


Semestre:	#1 Septiembre 2020 - Enero 2021	 Universitat Oberta de Catalunya
Asignatura:	Tipología y ciclo de vida de los datos	
Práctica:	#1	
Estudiante:	Félix Plúas Navarrete	

Descripción de la Práctica a realizar


El objetivo de esta actividad será la creación de un dataset a partir de los datos contenidos en una web. Para su realización, se deben cumplir los siguientes puntos:

1. Contexto. Explicar en qué contexto se ha recolectado la información. Explique por qué el sitio web elegido proporciona dicha información.
2. Definir un título para el dataset. Elegir un título que sea descriptivo.
3. Descripción del dataset. Desarrollar una descripción breve del conjunto de datos que se ha extraído (es necesario que esta descripción tenga sentido con el título elegido).
4. Representación gráfica. Presentar una imagen o esquema que identifique el dataset visualmente
5. Contenido. Explicar los campos que incluye el dataset, el periodo de tiempo de los datos y cómo se ha recogido.
6. Agradecimientos. Presentar al propietario del conjunto de datos. Es necesario incluir citas de investigación o análisis anteriores (si los hay).
7. Inspiración. Explique por qué es interesante este conjunto de datos y qué preguntas se pretenden responder.
8. Licencia. Seleccione una de estas licencias para su dataset y explique el motivo de su selección:
 - Released Under CC0: Public Domain License
 - Released Under CC BY-NC-SA 4.0 License
 - Released Under CC BY-SA 4.0 License
 - Database released under Open Database License, individual contents under Database Contents License
 - Other (specified above)
 - Unknown License
9. Código. Adjuntar el código con el que se ha generado el dataset, preferiblemente en Python o, alternativamente, en R.
10. Dataset. Publicación del dataset en formato CSV en Zenodo (obtención del DOI) con una breve descripción.

Recursos

Los siguientes recursos son de utilidad para la realización de la PEC:

- Subirats, L., Calvo, M. (2018). Web Scraping. Editorial UOC.
- Masip, D. El lenguaje Python. Editorial UOC.
- Lawson, R. (2015). Web Scraping with Python. Packt Publishing Ltd. Chapter 2. Scraping the Data.
- Simon Munzert, Christian Rubba, Peter Meißner, Dominic Nyhuis. (2015). Automated Data Collection with R: A Practical Guide to Web Scraping and Text Mining. John Wiley & Sons.
- Tutorial de Github <https://guides.github.com/activities/hello-world>.

Semestre:	#1 Septiembre 2020 - Enero 2021	 Universitat Oberta de Catalunya
Asignatura:	Tipología y ciclo de vida de los datos	
Práctica:	#1	
Estudiante:	Félix Plúas Navarrete	

Entregables:

URL GitHub:

URL: <https://github.com/fpluasn/Web-scraping-python>

Respuestas a preguntas planteadas:

1. Contexto. Explicar en qué contexto se ha recolectado la información. Explique por qué el sitio web elegido proporciona dicha información.

Hoy en día existe una alta demanda de información de empresas, personas, comercios como números de teléfono, direcciones, area de negocio, región, etc. que resulta de gran valor cuando se pretende establecer contacto en medios masivos. Gran parte de esta información se encuentra alojada en sitios que prestan servicios de páginas amarillas.

Muchos de estos sitios cobran el acceso a su información cuando se trata de consultas en grandes volúmenes ya sea por:

- periodos de tiempo
- región
- area de interés
- consulta filtrada por palabras

Sin embargo, es posible acceder a esta misma información en bloques más pequeños, lo que demanda una inversión mayor en recursos (personas, equipos) y tiempo a destinar para recolectar la información necesaria.


Tomando esto como premisa para este ejercicio hemos elegido al sitio web de “EDINA” <https://www.edina.com.ec> que es una de las páginas amarillas en Ecuador con acceso libre consultas limitadas para realizar un proceso de Web Scraping.

Imagen de sitio web de prueba:



2. Definir un título para el dataset. Elegir un título que sea descriptivo.

Catálogo de empresas

Semestre:	#1 Septiembre 2020 - Enero 2021	 Universitat Oberta de Catalunya
Asignatura:	Tipología y ciclo de vida de los datos	
Práctica:	#1	
Estudiante:	Félix Plúas Navarrete	

3. Descripción del dataset. Desarrollar una descripción breve del conjunto de datos que se ha extraído (es necesario que esta descripción tenga sentido con el título elegido).

