



**UNIVERSIDAD DE GUAYAQUIL
FACULTAD DE INGENIERÍA INDUSTRIAL
CARRERA DE INGENIERÍA EN TELEINFORMÁTICA**

**TRABAJO DE TITULACIÓN
PREVIO A LA OBTENCIÓN DEL TÍTULO DE INGENIERA
EN TELEINFORMÁTICA**

**ÁREA
TECNOLOGÍA DE LOS ORDENADORES**

**TEMA
“SOPORTE AL MICROEMPENDIMIENTO POR MEDIO DE
PROTOTIPO DE DASHBOARD DE DATOS OBTENIDOS
MEDIANTE WEB SCRAPING EN MEDIOS DIGITALES.”**

**AUTORA
RUIZ RONQUILLO RUTH ROXANA**

**DIRECTOR DEL TRABAJO
ING. COMP. PLAZA VARGAS ÁNGEL MARCEL, MSC.**

**GUAYAQUIL, ABRIL 2021
ANEXO XI.- FICHA DE REGISTRO DE TRABAJO**



DE TITULACIÓN
FACULTAD DE INGENIERÍA INDUSTRIAL
CARRERA INGENIERÍA EN TELEINFORMÁTICA

REPOSITORIO NACIONAL EN CIENCIA Y TECNOLOGÍA			
FICHA DE REGISTRO DE TRABAJO DE TITULACIÓN			
TÍTULO Y SUBTÍTULO:	Soporte al microemprendimiento por medio de prototipo de dashboard de datos obtenidos mediante web scraping en medios digitales		
AUTOR(ES) (apellidos/nombres):	Ruiz Ronquillo Ruth Roxana		
REVISOR(ES)/TUTOR(ES) (apellidos/nombres):	Ing. Comp. Castillo León Rosa Elizabeth, Mg. / Ing.Comp. Plaza Vargas Angel Marcel, MSc.		
INSTITUCIÓN:	Universidad de Guayaquil		
UNIDAD/FACULTAD:	Facultad de Ingeniería Industrial		
MAESTRÍA/ESPECIALIDAD:			
GRADO OBTENIDO:	Ingeniera en Teleinformática		
FECHA DE PUBLICACIÓN:	28 de septiembre del 2021	No. DE PÁGINAS:	114
ÁREAS TEMÁTICAS:	Tecnología de los ordenadores		
PALABRAS CLAVES/KEYWORDS:	Raspado web, Tablero de mando, Python, Página web, Power Bi / Web scraping, Dashboard, Python, Web page, Power Bi.		
RESUMEN/ABSTRACT (100-150 palabras):			
<p>Resumen</p> <p>Actualmente, la importancia de la minería de datos se concentra en la generación de conocimiento útil desde cúmulos de datos, por ello, en este trabajo de grado se pone en práctica el uso de la técnica Web Scraping a fin de recopilar automáticamente grandes cantidades de datos de emprendimientos de Guayaquil presentados en un prototipo de dashboard mediante el desarrollo de una metodología de 4 fases: Selección de fuente de datos; Extracción y filtrado de datos; Almacenamiento de datos raspados y Visualización de datos. Como herramienta para el desarrollo del raspador web se usó Python y Power Bi para visualizar datos. Como resultado, los objetos visuales del tablero de mando permiten a cualquier persona que desee desarrollar actividades de microemprendimiento, analizar el comportamiento de las actividades económicas de las industrias activas en la ciudad y así, generar conocimiento y proporcionar un instrumento de discernimiento para la toma de decisiones.</p> <p>Abstract</p> <p>Currently, the importance of data mining is focused on the generation of useful insights from data clusters, therefore, this degree work implements the use of Web Scraping techniques to automatically collect large amounts of data from entrepreneurship in Guayaquil presented</p>			

in a prototype dashboard through the development of a methodology of 4 phases: Selection of data source; Data extraction and filtering; Storage of scraped data and data visualization. Python was used as a tool for the development of the web scraper, and Power Bi to visualize data. As a result, the elements in the dashboard allow anyone willing to develop a micro-entrepreneurship activity to analyze the behavior of the economic activities of the active industries in the city and thus, generate insights and provide a discerning tool for decision making.

ADJUNTO PDF:	SI X	NO
CONTACTO CON AUTOR/ES:	Teléfono: 0991505660	E-mail: ruth.ruizron@ug.edu.ec
CONTACTO CON LA INSTITUCIÓN:	Nombre: Ing. Ramón Maquilón Nicola	
	Teléfono: 593-2658128	
	E-mail: direccionTi@ug.edu.ec	



**ANEXO XII.- DECLARACIÓN DE AUTORÍA Y DE
AUTORIZACIÓN DE LICENCIA GRATUITA INTRANSFERIBLE Y
NO EXCLUSIVA PARA EL USO NO COMERCIAL DE LA OBRA
CON FINES NO ACADÉMICOS**



**FACULTAD DE INGENIERÍA INDUSTRIAL
CARRERA INGENIERÍA EN TELEINFORMÁTICA**

LICENCIA GRATUITA INTRANSFERIBLE Y NO COMERCIAL DE LA OBRA CON
FINES NO ACADÉMICOS

Yo, **RUIZ RONQUILLO RUTH ROXANA**, con C.C. No. **0959126806**, certifico que los contenidos desarrollados en este trabajo de titulación, cuyo título es “**SOPORTE AL MICROEMPENDIMIENTO POR MEDIO DE PROTOTIPO DE DASHBOARD DE DATOS OBTENIDOS MEDIANTE WEB SCRAPING EN MEDIOS DIGITALES**” son de mi absoluta propiedad y responsabilidad, en conformidad al Artículo 114 del CÓDIGO ORGÁNICO DE LA ECONOMÍA SOCIAL DE LOS CONOCIMIENTOS, CREATIVIDAD E INNOVACIÓN*, autorizo la utilización de una licencia gratuita intransferible, para el uso no comercial de la presente obra a favor de la Universidad de Guayaquil.

A handwritten signature in blue ink, appearing to read "Ruth Ruiz R.", written over a horizontal line.

RUIZ RONQUILLO RUTH ROXANA
C.C. No. 0959126806



**ANEXO VII.- CERTIFICADO PORCENTAJE DE
SIMILITUD
FACULTAD DE INGENIERÍA INDUSTRIAL
CARRERA INGENIERÍA EN TELEINFORMÁTICA**



Habiendo sido nombrado ING. ANGEL MARCEL PLAZA VARGAS, tutor del trabajo de titulación certifico que el presente trabajo de titulación ha sido elaborado por RUIZ RONQUILLO RUTH ROXANA, con mi respectiva supervisión como requerimiento parcial para la obtención del título de INGENIERO EN TELEINFORMÁTICA.

Se informa que el trabajo de titulación: SOPORTE AL MICROEMPREDIMIENTO POR MEDIO DE PROTOTIPO DE DASHBOARD DE DATOS OBTENIDOS MEDIANTE WEB SCRAPING EN MEDIOS DIGITALES, ha sido orientado durante todo el periodo de ejecución en el programa Antiplagio URKUND quedando el 1 % de coincidencia.

Documento: RUIZ RONQUILLO RUTH ROXANA.docx (D112376358)

Presentado: 2021-09-10 11:30 (-05:00)

Presentado por: ruth.rulzron@ug.edu.ec

Recibido: angel.plazav.ug@analysis.orkund.com

1% de estas 37 páginas, se componen de texto presente en 3 fuentes.

predicivos futuros. Se identifica 3 pasos generales: Procesamiento y generación de un conjunto de datos en base a noticias criminales, Implementación y validación de algoritmos de extracción e información, Elaboración de una interfaz de programación de aplicaciones para el consumo del modelo desarrollado CITATION Bus19 \j 21514 (Bustamente, 2019). En síntesis, este proyecto es un tanto más elaborado que los mencionados previamente, utiliza diversas herramientas para cumplir con los objetivos propuestos incluyendo dentro de las técnicas principales la minería de contenido web y la inteligencia artificial. Se encontró un trabajo de pregrado de la Universidad de Lima con el tema de "Plataforma de recomendación de habilidades tecnológicas según puesto de trabajo para profesionales de TI, en función de la demanda en las bolsas de trabajo digitales". El proyecto propone una plataforma que contribuya a las personas a visualizar cuáles son las

<https://secure.orkund.com/view/107072786-775624-514725>



Firmado electrónicamente por:
**ANGEL MARCEL PLAZA
VARGAS**

ING. ANGEL MARCEL PLAZA VARGAS
DOCENTE TUTOR
C.C. 0915953665
FECHA: 10/9/2021



ANEXO VI. - CERTIFICADO DEL DOCENTE-TUTOR DEL TRABAJO DE TITULACIÓN

FACULTAD DE INGENIERÍA INDUSTRIAL CARRERA INGENIERÍA EN TELEINFORMÁTICA



Guayaquil, 13 de septiembre del 2021.

Sr (a).

Ing. Annabelle Lizarzaburu Mora, MG.

Director (a) de Carrera Ingeniería en Teleinformática / Telemática

**FACULTAD DE INGENIERÍA INDUSTRIAL DE LA UNIVERSIDAD DE
GUAYAQUIL**

Ciudad. -

De mis consideraciones:

Envío a Ud. el Informe correspondiente a la tutoría realizada al Trabajo de Titulación **“SOPORTE AL MICROEMPENDIMIENTO POR MEDIO DE PROTOTIPO DE DASHBOARD DE DATOS OBTENIDOS MEDIANTE WEB SCRAPING EN MEDIOS DIGITALES”** del estudiante **RUIZ RONQUILLO RUTH ROXANA**, indicando que ha (cumplido con todos los parámetros establecidos en la normativa vigente:

- El trabajo es el resultado de una investigación.
- El estudiante demuestra conocimiento profesional integral.
- El trabajo presenta una propuesta en el área de conocimiento.
- El nivel de argumentación es coherente con el campo de conocimiento.

Adicionalmente, se adjunta el certificado de porcentaje de similitud y la valoración del trabajo de titulación con la respectiva calificación.

Dando por concluida esta tutoría de trabajo de titulación, **CERTIFICO**, para los fines pertinentes, que la estudiante está apta para continuar con el proceso de revisión final.

Atentamente,



Firmado electrónicamente por:
**ANGEL MARCEL
PLAZA VARGAS**

ING. ANGEL MARCEL PLAZA VARGAS, MSC
CC: 0915953665

13 de septiembre del 2021



ANEXO VIII.- INFORME DEL DOCENTE REVISOR

FACULTAD DE INGENIERÍA INDUSTRIAL CARRERA INGENIERÍA EN TELEINFORMÁTICA



Guayaquil, 22 de septiembre de 2021

Sr (a).

Ing. Annabelle Lizarzaburu Mora, MG.

Director (a) de Carrera Ingeniería en Telemática / Telemática

FACULTAD DE INGENIERÍA INDUSTRIAL DE LA UNIVERSIDAD DE GUAYAQUIL

Ciudad. -

De mis consideraciones:

Envío a Ud. el informe correspondiente a la REVISIÓN FINAL del Trabajo de Titulación **“SOPORTE AL MICROEMPREDIMIENTO POR MEDIO DE PROTOTIPO DE DASHBOARD DE DATOS OBTENIDOS MEDIANTE WEB SCRAPING EN MEDIOS DIGITALES”** del estudiante **RUIZ RONQUILLO RUTH ROXANA**. Las gestiones realizadas me permiten indicar que el trabajo fue revisado considerando todos los parámetros establecidos en las normativas vigentes, en el cumplimiento de los siguientes aspectos:

Cumplimiento de requisitos de forma:

El título tiene un máximo de 16 palabras.

La memoria escrita se ajusta a la estructura establecida.

El documento se ajusta a las normas de escritura científica seleccionadas por la Facultad.

La investigación es pertinente con la línea y sublíneas de investigación de la carrera.

Los soportes teóricos son de máximo 5 años.

La propuesta presentada es pertinente.

Cumplimiento con el Reglamento de Régimen Académico:

El trabajo es el resultado de una investigación.

El estudiante demuestra conocimiento profesional integral.

El trabajo presenta una propuesta en el área de conocimiento.

El nivel de argumentación es coherente con el campo de conocimiento.

Adicionalmente, se indica que fue revisado, el certificado de porcentaje de similitud, la valoración del tutor, así como de las páginas preliminares solicitadas, lo cual indica el que el trabajo de investigación cumple con los requisitos exigidos.

Una vez concluida esta revisión, considero que el estudiante está apto para continuar el proceso de titulación. Particular que comunicamos a usted para los fines pertinentes.

Atentamente,



Firmado electrónicamente por:

ROSA

ELIZABETH

CASTILLO LEON

ING. ROSA ELIZABETH CASTILLO LEÓN, MG.

C.C: 0922372610

FECHA: 22 DE SEPTIEMBRE DE 2021

Dedicatoria

A mis padres, hermanas y abuelita con inmesurable amor por ser el estímulo de todos mis logros.

Agradecimiento

Mi agradecimiento personal se dirige a mi familia, por su amor y apoyo incondicional, a mis amigos, por la motivación y fortaleza, y a mis docentes, por forjar los cimientos de mi desarrollo profesional, estoy hecha de pedacitos de ustedes. Gracias infinitas.

Índice General

N°	Descripción	Pág.
	Introducción	1

Capítulo I El Problema

N°	Descripción	Pág.
1.1.	Planteamiento del problema	2
1.2.	Formulación del problema	3
1.3.	Justificación	3
1.4.	Objetivos	4
1.4.1.	Objetivo general.	4
1.4.2.	Objetivos específicos.	4
1.5.	Delimitación del problema	4

Capítulo II Marco Teórico

N°	Descripción	Pág.
2.1.	Antecedentes del estudio	6
2.1.1.	Antecedentes de Web scraping	6
2.1.2.	Antecedentes de Dashboard	10
2.2.	Fundamentación teórica	13
2.2.1.	Minería de datos	13
2.2.1.1.	La relación de la minería de datos con otras disciplinas	14
2.2.2.	Metodologías de minería de datos	15
2.2.2.1.	Metodología KDD	15
2.2.2.2.	Metodología CRISP-DM	17
2.2.2.3	Metodología SEMMA	18
2.2.3.	Comparación de los procesos de las metodologías de minería de datos.	19
2.2.4.	Medios digitales	20

N°	Descripción	Pág.
2.2.5.	Sitios Web	20
2.2.5.1.	Tipos de sitios web	21
2.2.6.	Códigos de estado de respuesta HTTP	22
2.2.7.	Certificado digital	23
2.2.8.	Data Scraping	23
2.2.9.	Web scraping	24
2.2.9.1.	Web Crawler	25
2.2.9.2.	Extractor de datos	26
2.2.10.	Usos de web scraping	26
2.2.11.	Impedimentos para realizar web scraping	27
2.2.12.	Precauciones antes de hacer web scraping	27
2.2.13.	Técnicas de web scraping	28
2.2.14.	Herramientas de web scraping	29
2.2.14.1.	Herramientas de programación	29
2.2.14.2.	Aplicaciones de escritorio	29
2.2.14.3.	Aplicaciones web	30
2.2.14.4.	Extensiones para navegador	30
2.2.15.	Lenguajes empleados en Web scraping	31
2.2.15.1.	Python	33
2.2.15.2.	Node.js	33
2.2.15.3.	Ruby	33
2.2.15.4.	PHP	34
2.2.15.5.	C&C++	34
2.2.16.	Dashboard	34
2.2.17.	Capas de un dashboard	34
2.2.18.	Errores frecuentes de diseño	35
2.2.19.	Beneficios del desarrollo de un dashboard	35
2.2.20.	Herramientas para construir un tablero de mando	36
2.2.20.1.	Tableau	36
2.2.20.2.	Power BI	36

