Análisis_Sunedu

March 13, 2025

1 Análisis de Datos Exploratorio - Web Scraping de SUNEDU

Mi proyecto de análisis exploratorio de datos se enfocó en la obtención y análisis de datos desde la web de SUNEDU utilizando técnicas de web scraping. Empleé bibliotecas como BeautifulSoup, Requests y Pandas para extraer, procesar y analizar la información disponible en la plataforma de SUNEDU.

El proceso comenzó con la identificación de las URLs relevantes y la extracción de datos estructurados como tablas e información textual importante. Durante la extracción, se realizó un preprocesamiento de los datos para limpiar y estructurar la información, eliminando duplicados y completando valores faltantes cuando fue necesario. Los datos limpios se almacenaron en un DataFrame para facilitar su manipulación y análisis posterior.

Posteriormente, se procedió a realizar un análisis descriptivo con el objetivo de identificar tendencias, patrones y relaciones significativas en la información recopilada. Este proceso incluyó la creación de gráficos y tablas que resumen las características principales de los datos obtenidos.

Finalmente, los resultados se presentaron en un archivo limpio y estructurado, adecuado para análisis adicionales o futuros proyectos relacionados con la calidad educativa y las instituciones académicas reguladas por SUNEDU.

https://www.sunedu.gob.pe/

[1]:
```
#Importamos las librerías
import requests
from bs4 import BeautifulSoup
import pandas as pd
import openpyxl
```

[2]:
```
#URL de universidades licenciadas
url_lic = "https://www.sunedu.gob.pe/lista-de-universidades-licenciadas/"

#URL de universidades no licenciadas
url_no_lic = "https://www.sunedu.gob.pe/lista-de-universidades-denegadas/"
```

[3]:
```
#Modificamos el user-agent para que no detecte que es un robot
headers = {
    "user-agent": "Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36␣
 ↪(KHTML, like Gecko) Chrome/91.0.4472.124 Safari/537.36 Edg/91.0.864.59"
}
```

[4]:
```
#Solicitar datos de URL de universidades licenciadas
response_lic = requests.get(url_lic, headers=headers)

#Solicitar datos de URL de universidades no licenciadas
response_no_lic = requests.get(url_no_lic, headers=headers)
```

[5]:
```
#Parsear los datos solicitados de universidad licenciadas
soup_lic = BeautifulSoup(response_lic.text, "html.parser")

#Parsear los datos solicitados de universidad no licenciadas
soup_no_lic = BeautifulSoup(response_no_lic.text, "html.parser")
```

[6]:
```
#Tabla de universidad licenciadas
table_lic = soup_lic.find("table", id="tablepress-114")

#Tabla de universidad licenciadas
table_no_lic = soup_no_lic.find("table", id="tablepress-98")
```

[7]:
```
departamentos_provincias_no_lic = {
    "Universidad Católica Los Ángeles de Chimbote": ("Ancash", "Santa"),
    "Universidad Científica del Perú": ("Loreto", "Maynas"),
    "Universidad José Carlos Mariátegui": ("Moquegua", "Mariscal Nieto"),
    "Universidad Andina Néstor Cáceres Velásquez": ("Puno", "San Román"),
    "Universidad Autónoma San Francisco": ("Arequipa", "Arequipa"),
    "Universidad Privada Antonio Guillermo Urrelo": ("Cajamarca", "Cajamarca"),
    "Universidad Privada San Carlos": ("Puno", "San Román"),
    "Universidad Seminario Evangélico de Lima": ("Lima", "Lima"),
    "Universidad Latinoamericana CIMA": ("Lima", "Lima"),
    "Universidad Politécnica Amazónica": ("Loreto", "Maynas"),
    "Universidad Privada Líder Peruana": ("Lima", "Lima"),
    "Universidad Peruana Las Américas": ("Lima", "Lima"),
    "Universidad Santo Domingo de Guzmán": ("Lima", "Lima"),
    "Universidad Alas Peruanas": ("Lima", "Lima"),
    "Universidad Privada Leonardo Da Vinci": ("Lima", "Lima"),
    "Universidad Interamericana para el Desarrollo": ("Lima", "Lima"),
    "Universidad Peruana de Ciencias e Informática": ("Lima", "Lima"),
    "Universidad Peruana Santo Tomás de Aquino de Ciencia e Integración":␣
 ↪("Lima", "Lima"),
    "Universidad San Pedro": ("Ancash", "Santa"),
    "Universidad Seminario Bíblico Andino": ("Cusco", "Cusco"),
    "Escuela Internacional de Posgrado S.A.C.": ("Lima", "Lima"),
    "Universidad Privada Juan Mejía Baca": ("Lambayeque", "Chiclayo"),
    "Universidad Privada Autónoma del Sur": ("Arequipa", "Arequipa"),
    "Universidad Peruana Austral del Cusco": ("Cusco", "Cusco"),
    "Escuela de Postgrado San Francisco Xavier": ("Lima", "Lima"),
    "Universidad Ciencias de la Salud": ("Lima", "Lima"),
    "Universidad Privada SISE": ("Lima", "Lima"),
```

```
    "Universidad Inca Garcilaso de la Vega": ("Lima", "Lima"),
    "Universidad Peruana de Oriente": ("Loreto", "Maynas"),
    "Universidad Global del Cusco": ("Cusco", "Cusco"),
    "Universidad Privada Juan Pablo II": ("Lima", "Lima"),
    "Universidad Privada de la Selva Peruana": ("Huánuco", "Huánuco"),
    "Universidad de Ayacucho Federico Froebel": ("Ayacucho", "Huamanga"),
    "Universidad Privada de Ica": ("Ica", "Ica"),
    "Universidad Privada de Pucallpa": ("Ucayali", "Coronel Portillo"),
    "Universidad Particular de Chiclayo": ("Lambayeque", "Chiclayo"),
    "Universidad San Andrés": ("Lima", "Lima"),
    "Universidad Privada Arzobispo Loayza": ("Lima", "Lima"),
    "Universidad Privada Telesup": ("Lima", "Lima"),
    "Universidad Privada Sergio Bernales": ("Lima", "Lima"),
    "Universidad Peruana Simón Bolívar": ("Lima", "Lima"),
    "Universidad Marítima del Perú": ("Callao", "Callao"),
    "Universidad de Lambayeque": ("Lambayeque", "Chiclayo"),
    "Universidad Peruana de Investigación y Negocios": ("Lima", "Lima"),
    "Universidad Peruana de Integración Global S.A.C.": ("Lima", "Lima"),
    "Universidad Peruana de Arte Orval S.A.C.": ("Lima", "Lima")
}
```

