DESARROLLO PRACTICA 1 (WEB SCRAPING)

ROSEMBERG ALVAREZ DIAZ

UNIVERSITAT OBERTA DE CATALUNYA
TIPOLOGÍA Y CICLO DE VIDA DE LOS DATOS
BOGOTA D.C.
2020

TABLA DE CONTENIDO

TABLA DE ILUSTRACIONES

CONTEXTO

En algún momento de nuestras vidas como padres, nos enfrentamos a la búsqueda de colegio para nuestros hijos, no una, si no varias veces y en diferentes etapas de nuestras vidas. En este proceso de búsqueda nos enfrentamos con no poder acceder a información veraz y completa de instituciones educativas que apoyara el proceso de selección. A su vez, instituciones estatales desean conocer la oferta de los mejores colegios privados para definir políticas orientadas a mejorar la calidad en la prestación del servicio de las instituciones públicas. Finalmente, la importancia monitorizar el impacto del Coronavirus en educación en las instituciones privadas de Colombia.

En base a estas necesidades la empresa CIPRES MERCADEO EDUCATIVO se ha dedicado a contactar las instituciones educativas colombianas para que hagan parte de la comunidad de los mejores colegios privados. Aquellas instituciones que esten interesadas deberán facilitar datos de caracterización; como por ejemplo, si es: privado, bilingue, mixto, femenino, masculino, catolico, ciudad de ubicación, etc.

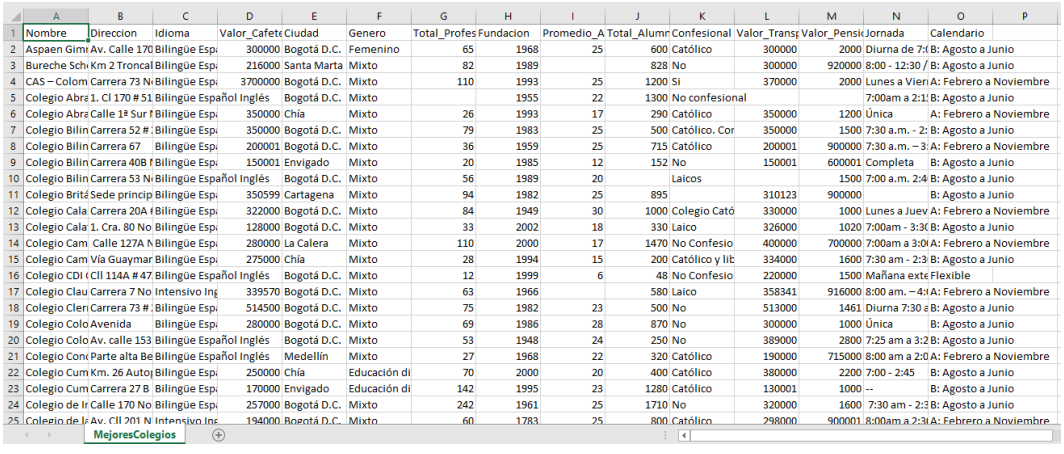
TÍTULO

OFERTA DE LOS MEJORES COLEGIOS DE COLOMBIA

DESCRIPCIÓN

Este conjunto de datos provee la oferta de las mejores instituciones educativas privadas en Colombia a través de un conjunto de datos de caracterización completo, veraz y actualizado.

