INSTITUTO FEDERAL
DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA
Brasília
Campus Brasília

TECNOLOGIA EM SISTEMAS PARA INTERNET

Matheus Nícollas de Souza Mota
Thiago Marinho da Silva Campos

RELATÓRIO DE PRÁTICA INTEGRADA
DE
CIÊNCIA DE DADOS E INTELIGÊNCIA ARTIFICIAL

Brasília - DF
18/09/2020

Sumário

1. Objetivos

O objetivo é fazer uma limpeza nos dados obtidos anteriormente e prepara-los para que possamos descobrir valores através deles.

2. Descrição do problema

O projeto consiste em reunir fatos interessantes relacionados a OVINIs, a partir de relatos realizados dentro de um período de vinte anos usando o site Nuforc. O desafio consiste em fazer a extração de dados de forma tabular, afinal, para que os dados possam ser analisados eles acabam se tornando tabelas.

O WebScraping consiste em extrair os dados formatados com tag's da linguagem HTML. Como iremos extrair vinte anos de dados, consultaremos 240 páginas web, uma por cada mês, por vinte anos, entre setembro 1997 e agosto de 2017.

Para descobrir valores através destes dados, devemos carregar o csv gerado anteriormente em um dataframe, remover os registros que tenham valores vazios, remover registros diferentes dos 51 estados americanos, remover variáveis irrelevantes para a análise, manter os registros de shapes mais populares no caso com mais de 1000 ocorrencias e salvar estes dados tratados em um arquivo csv.

3. Desenvolvimento

O desenvolvimento do algoritmo foi feito na plataforma Google Collab, esta plataforma foi escolhida pois ao iniciar um notebook na mesma, as bibliotecas e dependências do Python são todas da nuvem.

Carregamos nosso arquivo óvnis.csv em um dataframe. Fizemos uma limpeza nos dados removendo os registros que tenham valores vazios. Em seguida filtramos os dados para manter somente registros referentes aos 51 estados americanos. Removemos variáveis que não serão usadas na análise de dados. Filtramos os registros de shapes para manter somente os mais populares com mais de 1000 ocorrencias. E por último salvamos o resultado em um dataframe com nome df_ovni_limpo.csv.

4 Código implementado

```
import pandas as pd

#1. Carregar o seu arquivo OVNIS.csv em um dataframe
df_ovnis = pd.read_csv("ovnis.csv", index_col=[0])
df_ovnis
```

```
#2. Remover registros que tenham valores vazios (None, Unknown, ...) p
ara City, State e Shape
df_notnull = df_ovnis.dropna(subset=["State"])
df_notnull = df_notnull.dropna(subset=["City"])
df_notnull = df_notnull.dropna(subset=["Shape"])
df_remove = df_notnull.loc[(df_notnull['Shape'] == 'Unknown')
                         | (df_notnull['Shape'] == 'None')
                         | (df_notnull['City'] == 'Unknown')
                         | (df_notnull['City'] == 'None')
                         | (df_notnull['State'] == 'Unknown')
                         | (df_notnull['State'] == 'None')]
df_notnull = df_notnull.drop(df_remove.index)
df_notnull
```

```
#3. Manter somente os registros referentes aos 51 estados dos Estados
Unidos:
states = ["AL", "AK", "AZ", "AR", "CA", "CO", "CT", "DC", "DE", "FL",
"GA",
          "HI", "ID", "IL", "IN", "IA", "KS", "KY", "LA", "ME", "MD",
          "MA", "MI", "MN", "MS", "MO", "MT", "NE", "NV", "NH", "NJ",
          "NM", "NY", "NC", "ND", "OH", "OK", "OR", "PA", "RI", "SC",
          "SD", "TN", "TX", "UT", "VT", "VA", "WA", "WV", "WI", "WY"]

selection = df_notnull['State'].isin(states)
df_usa = df_notnull[selection]
df_usa
```

```
#4. Remover variáveis irrelevantes para a análise (Duration, Summary e
Posted);
df_colunas = df_usa.drop(columns=['Duration'])
df_colunas = df_colunas.drop(columns=['Summary'])
df_colunas = df_colunas.drop(columns=['Posted'])
df_colunas
```

```
#5. Manter somente os registros de Shapes mais populares (com mais de
1000 ocorrências);

df_colunas['Quantidade'] = df_colunas.groupby('Shape')['Shape'].transf
orm('count')
df_final = df_colunas.query('Quantidade >= 1000')
df_final = df_final.drop(columns=['Quantidade'])
df_final
```

```
#6. Salvar o dataframe final em um arquivo CSV com o nome
"df_OVNI_limpo". df_final.to_csv("df_OVNI_limpo.csv")
```

5 Considerações Finais

Com este projeto fizemos uma limpeza de dados preparando-os para uma etapa de análise onde será possível mensurar e obter respostas claras a determinadas questões. Abaixo segue o resultado obtido.

	Date / Time	City	State	Shape
1	9/22/97 20:00	Solomons Island	MD	Disk
2	9/19/97	Garden Grove	CA	Rectangle
4	9/15/97 00:00	Houston	TX	Disk
5	9/15/97 20:00	Santa Fe	NM	Light
6	9/15/97 20:00	Kent	WA	Sphere
...	...	...	...	...
71895	8/1/17 06:15	Columbus (North)	GA	Fireball
71896	8/1/17 02:45	Corcoran	MN	Light
71897	8/1/17 02:00	Moreno Valley	CA	Other
71898	8/1/17 01:00	Bradenton	FL	Other
71900	8/1/17	Laurel	MD	Other

6 Referencias

PANDAS. Https://pandas.pydata.org/docs/. Disponível em: https://pandas.pydata.org/docs/. Acesso em: 19 set. 2020.

REQUESTS. Requests: HTTP for Humans™. Disponível em: https://requests.readthedocs.io/en/master/. Acesso em: 19 set. 2020.

SOUP, Beautiful. Beautiful Soup Documentation. Disponível em: https://www.crummy.com/software/BeautifulSoup/bs4/doc/. Acesso em: 19 set. 2020.

SILVA, Réulison. Como fazer um Web Scraping com Python. Disponível em: https://goomore.com/blog/web-scraping-python/. Acesso em: 19 set. 2020