N°	Descripción	Pág.
2.3.	Fundamentación legal	37

Capítulo III

Propuesta

N°	Descripción	Pág.
3.1.	Métodos de investigación	40
3.1.1.	Diseño de investigación	40
3.1.1.1.	Enfoque Mixto	40
3.1.2.	Metodología de investigación	41
3.1.2.1.	Documental	41
3.1.2.2.	Cuasi experimental	42
3.1.2.3.	Descriptiva	42
3.1.3.	Técnica	42
3.1.3.1.	Análisis de la entrevista	42
3.2.	Metodología a desarrollar	50
3.2.1.	Factibilidad técnica	51
3.2.2.	Selección de la fuente de datos	52
3.2.3.	Extracción y filtrado de datos	55
3.2.3.1.	Inspeccionar los elementos del lenguaje de marcado	55
3.2.3.2.	Extracción de los datos	56
3.2.3.3.	Código	57
3.2.4.	Almacenamiento de los datos raspados	63
3.2.5.	Visualización de los datos	66
3.2.5.1.	Conexión con origen de datos	67
3.2.5.2.	Diseño del Dashboard	68
3.3.	Conclusiones y recomendaciones	72
3.3.1.	Conclusiones	72
3.3.2.	Recomendaciones	73
	ANEXOS	74

N°	Descripción	Pág.
	Bibliografía	89

Índice de Tablas

N°	Descripción	Pág.
1.	Resumen de antecedentes de Web Scraping	8
2.	Resumen de antecedentes de Dashboard	12
3.	Códigos HTTP	22
4.	Procedimiento para hacer web scraping	26
5.	Tabla comparativa de herramienta de web scraping.	31
6.	Tipos de enfoques mixtos	40
7.	Perfil académico del profesional consultado.	43
8.	Nivel de conocimiento de los entrevistados en Web Scraping.	44
9.	Lenguaje de programación.	44
10.	La ética del web scraping en páginas con medidas de seguridad antibot.	45
11.	Argumentación de la ética en el web scraping	46
12.	Nivel de conocimiento de los entrevistados en Dashboard.	47
13.	Opinión de expertos sobre el aporte del dashboard para los microemprendedores.	48
14.	Herramienta gratuita para Dashboard.	48
15.	Tipos de gráficos estadísticos para dashboard.	49

Índice de Figuras

Nº	Descripción	Pág.
1.	Disciplinas que contribuyen a la minería de datos.	14
2.	Desglose jerárquico de CRISP-DM.	17
3.	Fases de la metodología SEMMA.	18
4.	Esquema básico del servicio web.	21
5.	Marco general del Web Scraping.	25
6.	Proceso de Web Crawler.	25
7.	Parámetros para escoger un lenguaje de programación de web scraping.	32
8.	Diseño de triangulación concurrente.	41
9.	Nivel de conocimiento de los entrevistados en Web Scraping.	44
10.	Lenguaje de programación.	45
11.	La ética del web scraping en páginas con medidas de seguridad antibot.	46
12.	Nivel de conocimiento de los entrevistados en Dashboard.	47
13.	Opinión de expertos sobre el aporte del dashboard para los microemprendedores	48
14.	Herramienta gratuita para dashboard.	49
15.	Tipos de gráficos estadísticos para Dashboard.	50
16.	Esquema general de la metodología.	51
17.	Medidas de seguridad en SRI para consultar de RUC.	53
18.	RUC consultado no encontrado.	53
19.	Página de emprendimientos.	54
20.	Parte del HTML de la página de la sección de Emprendimientos.	55
21.	Parte del HTML de la página de un emprendimiento.	56
22.	Librerías importadas para extracción de url.	58
23.	Función para extraer url's de varias páginas.	58
24.	Presentación de URL filtradas.	59
25.	Imprimir la lista enlacesgye() para visualizar las url's.	60
26.	Librerías importadas para extraer los datos de emprendimientos.	60
27.	Extracción de información de emprendimientos de Guayaquil.	61
28.	Error de la clase dictReader.	62
29.	Imprimir datos de emprendimientos.	62

N°	Descripción	Pág.
30.	Código para almacenar las url's en .CSV.	63
31.	Código para almacenar los datos de cada emprendimiento.	63
32.	Carpeta PycharmPrjects con archivos CSV.	64
33.	Archivos CSV en Excel.	65
34.	Columnas agregadas con significado de código CIIU.	66
35.	Registro con cuenta institucional en Power Bi.	67
36.	Origen de datos.	67
37.	Selección de tabla dentro del archivo Directorio de emprendimientos.	68
38.	Ejemplo de objeto visual Segmentación de datos con varios campos.	69
39.	Dashboard final.	70
40.	Filtro 1 del dashboard.	71
41.	Filtro 2 del dashboard.	71
42.	Negar reciprocidad de filtrado entre objetos.	72

Índice de anexos

N°	Descripción	Pág.
1.	Decreto Ejecutivo 101	75
2.	Código Orgánico de la Economía Social	76
3.	Ley Orgánica de Transparencia y Acceso a la Información Pública	78
4.	La Entrevista	79
5.	Instalación de Python	81
6.	Instalación de Pycharm	83
7.	Crear un nuevo proyecto.	85
8.	Instalar librerías desde la terminal de Pycharm	86



**ANEXO XIII.- RESUMEN DEL TRABAJO DE
TITULACIÓN (ESPAÑOL)
FACULTAD DE INGENIERÍA INDUSTRIAL
CARRERA INGENIERÍA EN TELEINFORMÁTICA**



**“SOPORTE AL MICROEMPENDIMIENTO POR MEDIO DE PROTOTIPO DE
DASHBOARD DE DATOS OBTENIDOS MEDIANTE WEB SCRAPING EN MEDIOS
DIGITALES”**

Autor: Ruiz Ronquillo Ruth Roxana

Tutor: Ing. Comp. Plaza Vargas Ángel Marcel, MSc.

Resumen

Actualmente, la importancia de la minería de datos se concentra en la generación de conocimiento útil desde cúmulos de datos, por ello, en este trabajo de grado se pone en práctica el uso de la técnica Web Scraping a fin de recopilar automáticamente grandes cantidades de datos de emprendimientos de Guayaquil presentados en un prototipo de dashboard mediante el desarrollo de una metodología de 4 fases: Selección de fuente de datos; Extracción y filtrado de datos; Almacenamiento de datos raspados y Visualización de datos. Como herramienta para el desarrollo del raspador web se usó Python y Power Bi para visualizar datos. Como resultado, los objetos visuales del tablero de mando permiten a cualquier persona que desee desarrollar actividades de microemprendimiento analizar el comportamiento de las actividades económicas de las industrias activas en la ciudad y así, generar conocimiento y proporcionar un instrumento de discernimiento para la toma de decisiones.

Palabras Claves: Raspado web, Tablero de mando, Python, Página web, Power Bi.



**ANEXO XIV.- RESUMEN DEL TRABAJO DE
TITULACIÓN (INGLÉS)
FACULTAD DE INGENIERÍA INDUSTRIAL
CARRERA INGENIERÍA EN TELEINFORMÁTICA**



**“MICRO-ENTREPRENEURSHIP SUPPORT THROUGH A PROTOTYPE
DASHBOARD OF DATA OBTAINED THROUGH WEB SCRAPING IN DIGITAL
MEDIA”**

Author: Ruiz Ronquillo Ruth Roxana

Advisor: Ing. Comp. Plaza Vargas Ángel Marcel, MSc.

Abstract

Currently, the importance of data mining is focused on the generation of useful insights from data clusters, therefore, this degree work implements the use of Web Scraping techniques to automatically collect large amounts of data from entrepreneurship in Guayaquil presented in a prototype dashboard through the development of a methodology of 4 phases: Selection of data source; Data extraction and filtering; Storage of scraped data and data visualization. Python was used as a tool for the development of the web scraper, and Power Bi to visualize data. As a result, the elements in the dashboard allow anyone willing to develop a micro-entrepreneurship activity to analyze the behavior of the economic activities of the active industries in the city and thus, generate insights and provide a discerning tool for decision making.

Keywords: Web scraping, Dashboard, Python, Web page, Power Bi.

Introducción

Hace un par de décadas, el desarrollo tecnológico experimentó el cambio más drástico en el mundo de la informática y la información, el deseo de adquirir conocimiento sin límite dio lugar a nuevos hallazgos y artificios que hicieron posible que miles de millones de datos estén disponibles a través de una misma pantalla con tan solo dar un clic, así nació la web, el sistema de gestión de información con mayor difusión en la historia humana.

Esta avalancha de datos que crece constantemente se presenta en medios digitales como paginas o sitios web distribuidos en Internet de manera pública y accesible, generando un fuerte interés en el área del análisis de datos que en conjunto con técnicas de minería de datos construyen instrumentos de utilidad desde enormes cantidades de datos para encontrar minas de oro o conocimiento.

El web scraping es una técnica programable de data mining que permite extraer de manera estructurada los datos contenidos en una serie de páginas web construidas de forma similar. A través del análisis del lenguaje de contenido de la página web se buscan patrones claros que permitan automatizar la obtención de la información que se desea, darle una estructura y almacenarla.

Dentro del proceso de data mining los componentes visuales, también llamados Tableros de mando (Dashboard), son de gran ayuda para presentar grandes cantidades de datos recabados o extraídos de un sitio web que, por medio de gráficos aportan significativamente en la visualización de la información para el usuario final.

El documento estará organizado de la siguiente manera:

Capítulo I estará compuesto por: Planteamiento del problema, formulación de problema, justificación e importancia, objetivo general y específico y el alcance.

Capitulo II también conocido como marco teórico, se incorporará lo siguiente: Antecedentes de investigación, fundamentación teórica y fundamentación legal.

Capitulo III es el último capítulo, se centrará en el desarrollo de la propuesta de investigación, tipo de investigación, técnicas e instrumentos, metodología a emplearse y finaliza con las conclusiones y recomendaciones del proyecto.

Capítulo I

El Problema

1.1. Planteamiento del problema

Las mipymes son la espina dorsal de la economía mundial, figuran el 95% de las empresas y desencadenan el 60% del empleo a nivel mundial, además, para los países en vías de desarrollo las mipymes aportan aproximadamente el 35% del PIB (Organización Mundial del Comercio, 2016). Siendo fundamentales para salvaguardar la productividad y el empleo.

Actualmente, en el Ecuador, el microemprendimiento es uno de los principales motores de estímulo en la economía, por contribuir en el desarrollo social y crecimiento económico y, al mismo tiempo, ser una medida de auxilio ante el desempleo y la necesidad que se ha estado viviendo con la abrumadora realidad subsistente debido a la crisis sanitaria causada por la covid-19.

Dentro del informe del Global Entrepreneurship Monitor, se determinó que la tasa de Actividad Emprendedora Temprana (TEA) en el Ecuador es de 36,2%, considerándose como uno de los países con la tasa más alta en comparación con Guatemala, Brasil, Panamá y Paraguay dentro de América Latinoamérica y el Caribe (Lasio, Amaya, Zambrano, & Ordeñana, 2020). Pero también es uno de los países en donde más emprendimientos fracasan y las propuestas de nuevas ideas de negocio no maduran en empresas rentables y sostenibles.

Las principales razones del cierre de un emprendimiento se centran en problemas personales (32.7%) y falta de rentabilidad (25.2%). La falta de rentabilidad es una de las razones que se mantiene desde estudios realizados previamente (Lasio et al., 2020). Siendo la causa de este efecto, la omisión de todas aquellas investigaciones relacionadas con la generación de ingresos en un emprendimiento, la falta de conocimiento del entorno y la subestimación de la competencia del mercado (Korporate Technologies Group, 2017).

En todo caso, en el Ecuador, existen datos de acceso público sobre emprendimientos que pueden ser consultadas en páginas como el SRI, Superintendencia de compañías o Directorios web de emprendimientos, pero es escaso o casi nulo encontrar entornos que integren estos datos para generar conocimiento e información inteligible que soporte y fomente factores que promuevan al microemprendimiento.

1.2. Formulación del problema

Tomando en cuenta la problemática planteada surgen las siguientes preguntas:

- ¿Es posible efectuar web scraping en medios digitales ecuatorianos?
- ¿Es posible desarrollar un tablero de mando con datos de información pública obtenidos de medios digitales ecuatorianos, donde se visualicen variables que puedan servir de interés dentro de una posible propuesta de emprendimiento?

1.3. Justificación

Dentro del mundo digital los datos se crean constantemente, y este crecimiento vertiginoso que procede de distintas partes del mundo ha sido posible con la llegada de Internet, una red de redes de ordenadores interconectadas que cambió nuestra forma de comunicar e interactuar con otras personas a través de un canal de comunicación cuyas premisas se basan en la rapidez, simultaneidad y en grandes volúmenes de datos permitiendo con ella el surgimiento de uno de los sistemas de distribución y gestión de información más popular a nivel global conocido como la World Wide Web.

La web alberga una vasta cantidad de información, hoy en día estos grandes volúmenes de datos están marcando diferencia y siendo aprovechados por distintos entes para mejorar la toma decisiones, destacar en campañas de marketing, optimizar procesos de negocio, controlar el cumplimiento de leyes, aumentar la seguridad, avanzar en nuevas tecnologías e investigación, entre otros.

El Data Scraping es una de las técnicas de recolección de datos más distintivas entre los actuales métodos digitales dado que hace posible la investigación fundamentada en datos obtenidos de medios digitales (Marres & Weltevrede, 2013). Y es precisamente la herramienta tecnológica que contextualiza este trabajo de tesis, más específicamente el Web Scraping, que consiste en la extracción de información en medios digitales como sitios y páginas web con el uso de técnicas y herramientas, una vez obtenido y procesado el código HTML de una página o páginas web extraerá y recolectará datos de forma automática que podrán ser reutilizados para fines o necesidades como: Comparación de precios; Investigación; Monitoreo de datos cambiantes; Recolección de contactos; Periodismo de datos.

La visualización de los datos también juega un papel importante dentro de cualquier proceso de análisis de datos, ya que permite generar conocimiento y mostrar información de manera

concreta, oportuna y organizada. Un tablero de mando (Dashboard) representa un resumen y una visión general de la convergencia y unificación de un recurso o varios recursos contenedores de datos en una sola vista que ayuda a comprender la complejidad que existe detrás de la relación de dichos datos, siendo por ejemplo utilizados para reportar datos, monitorear sistemas o hacer seguimiento de métricas importantes de negocios.

Con relación a la problemática expuesta, el fracaso o el éxito de un emprendimiento depende de varios factores, en el informe GEM, que aporta con una radiografía del emprendimiento cada año, se determinó, gracias a la contribución de la apreciación de expertos consultados que factores como Educación y Capacitación, Transferencia I+D e Información promueven y fomentan el emprendimiento (Lasio, Ordeñana, Caicedo, Samaniego, & Izquierdo, 2018). Permitiendo sobre todo esta última identificar oportunidades de mercado, reconocer sectores exitosos e integrar TICs como ventaja de discernimiento.

Dentro de este orden ideas, se pretende hacer uso de la técnica Web Scraping utilizando una herramienta de desarrollo que permita construir un bot de scraping para automatizar la búsqueda y extraer información de páginas web sobre emprendimientos que será estructurada, almacenada y finalmente presentada mediante un dashboard que representará un medio visual para el soporte al microemprendimiento.

1.4. Objetivos

1.4.1. Objetivo general.

Desarrollar un prototipo de dashboard para la presentación de datos obtenidos de medios digitales mediante Web Scraping.