REPRESENTACIÓN GRÁFICA



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	Nombre	Direccion	Idioma	Valor_Cafete	Ciudad	Genero	Total_Profes	Fundacion	Promedio_A	Total_Alumn	Confesional	Valor_Transp	Valor_Pensi	Jornada	Calendario	
2	Aspaen Gimi	Av. Calle 170	Bilingüe Espa	300000	Bogotá D.C.	Femenino	65	1968	25	600	Católico	300000	2000	Diurna de 7:(	B: Agosto a Junio	
3	Bureche Sch	Km 2 Troncal	Bilingüe Espa	216000	Santa Marta	Mixto	82	1989		828	No	300000	920000	8:00 - 12:30 /	B: Agosto a Junio	
4	CAS – Colom	Carrera 73 N	Bilingüe Espa	3700000	Bogotá D.C.	Mixto	110	1993	25	1200	Si	370000	2000	Lunes a Vier	A: Febrero a Noviembre	
5	Colegio Abra	1. Cl 170 # 51	Bilingüe Español Inglés		Bogotá D.C.	Mixto		1955	22	1300	No confesional			7:00am a 2:1	B: Agosto a Junio	
6	Colegio Abra	Calle 1ª Sur	Bilingüe Espa	350000	Chía	Mixto	26	1993	17	290	Católico	350000	1200	Única	A: Febrero a Noviembre	
7	Colegio Bilin	Carrera 52 #	Bilingüe Espa	350000	Bogotá D.C.	Mixto	79	1983	25	500	Católico. Cor	350000	1500	7:30 a.m. - 2:	B: Agosto a Junio	
8	Colegio Bilin	Carrera 67	Bilingüe Espa	200001	Bogotá D.C.	Mixto	36	1959	25	715	Católico	200001	900000	7:30 a.m. – 3	A: Febrero a Noviembre	
9	Colegio Bilin	Carrera 40B	Bilingüe Espa	150001	Envigado	Mixto	20	1985	12	152	No	150001	600001	Completa	B: Agosto a Junio	
10	Colegio Bilin	Carrera 53 N	Bilingüe Español Inglés		Bogotá D.C.	Mixto	56	1989	20		Laicos		1500	7:00 a.m. 2:4	B: Agosto a Junio	
11	Colegio Britá	Sede princip	Bilingüe Espa	350599	Cartagena	Mixto	94	1982	25	895		310123	900000		B: Agosto a Junio	
12	Colegio Cala	Carrera 20A	Bilingüe Espa	322000	Bogotá D.C.	Mixto	84	1949	30	1000	Colegio Cató	330000	1000	Lunes a Juev	A: Febrero a Noviembre	
13	Colegio Cala	1. Cra. 80 No	Bilingüe Espa	128000	Bogotá D.C.	Mixto	33	2002	18	330	Laico	326000	1020	7:00am - 3:3(	B: Agosto a Junio	
14	Colegio Cam	Calle 127A N	Bilingüe Espa	280000	La Calera	Mixto	110	2000	17	1470	No Confesio	400000	700000	7:00am a 3:0(	A: Febrero a Noviembre	
15	Colegio Cam	Vía Guaymar	Bilingüe Espa	275000	Chía	Mixto	28	1994	15	200	Católico y lib	334000	1600	7:30 am - 2:3	B: Agosto a Junio	
16	Colegio CDI (	Cll 114A # 47	Bilingüe Español Inglés		Bogotá D.C.	Mixto	12	1999	6	48	No Confesio	220000	1500	Mañana exte	Flexible	
17	Colegio Clau	Carrera 7 No	Intensivo Ing	339570	Bogotá D.C.	Mixto	63	1966		580	Laico	358341	916000	8:00 am. – 4:(	A: Febrero a Noviembre	
18	Colegio Cler	Carrera 73 #	Bilingüe Espa	514500	Bogotá D.C.	Mixto	75	1982	23	500	No	513000	1461	Diurna 7:30 a	B: Agosto a Junio	
19	Colegio Colo	Avenida	Bilingüe Espa	280000	Bogotá D.C.	Mixto	69	1986	28	870	No	300000	1000	Única	B: Agosto a Junio	
20	Colegio Colo	Av. calle 153	Bilingüe Español Inglés		Bogotá D.C.	Mixto	53	1948	24	250	No	389000	2800	7:25 am a 3:2	B: Agosto a Junio	
21	Colegio Con(	Parte alta Be	Bilingüe Español Inglés		Medellín	Mixto	27	1968	22	320	Católico	190000	715000	8:00 am a 2:0	A: Febrero a Noviembre	
22	Colegio Cum	Km. 26 Autop	Bilingüe Espa	250000	Chía	Educación di	70	2000	20	400	Católico	380000	2200	7:00 - 2:45	B: Agosto a Junio	
23	Colegio Cum	Carrera 27 B	Bilingüe Espa	170000	Envigado	Educación di	142	1995	23	1280	Católico	130001	1000	--	B: Agosto a Junio	
24	Colegio de Ir	Calle 170 No	Bilingüe Espa	257000	Bogotá D.C.	Mixto	242	1961	25	1710	No	320000	1600	7:30 am - 2:3	B: Agosto a Junio	
25	Colegio de la	Av. Cll 201 N	Intensivo Ing	194000	Bogotá D.C.	Mixto	60	1783	25	800	Católico	298000	900001	8:00am a 2:3(	A: Febrero a Noviembre	