[8]:
```
#Creamos las listas donde irán los datos
nombre_universidad = []
fecha_licen_resolución = []
departamento_provincia = []
tipo_gestion = []
status = []
```

[10]:
```
#Bucle de universidades licenciadas
for row in table_lic.find_all("tr")[1:]:
    universidad = row.find_all("td")[0].text.strip()
    fecha = row.find_all("td")[1].text.strip()
    depart = row.find_all("td")[3].text.strip()
    tipo = row.find_all("td")[4].text.strip()
    nombre_universidad.append(universidad)
    fecha_licen_resolución.append(fecha)
    departamento_provincia.append(depart)
    tipo_gestion.append(tipo)
    status.append("Licenciada")

#Bucle de universidades no licenciadas
for row in table_no_lic.find_all("tr")[1:]:
    universidad = row.find_all("td")[0].text.strip()
    fecha = row.find_all("td")[1].text.strip()
    tipo = row.find_all("td")[3].text.strip()
    departamento, provincia = departamentos_provincias_no_lic.get(universidad,␣
 ↪("Desconocido", "Desconocido"))
```

```
    nombre_universidad.append(universidad)
    fecha_licen_resolución.append(fecha)
    departamento_provincia.append(f"{departamento}/{provincia}")  #␣
 ↪Concatenamos el departamento y provincia
    tipo_gestion.append(tipo)
    status.append("No Licenciada")

df = pd.DataFrame({
    "NOMBRE DE LA UNIVERSIDAD": nombre_universidad,
    "FECHA LIC. / FECHA RESOL.": fecha_licen_resolución,
    "DEPARTAMENTO/PROVINCIA": departamento_provincia,
    "TIPO DE GESTIÓN": tipo_gestion,
    "STATUS" : status
})

df[['DEPARTAMENTO', 'PROVINCIA']] = df['DEPARTAMENTO/PROVINCIA'].str.split('/',␣
 ↪expand=True)

df = df.drop('DEPARTAMENTO/PROVINCIA', axis=1)

df
```

[10]:

	NOMBRE DE LA UNIVERSIDAD	FECHA LIC. / FECHA RESOL.	TIPO DE GESTIÓN	STATUS	DEPARTAMENTO
0	Universidad Privada de Trujillo	11-11-2024	Privada	Licenciada	Laredo
1	Universidad Peruana del Centro	18-07-2024	Privada	Licenciada	Junín
2	Universidad Nacional Ciro Alegría	06-10-2023	Pública	Licenciada	La Libertad
3	Universidad Nacional Pedro Ruiz Gallo	09-06-2023	Pública	Licenciada	Lambayeque
4	Universidad Nacional San Luis Gonzaga	15-01-2022	Pública	Licenciada	Ica
..	…	…	…	…	…
279	Universidad Marítima del Perú	18-01-2019	Privada	No Licenciada	Callao
280	Universidad de Lambayeque	21-12-2018	Privada	No Licenciada	Lambayeque
281	Universidad Peruana de Investigación y Negocios	29-11-2018	Privada	No Licenciada	Lima
282	Universidad Peruana de Integración Global S.A.C.	30-10-2018	Privada	No Licenciada	Lima
283	Universidad Peruana de Arte Orval S.A.C.	18-10-2018	Privada	No Licenciada	Lima

```
           PROVINCIA
0           Trujillo
1           Huancayo
2    Sánchez Carrión
3         Lambayeque
4                Ica
..               …
279          Callao
280        Chiclayo
281            Lima
282            Lima
283            Lima

[284 rows x 6 columns]
```

[11]:
```
#Exportamos los datos en archivo excel
df.to_excel("datos.xlsx", index=False)
```