1.4.2. Objetivos específicos.

- Determinar el medio digital del cual se extraerán los datos.
- Establecer la técnica y herramienta para la aplicación del raspado web en medios digitales.
- Construir un raspador web para la obtención de los datos.
- Desarrollar un dashboard para presentar los datos extraídos.

1.5. Delimitación del problema

En este proyecto se contempla trabajar sobre un medio digital nacional que contenga información relacionada al emprendimiento, este sitio web es un directorio de emprendimientos. La extracción de los datos en la web solo se ejecutará una única vez en el medio mencionado

para alimentar un tablero de mando que estará almacenado localmente y presentará de manera ordenada y organizada los datos obtenidos con el raspado web.

El proyecto se encuentra delimitado geográficamente en la ciudad de Guayaquil, por lo tanto, del directorio de emprendimiento que se escoja solo se extraerá información de emprendimientos de la ciudad escogida y serán recolectados a partir del año 2000. Según (Tas & Togay, 2014) la dolarización ejerció un efecto positivo y un cambio rotundo en la economía y (Burgos & Villar, 2016) señala que la revolución tecnológica a partir del siglo XX cambió a nivel mundial la forma de emprender, transformando la sociedad, la economía y la cultura.

Capítulo II

Marco Teórico

2.1. Antecedentes del estudio

La construcción de esta investigación dio lugar a consultar diferentes fuentes de información bibliográfica obteniendo como fruto de la búsqueda artículos y trabajos de grado y máster similares, pero aplicados en diversos entornos. En el apartado de Web Scraping la convergencia de estos trabajos reside en la aplicación de técnicas para extraer datos desde la web, y en el apartado de Dashboard, en el desarrollo de visualizadores de datos para generar conocimiento y contribuir en la toma decisiones.

A continuación, se expondrá los temas encontrados que aportan conocimiento al presente proyecto con la finalidad de conocer herramientas oportunas para el desarrollo del tema propuesto.

2.1.1. Antecedentes de Web scraping

Los autores Sanchez, Durán, Ballesteros-Ricaurte y Gonzáles-Amarillo, hacen público en el 2020 con la ayuda de la revista científica internacional de investigaciones en el campo de ingeniería INGE CUC, el artículo “ScraCOVID-19: Plataforma informática de contenido digital mediante Scraping y almacenamiento NoSQL” publicado en Colombia, Barranquilla, en el cual aplican la técnica de Web Scraping en cuatro medios digitales de medios de comunicación verídicos y confiables para extraer las noticias asociadas al tema de la pandemia causada por COVID-19 e integrar dichos datos en una única plataforma informativa con información que puede ser filtrada en función del interés del usuario como: desempleo, educación, maltrato, vacunas, pobreza, salud mental, evadiendo las noticias falsas y manteniendo informada a toda una comunidad. El enfoque de esta investigación hace referencia al marco de trabajo para la extracción de información, se resume en cuatro etapas: Selección de fuentes de datos: Medios digitales de Colombia de la categoría de radio, revista y periódico; Aplicación de la técnica de scraping: Python con la librería de BeautifulSoup y Request entre otras; Almacenamiento orientado a documentos: MongoDB y Diseño de la plataforma: Infraestructura Express del ambiente de trabajo Nodejs (Sanchez, Durán, Ballesteros-Ricaurte, & Gonzáles-Amarillo, 2020).

Este artículo se apoya del uso de la minería de contenido web para conseguir información desde diferentes fuentes e integrarlas en una sola interfaz de usuario que permitió crear un entorno con información sucinta y esencial sobre el tema de interés.

Daniel Espinoza de la Universidad de Chile en el año 2020 publica la tesis de pregrado “Diseño y Ejecución de Arquitectura de descarga, modelamiento y análisis de datos para ampliar servicios en una empresa de tecnología”, en donde construye un sistema de análisis de datos que se alimenta de la ejecución del raspado automático en la información tributaria de los clientes de una empresa dedicada al servicio de boletas electrónicas, esta herramienta se la destina para cada uno de los clientes de dicha empresa como mejora de la propuesta de valor del servicio y al mismo tiempo, proveer a dicha empresa de una base de datos actualizada periódicamente para idear nuevos productos. El diseño consta de 5 capas: Fuente de datos: Servicio de Impuesto Interno (SII); Extracción, transformación y carga de datos: Python con Selenium y PostgreSQL entre otras; Modelado de datos: Tipo estrella; Análisis de datos mediante dashboard: Power BI; Propuesta de servicio: Decisiones del cliente (Espinoza, 2020).

Con el uso de la técnica de minería de datos, web scraping, logró automatizar el proceso de extracción de información tributaria de los clientes de la empresa para añadir dentro del servicio que brinda esta empresa un tablero de mando que muestre indicadores de gestión para la toma de decisiones y así mejorar la promesa de valor.

Bustamante Gina, resalta en el tema de tesis “Extracción de información para la generación de reportes estructurados a partir de noticias peruanas relacionadas a crímenes” que la recolección de grandes volúmenes de datos con herramientas de la ciencia computacional del nuevo siglo ha permitido dentro del análisis criminal abarcar desde la predicción del lugar de un delito con datos numéricos hasta conocer nombres y agencias con descripciones textuales en el medio digital. Bajo este contexto, el autor utiliza técnicas de web scraping y procesamiento de lenguaje natural para extraer información de noticias vinculadas al dominio criminal de Perú, con la intención de identificar automáticamente culpables, víctimas y lugares que abordan un acontecimiento delictivo, presentándola dentro de un reporte estructurado para el análisis del público en general y para el desarrollo de modelos predictivos futuros. Se identifica 3 pasos generales: Procesamiento y generación de un conjunto de datos en base a noticias criminales; Implementación y validación de algoritmos de extracción e información; Elaboración de una

interfaz de programación de aplicaciones para el consumo del modelo desarrollado (Bustamente, 2019).

En síntesis, este proyecto es un tanto más elaborado que los mencionados previamente, utiliza diversas herramientas para cumplir con los objetivos propuestos incluyendo dentro de las técnicas principales la minera de contenido web y la inteligencia artificial.

Se encontró un trabajo de pregrado de la Universidad de Lima con el tema de “Plataforma de recomendación de habilidades tecnológicas según puesto de trabajo para profesionales de TI, en función de la demanda en las bolsas de trabajo digitales”. El proyecto propone una plataforma que contribuya a las personas a visualizar cuales son las habilidades técnicas en el área de TI que requiere el mercado con mayor demanda para un puesto de trabajo y al mismo tiempo obtener la sugerencia de instituciones para estudiar dichas técnicas. La base de esta plataforma es el web scraping y el procesamiento de lenguaje natural que permitió extraer y analizar los anuncios de trabajo relacionados a la tecnología desde la red social orientada al uso empresarial LinkedIn, con la ayuda de una lista de roles de trabajo definida (Sanchez D. , 2020).

Se puede concluir que el uso de técnicas de minería de contenido web, es eficaz para la extracción de datos.

Tabla 1.

Resumen de antecedentes de Web Scraping

Tema	Año	Propósito	Población objetivo	Fuente de datos	Herramientas
ScraCOVID-19: Plataforma informática de contenido digital mediante Scraping y almacenamiento NoSQL	2020	Mantener informada a la comunidad sobre temas relacionados a la pandemia causada por el covid-19, a través de una plataforma web dedicada a acceder a noticias actualizadas y de	Público en general	Diarios de noticias de Colombia	Python: Beautiful Soup, Requests, pymongo, JSON; MongoDB; Node.js.

		manera rápida desde 4 medios digitales del país.		
“Diseño y ejecución de arquitectura de descarga, modelamiento y análisis de datos para ampliar servicios en una empresa de tecnología”	2020	Construcción de un tablero de mando para el análisis de datos extraídos de facturas tributarias en la web para mostrar indicadores de gestión.	Cientes de una empresa de servicio de boletas electrónicas y para la misma empresa.	Servicio de fuentes internos (SII) Python: Selenium, Requests, Xmltodict, Boto3, Psycopg2; Aurora PostgreSQL; PowerBI
“Extracción de información para la generación de reportes estructurados a partir de noticias peruanas relacionadas a crímenes”	2019	Bajo el contexto de análisis criminal, se propone la construcción de una herramienta de extracción de datos de diarios digitales con la finalidad de identificar culpables, víctimas y lugares de un hecho delictivo.	Público en general y para el desarrollo de modelos predictivos futuros	Diarios peruanos: El Comercio, La República, Radio programas de Perú. Jupyter; Python: BeautifulSoup, Flask, Django; MongoDB; Stanford CoreNLP; Protégé; Hermit; Wordnet
“Plataforma de recomendación de habilidades tecnológicas”	2020	Proveer a los profesionales de TI una plataforma web con	Profesionales TI entre 22 y 30 años	LinkedIn Python: BeautifulSoup, NLTK, jaro-winkler;

según puesto de trabajo para profesionales de TI en función de la demanda en las bolsas de trabajo digitales”	información centralizada sobre las habilidades más demandadas por la mayoría de las empresas y así mismo, recomendar cursos en donde pueden adquirir estas habilidades.	MongoDB; Servicio de Google db-n1-standard-1; Marvel
--	---	--

Información adaptada de los proyectos y artículos investigados. Elaborado por Ruth Ruiz

2.1.2. Antecedentes de Dashboard

Se halló un trabajo de maestría de la Universidad Politécnica de Valencia, con el tema “Diseño e implementación de un dashboard de soporte académico basado en datos de entornos virtuales de aprendizaje”, en el que proponen el diseño e implementación de un prototipo de dashboard bajo el marco de Learning Analytics, el tablero de mando permitirá analizar los datos que generan los estudiantes en las plataformas de aprendizaje y otros medios de la Universidad Politécnica de Valencia, con la finalidad de proporcionar a los docentes indicadores para mejorar la metodología de enseñanza, identificar patrones de abandono de carrera o registrar los recursos digitales de aprendizaje más utilizados y, a los estudiantes, indicadores de patrones de aprendizaje que ayuden a crear conciencia social y habilidades de autorregulación para mejorar las estrategias de aprendizaje (Haro, 2018)

El desarrollo de un tablero de mando permitió brindar soporte académico, tanto a docentes como estudiantes con la visualización de la integración de los datos desde diversas fuentes en una sola pantalla mediante la combinación de diferentes tipos de gráficos.

En el año 2017, Sim Livvi; Ban, Kenneth y Tan Tin de la National University of Singapore y Sethi Sunil; Loh Tze de la National University Hospital presentaron un estudio piloto en el artículo “Development of a clinical decision support system for diabetes care: A pilot study” donde desarrollan un tablero de mando como sistema de apoyo a las decisiones clínicas específicas de la diabetes, mostrando resultados del estado glucémico, lípido y de la función

renal resumida a través de gráficos y de un sistema codificado de colores que permitió un reconocimiento rápido del control metabólico de los pacientes. El tablero de mando cuenta con 4 módulos; módulo de información del paciente, módulo de codificación de colores que sirve para indicar diferentes niveles de control en base a los marcadores de laboratorio de un examen, si el nivel es óptimo el color estaría en verde, caso contrario, rojo; módulo de alerta que sirve para indicar que el intervalo de prueba recomendado para realizar algún examen en particular para el monitoreo de un paciente ha sido superado y gráficos interactivos contruidos de los resultados históricos de los exámenes (Sim, Ban, Tan, Sethi, & Loh, 2017).

En resumen, el sistema tradicional de los registros electrónicos de salud tiende a mostrar información segregada y fragmentada, tornando desafiante interpretar los resultados relacionados con el cuidado de la diabetes de un paciente. Según los expertos consultados en este estudio piloto, el desarrollo de un sistema de apoyo en comparación con el sistema tradicional mejoró significativamente la interpretación de los resultados de los pacientes permitiendo identificar resultados anormales de laboratorio, alertar el plazo máximo para realizar un siguiente examen de control y obtener una comparativa en base a los resultados históricos de pruebas o exámenes.

Se encontró una tesis de grado en el repositorio de la Universidad Católica de Colombia con el tema “Ciudades inteligentes y datos abiertos: Un dashboard basado en minería de datos”. El proyecto de grado, mediante el uso de datos abiertos suministrados por el SECOP (Sistema electrónico de contratación pública) en el cual se encuentran diversos conjuntos de datos invaluables que incluye información de contratación pública, realizó un desarrollo progresivo que culmina en el desarrollo de un dashboard para publicar los datos analizados que contribuyan en la búsqueda de patrones que ayuden a determinar comportamientos fuera de lo común en el sector público y a evidenciar los problemas de transparencia de los procesos de licitación en el país colombiano y así, gobernantes, pensadores públicos o miembros de movimientos ciudadanos puedan influir, informar o tomar decisiones basándose en datos (Estevez, 2017).

El análisis de datos abiertos aporta en la tendencia de toma de decisiones con el fin de ayudar a resolver problemas.

En el proyecto de grado “Implementación de una solución business intelligence para apoyar en la toma de decisiones en la empresa Agro Micro Biotech Sac” se propone la implementación de un Tablero de mando bajo el concepto de Business Intelligence para el área comercial de la

empresa Agro Micro Biotech Sac que integre islas de datos que están almacenadas en distintas fuentes de la empresa para brindar información útil (que en este caso es: solicitud del usuario comercial y propuestas económicas y cotizaciones) y apoyar a la toma de decisiones como alternativa a un sistema manual de consulta de información comercial de alta dependencia en otras áreas, cuellos de botellas e información errónea (Lozano, 2020).

Con esta solución se consiguió información actualizada, comprensible, confiable y de fácil acceso para el usuario final.

Tabla 2.

Resumen de antecedentes de Dashboard

Tema	Año	Propósito	Población objetivo	Fuentes de información	Herramientas
“Diseño e implementación de un dashboard de soporte académico basado en datos de entornos virtuales de aprendizaje”	2018	Promover una cultura de autorregulación facilitando a los usuarios una herramienta para dar soporte en la toma de decisiones basada en datos.	Profesores y alumnos de la Universidad Politécnica de Valencia	Poliformat: Estadísticas; Ficheros de calificaciones, asistencias y eventos.	Pentaho Data Integration; Tableau y Qlik-sense; Amazon Cloud Watch; Java y Apache TomCat7; Plantilla de GitHub.
“Development of a clinical decision support system for diabetes care: A pilot study”	2017	Desarrollar un Dashboard que incorpore varias funciones de apoyo a la toma de decisiones para mejorar el manejo de la diabetes.	Médicos	Sistema de registros de salud electrónico tradicionales	Python, Bokeh; MySQL, HTML; CSS; Bootstrap.

“Ciudades inteligentes y datos abiertos: Un dashboard basado en minería de datos”	2017	Predecir el comportamiento de la contratación pública y contribuir en la transparencia de los procesos.	Gobernantes, pensadores públicos o miembros de movimiento s ciudadanos.	Java EE; HTML; JavaScript; MongoDB; RapidMiner y Tableau
“Implementación de una solución Business Intelligence para apoyar en la toma de decisiones en la empresa Agro Micro Biotech Sac”	2020	Proporcionar al usuario del área comercial información de venta y cotizaciones para que generen reportes y análisis de las estadísticas de ventas del mes	Área comercial	Starsoft y Humasoft
				SQL Server 2017 SE; SQL Server Integration Services y Analysis Services; Visual Studio; Power BI

Información adaptada de los proyectos y artículos investigados. Elaborado por Ruth Ruiz

2.2. Fundamentación teórica

2.2.1. Minería de datos

Escuchar el termino “minería de datos” podría ser considerado para muchos como “nada nuevo bajo el sol”. Efectivamente, la definición de minería de datos no surgió por el desarrollo de nuevas tecnologías, sino por la aparición de nuevas necesidades y por el reconocimiento del valor y del gran potencial que posee la ingente cantidad de datos informáticos que se encuentran difundidos en distintos formatos y que provienen de sistemas de información de gobiernos, empresas, instituciones y otras fuentes particulares.