MejoresColegios

Ilustración 1 - DataSet Mejores Colegios Privados de Colombia

CONTENIDO

VARIABLE	TIPO	DESCRIPCIÓN
Nombre	Texto	Razón social de la institución educativa.
Dirección	Texto	Matricula catastral donde se encuentra ubicada la institución educativa.
Idioma	Texto	Lenguas en las cual se imparte las clases en la institución educativa
Valor_Cafeteria	Numérico	Valor mensual en pesos colombianos del servicio de cafetería.
Ciudad	Texto	Nombre de la ciudad en donde se encuentra ubicado la institución educativa.
Genero	Texto	Sexo(s) admitidos por la institución educativa.
Total_Profesores	Numérico	Cantidad de profesores que hacen parte de la institución educativa.
Fundacion	Numérico	Año en que fue creado la institución educativa.
Promedio_Alumnos	Numérico	Promedio de estudiantes por clase.
Total_Alumnos	Numérico	Cantidad de alumnos que hacen parte de la institución educativa.
Confesional	Texto	Culto religioso que se profesa en la institución educativa.
Valor_Transporte	Numérico	Valor mensual en pesos colombianos de la ruta escolar.
Valor_Pension	Numérico	Valor mensual en pesos colombianos de la pensión.
Jornada	Texto	tiempo que dedica la institución educativa a sus estudiantes en la prestación directa del servicio educativo.
Calendario	Texto	meses de iniciación y finalización de las actividades académicas.

El proceso de recolección se inicia ingresando a la url https://losmejorescolegios.com/colegios en donde podemos encontrar un listado de las instituciones educativas que hacen parte de la comunidad de los mejores colegios (Ver Ilustración 2)

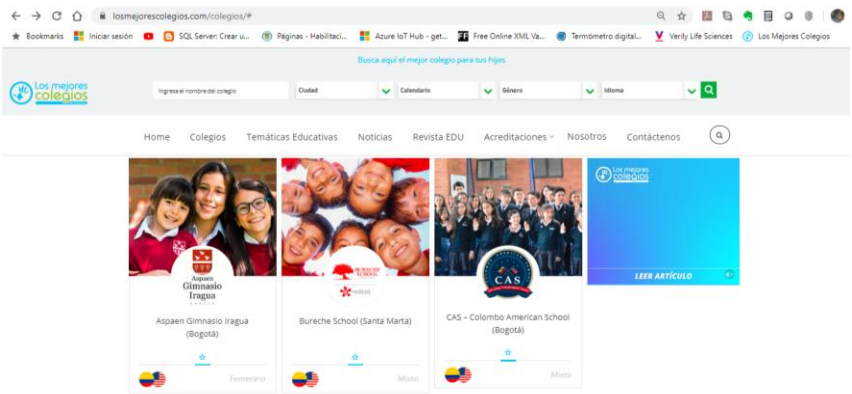


Ilustración 2 - https://losmejorescolegios.com/colegios

Se procede a recorrer el correspondiente listado para extraer las URLS que nos permitiran acceder a los datos de caracterización de cada una las instituciones educativas. El nodo que contiene la URL se encuentra identificado con el selector XPath "h2.course-title a" del cual extraemos el enlace a través en el atributo "href" (Ver Ilustración 3).

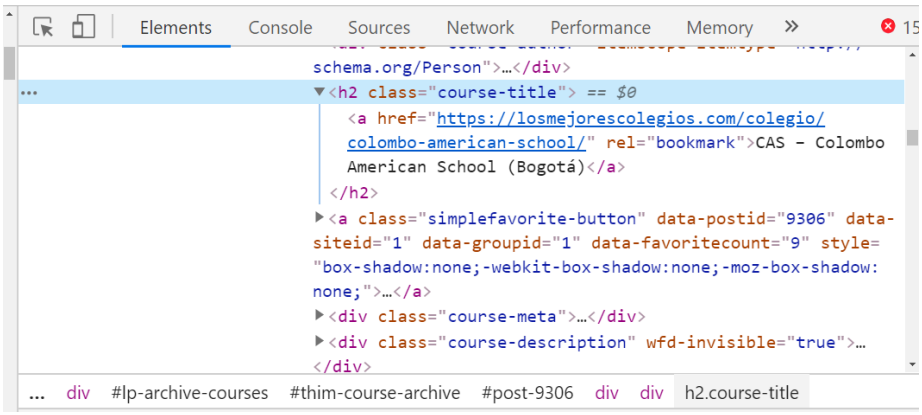


Ilustración 3 - Selector XPath URL

Seguidamente procedemos acceder a cada uno de los documentos HTML que contiene la información de caracterización de nuestro interés recorriendo el listado de URLS previamente creado. Una vez se accede a la página web de la institución educativa procedemos a extraer la siguiente información:

El nodo que contiene la razón social de la institución educativa se encuentra identificado con el selector XPath " div.contenedor-titulo-colegio h2.entry-title" del cual extraemos el texto (Ver Ilustración 4).