El dataset contiene información general de empresas del Ecuador, información que es constantemente actualizada por los administradores de EDINA.

EDINA recolecta información de distintos entes públicos y privados y los pone a disposición en modalidad paga y no paga desde su plataforma.

En este caso nos es de mucha utilidad ya que, en lugar de consultar distintas fuentes para obtener el mismo resultado, nos enfocamos en un solo sitio que concentra la información requerida.

En esta práctica hemos abarcado los datos generales de cada compañía, pero es posible obtener información más detallada.

4. Representación gráfica. Presentar una imagen o esquema que identifique el dataset visualmente


Las empresas que recolectan datos para su posterior análisis, alcanzan sus objetivos en base a las decisiones que toman producto de los análisis previamente ejecutados.

Muchas de estas empresas contratan servicios en agendas digitales para ser localizados con mayor facilidad.

La siguiente imagen detalla cómo se forma el dataset y como se monetiza.



Es un ciclo donde el cliente que en primera instancia cancela para tener presencia en un medio digital, puede convertirse en un consumidor de la misma información para alcanzar sus objetivos de negocio.

Semestre:	#1 Septiembre 2020 - Enero 2021	 Universitat Oberta de Catalunya
Asignatura:	Tipología y ciclo de vida de los datos	
Práctica:	#1	
Estudiante:	Félix Plúas Navarrete	

5. Contenido. Explicar los campos que incluye el dataset, el periodo de tiempo de los datos y cómo se ha recogido.

El programa de Web Scraping realizado en esta práctica, fue escrito en Python en su versión 3.8 con el apoyo de algunas librerías que serán detalladas en un punto posterior.

La información almacenada en EDINA es actualizada constantemente por lo que si se necesitase contar con esta información de forma regular la recomendación sería ejecutar el proceso en periodos mensuales.

A continuación, se detallan los campos obtenidos para este dataset.

- a) Company: Nombre de la compañía.
- b) Description: Descripción o reseña de la empresa.
- c) Area: Tipo de negocio / servicio.
- d) Region: Provincia / Ciudad a la que pertenece.
- e) Address: Dirección de la oficina / Matriz

Es posible obtener otro tipo de detalle como números telefónicos y sucursales realizando subconsultas en el sitio por cada registro encontrado.


6. Agradecimientos. Presentar al propietario del conjunto de datos. Es necesario incluir citas de investigación o análisis anteriores (si los hay).

El representante legal de EDINA es el Sr. Chiriboga Guevara Ramiro, intenté contactarme con ellos mediante la sección de contactos del sitio web, pero no obtuve respuesta. En el sitio web no se encuentra información referente a política de privacidad de datos o procesos de extracción.

Por ética profesional el Web Scraping solo se ejecutó durante la ejecución de esta práctica en una sola página.

Para el desarrollo de esta práctica he tomado como referencia los siguientes trabajos:

- <https://jarroba.com/scraping-python-beautifulsoup-ejemplos/>
- <https://hackernoon.com/web-scraping-con-python-guia-paso-a-paso-1p1l33vu>
- <https://likegeeks.com/es/web-scraping-beautiful-soup-y-selenium/>

Semestre:	#1 Septiembre 2020 - Enero 2021	 Universitat Oberta de Catalunya
Asignatura:	Tipología y ciclo de vida de los datos	
Práctica:	#1	
Estudiante:	Félix Plúas Navarrete	

7. Inspiración. Explique por qué es interesante este conjunto de datos y qué preguntas se pretenden responder.

Este set de datos como tal nos ayuda como complemento en otros análisis como, por ejemplo, participación de empresas guayaquileñas en las exportaciones.

El poder contar con información actualizada ya sea general o detallada de las diferentes instituciones del país nos ayudaría a segmentar de forma asertiva y no en base a un supuesto.

Preguntas que podríamos responder con este set de datos pueden ser:

- ¿Cuántas empresas pertenecen a la region “X” del Ecuador?
- ¿Cuáles son las empresas más antiguas dedicadas a la actividad “X”?
- ¿Qué áreas o segmentos cubren las empresas de la ciudad “X” en Ecuador?