“Los datos pasan de ser un “producto” a ser una “materia prima” que hay que explotar para obtener el verdadero “producto elaborado”, el conocimiento; un conocimiento que ha de ser

especialmente valioso para la ayuda en la toma de decisiones sobre el ámbito en el que se han recopilado o extraído los datos”. (Ferri & Ramírez, 2004)

La revolución digital y la aceptación de esta por parte de los usuarios, facilita la generación, recuperación, procesamiento, almacenamiento, distribución y transmisión de información digitalizada. De este modo, y gracias al avance tecnológico en la administración de bases de datos (que integra datos como imágenes, videos, textos en una sola base) se puede definir el Data Mining como el proceso de descubrir conocimiento entendible y útil desde grandes volúmenes de datos de diversos formatos almacenados en repositorios; con la finalidad de representar un recurso valioso de apoyo para la toma de decisiones.

En la minería de datos hay dos actividades principales: descripción y predicción. Las actividades descriptivas se enfocan en descubrir patrones o relaciones interesantes que describen los datos y las actividades predictivas infieren posibles escenarios futuros basados en datos históricos (Riquelme, Ruiz, & Gilbert, 2006).

2.2.1.1. La relación de la minería de datos con otras disciplinas

La minería de datos es un campo interdisciplinario que ha evolucionado en conjunto o como una extensión de otras tecnologías. Por esta razón, la investigación y el progreso de la minería de datos se sustenta de lo que generan estas áreas relacionadas.

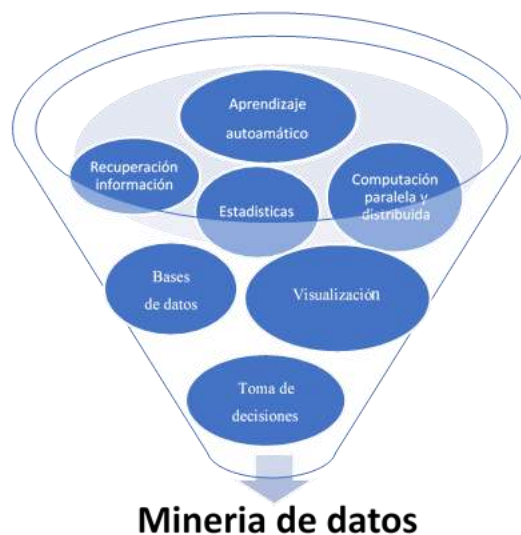


Figura 1. Disciplinas que contribuyen a la minería de datos. Información adaptada del libro *Introducción a la minería de datos*. Elaborado por Ruth Ruiz.

Bases de datos: Las técnicas eficientes de indexación y acceso a datos son importantes para diseñar algoritmos de minería de datos eficientes.

Estadística: Esta doctrina ha suministrado una variedad de conceptos, teorías, técnicas y algoritmos muy útiles para la minería de datos que incluye la media, análisis variante y univariante, teoría de muestreo, etc. Muchos paquetes que sirven para el análisis de datos se venden como mineros de datos.

Sistema de toma de decisiones: Estos sistemas informáticos se contextualizan en el uso de datos y modelos para crear, estimar, evaluar y comparar sistemáticamente alternativas, y así ayudar a los encargados a tomar una decisión.

Aprendizaje automático: Este campo es una mezcla entre inteligencia artificial y estadística, la primera desarrolla programas capaces de aprender y en conjunto con la segunda forman el centro del análisis inteligente de datos, es decir, la máquina toma como referencia un modelo y en base a esos datos, solucionan un problema.

Visualización de datos: Las técnicas de visualización permiten a los usuarios descubrir o comprender intuitivamente patrones que a partir de descripciones escritas o matemáticas resultarían desconcertantes, por ejemplo, gráficos de barras, figuras icónicas, representaciones jerárquicas, etc.

Computación paralela y distribuida: Las bases de muchos sistemas se cimentan desde la computación paralela y distribuida, por disminuir el coste computacional al ejecutar una tarea compleja en diferentes computadores y permitir la escalabilidad de algoritmos.

2.2.2. Metodologías de minería de datos

Dentro de la minería de datos existe un conjunto de metodologías que permiten ejecutar de manera sistemática y no trivial un proyecto data mining de principio a fin. Estas metodologías ayudan a comprender el proceso de descubrimiento de conocimiento y brindan orientación para planificar y materializar un proyecto en mente.

A continuación, se describirá a breves rasgos 3 de las metodologías tradicionales más dominadas en la minería de datos, entre ellas se encuentran KDD, CRISP-DM y SEMMA.

2.2.2.1. Metodología KDD

Metodología KDD (Knowledge Discovery in Data bases) o lo que se traduce como Descubrimiento de conocimiento en bases de datos, fue el primer ejemplar aceptado por la comunidad científica para establecer las principales fases de un proyecto de minería de

información. Hoy en día, los términos KDD y minería de datos se utilizan para referirse al proceso de descubrimiento de conocimiento.

KDD tiene como objetivo darle sentido a los datos mediante el desarrollo de técnicas y métodos que permiten tratar grandes volúmenes de datos que puedan interpretarse de manera más compacta (ej. informe breve), abstracto (ej. descriptivo de aproximación) y útil (ej. Modelo predictivo).

Fayyad, Piatetsky-Shapiro, & Smyth (1996) en el artículo “From data mining to knowledge discovery in databases” describe a grandes rasgos las 9 fases que nacen de la visión práctica de Brachman y Anand:

Comprensión del dominio de aplicación: Consiste en comprender el dominio de aplicación, la información previa importante e identificar los objetivos que el cliente desea alcanzar con el proceso KDD.

Creación del conjunto de datos: Elegir un conjunto o un subconjunto o una muestra de datos desde el cual se obtendrá el nuevo conocimiento.

Limpieza y pre-procesamiento: En esta etapa se ejecutan acciones como: eliminar ruido de datos si es oportuno, determinar estrategias para tratar con la falencia de datos y datos anormales.

Reducción y proyección de los datos: Identificar peculiaridades de utilidad en los datos para representarlos en función de los objetivos de la tarea utilizando métodos de reducción y de transformación de datos.

Determinación de la minería de datos: Se elige la técnica de minería de datos con la cual trabajar como por ejemplo, regresión, agrupación, clasificación. Considerando en todo momento los objetivos planteados en la primera fase.

Determinación de algoritmo de minería: La fase 6 abarca temas relacionados con el análisis exploratorio y la selección del algoritmo que se utilizará de acuerdo al problema y a los datos con el propósito de encontrar patrones.

Minería de datos: Se aplica la técnica y el algoritmo a los datos reducidos y transformados.

Interpretación: Interpretar los patrones encontrados, en fase puede implicar medios de visualización.

Utilización del nuevo conocimiento: Actuar sobre el descubrimiento del nuevo conocimiento e incorporarlo en otros sistemas o simplemente documentarlo y reportarlo a las partes interesadas para que influya en la toma de decisiones (p. 42).

2.2.2.2. Metodología CRISP-DM

El proceso CRISP-DM se desarrolló originalmente a finales de 1996 gracias a los esfuerzos de un consorcio conformado por NCR, SPSS y DiamlerChrysler. En un nivel abstracto, se describe como un desglose jerárquico o un modelo de procesos jerárquicos que van desde lo general hacia lo específico: fases, tareas genéricas, tareas especializadas e instancias de procesos.

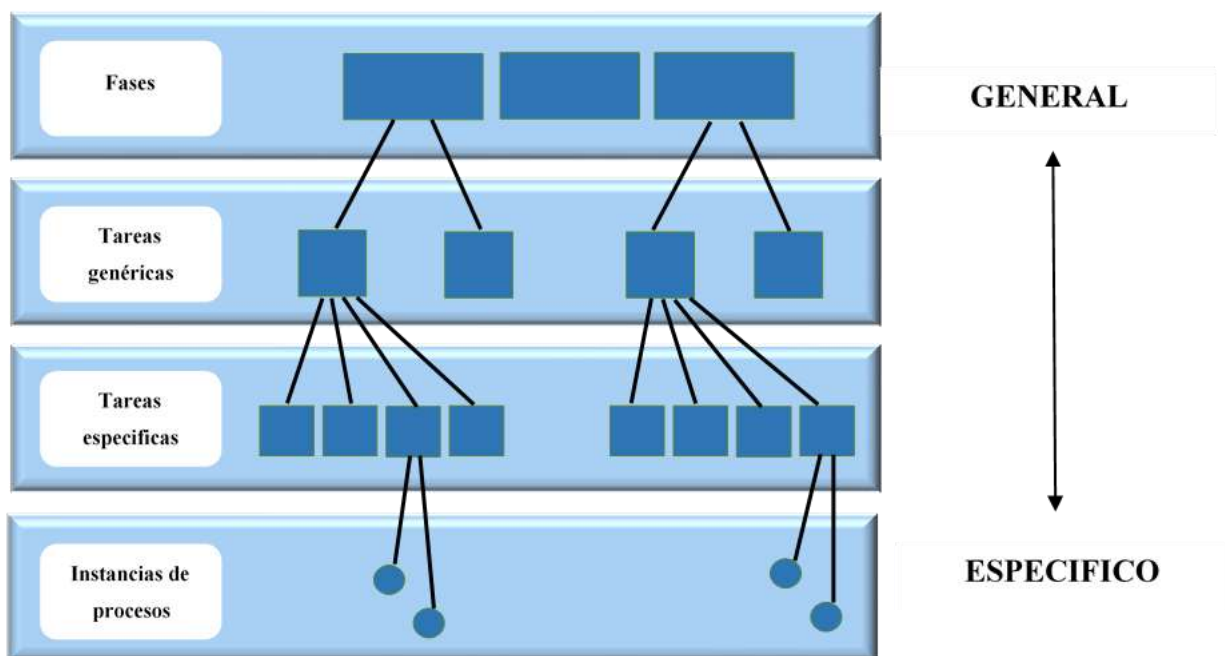


Figura 2. Desglose jerárquico de CRISP-DM. Información adaptada del libro *CRISP-DM 1.0*. Elaborado por Ruth Ruiz

El primer nivel del ciclo de vida de esta metodología consta de 6 fases cíclicas, interactivas, bidireccionales y flexibles, estas fases son: Comprensión del negocio, comprensión de los datos, preparación de datos, modelado, evaluación e implantación. Cada una de las fases mencionadas se dividen en tareas generales de segundo nivel y estas en tareas específicas de tercer nivel de las cuales no siempre hay sugerencias sobre cómo realizarlas, en otras palabras, CRISP-DM proporciona un conjunto de tareas y actividades para cada una de las fases, pero no especifica como efectuarlas, por último, están las instancias de procesos que registran las acciones,

decisiones y resultados de la minería de datos, que se definen en base a las tareas de los niveles superiores (Chapman, y otros, 2000).

Comprensión del negocio: Se enfoca en comprender los objetivos y requisitos del proyecto desde el punto de vista del negocio, con el fin de desarrollar un plan preliminar para lograr los objetivos.

Comprensión de los datos: Esta fase incluye actividades de recopilación de datos, identificación de problemas de calidad de los datos, descubrimiento de los primeros indicios de conocimiento en los datos o detección de un subconjunto de datos que permite formular hipótesis interesantes sobre información oculta.

Preparación de los datos: Transforma y limpia los datos recopilados para construir un nuevo conjunto de datos que se convertirán en la entrada de las herramientas de modelado.

Modelado: Esta fase abarca la selección de la técnica de modelado para generar un diseño de pruebas que permita procesar el conjunto de datos seleccionados a través de herramientas de minería de datos.

Evaluación: El modelo resultante se evalúa a fondo con la finalidad de verificar que los resultados sean correctos y garantizar que se hayan alcanzado los objetivos planteados al inicio del proyecto.

Despliegue: En esta fase, que no necesariamente será la última, sobre todo si el propósito es aumentar el conocimiento gradualmente, los conocimientos adquiridos deberán organizarse y presentarse de una manera comprensible para que el usuario final pueda hacer uso de la herramienta, ej.: informes.

2.2.2.3. Metodología SEMMA

SAS Institute es el desarrollador de la metodología SEMMA quien la define como el proceso de muestreo, exploración, modificación, modelado y evaluación que se aplica a cantidades significativas de datos para descubrir patrones de interés previamente desconocidos que se pueden utilizar como herramienta de ventaja comercial para el negocio.



Figura 3. Fases de la metodología SEMMA. Elaborado por Ruth Ruiz.

Según Sumathi & Sivanandam (2006), menciona que la metodología SEMMA facilita la aplicación de estadística exploratoria y técnicas de visualización, así como la selección y transformación de las variables más importantes, con el objetivo de generar modelos predictivos y evaluar resultados para apoyar una mejor toma de decisiones (p. 233) .

(Hernandez & Dueñas, 2009) describe las 5 fases de la metodología:

Muestreo: Durante esta fase se extrae una muestra de datos para que se pueda iniciar el análisis de la representación de las características generales de la población. La muestra debe ser lo suficientemente grande para contener información significativa, pero lo suficientemente pequeña para procesarse rápidamente.

Exploración: La exploración de datos estadísticos permite monitorear, administrar, detectar, identificar y eliminar los datos que representan tendencias imprevistas, anomalías u omisiones durante el próximo periodo de recopilación de información.

Modificación: En este paso, la selección y transformación de los datos se realiza en base a las variables seleccionadas para el proceso de minado. Esto permite adaptar la elección de modelos y el enfoque de diseño en función de ellos.

Modelado: Esta fase se sirve de herramientas de software que permitan ejecutar técnicas propias de la minería de datos como métodos estadísticos, agrupamiento, reglas de asociación, etc., las cuales despliegan hacia el descubrimiento de asociaciones de datos que ayudan a predecir de manera confiable un resultado deseado.

Evaluación: Se trata de evaluar los datos de acuerdo con la utilidad de los datos y la confiabilidad de las conclusiones del proceso.

La metodología SEMMA se diseñó para trabajar exclusivamente con el software SAS Enterprise Miner y está destinado a guiar a los usuarios a través de la minería de datos.

2.2.3. Comparación de los procesos de las metodologías de minería de datos.

Las metodologías presentadas comparten una teoría en común, esta es, estructurar el proceso de data mining en varias fases interrelacionadas entre sí, lo cual proporciona una idea más amplia de la realización del proyecto, guía el desarrollo del proceso y facilita la adaptación de la metodología acorde a los objetivos de un proyecto de minería de datos. No obstante, no todas las metodologías comparten la misma esencia. SEMMA se enfoca más en las características técnicas del desarrollo de un proyecto, esta diferencia se nota desde la primera fase de la metodología al empezar con el muestreo de datos, mientras que CRISP-DM y KDD procuran

mantener un punto de vista más general, enfocándose en el análisis de requerimientos y en la comprensión del negocio. Cada una de las metodologías cuentan con fases dedicadas a tareas de selección, preparación y modelado de los datos para el descubrimiento de patrones. Otra diferencia entre las metodologías mencionadas radica en la dependencia de herramientas de software, SEMMA está muy ligada a los productos SAS, solo los aspectos generales son abiertos mientras que KDD y CRISP-DM son metodologías más neutras respecto al uso de herramientas de software para el desarrollo del proyecto de data mining, su distribución es libre y sin costo alguno. Respecto a la etapa de evaluación y validación de los resultados obtenidos, la metodología SEMMA evalúa el sistema en base al desempeño del modelo mientras que KDD evalúa en base al cumplimiento de los objetivos planteados y CRISP-DM evalúa el sistema en base al cumplimiento de los objetivos planteados y al desempeño del modelo.