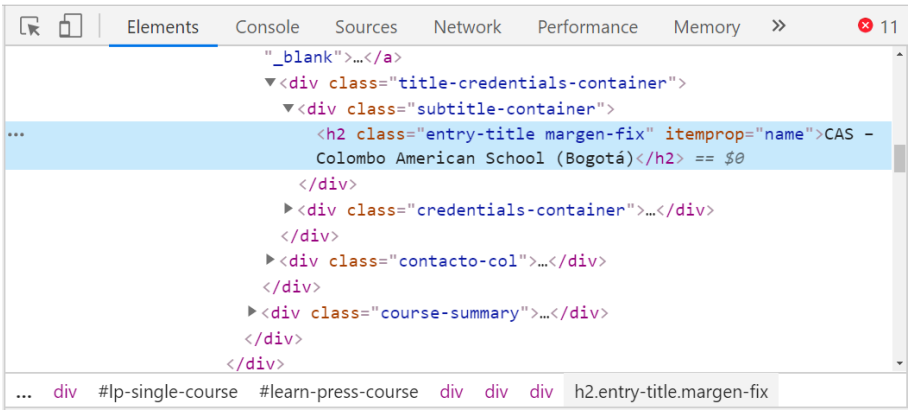


Ilustración 4 - Selector XPath Nombre

El nodo que contiene la dirección se encuentra identificado con el selector XPath " div.item-direccion div.field-value" del cual extraemos el texto (Ver Ilustración 5).

Ilustración 5 - Selector XPath Dirección

El nodo que contiene el idioma se encuentra identificado con el selector XPath " div.item-Idioma div.field-value" del cual extraemos el texto (Ver Ilustración 6).

Ilustración 6 - Selector XPath Idioma

El nodo que contiene el valor de la cafetería se encuentra identificado con el selector XPath " div.item-costo_cafeteria div.field-value" del cual extraemos el texto (Ver Ilustración 7).

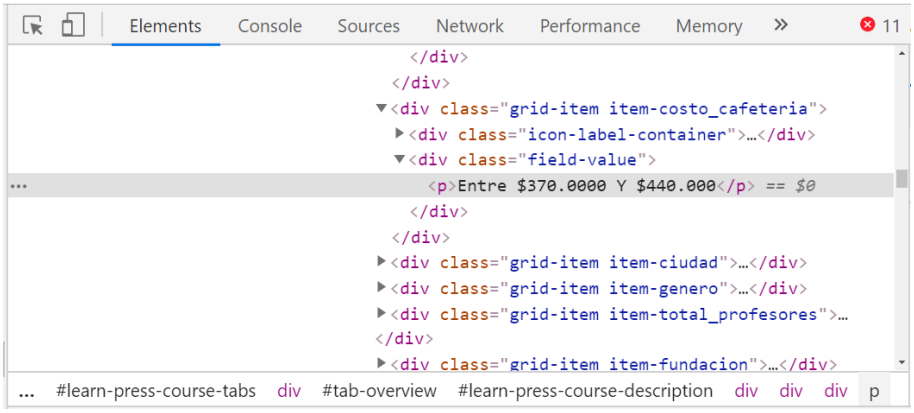


Ilustración 7 - Selector XPath Valor de la cafetería

El nodo que contiene la ciudad se encuentra identificado con el selector XPath " div.item-ciudad div.field-value" del cual extraemos el texto (Ver Ilustración 8).

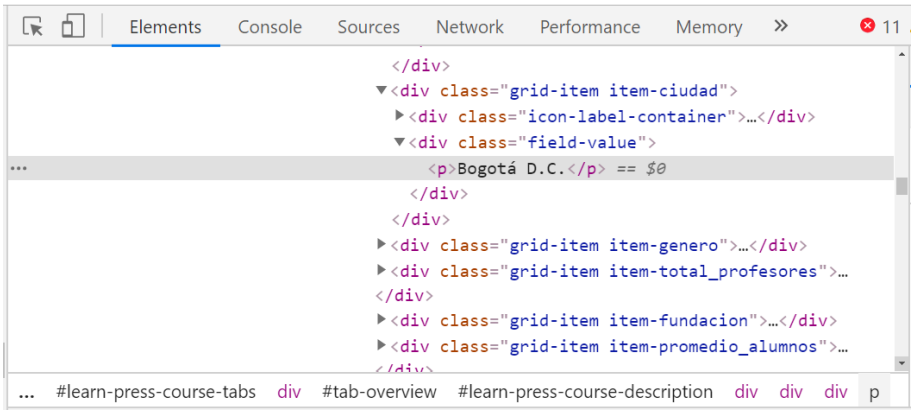


Ilustración 8 - Selector XPath Ciudad

El nodo que contiene el género se encuentra identificado con el selector XPath " div.item-genero div.field-value" del cual extraemos el texto (Ver Ilustración 9).