8. Licencia. Seleccione una de estas licencias para su dataset y explique el motivo de su selección:

- Released Under CC0: Public Domain License
- Released Under CC BY-NC-SA 4.0 License
- Released Under CC BY-SA 4.0 License
- Database released under Open Database License, individual contents under Database Contents License
- Other (specified above) o Unknown License


Para este dataset considero apropiado la licencia “Released Under CC BY-SA 4.0 License” ya que nos permite:

- Compartir: copia y redistribuye el material en cualquier medio o formato.
- Adaptarse: remezclar, transformar y construir sobre el material para cualquier propósito, incluso comercial.

Respetando las siguientes consideraciones

- Atribución: debe otorgar el crédito correspondiente, proporcionar un enlace a la licencia e indicar si se realizaron cambios. Puede hacerlo de cualquier manera razonable, pero no de ninguna manera que sugiera que el licenciante lo respalda a usted o su uso.
- ShareAlike: si remezcla, transforma o construye sobre el material, debe distribuir sus contribuciones bajo la misma licencia que el original.

De esta forma impulsamos a personas y organizaciones que pudiesen estar interesados en este set de datos a desarrollar nuevos proyectos respetando la autoría mediante el reconocimiento de su fuente original

Semestre:	#1 Septiembre 2020 - Enero 2021	 Universitat Oberta de Catalunya
Asignatura:	Tipología y ciclo de vida de los datos	
Práctica:	#1	
Estudiante:	Félix Plúas Navarrete	

9. Código. Adjuntar el código con el que se ha generado el dataset, preferiblemente en Python o, alternativamente, en R.

El código fue hecho en lenguaje Python con la implementación de la librería Selenium. El código fuente se encuentra en el repositorio de GitHub


Url: <https://n9.cl/blel5>

10. Dataset. Publicación del dataset en formato CSV en Zenodo (obtención del DOI) con una breve descripción.

El dataset se encuentra en el repositorio de GitHub

Url: <https://n9.cl/yeiu2>

Contribuciones	Firma
Investigación previa	FP
Redacción de las respuestas	FP
Desarrollo código	FP

Semestre:	#1 Septiembre 2020 - Enero 2021	 Universitat Oberta de Catalunya
Asignatura:	Tipología y ciclo de vida de los datos	
Práctica:	#1	
Estudiante:	Félix Plúas Navarrete	

Código Python

#INSTALACION LIBRERIAS

!pip install Selenium

```
Requirement already satisfied: selenium in c:\users\fpluas\.conda\envs\tipologi
adatos\lib\site-packages (3.141.0)
Requirement already satisfied: urllib3 in c:\users\fpluas\.conda\envs\tipolog
iadatos\lib\site-packages (from selenium) (1.25.11)
```

!pip install BeautifulSoup4

```
Requirement already satisfied: BeautifulSoup4 in c:\users\fpluas\.conda\envs
\tipologiadatos\lib\site-packages (4.9.3)
Requirement already satisfied: soupsieve>1.2; python_version >= "3.0" in c:\u
sers\fpluas\.conda\envs\tipologiadatos\lib\site-packages (from BeautifulSoup
4) (2.0.1)
```

!pip install pandas

```
Requirement already satisfied: pandas in c:\users\fpluas\.conda\envs\tipologi
adatos\lib\site-packages (1.1.3)
Requirement already satisfied: pytz>=2017.2 in c:\users\fpluas\.conda\envs\ti
pologiadatos\lib\site-packages (from pandas) (2020.1)
Requirement already satisfied: numpy>=1.15.4 in c:\users\fpluas\.conda\envs\t
ipologiadatos\lib\site-packages (from pandas) (1.19.3)
Requirement already satisfied: python-dateutil>=2.7.3 in c:\users\fpluas\.con
da\envs\tipologiadatos\lib\site-packages (from pandas) (2.8.1)
Requirement already satisfied: six>=1.5 in c:\users\fpluas\.conda\envs\tipolo
giadatos\lib\site-packages (from python-dateutil>=2.7.3->pandas) (1.15.0)
```

#IMPORTACION DE LIBRERIAS


```
from selenium import webdriver
```

```
from bs4 import BeautifulSoup
```

```
import pandas as pd
```