2.2.4. Medios digitales

Los medios digitales componen una forma innovadora de ver y enseñar el mundo, antes se decía que la tecnología era intrínseca a nuestras prácticas, como la piedra que generó fuego. Hoy se puede decir que el “impacto específico” de la tecnología en la naturaleza, cultura, religión, economía, política, educación, etc., ya no existe. Todos estos aspectos del mundo están presentes en la tecnología, permitiendo la evolución y desarrollo en conjunto (Educ.ar portal, 2017).

Los medios digitales son imágenes digitales, videos digitales, audio digital, videojuego, redes sociales, página web y sitios web, archivos digitales, bases de datos, y se codifican en un formato legible para ser interpretados por maquinas. Los medios digitales se pueden crear, visualizar, editar, distribuir y almacenar en un dispositivo electrónico digital e involucran el manejo y la interacción con gráficos, textos, sonidos, imágenes y videos (Instituto La Salle Florida, 2015) .

2.2.5. Sitios Web

Un sitio web consiste en una serie de páginas web enlazadas entre sí mediante hipervínculos alojadas en un servidor web que permite a los usuarios obtener los servicios que ofrezca a través de internet o de una intranet desde un navegador web (Ferrer, 2014). Toda esta comunicación se realiza a través del protocolo HTTP, el código de la página es básicamente HTML, CSS y JavaScript.

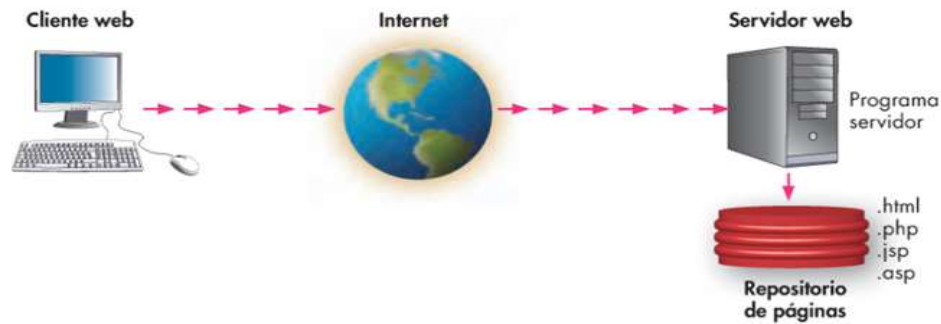


Figura 4. Esquema básico del servicio web. Tomado del libro *Aplicaciones Web*. Elaborado por Lerma-Blasco Raúl.

2.2.5.1. Tipos de sitios web

Se puede encontrar sitios web como blogs, wikis, redifusión web, redes sociales, periódicos.

- **Blogs:** Los blogs son sitios web que permiten compartir conocimientos y experiencias de manera periódica, puede ser diaria, semanal o mensual, los autores son libres de mantener, eliminar o modificar las publicaciones que realizan. Una de las principales características es que todas las entradas publicadas en el blog se archivan y se muestra en orden cronológico (Zofio, 2013, pág. 65). Entre las aplicaciones más destacadas para crear blogs constan *WordPress* y *Blogger*.
- **Wikis:** Las wikis se caracterizan por ser herramientas colaborativas (que pueden estar sujetas a autorización otorgada), sencillas, flexibles y potentes. Cada página dentro de la wiki tiene un título único para ayudar a distinguirla de otras páginas y pueden tener un historial de modificaciones para recuperar el contenido si es necesario. Algunos sitios para crear wikis fácilmente son *Wikispaces* y *Wetpaint*.
- **Redifusión web:** También conocido como sindicación web, permite acceder y distribuir información actualizada entre sitios web en una sola interfaz, en donde las páginas fuentes o canales web suministran a terceros de información (ej. noticias, nuevas entradas de blogs, comentarios de foro) ahorrando la necesidad de tener que iniciar una sesión o visitar cada uno de los sitios web (Zofio, 2013, pág. 63). Los dos formatos de distribución principales son *RSS* y *ATOM*.
- **Redes sociales:** Estos sitios web denominados redes sociales ofrecen la posibilidad a personas y organizaciones de relacionarse con otros a través de un mundo digital, permitiendo crear comunidades a partir de intereses y valores comunes, expresar intereses, compartir historias, preferencias y experiencias. En la actualidad las redes

sociales se aplican en diversos campos como negocios, educación, investigación, entre otros. Algunos ejemplos de las redes sociales más representativas son *Linkedin*, *Facebook*, *Instagram*, *Twitter*.

- **Periódicos digitales:** Los periódicos digitales son medios de comunicación masiva que tienen como función principal brindar información detallada sobre los hechos más representativos, comunicar opiniones, proporcionar entretenimiento y contenido de interés general. Algunos de los periódicos digitales ofrecen contenido de forma gratuita y otros de pago. Son populares por generar actualizaciones instantáneas de información y noticias valiéndose de recursos audiovisuales como imágenes, textos y videos (Raffino, 2021). Ejemplos de periódicos digitales *el universo*, *el comercio*, *el telégrafo*, entre otros.

2.2.6. Códigos de estado de respuesta HTTP

Cuando se establece una comunicación cliente web – servidor web existen estados de respuesta que indican si una solicitud HTTP en particular se completó correctamente. Las respuestas se clasifican en cinco clases:

Tabla 3.

Códigos HTTP

Clase	Descripción
Informativas	Esta respuesta esperada por el servidor indica que la solicitud ha sido recibida y el proceso puede continuar. Entre los códigos más comunes están 100: continue, 102: processing, 103: checkpoint.
Satisfactorias	Con esta respuesta el servidor recibe, procesa y comprende correctamente la solicitud. El código HTTP 200 comunica que la petición ha sido verificada correctamente.
Redirecciones	Usando esta respuesta el servidor da a entender que el cliente debe realizar una nueva acción. En la mayoría de los casos, sin la participación del usuario, solo se carga una URL. Por ejemplo, cuando un recurso dejó de existir permanentemente en una dirección específica, rebota el código 301 y cuando es temporalmente, el 302.
Errores de cliente	Los códigos correspondientes a este tipo de respuesta señalan que el servidor no pudo responder a la solicitud recibida, debido a un error del

cliente. 404 es uno de los códigos más famosos y aparece cuando el recurso no está disponible en el servidor.

Las respuestas de este tipo se presentan cuando el servidor identifica un problema, pero desconoce el por qué ni cómo manejarlo. Los errores típicos son el 502: Bad Gateway y 503: Service Unavailable.

Información adaptada de Xataka Basics. Elaborada por Ruth Ruiz.

2.2.7. Certificado digital

Un certificado digital es un tipo especial de documento que está firmado digitalmente por una Autoridad Certificante (AC), con la finalidad de acreditar electrónicamente la autenticidad de una entidad asociándola con datos de verificación de una firma para validar la identidad.

Estos certificados o credenciales pueden ser de diferentes tipos, entre ellos se encuentran a los que son dirigidos:

- **Para usuarios**, que están diseñados para verificar la identidad de un individuo y vincular un nombre específico con una clave pública y son emitidos por una AC.
- **Para servidores**, que están diseñados para comprobar la autenticidad de un sitio web o páginas web y garantiza la identidad de un servidor mediante la distribución de una clave pública.
- **Para software**, que tienen como propósito proporcionar un sistema de descarga de código confiable y mitigar los efectos de los programas hostiles a través un certificado de código.
- **Para autoridades certificadoras**, que contienen el nombre y la llave publica de una AC que pueden ser autofirmadas o puede firmarla otra AC y se suelen distribuir entre los propios navegadores.

El uso más común de certificados digitales es para servidores web que usan HTTPS, esto permite que un navegador verifique la autenticidad de una entidad ante el usuario, siendo un proceso fundamental sobre todo en el comercio electrónico o en aplicaciones que manipulan datos financieros (Pacheco, 2014).

2.2.8. Data Scraping

La data scraping es una de las técnicas de adquisición de información que se puede interpretar como el conjunto de herramientas informáticas aplicadas de diversas formas para adquirir información legible por el ser humano a partir de una cadena de datos en bruto. Estas

herramientas informáticas pueden ser de diferentes tipos como por ejemplo herramientas web, librerías de programación, aplicaciones de escritorio y otras extensiones de programas.

Los data scrapers han ganado popularidad y se usan comúnmente en diferentes ámbitos en comparación con las tecnologías de recopilación de información tradicional por una de las ventajas más distintivas como lo es el ahorro esencial de tiempo, que no es la única ventaja, pero si la más importante, debido a que con la construcción de un data scraper automatizado lo que se desea conseguir será alcanzable en una fracción de tiempo de lo que tomaría a una persona realizar dicha tarea, y así mismo ejecutarlo repetidas veces (Lopez, 2015) .

Según la fuente de información, los data scrapers pueden dividirse en dos categorías, “screen scrapers” que son particularmente programas que utilizan técnicas de extracción de información para conseguir información relevante de la interface de otra aplicación; tal es el caso de los programas de reconocimiento óptico de caracteres y de conexión remota; y luego están los “Web scrapers” que permiten recolectar datos de la web a través de técnicas como, por ejemplo, analizadores HTML, es en este segundo tipo de scraper en el cual se centra este trabajo.

2.2.9. Web scraping

El web scraping es una de las herramientas más poderosas para la extracción de datos no estructurados de la web en forma automática; traducido del inglés, significa lo mismo que “cavar en la web” o “raspado web” y representa una forma de hacer minería de datos. El web scraping es la solución intermedia entre la recopilación manual de datos (copy - paste) y la recopilación automática basado en protocolos predefinidos (API). La aplicación de esta técnica prima cuando dichos protocolos no están disponibles o cuando la cantidad de datos que se desea extraer es demasiado grande como para realizarlo manualmente (López, 2018).

Generalmente, se hace utilizando un software que simula la navegación humana y recopila información específica desde distintos sitios o páginas web. En tal sentido, cada vez que alguien realiza la acción “copiar y pegar” en diferentes páginas web para extraer datos y utilizarlos para un fin concreto, lo que realizó no es otra cosa más que un “scrapeo”. Sin embargo, este método puede resultar costoso debido a que puede llevar tiempo recuperar grandes cantidades de datos desde diferentes sitios, organizarlos y estructurarlos para el uso posterior (Hanretty, 2013).

Un raspador web consta comúnmente de dos componentes principales: Navegar por los enlaces de un sitio web y extraer datos de los enlaces visitados, lo que se conoce como *web crawler* y *extractor de datos* respectivamente (Vicente, 2018).

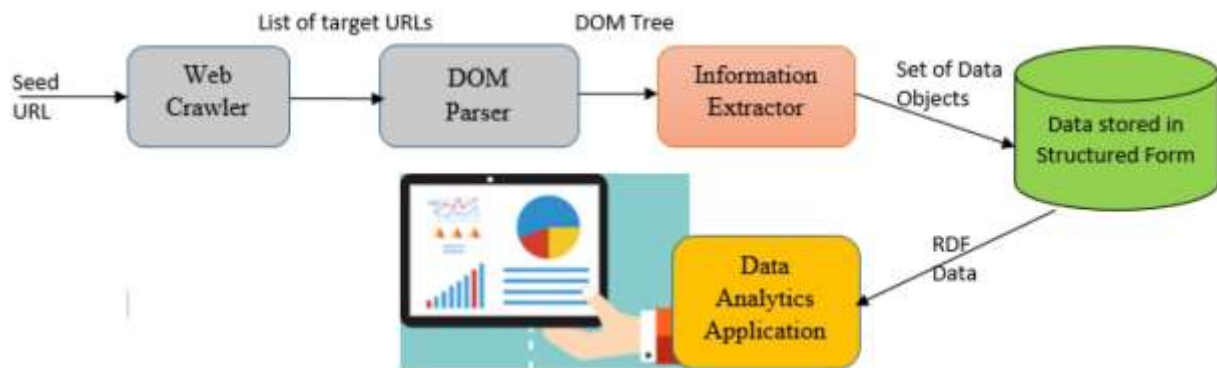


Figura 5. Marco general del Web Scraping. Tomado de *Automated scraping of structured data record from health discussion forums using semantic analysis*. Elaborado por Baskaran y Ramanujam.

2.2.9.1. Web Crawler

El punto de partida para la ejecución de un Web crawler comienza con la selección de una URL semilla o también conocida como URL raíz. El web crawler examina la URL semilla con el objetivo de identificar, indexar, guardar y almacenar los hiperenlaces contenidos en el sitio web para el análisis posterior de acuerdo con un conjunto determinado de reglas.

Estas arañas web son la razón principal por la que buscadores como Google, Bing, DuckDuckGo, Yahoo proporcionan resultados actualizados y al día debido a que deambulan por Internet de forma puntual por todas las ramificaciones en busca de información, examinan el contenido y guardan la información en índices y bases de datos para incrementar la productividad de los motores de búsqueda (Digital Guide IONOS, 2020).



Figura 6. Proceso de Web Crawler. Tomado de *Feeding de Maching*. Elaborado por Felipe Veloso

2.2.9.2. *Extractor de datos*

El extractor de datos es un software interactivo que consigue la información necesaria de una página web, aprovechando la estructura HTML de la página web y utilizando las expresiones regulares, palabras claves, elementos y atributos contenidos en la estructura propia del lenguaje de contenido para extraer la información encapsulada que existe en estos selectores. En la **Tabla 3.** se resume el proceso de un extractor de datos (Vicente, 2018, pág. 4).

Tabla 4.

Procedimiento para hacer web scraping

	Descripción
Inspeccionar código fuente	Identificar dentro del lenguaje de contenido los elementos, atributos, expresiones regulares o palabras claves que contienen los datos que se desean.
Extracción de datos	Escribir un programa y diseñar las funciones que sean necesarias para conseguir los datos que se necesitan. Por ejemplo: parseando el código HTML
Almacenar o procesar los datos extraídos	Trasladar los datos extraídos en una estructura útil para el uso y la visualización posterior.

Información adaptada de Fahrenheit Freiheit. Elaborado por Ruth Ruiz.

2.2.10. Usos de web scraping

El uso del web scraping va acorde a la necesidad de cada individuo, esta técnica permite aprovechar datos de la web que son cada vez más valiosos y generar aplicaciones prácticas como:

- Comparar precios en línea
- Descubrir cambios en la web
- Recolectar noticias, artículos, etc.
- Recopilar contactos de interés
- Aplicaciones de Inteligencia Empresarial
- Aumentar significativamente la cantidad de datos para investigación

- Análisis del competidor
- Periodismo de datos
- Análisis de sentimiento de clientes
- Monitoreo de datos meteorológicos

2.2.11. Impedimentos para realizar web scraping

Hay varios factores que entorpecen la recuperación de datos con herramientas destinadas a realizar web scraping, llegando a complicar o impedir que estos programas cumplan con el objetivo para el cual fueron desarrollados. Cuanto más valor tenga una página web, más seguras serán las medidas que adoptarán los operadores. Según Imperva (2016), algunos ejemplos de impedimentos para realizar web scraping son:

- Sistemas de autenticación como CAPTCHA contenidos en páginas web, en donde no es posible programar un scraper totalmente automatizado para extraer información ya que, en cada solicitud requiere una validación totalmente diferente a la anterior.
- Páginas antiguas que no han actualizado ni modificado desde hace mucho tiempo la estructura simple del lenguaje de contenido lo que imposibilita la extracción organizada desde las etiquetas HTML.
- Bloqueadores de dirección IP que se activan como medida de prevención ante ataques de denegación de servicio que se puede producir al recibir múltiples peticiones de un bot en un mismo servidor.
- Uso de cookies para dar seguimiento a la interacción de los usuarios con el recurso web para rastrear patrones de navegación anormales y tasas de solicitudes sospechosamente agresivas lo que ayuda a identificar bots.