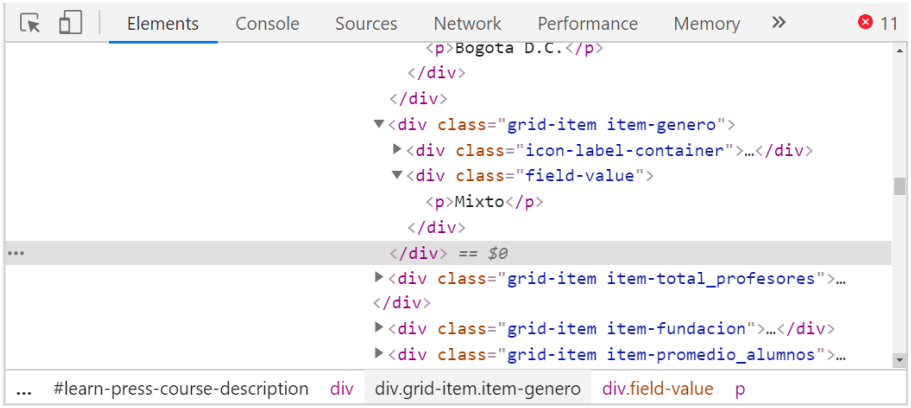


Ilustración 9 - Selector XPath Género

El nodo que contiene la cantidad total de profesores se encuentra identificado con el selector XPath " div.item-total_profesores div.field-value" del cual extraemos el texto (Ver Ilustración 10).

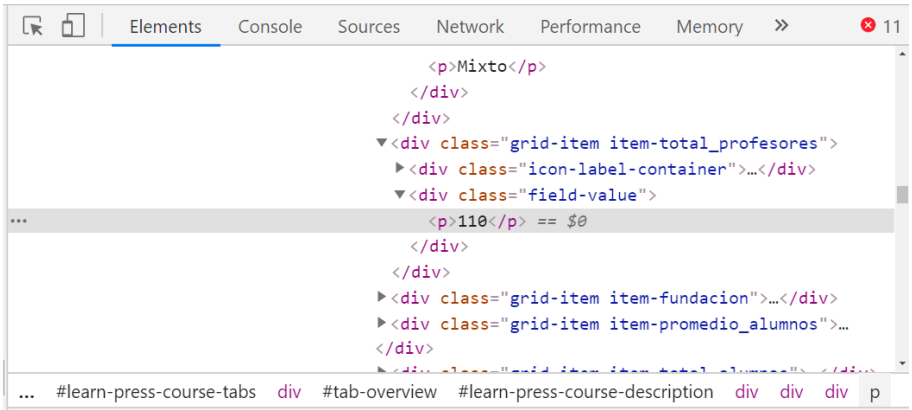


Ilustración 10 - Selector XPath Total Profesores

El nodo que contiene el año de fundación se encuentra identificado con el selector XPath " div.item-fundacion div.field-value" del cual extraemos el texto (Ver Ilustración 11).

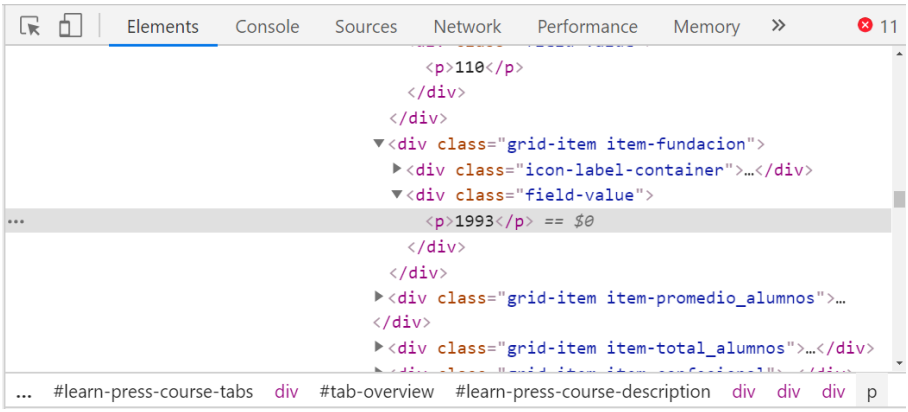


Ilustración 11 - Selector XPath Fundación

El nodo que contiene el promedio de alumnos por salón de clases se encuentra identificado con el selector XPath " div.item-promedio_alumnos div.field-value" del cual extraemos el texto (Ver Ilustración 12).