#DEFINICION DE CHROMEDRIVER

```
driver =
webdriver.Chrome(executable_path=r'C:/chromedriver_win32/chromedriver.e
xe')
```

Semestre:	#1 Septiembre 2020 - Enero 2021	 Universitat Oberta de Catalunya
Asignatura:	Tipología y ciclo de vida de los datos	
Práctica:	#1	
Estudiante:	Félix Plúas Navarrete	

#DEFINICION DE ARREGLOS Y URL PARA WEBSCRAPING

```
company=[]
```

```
area=[]
```

```
description=[]
```

```
address=[]
```

```
region=[]
```

```
driver.get('https://www.edina.com.ec/Buscador?b=empresas&c=ecuador&pagina=1')
```

#OBTENEMOS EL CODIGO FUENTE DE LA VISTA

```
content = driver.page_source
```

#ALMACENAMOS LA ESTRUCTURA HTML

```
soup = BeautifulSoup(content)
```


#VISUALIZAMOS EL CODIGO FUENTE

```
print(content)
```

```
<html lang="en" class="no-js"><head>
  <meta charset="utf-8">
  <meta http-equiv="X-UA-Compatible" content="IE=edge">
  <meta name="viewport" content="width=device-width, initial-scale=1, maximum-scale=1">

  <!-- Chrome, Firefox OS, Opera and Vivaldi -->
  <meta name="theme-color" content="#FACC2E">
  <!-- Windows Phone -->
  <meta name="msapplication-navbutton-color" content="#FACC2E">
  <!-- iOS Safari -->
  <meta name="apple-mobile-web-app-status-bar-style" content="#FACC2E">

  <title>empresas en Ecuador | PáginasAmarillas EC</title>
  <meta name="description" content="Buscar empresas en Ecuador, Páginas Amarillas Ecuador. Encuentre información útil, dirección, horarios, y el número de teléfono de las empresas, servicios más importantes del Ecuador.">
  <meta name="keywords" content="empresas,Ecuador,paginas amarillas ecuador,guía telefonica,guía telefonica ecuador,paginas amarillas ecuador">
  <meta name="robots" content="index, follow">
```


Semestre:	#1 Septiembre 2020 - Enero 2021	 Universitat Oberta de Catalunya
Asignatura:	Tipología y ciclo de vida de los datos	
Práctica:	#1	
Estudiante:	Félix Plúas Navarrete	

#ALMACENAMOS LOS DIFERENTES DATOS QUE NOS PROPORCIONA LA PAGINA

```
for a in soup.findAll('div', attrs={'class':'place-post__content'}):

    companyName=a.find('a',href=True, attrs={'itemprop':'name'})

    areaCompany=a.find('p',attrs={'class':'place-post__description'})

    descriptionCompany=a.find('p',attrs={'itemprop':'streetAddress'})

    addressCompany=a.find('span',attrs={'itemprop':'addressLocality'})

    regionCompany=a.find('span',attrs={'itemprop':'addressRegion'})


    company.append(companyName.text)

    area.append(areaCompany.text)

    description.append(descriptionCompany.text)

    address.append(addressCompany.text)

    region.append(regionCompany.text)
```


#ALMACENAMOS EN UN DATAFRAME LOS DATOS RECOLECTADOS

```
df=pd.DataFrame({'Company':company,'Description':description,'Area':area,'Region':region,'Address':address})
```

#IMPRIMIMOS DATAFRAME

```
print(df)
```

	Company	Description
0	Empresa Eléctrica Regional Centro Sur C.A.	
1	Actuaria Cía. Ltda.	
2	Emac Empresa Municipal De Aseo De Cuenca	
3	Corporación Nacional de Electricidad CNEL EP ...	
4	CACPEG Cooperativa de Ahorro y Crédito de La ...	
5	EMUCE - Empresa Municipal de Cementerios y Ex...	
6	Empresa Municipal de Agua Potable, Alcantaril...	
7	Elecaustro	
8	Empresa Pública Agropzaching	
9	Af Control de Plagas	
10	FUMIGSA S.A.	
11	Solconplag	
12	Seguros del Pichincha	
13	Zurich Seguros Ecuador S.A.	
14	ALSECA	