2.2.12. Precauciones antes de hacer web scraping

Para buscadores como Bing, Google, DuckDuckGo, el análisis de millones de páginas y la indexación de estas son tareas cotidianas del día a día, sin embargo, para algunos sitios web el realizar estas acciones programadas y consecutivas pueden considerarse maliciosas, debido a que todos los bots usan el mismo sistema para acceder a los datos del sitio y resulta complicado distinguir entre quienes son y quienes no, por lo tanto, antes de realizar web scraping es importante tener en cuenta ciertos aspectos (Mousinho, 2019):

- Verificar las restricciones de acceso a través del archivo robots.txt que se encuentra ubicado en el directorio principal de la web. Ej: User-agent: * Disallow: /

- Revisar políticas de privacidad y navegación de la página web a la cual se desea scrapear.
- Tener en cuenta las leyes locales de protección de datos del país en donde aloja el sitio web.
- No realizar tareas repetitivas en corto tiempo para evitar bloqueos.
- Verificar que los enlaces a extraer no se hayan ocultado mediante CSS ya que es una técnica de honeypot de los sitios anti-scraping para atrapar arañas web.

2.2.13. Técnicas de web scraping

La tecnología de extracción de datos web ha revolucionado la interrelación usuario-computadora por facilitar el acceso a una gran cantidad de información para beneficio del usuario. A través del tiempo varias técnicas para recopilar información se han venido desarrollando como las que se describen a continuación (Saurkar, Pathare, & Gode, 2018):

Copiar y pegar: Es una de las soluciones más simples de raspado web, pero también uno de las más primitivos y lenta, por lo general, solo se realiza una vez.

Programación HTTP: Facilita el acceso y la recuperación de información contenida en páginas web estáticas y dinámicas a través de la expedición de peticiones HTTP al servidor web remoto.

Parser HTML: Es un analizador sintáctico que permite comprender la estructura y contenido de una página o sitio web para extraer datos de preferencia. Generalmente se construyen mediante lenguajes de programación como Python, PHP y R.

Web Scraping software: El uso de esta técnica de web scraping es menos compleja que la técnica anterior, con la ayuda de una interfaz gráfica permite reconocer automáticamente la estructura de una página web y simular la navegación humana, eliminando la necesidad de redactar manualmente el código de un bot para conseguir los datos deseados y transformarlos en un formato útil y descifrable.

Análisis DOM (Document object model): Con los objetos DOM que representan documentos HTML y XML a través de un conjunto estándar de objetos, es posible acceder, añadir y modificar el contenido estructurado del HTML y XML a través de un lenguaje de programación como JavaScript y recuperar el contenido o parte del contenido dinámico generado por los scripts del cliente web.

Uso de APIs: A través de una interfaz de programación de aplicación, ya sea privada o pública, también es posible acceder y recuperar datos estructurados mediante programación, algunos ejemplos de sitios web que proporcionan APIs son LinkedIn, Twitter y Facebook (Barría, 2020).

2.2.14. Herramientas de web scraping

Existen diversas herramientas dedicadas a realizar web scraping que facilitan la comprensión del usuario y el ahorro de tiempo cuando se desea conseguir grandes cantidades de datos entre las cuales se encuentran herramientas de programación, aplicaciones de escritorio; aplicaciones web y extensiones para navegador.

2.2.14.1. Herramientas de programación

- **BeautifulSoup** es una biblioteca de Python que realiza interacciones con los elementos de una o varias páginas web. Similar a utilizar la opción “Inspeccionar elemento” en el navegador, esta librería es capaz de extraer o recabar datos del contenido en HTML o XML del sitio mediante un analizador o parser para transformarlo en un árbol de objetos de Python, es completamente configurable y personalizable, además de ser gratuita y de uso ilimitado, la desventaja de esta herramienta es no poder efectuar la extracción de contenido dinámico (Lozano Gomez, 2020).
- **Scrapy** es el marco de trabajo de Python de código abierto y multipropósito más popular en temas relacionados a crawling y scraping en la web, aplicado principalmente en minería de dato por ser rápido y concurrente, lo que resulta muy útil al momento de descargar datos de manera masiva (Scrapy, 2021) .

2.2.14.2. Aplicaciones de escritorio

- **Parsehub** es la herramienta de escritorio de web scraping ideal para ser utilizada por personas que carecen de conocimientos técnicos de extracción de datos, tiene una función de “selección rápida”, la cual determina automáticamente la estructura de una página web y agrupa los datos que poseen relación para su posterior almacenamiento, al mismo tiempo, para los desarrolladores cuenta con funciones avanzadas que brindan el control total de la recuperación de datos. ParseHub está diseñado para abordar casos complejos y mal diseñados de páginas web que usan AJAX y JavaScript (ParseHub, 2015).

- **Octoparse**, también es una herramienta simple y poderosa que permite el raspado automático de datos contenidos en sistemas web públicamente perceptibles en Internet estáticos y dinámicos, ahorrando la necesidad de generar código para extraer una gran cantidad información del sitio web. Una de las características particulares de Octoparse es proporcionar cientos de plantillas de rastreo web para personas sin conocimiento de programación, solo basta poner el URL y una palabra clave. Posee 2 versiones de uso, una gratuita muy limitada y una pagada (Octoparse, 2021).

2.2.14.3. Aplicaciones web

- **Dexi.io** es una poderosa herramienta de web scraping con características superiores a Octoparse o Parsehub dado que, posee diversos tipos de bots para realizar varias tareas de raspado web como por ejemplo extractores, rastreadores y tuberías, esta última, orquesta el trabajo de los otros bots uniéndolos en un mismo proyecto, sin embargo, esta herramienta no es para todos, está dirigido principalmente para usuarios con habilidades de programación, a quienes les da la opción, a través de un editor visual, diseñar un robot potente que realice tareas concisas (Dexi.io, 2020).
- **Import.io** considerada como una de las mejores herramientas de rastreo web, por ser fácil de usar y por poseer una interfaz sencilla que elimina toda necesidad de escribir código haciendo uso del principio apuntar y clicar para extraer datos y organizarlos en conjuntos de datos, es la más utilizada para ejemplificar lo que significa web scraping, maneja paginas JavaScript, AJAX e inclusive puede extraer datos detrás de un inicio de sesión. Actualmente no cuenta con planes gratuitos y, a medida que se trabaja con más paginas el precio aumenta, lo puede convertir a esta herramienta en una de las más costosas (Import.io, 2020).

2.2.14.4. Extensiones para navegador

- **Web Scraper.io** es un pequeño programa complementario o extensión que se integra dentro de la herramienta para desarrolladores del navegador y permite extraer datos de una página web seleccionando los elementos que contenga para construir un mapa del sitio adaptando la extracción de datos a partir de dicha estructura. Dispone de pocas funcionalidades y tiene una interfaz no tan amigable para el usuario (Web Scraper, 2016).

Tabla 5.

Tabla comparativa de herramienta de web scraping.

Herramienta	Entorno	Dificultad	Gratis/Pago	Formato de documentación
BeautifulSoup	Windows, Linux, Mac, BSD.	Requiere de conocimientos de programación y desarrollo web	Gratis e ilimitado	Múltiples formatos
Scrapy	Windows, Linux, Mac, BSD.	Requiere de conocimientos de programación	Gratis e ilimitado	Múltiples formatos
ParseHub	Windows, Linux, Mac	Interfaz flexible, amigable e interactiva	Plan gratuito limitado	CSV, JSON
Octoparse	Windows, Mac, Nube	Interfaz intuitiva y fácil de usar	Plan gratuito limitado	API, CSV, XLSX, bases de datos.
Dexi.io	Navegador web, Nube.	Interfaz compleja y poco amigable.	Plan gratuito limitado	CSV, JSON, HTML, XLSX
Import.io	Navegador web, Nube	Interfaz agradable y guiada	Precios a solicitud	HTML, CSV, JSON, XLSX
Web Scraper.io	Extensión de Chrome y Firefox, Nube	Interfaz poco amigable	Plan gratuito limitado	CSV, JSON, XLSX

*Información tomada de la investigación directa. Elaborado por Ruth Ruiz***2.2.15. Lenguajes empleados en Web scraping**

Cuando se habla de web scraping, como se observó en el punto anterior, es posible encontrar un conjunto de herramientas y aplicaciones que permiten aplicar esta técnica para abordar problemáticas de diversos indoles. Algunas de estas herramientas son capaces de operar sin necesidad de codificar absolutamente nada, pero cuando el raspado de datos presenta aspectos como, por ejemplo, un código fuente HTML mal estructurado, datos deseados a extraer

dispersos en más de una página debido a la estructura de presentación de los datos o cuando la frecuencia de extracción y el volumen de los datos aumenta, será necesario considerar el uso de herramientas que permitan construir y personalizar un raspador web puesto que, debido a la complejidad del escenario, herramientas prediseñadas comienzan a parecer muy limitadas, no cumplen con los objetivos e inclusive se vuelven costosas, haciendo que el mejor camino a seguir sea utilizar un lenguaje de programación para crear una solución propia de raspado web.

Hacer uso de un lenguaje de programación trae beneficios como: Utilizar bibliotecas de terceros que facilitan el proceso de análisis y extracción de datos; contar con toda una comunidad que brinda soporte para desarrolladores mediante sitios web como Stackoverflow; simular la navegación humana en intervalos de tiempo moderados para evitar ser bloqueado o entrar a la lista negra; tener libre albedrio en el enfoque del raspado web sin estar limitado por contrato con proveedores externos.

Muchos lenguajes de programación ocupan el primer lugar en proyectos relacionados a minería de datos, lo que dificulta seleccionar uno en particular, si se toma la decisión equivocada terminaría costando tiempo y energía en algo que no proporcionará los resultados esperados.

Arsalan (2021) menciona algunos parametros a tener en cuenta para seleccionar un lenguaje de programación mas conveniente.

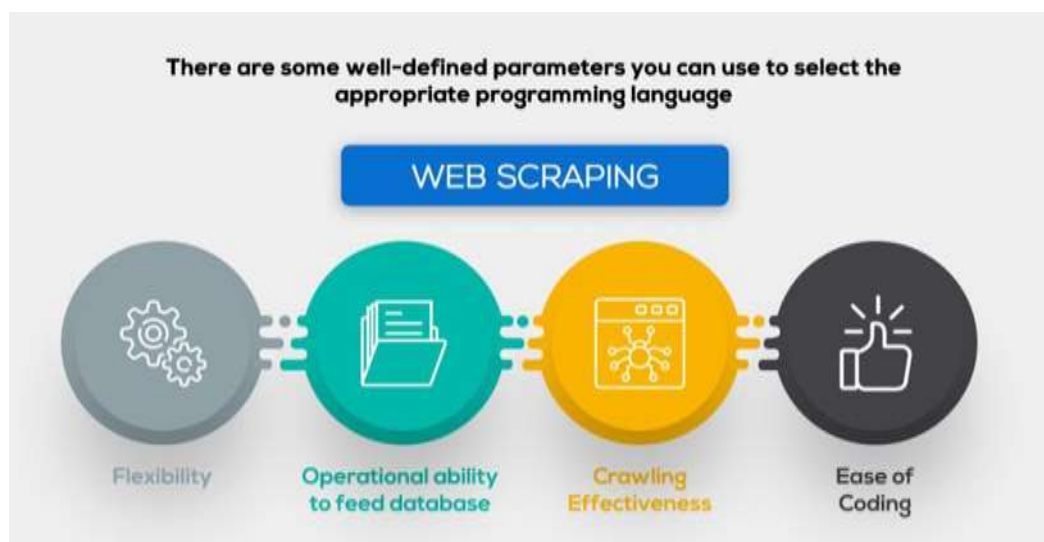


Figura 7. Parámetros para escoger un lenguaje de programación de web scraping. Tomado de Information Transformation Services. Elaborado por Arsalan.

Y a la misma vez, presenta los 5 mejores lenguajes de programación para realizar web scraping, y estos son: Python, Node.js, Ruby, PHP y C&C++.

2.2.15.1. Python

Python es el lenguaje de programación de alto nivel más popular para la creación de programas de web scraping que se conoce en la actualidad, es flexible, fácil de aprender, sencillo y open source y garantiza un proceso de raspado web sin errores. Posee un ecosistema bastante completo y puede manejar la mayoría de los procesos relacionados con la extracción de datos sin inconvenientes, una de las bibliotecas más utilizadas es BeautifulSoup, diseñada para realizar una extracción de datos rápida y eficiente, permite, por ejemplo, analizar paginas HTML o XML a través de analizadores como html5lib o lxml para identificar elementos de los cuales se planea obtener los datos y extraer los datos o convertir el formato de documentación de un documento entrante de Unicode a UTF-8. Python es un lenguaje de amplio uso, simple y natural.

2.2.15.2. Node.js

Node.js es en particular el lenguaje de programación más oportuno para rastreo de datos en sitios o páginas web que aplican el manejo de prácticas de codificación dinámica para representar la interfaz del usuario. A pesar de que admita el concepto de rastreo distribuido, tiene una estabilidad de comunicación relativamente baja y no se recomienda para grandes proyectos que requieran extraer datos en más de una página y más de una vez. Cada proceso que se ejecuta en node.js requiere de un núcleo de la unidad central de procesamiento (CPU), muchos usuarios aprovechan esta cualidad e ingresan múltiples instancias del mismo proyecto de raspado dentro un solo proceso. Cheerio y ExpressJS son bibliotecas de terceros que trabajan en node.js para ejecutar técnicas de crawling y scraping en servidores y aplicaciones web respectivamente.

2.2.15.3. Ruby

Ruby es uno de los lenguajes de programación a mirar con buenos ojos, representa el equilibrio entre la programación funcional y la programación imperativa. En comparación con otros lenguajes de programación, Ruby, es ampliamente empleado debido a la sencillez y naturalidad productiva. Ruby posee un marco web de escritura de código simple llamado Ruby on rails que permite desarrollar aplicaciones de web scraping y proporciona características que reducen el tiempo de desarrollo y los tiempos de respuesta. Los Rubygems (gestor de paquetes

de Ruby) utilizados para este tipo de técnica son HTTParty, NokoGiri y Pry, el primero permite la transferencia de solicitudes HTTP, el segundo analiza HTML, XML, CSS y el tercero posibilita depurar programas para facilitar el raspado web.

2.2.15.4. PHP

PHP es el lenguaje menos popular para crear raspadores web, esto se puede deber al débil soporte con async y multi-threading. No se recomienda el uso de PHP para la búsqueda web, ya que los problemas relacionados con colas y programación de tareas pueden estar relacionados con el uso de este lenguaje. Sin embargo, con la librería cURL, es posible extraer imágenes, texto, videos y gráficos de numerosos sitios web, consigue transferir archivos de manera eficiente utilizando una lista de protocolos que incluye HTTP y FTP, tener todo esto, permite crear arañas web para descargar información de cualquier tipo en un medio digital.

2.2.15.5. C&C++

C&C++, ambos son una excelente opción para la construcción de un raspador web único, capaz de proporcionar los mejores resultados cuando se trata de extraer datos de la web. A diferencia de los otros lenguajes de programación, este podría resultar costoso. En muchas ocasiones, el uso de este lenguaje de programación no se recomienda a menos que, sea una organización centrada en el análisis de datos. Es sencillo y cómodo de entender, paralelizar un raspador web con C&C++ es una opción. Libcurl es una de las librerías útiles para el proceso que permite buscar el URL para luego desarrollar los requerimientos, debe considerarse como la última opción dentro de la minería de datos.