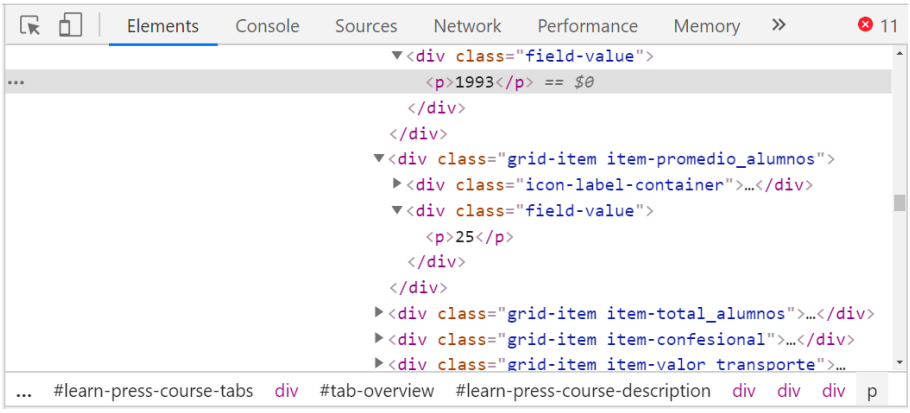


Ilustración 12 - Selector XPath Promedio Alumnos

El nodo que contiene el total de alumnos se encuentra identificado con el selector XPath " div.item-total_alumnos div.field-value" del cual extraemos el texto (Ver Ilustración 13).

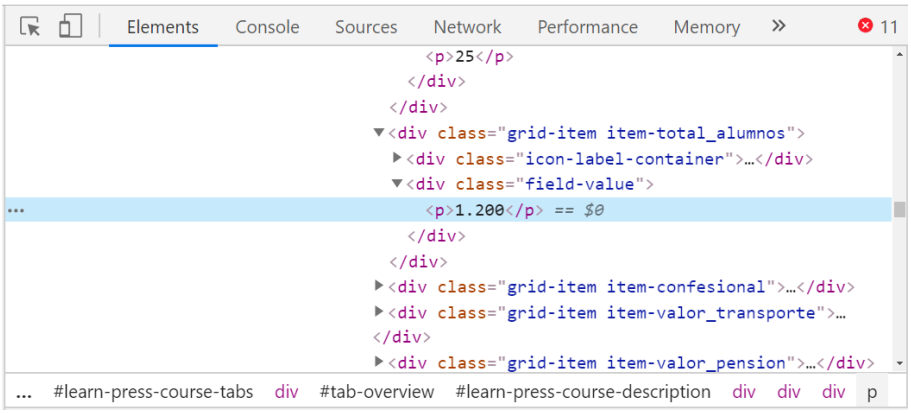


Ilustración 13 - Selector XPath Total Alumnos

El nodo que contiene el culto religioso que se profesa en la institución educativa se encuentra identificado con el selector XPath " div.item-confesional div.field-value" del cual extraemos el texto (Ver Ilustración 14).

Ilustración 14 - Selector XPath Total Confesional

El nodo que contiene el valor del transporte se encuentra identificado con el selector XPath " div.item-valor_transporte div.field-value" del cual extraemos el texto (Ver Ilustración 15).

Ilustración 15 - Selector XPath Valor Transporte

El nodo que contiene el valor de la pensión se encuentra identificado con el selector XPath " div.item-valor_pension div.field-value" del cual extraemos el texto (Ver Ilustración 16).