Semestre:	#1 Septiembre 2020 - Enero 2021	 Universitat Oberta de Catalunya
Asignatura:	Tipología y ciclo de vida de los datos	
Práctica:	#1	
Estudiante:	Félix Plúas Navarrete	

#DEFINIMOS UNA FUNCION PARA LIMPIAR LOS ESPACIOS EN EL DATAFRAME

```
def trim_all_columns(df):
    trim_strings = lambda x: x.strip() if isinstance(x, str) else x
    return df.applymap(trim_strings)
```

#LIMPIAMOS LOS ESPACIOS EN EL DATAFRAME

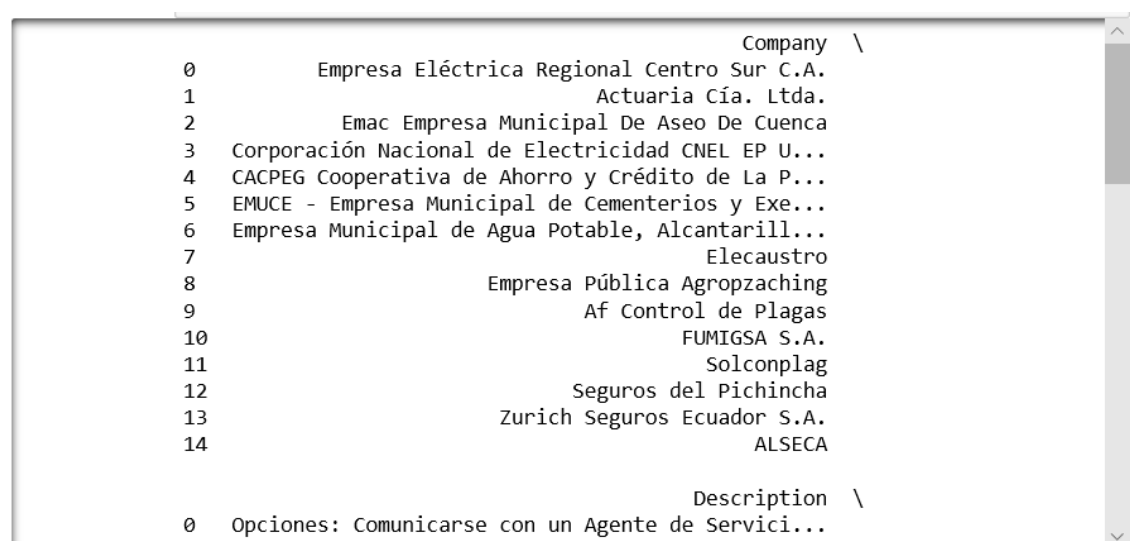
```
df = trim_all_columns(df)
```

#REMOVEMOS COMILLAS Y COMAS DE LAS COLUMNAS DEL DATAFRAME

```
df["Company"].replace({' ': ' ', '"': ' '}, inplace=True)
df["Description"].replace({' ': ' ', '"': ' '}, inplace=True)
df["Area"].replace({' ': ' ', '"': ' '}, inplace=True)
df["Region"].replace({' ': ' ', '"': ' '}, inplace=True)
df["Address"].replace({' ': ' ', '"': ' '}, inplace=True)
```

#IMPRIMIMOS DATAFRAME

```
print(df)
```



	Company \	Description \
0	Empresa Eléctrica Regional Centro Sur C.A.	Opciones: Comunicarse con un Agente de Servi...
1	Actuaria Cía. Ltda.	Empresa creada en el año 1996, ofrece a sus cl...
2	Emac Empresa Municipal De Aseo De Cuenca	
3	Corporación Nacional de Electricidad CNEL EP U...	
4	CACPEG Cooperativa de Ahorro y Crédito de La P...	
5	EMUCE - Empresa Municipal de Cementerios y Exe...	
6	Empresa Municipal de Agua Potable, Alcantarill...	
7	Elecaustro	
8	Empresa Pública Agropzaching	
9	Af Control de Plagas	
10	FUMIGSA S.A.	
11	Solconplag	
12	Seguros del Pichincha	
13	Zurich Seguros Ecuador S.A.	
14	ALSECA	

#ALMACENAMOS EL DATAFRAME EN UN CSV

```
df.to_csv('Company.csv', index=False, encoding='utf-8')
```