2.2.16. Dashboard

Los tableros de mando y la visualización son herramientas cognitivas que brindan un mayor “control” sobre grandes cantidades de datos. Estas herramientas ayudan a las personas o entidades a identificar visualmente tendencias, patrones y anomalías, reflexionar sobre lo que ven y guiarlas para tomar decisiones efectivas (Brath & Peters, 2004) .

Los tableros de mando son herramientas que se utilizan en temas de gestión de la información e inteligencia empresarial para organizar, almacenar y visualizar gráficamente datos de una variedad de fuentes de una manera que es fácil de entender y analizar.

2.2.17. Capas de un dashboard

Desde el punto de vista de Lema (2016), establece que las mejores prácticas para desarrollar cuadros de mando requieren tres niveles o vistas de información:

- **Capa o vista superior:** Los usuarios realizan un seguimiento de la información que a menudo, se resume y muestra gráficamente, señalando los indicadores más relevantes y el estado de las condiciones anormales.
- **Capa o vista intermedia:** Capa en la que se extiende la información que sustenta a los indicadores y alertas del nivel superior. Navegando y profundizando entre tablas y gráficos, y realizando cálculos complejos, es posible encontrar condiciones y tendencias especiales desde todos los ángulos.
- **Capa o vista inferior:** Los usuarios tienen datos de vital importancia que les permite encontrar la raíz de cualquier problema mediante los informes y registros de transacciones detallados.

2.2.18. Errores frecuentes de diseño

El diseño eficiente es esencial para los dashboards. Un buen diseño de información transmite claramente información importante a los usuarios y hace que la información de soporte sea fácilmente accesible, por lo tanto, se debe tener en cuenta ciertas consideraciones al momento de diseñar un dashboard y evitar errores como:

1. Excederse en el uso de colores y manejar interfaces poco armoniosas.
2. Saturar la pantalla y sobredecorar.
3. Proveer información que está fuera de contexto.
4. Dispersar los datos en más de una pantalla impidiendo la interpretación de una sola vista.
5. Diseñar una visualización ilegible, desagradable y simple.
6. Escoger gráficos inapropiados.
7. No enfatizar lo suficiente los datos importantes o nada en absoluto.

2.2.19. Beneficios del desarrollo de un dashboard

- Proporcionar de manera general, el estado de una empresa, institución o una situación para guiar al/los interesado/s a tomar decisiones.
- Presentar información fácil de interpretar, oportuna, fiable y verificable.
- Detectar cambios en planes estratégicos u operativos y descubrir la razón de origen para corregir a tiempo.
- Monitoreo en tiempo real.
- Explicar las acciones que se efectúan a corto y largo plazo.

- Generar confianza y colaboración.

2.2.20. Herramientas para construir un tablero de mando

Cuando se habla de Data mining, Big data o Business Intelligence, se entiende de buenas a primeras que hay que lidiar con el procesamiento de incesantes cantidades de datos independientes y dispersos para convertirlos en gráficos o mapas que sean sencillos de interpretar y de utilidad para una entidad, persona o individuo en cuestión (García , 2020).

Por suerte, existen herramientas que permiten realizar este trabajo; Tableau y Power BI son 2 de las más conocidas en este ámbito, ambas herramientas hacen que sea fácil la creación de tableros de mando inteligibles y visualmente atractivo sin ningún conocimiento de programación.

2.2.20.1. Tableau

Tableau es una poderosa herramienta de visualización y análisis de datos que facilita a personas y empresas comprender y tomar de decisiones a través paneles visuales que optimizan, simplifican y procesan datos en bruto. Posee funciones simples como arrastrar y soltar lo que permite que cualquier persona pueda acceder, analizar, generar

s y distribuir esa información con otros.

Existen varias distribuciones Tableau: Tableau Desktop, Tableau Server, Tableau Mobile, Tableau Public. Todas estas son de plan pagado \$70/usuario/mes a excepción de Tableau Public que es la versión gratuita, aunque no es del todo generosa, es muy limitada en cuanto a diseño.

Tableau puede conectarse con archivos de Excel, Access, Texto y bases de datos como Microsoft SQL Server, MySQL, Oracle, Big Query por lo que funciona con cualquier tipo y tamaño de datos, además, Tableau posee la capacidad de trabajar con grandes magnitudes de datos.

2.2.20.2. Power BI

Power BI es el generador de tableros de mando de Microsoft con un alto grado de personalización, es probablemente una de las herramientas más versátiles y poderosas para crear cuadros de mando completos capaces de administrar un gran volumen de datos. Power BI puede integrarse con la caja de herramientas de Dynamics 365 que incluye Office 365 y el almacenamiento en Azure. La interfaz puede resultar familiar, pues posee similitudes con Excel o PowerPoint con muchos botones y herramientas de creación.

Power BI admite múltiples fuentes de datos que incluyen la Suite Office como Excel, Access, SQL, Azure SQL Database, Salesforce, SharePoint, puede conectarse con datos importados y resultados en caliente (en tiempo real), además, se considera una herramienta flexible ya que permite reprocesar dinámicamente bases de datos que tengan o no los datos modelados, ej.: tablas planas o modelo de estrella.

En lo económico, Power BI posee versiones de Escritorio, móvil, pro y premium, el plan pagado tiene un precio de \$10, también posee un plan gratuito con una duración de un mes o 2 meses si posee una cuenta empresarial o institucional que incluye las funcionalidades Power BI PRO.

2.3. Fundamentación legal

Como ocurre en muchas áreas de la tecnología, el precedente legal para el raspado web es pobre. Una buena regla para evitar problemas es atenerse siempre a los términos, condiciones y documentos de derechos de autor del sitio web si corresponde (Mitchell, 2013).

De modo que, la legalidad del web scraping recae en la forma de utilizar esta técnica, por lo tanto, es necesario actuar con cautela y no vulnerar los derechos de propiedad intelectual, violar los términos legales y condiciones de uso o incumplir con las normativas de protección de datos personales (Ron, 2019).

Según la Constitución de la Republica del Ecuador (2015):

Art. 18.- “Todas las personas, en forma individual o colectiva, tienen derecho a:

1. Buscar, recibir, intercambiar, producir y difundir información veraz, verificada, oportuna, contextualizada, plural, sin censura previa acerca de los hechos, acontecimientos y procesos de interés general, y con responsabilidad ulterior.
2. Acceder libremente a la información generada en entidades públicas, o en las privadas que manejen fondos del Estado o realicen funciones públicas. No existirá reserva de información excepto en los casos expresamente establecidos en la ley. En caso de violación a los derechos humanos, ninguna entidad pública negará la información” (pág. 15).

Art. 350.- “El sistema de educación superior tiene como finalidad la formación académica y profesional con visión científica y humanista; la investigación científica y tecnológica; la innovación, promoción, desarrollo y difusión de los saberes y las culturas; la construcción de

soluciones para los problemas del país, en relación con los objetivos del régimen de desarrollo” (Constitución de la Republica del Ecuador, 2015, pág. 157).

De acuerdo con el Decreto Ejecutivo 1014 de software libre Ecuador (2011):

Art. 2.- Se entiende por software libre, a los programas de computación que se pueden utilizar y distribuir sin restricción alguna, que permitan el acceso a los códigos fuentes y que sus aplicaciones puedan ser mejoradas. Estos programas de computación tienen las siguientes libertades: Utilización de programa con cualquier propósito de uso común; Distribución de copias sin restricción alguna; Estudio y modificación de programa; Publicación del programa mejorado (pág. 1)

Según el **Art. 134.-** Actividades permitidas sin autorización del Código Orgánico de la Economía Social de los Conocimientos señala que, se permite las actividades relativas a un software de licita circulación, sin que se requiera autorización del autor o titular, ni pago de valor alguno, en los siguientes casos (Libia, 2015) :

3. Las actividades de ingeniería inversa sobre una copia legítimamente obtenida de un software que se realicen con el único propósito de lograr la compatibilidad operativa entre programas o para fines de investigación y educativos;

Según Vergara (2017) en la Ley Orgánica del Sistema Nacional de Registro de Datos Públicos los datos de carácter personal se definen en el **Art. 6.- Accesibilidad y confidencialidad** que dice lo siguiente:

“Son confidenciales los datos de carácter personal, tales como: ideología, afiliación política o sindical, etnia, estado de salud, orientación sexual, religión, condición migratoria y los demás atinentes a la intimidad personal y en especial aquella información cuyo uso público atente contra los derechos humanos consagrados en la Constitución e instrumentos internacionales. El acceso a estos datos sólo será posible con autorización expresa del titular de la información, por mandato de la ley o por orden judicial” (pág. 4).

El **Art. 234.-** Acceso no consentido a un sistema informático, telemático o de telecomunicaciones dentro del Código Orgánico Integral Penal menciona lo siguiente:

“La persona que sin autorización acceda en todo o en parte a un sistema informático o sistema telemático o de telecomunicaciones o se mantenga dentro del mismo en contra de la voluntad de quien tenga el legítimo derecho, para explotar ilegítimamente el acceso logrado, modificar un portal web, desviar o redireccionar de tráfico de datos o voz u ofrecer

servicios que estos sistemas proveen a terceros, sin pagarlos a los proveedores de servicios legítimos, será sancionada con la pena privativa de la libertad de tres a cinco años” (Ordoñez Rivas, 2021).

Capítulo III

Propuesta

3.1. Métodos de investigación

3.1.1. Diseño de investigación

3.1.1.1. Enfoque Mixto

“La meta de la investigación mixta no es reemplazar a la investigación cuantitativa ni a la cualitativa, sino utilizar las fortalezas de ambos tipos de indagación combinándolas y tratando de minimizar sus debilidades potenciales” (Hernandez Sampieri, 2014, pág. 532).

De acuerdo con Castro & Godino (2021), las características del enfoque mixto son:

- Representar procesos sistemáticos, empíricos y críticos de investigación.
- Recolectar y analizar datos cualitativos y cuantitativos.
- Realizar conclusiones de la información recolectada para obtener desde varios puntos de vistas mayor objetividad.
- Aplicar y rescatar lo más destacado de cada método.

Tabla 6.

Tipos de enfoques mixtos

Explicativo secuencial	En primer lugar, recopila y analiza datos cuantitativos para proceder a realizar el mismo proceso con los cualitativos, priorizando los datos cuantitativos para la posterior interpretación en los resultados de la encuesta
Exploratorio secuencial	Los datos cualitativos se priorizan en este tipo de enfoque mixto, con el proceso de recopilación y análisis de los datos cualitativos y por último los cuantitativos.
Triangulación concurrente	Realiza por separado el proceso de recopilación y análisis de datos cualitativos y cuantitativos, pero al mismo tiempo, con el mismo grado de prioridad para la interpretación de los resultados.
Anidado o incrustado concurrente	Efectúa un solo proceso de recopilación de datos, en el cual se prioriza al método más dominante y se anida con uno de baja prioridad.

Información adaptada de Lifeder. Elaborado por Ruth Ruiz

El diseño de triangulación concurrente es el diseño más popular y utilizado por los investigadores que desean corroborar los resultados, minimiza las debilidades de los métodos y realiza una validación mutua entre los datos cualitativos y cuantitativos.

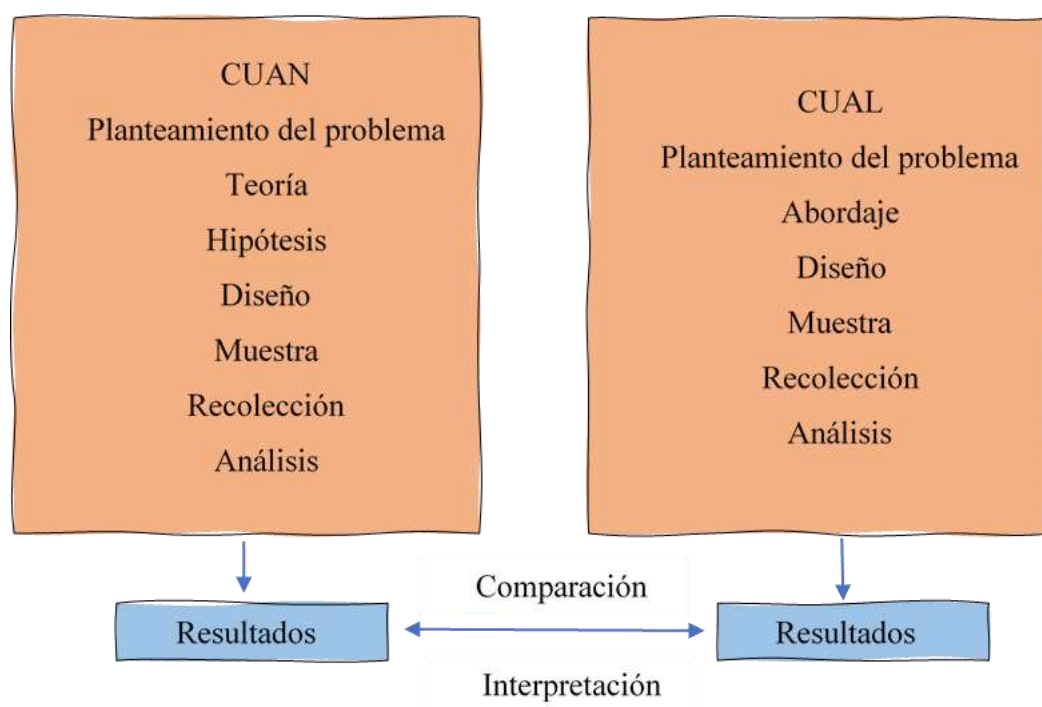


Figura 8. Diseño de triangulación concurrente. Información adaptada de libro *Metodología de la investigación*. Elaborado por Ruth Ruiz.

3.1.2. Metodología de investigación

3.1.2.1. Documental

“La investigación documental es un procedimiento científico, un proceso sistemático de indagación, recolección, organización, análisis e interpretación de información o datos en torno a un determinado tema. Al igual que otros tipos de investigación, éste es conducente a la construcción de conocimientos”. (Alfonzo, 1995)

La investigación documental permite elaborar de forma correcta los temas que intervienen dentro del proceso teórico del trabajo de titulación, por medio de la recolección de datos tanto por libros, artículos, repositorios y otros sitios con información verídica para sustentar de donde provienen y comprender las técnicas y herramientas utilizadas, además, ayuda a encontrar la información legal para cumplir las normativas establecidas por el gobierno ecuatoriano.

3.1.2.2. Cuasi experimental

Esta investigación ayuda a obtener los datos de un ambiente específico donde no se puede controlar todas variables involucradas como en el caso de la extracción de datos de un sitio web mediante el procesamiento de datos.

3.1.2.3. Descriptiva

“Busca especificar las propiedades, las características y los perfiles de personas, grupos, comunidades, procesos, objetos o cualquier otro fenómeno que se someta a un análisis. Es decir, únicamente pretenden medir o recoger información de manera independiente o conjunta sobre los conceptos o las variables a las que se refieren, esto es, su objetivo no es indicar como se relacionan estas”. (Hernandez Sampieri, 2014, p. 92)

Este tipo de investigación tiene como objetivo obtener información precisa y real mediante técnicas de recolección de datos, por tal motivo, se escoge la entrevista dirigida a profesionales expertos, que ayuda a sustentar la validez de las decisiones que se optarán para el desarrollo del prototipo de dashboard, todo esto es posible por los métodos que contiene en enfoque mixto.

3.1.3. Técnica

La técnica de investigación proporciona criterios e instrumentos destinados a garantizar la operatividad del proceso de investigación para obtener más conocimiento e información y así resolver dudas.