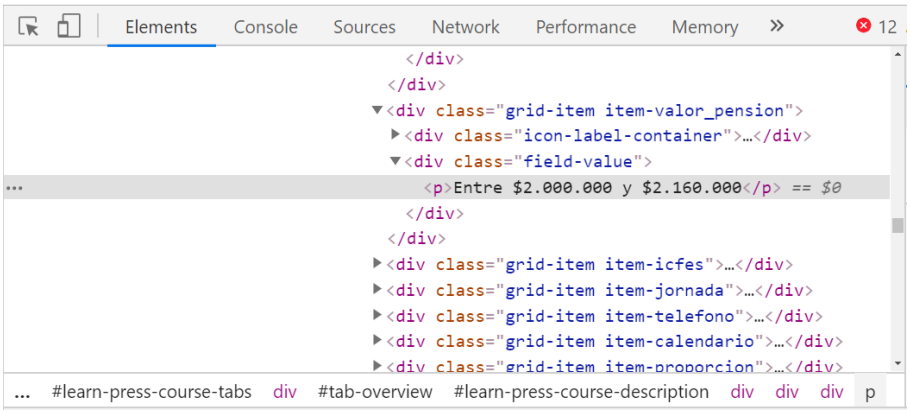


Ilustración 16 - Selector XPath Valor Pensión

El nodo que contiene el tiempo que dedica la institución educativa a sus estudiantes en la prestación directa del servicio educativo, se encuentra identificado con el selector XPath " div.item-jornada div.field-value" del cual extraemos el texto (Ver Ilustración 17).

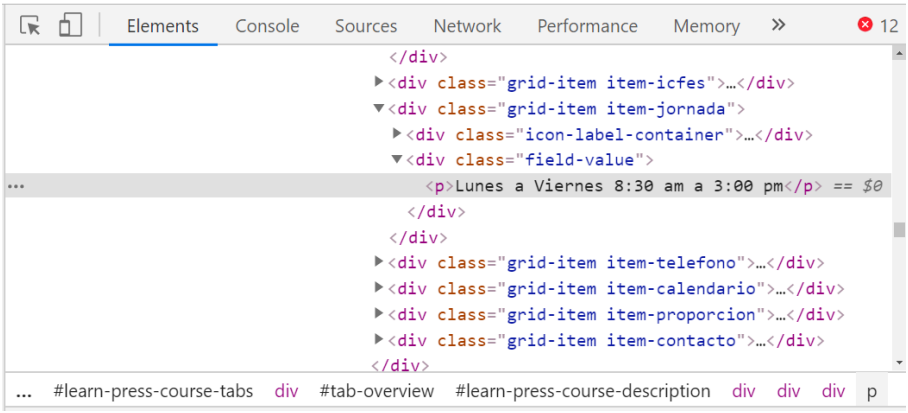


Ilustración 17 - Selector XPath Jornada

El nodo que contiene el mes de iniciación y finalización de las actividades académicas., se encuentra identificado con el selector XPath " div.item-calendario div.field-value" del cual extraemos el texto (Ver Ilustración 18).

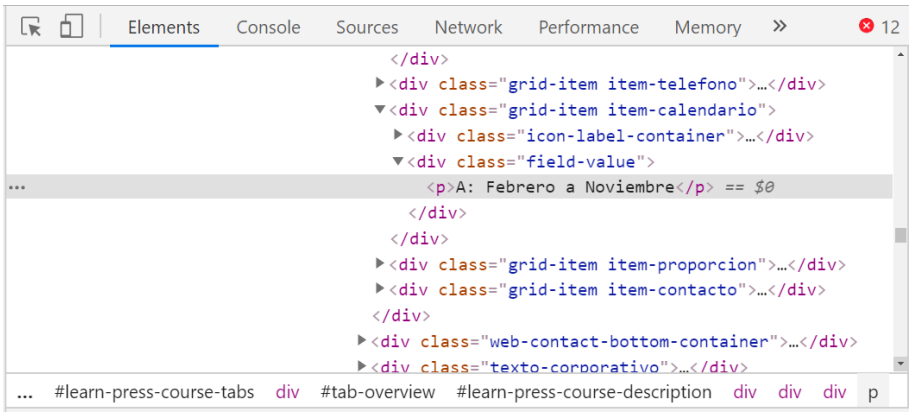


Ilustración 18 - Selector XPath Calendario

Finalmente obtenemos el dataset con la oferta de las mejores instituciones educativas de Colombia.

INSPIRACIÓN

En algún momento de nuestras vidas como padres, nos enfrentamos a la búsqueda de colegio para nuestros hijos, no una, si no varias veces y en diferentes etapas de nuestras vidas. En este proceso de búsqueda nos enfrentamos con no poder acceder a información veraz y completa de instituciones educativas que apoyara el proceso de selección. A su vez, instituciones estatales desean conocer la oferta de los mejores colegios privados para definir políticas orientadas a mejorar la calidad en la prestación del servicio educativo por parte de las instituciones públicas. Finalmente, permite resolver preguntas como:

¿En dónde se da la mayor concentración de estas instituciones educativas?

¿Cuál es el costo promedio mensual de los servicios educativos?

¿Cuál es la proporción de instituciones educativas que ofertan el bilingüismo?

¿Cuál es la relación entre profesores y alumnos?

