos é bastante maleável, o grupo está utilizando a plataforma Google Colaboratory, assim todos podem modificar e acrescentar o código quando necessário.

3.1. Código implementado - Gráficos

- Importação de bibliotecas e carregamento de arquivo:

```
#importantando Bibliotecas
!pip install -U pandasql
import pandas as pd
import numpy as np
import pandasql
#Carregando o csv OVNIS.csv em um dataframe
ovnis = pd.read_csv('OVNIS.csv')
ovnis
```

- Remoção dos registros vazios:

```
#Removendo registros vazios
ovnis.drop(ovnis.index[ovnis['City'] == 'None'], inplace = True)
ovnis.drop(ovnis.index[ovnis['City'] == None], inplace = True)
ovnis.drop(ovnis.index[ovnis['State'] == None], inplace = True)
ovnis.drop(ovnis.index[ovnis['Shape'] == None], inplace = True)
ovnis.drop(ovnis.index[ovnis['Shape'] == "Unknown"], inplace = True)
ovnis['State'].dropna()
ovnis['Shape'].dropna()
ovnis['City'].dropna()

ovnis
```

- Consulta SQL para criar um *dataframe* somente com avistamentos em Estados pertencentes aos Estado Unidos:

```
sql = """
  SELECT * FROM ovnis WHERE
  STATE LIKE 'AL'
  OR STATE LIKE 'AK'
  OR STATE LIKE 'AZ'
  OR STATE LIKE 'AR'
  OR STATE LIKE 'CA'
  OR STATE LIKE 'CO'
  OR STATE LIKE 'CT'
  OR STATE LIKE 'DE'
  OR STATE LIKE 'DC'
```

```
  OR STATE LIKE 'FL'
  OR STATE LIKE 'GA'
  OR STATE LIKE 'HI'
  OR STATE LIKE 'ID'
  OR STATE LIKE 'IL'
  OR STATE LIKE 'IN'
  OR STATE LIKE 'IA'
  OR STATE LIKE 'KS'
  OR STATE LIKE 'KY'
  OR STATE LIKE 'LA'
  OR STATE LIKE 'ME'
  OR STATE LIKE 'MT'
  OR STATE LIKE 'NE'
  OR STATE LIKE 'NV'
  OR STATE LIKE 'NH'
  OR STATE LIKE 'NJ'
  OR STATE LIKE 'NM'
  OR STATE LIKE 'NY'
  OR STATE LIKE 'NC'
  OR STATE LIKE 'ND'
  OR STATE LIKE 'OH'
  OR STATE LIKE 'OK'
  OR STATE LIKE 'OR'
  OR STATE LIKE 'MD'
  OR STATE LIKE 'MA'
  OR STATE LIKE 'MI'
  OR STATE LIKE 'MN'
  OR STATE LIKE 'MS'
  OR STATE LIKE 'MO'
  OR STATE LIKE 'PA'
  OR STATE LIKE 'RI'
  OR STATE LIKE 'SC'
  OR STATE LIKE 'SD'
  OR STATE LIKE 'TN'
  OR STATE LIKE 'TX'
  OR STATE LIKE 'UT'
  OR STATE LIKE 'VT'
  OR STATE LIKE 'VA'
  OR STATE LIKE 'WA'
  OR STATE LIKE 'WV'
  OR STATE LIKE 'WI'
  OR STATE LIKE 'WY'
"""
```

```
eua_registros = pandasql.sqldf(sql.lower(), locals())
eua_registros
```

- Retirar espaços dos nomes das colunas
- Renomear Coluna DateTime
- Por meio de pesquisa SQL, criar um *dataframe* contendo somente as colunas da tabela que são relevantes para o presente trabalho.

```
#Removendo espaços das colunas
eua_registros.rename(columns = lambda x: x.replace(' ', '_'),
inplace=True)

#mudando o nome da coluna DateTime
eua_registros['DateTime'] = eua_registros['Date_/_Time']

#Filtrando dados
q = """
SELECT DateTime, City,  State,  Shape
FROM eua_registros
"""

df1 = pandasql.sqldf(q, locals())
df1
```

- Descobrir formas de OVNIS com mais de 1000 registros de avistamento.
- Salvar um novo *dataframe* contendo somente aparições com estas formas.

```
#Registros com mais de 1000 ocorrencias

n_shape = df1['Shape'].value_counts()
maioresque1000 = n_shape[n_shape > 1000]
maioresque1000

#Filtrando dados e agrupando
q = """
SELECT *
FROM df1
WHERE Shape in ('Light', 'Circle', 'Triangle', 'Fireball', 'Sphere',
'Other', 'Oval', 'Disk', 'Formation', 'Changing', 'Cigar', 'Flash',
'Rectangle')

"""
```

```
df2 = pandasql.sqldf(q, locals())
df2
```

- Gerar o arquivo .csv limpo.

```
#Gerarando um arquivo csv limpo
df2.to_csv('df_OVNIS_limpo.csv')
```

4. Considerações Finais

Com um *dataset* limpo, não somente obteremos resultados mais precisos e condizentes com a realidade, como também estaremos economizando bastante processamento por estarmos utilizando somente os dados relevantes ao presente trabalho.

Referências

- Dividir data (dia, mês, ano) em novas colunas - DataFrame Pandas. 2021. Disponível em:
    - <https://pt.stackoverflow.com/questions/428236/dividir-data-dia-m%C3%AAs-ano-em-novas-colunas-dataframe-pandas>