¿Cual es la correlación entre el culto religioso y el sexo de admisión de los estudiantes?

¿Cuál es el costo promedio de los servicos de Transporte, Cafetería y Pensión?

¿Cuántas instituciones han cerra por la pandemia de la covid-19?

¿Qué proporción de variación a presentando los costos de pensión, transporte y cafetería por el coronavirus?

LICENCIA

La licencia que se le otorga al DataSet es “Released Under CC0: Public Domain License”. Se toma esta decisión partiendo de la premisa de que el portal https://losmejorescolegios.com es de acceso libre a toda la población, no existen

exclusiones en los directorios y/o paginas concretas, el acceso es completo a todos los robots y de que la información extraída no contiene datos sensibles que sean protegidos a través de las leyes relacionas con Habeas Data. Por lo tanto, se podrá copiar, modificar, distribuir e interpretar el DataSet, incluso para propósitos comerciales, sin pedir permiso.

CÓDIGO

Si se desea ver el código correspondiente podrá hacerlo revisando los documentos ALVAREZ_DIAZ_Web_Scraping.rmd o ALVAREZ_DIAZ_Web_Scraping.pdf. Es importante aclarar que esta práctica se desarrolló en R.

PUBLICACIÓN ZENODO

Se lleva a cabo la publicación del DataSet “OFERTA DE LOS MEJORES COLEGIOS DE COLOMBIA” en la plataforma ZENODO. Se puede llevar a cabo la búsqueda del correspondiente DataSet, ingresando en el buscador de la plataforma la frase “MEJORES COLEGIOS”. (Ver Ilustración 19).

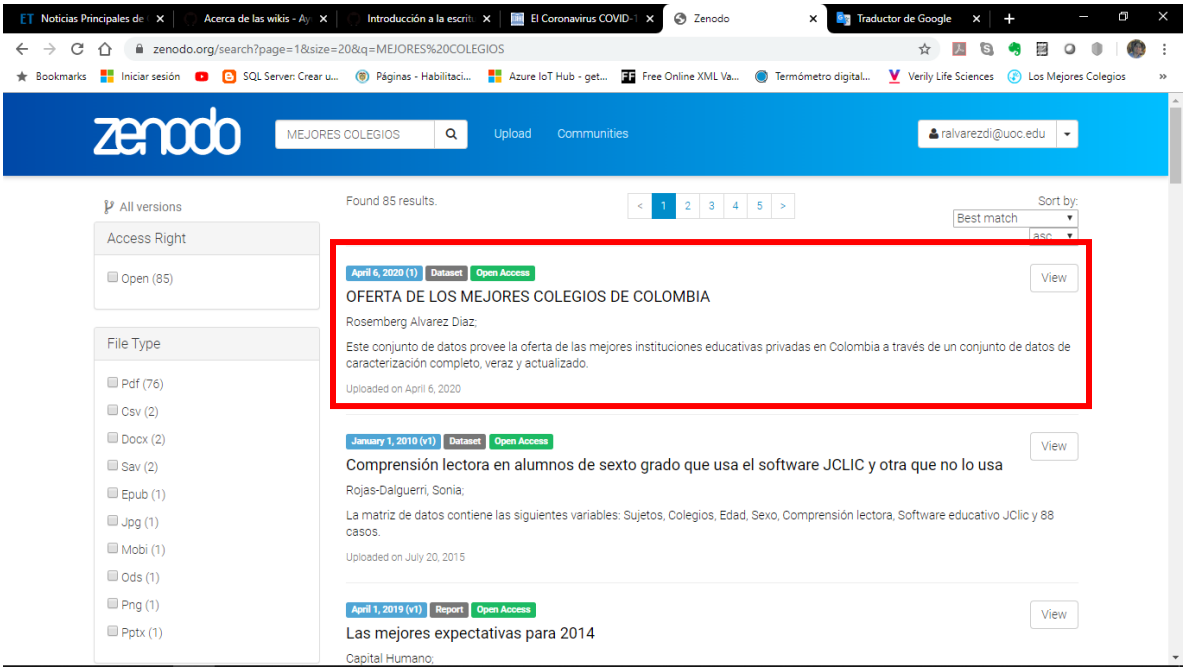


Ilustración 19 - Busqueda Zenodo

CONTRIBUCIÓN

COMPONENTE	FIRMA
Investigación previa	RAD (Rosemberg Álvarez Diaz)
Redacción de las respuestas	RAD (Rosemberg Álvarez Diaz)
Desarrollo código	RAD (Rosemberg Álvarez Diaz)