El instrumento por utilizar será una entrevista tipo encuesta que permitirá obtener datos cualitativos y cuantitativos con el objetivo de recopilar información que aporte al trabajo de titulación con relación a la parte técnica (herramientas y software) que involucra la creación de un tablero de mando y un raspador web como soporte para los microemprendedores desde el punto de vista de profesionales en las diferentes áreas.

3.1.3.1. Análisis de la entrevista

Se obtuvo la colaboración de 4 profesionales con experiencia en el área de la Ciencia de Datos para realizar la entrevista tipo encuesta que se llevó a cabo mediante la herramienta Google Forms, consta de 4 preguntas para construir el perfil del profesional consultado y 8 preguntas técnicas que aportan con información para el desarrollo de los objetivos planteados, el resultado es el siguiente:

Tabla 7.

Perfil académico del profesional consultado.

Apellidos y Nombres	Formación académica y Área de estudio actual
Pilacuán Bonete Luis	Título de tercer nivel: Ingeniería Industrial
	Título de cuarto nivel: Máster en Administración de Empresas
	Área de estudio: Estadísticas
	Título de tercer nivel: Ingeniería en Computación
Plaza Vargas Ángel Marcel	Título de cuarto nivel: Máster en Modelado Computacional
	Área de estudio: Minería de Datos y Microcontroladores
	Título de tercer nivel: Licenciatura en Sistemas de Información
	Título de cuarto nivel: Máster en Sistemas de Información
García Juan Carlos	Área de estudio: Ciencia de datos
	Título de tercer nivel: Ingeniería en Electrónica y Telecomunicaciones
	Título de cuarto nivel: Máster Universitario en Ingeniería de Sistemas y Servicios para la Sociedad de la Información
	Área de estudio: Processing of Vocal Fold Motion, Vocal-Fold Dynamics, Data Mining

 Información tomada de la entrevista en Google Forms. Elaborado por Ruth Ruiz.

1. ¿Cuál es su nivel de experiencia con trabajos relacionados a web scraping?

Tabla 8.

Nivel de conocimiento de los entrevistados en Web Scraping.

Descripción	Frecuencia de respuesta	Porcentaje
Alto	2	50%
Medio	2	50%
Bajo	0	0%
Nada	0	0%

Información tomada de la entrevista realizada en Google Forms. Elaborado por Ruth Ruiz.

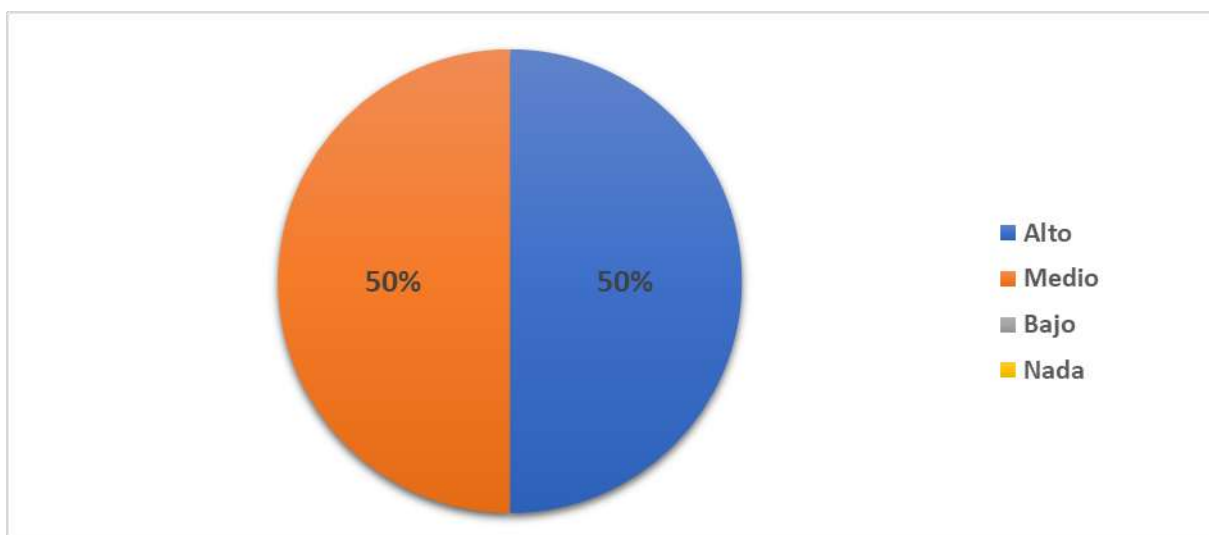


Figura 9. Nivel de conocimiento de los entrevistados en Web Scraping. Información tomada de la entrevista realizada en Google Forms. Elaborado por Ruth Ruiz.

Análisis: De acuerdo con los resultados que se reflejan en la figura 9. se observa que el 50% de los entrevistados poseen un nivel de experiencia alto en trabajos relacionados a Web Scraping mientras que el otro 50% poseen un nivel de experiencia medio. Obtener estas respuestas demuestra que los entrevistados son capaces de ofrecer un punto de vista crítico, objetivo y argumentativo para las preguntas posteriores.

2. ¿Qué lenguaje de programación cree que es adecuado para realizar web scraping?

Tabla 9.

Lenguaje de programación.

Descripción	Frecuencia de respuesta	Porcentaje
Python	3	75%
PHP	0	0%

Ruby	0	0%
C++	1	25%

Información tomada de la entrevista realizada en Google Forms. Elaborado por Ruth Ruiz.

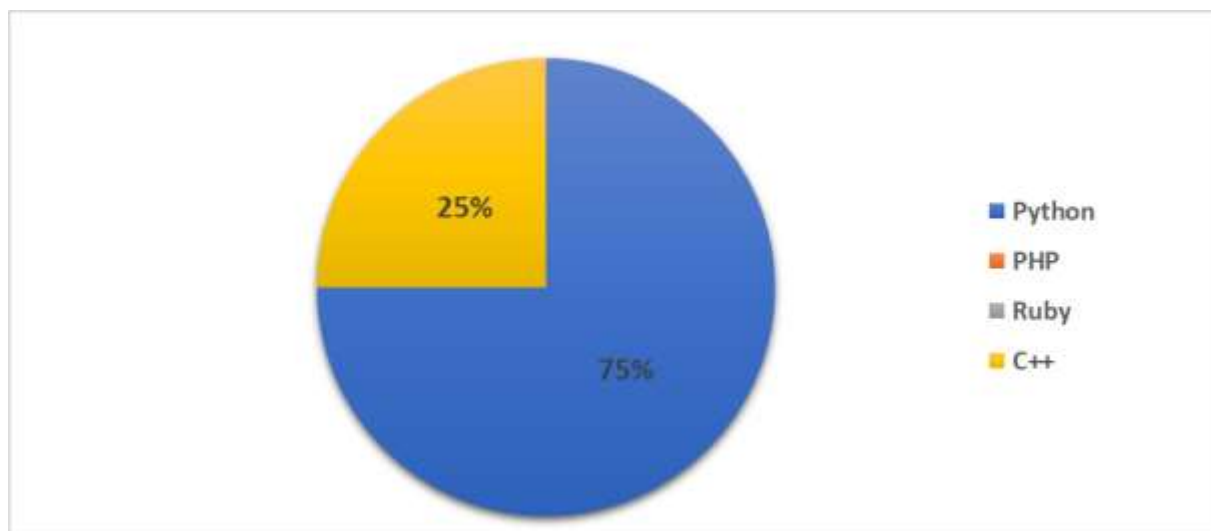


Figura 10. Lenguaje de programación. Información tomada de la entrevista realizada en Google Forms. Elaborado por Ruth Ruiz.

Análisis: De acuerdo con los resultados que se reflejan en la figura 10. Se observa que el 75% de los entrevistados se inclinan por recomendar el uso del lenguaje de programación Python para realizar web scraping mientras que el 25% recomienda C++.

3. ¿Considera ético realizar web scraping en páginas que contengan medidas de seguridad como captcha o robots.txt */disable?

Tabla 10.

La ética del web scraping en páginas con medidas de seguridad antibot.

Descripción	Frecuencia de respuesta	Porcentaje
Sí	2	50%
No	2	50%

Información tomada de la entrevista realizada en Google Forms. Realizado por Ruth Ruiz.

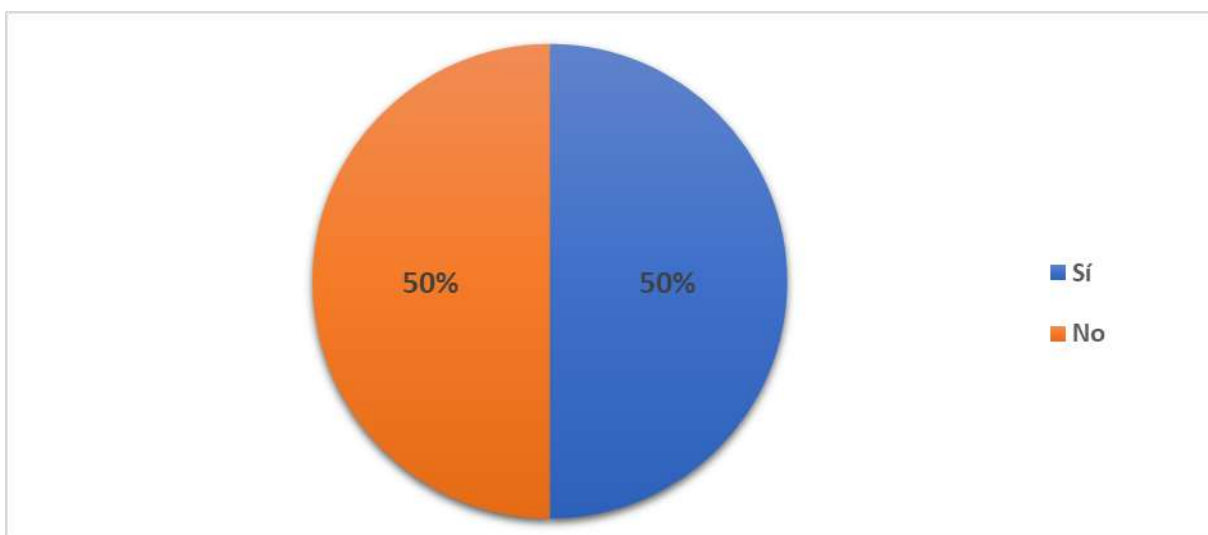


Figura 11. La ética del web scraping en páginas con medidas de seguridad antibot. Información tomada de la entrevista realizada en Google Forms. Elaborado por Ruth Ruiz.

Análisis: De acuerdo con los resultados que se reflejan en la figura 11, se observa que el 50% de los entrevistados considera ético realizar web scraping dentro de una página que contiene medidas de seguridad como captcha y archivo robots.txt */disable mientras que el otro 50% piensa lo contrario.

4. En base a la pregunta anterior, justifique su respuesta.

Tabla 11.

Argumentación de la ética en el web scraping.

Descripción

Sí, los datos son libres y más cuando están en la web.

No, el uso de estos recursos está destinado a ofrecer un nivel de seguridad adicional a un sitio web de ataques automatizados, y es parte del derecho del dueño del portal repeler el uso de aplicaciones que intenten reemplazar las interacciones humanas.

Sí, solamente para uso en la ciencia.

No, uso de recursos en la web

Información tomada de la entrevista realizada en Google Forms. Elaborado por Ruth Ruiz.

Análisis: El 50% de los entrevistados que respondieron SÍ en la pregunta anterior, justifican que ejecutar un web scraper dentro de una página web que contiene medidas de seguridad es totalmente ético debido a que los datos son libres, sobre todo cuando están disponibles en la web, además, se destaca que es una acción totalmente valida si se realiza para el uso de la

ciencia. El otro 50% de los entrevistados que respondieron NO en la pregunta anterior, justifican que no es ético realizar web scraping en una página que contiene medidas de seguridad, sí el dueño del portal web desea repeler las acciones que intentan reemplazar las interacciones humanas está en todo su derecho al igual que si desea evitar el consumo masivo de los recursos del sitio web.

5. ¿Cuál es su nivel de experiencia con trabajos relacionados a Dashboard?

Tabla 12.

Nivel de conocimiento de los entrevistados en Dashboard.

Descripción	Frecuencia de respuesta	Porcentaje
Alto	3	75%
Medio	1	25%
Bajo	0	0%
Nada	0	0%

Información tomada de la entrevista realizada en Google Forms. Elaborado por Ruth Ruiz.

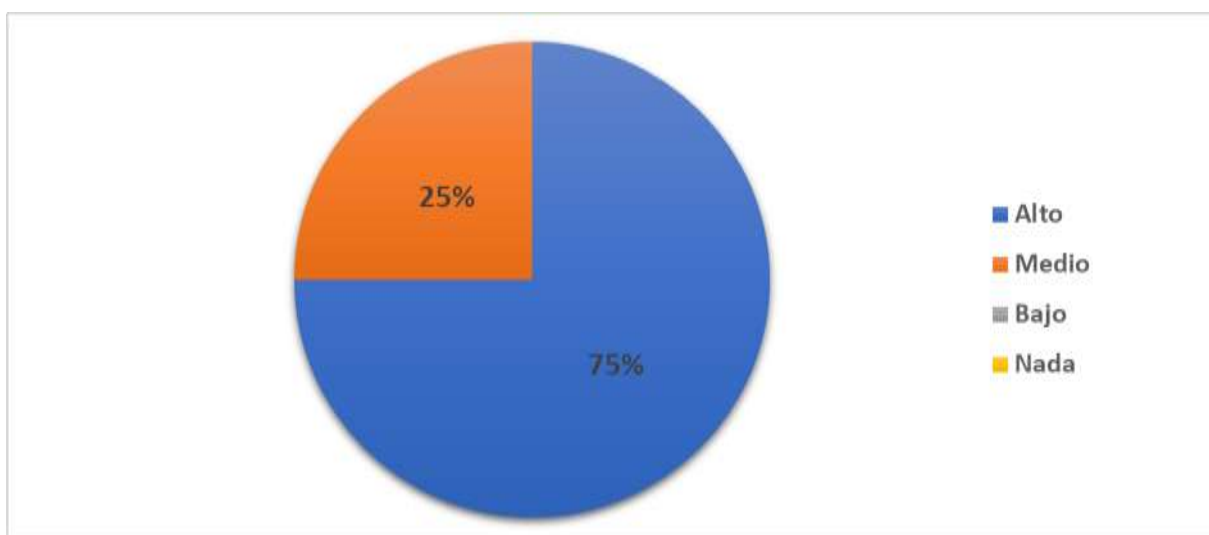


Figura 12. Nivel de conocimiento de los entrevistados en Dashboard. Información tomada de la entrevista realizada en Google Forms. Elaborado por Ruth Ruiz.

Análisis: De acuerdo con los resultados que se reflejan en la figura 12. se observa que el 75% de los entrevistados poseen un nivel de experiencia alto en trabajos relacionados a Dashboard mientras que el 25% posee un nivel de experiencia medio. Obtener estas respuestas demuestra que los entrevistados son capaces de ofrecer un punto de vista crítico, objetivo y argumentativo para las preguntas posteriores.

6. Según su opinión, ¿Considera que una base de datos construida con información pública de emprendimientos integrada en un dashboard aportaría con información de utilidad para los microemprendedores?

Tabla 13.

Opinión de expertos sobre el aporte del dashboard para los microemprendedores.

Descripción	Frecuencia de respuesta	Porcentaje
Sí	4	100%
No	0	0%

Información tomada de la entrevista realizada en Google Forms. Elaborado por Ruth Ruiz.

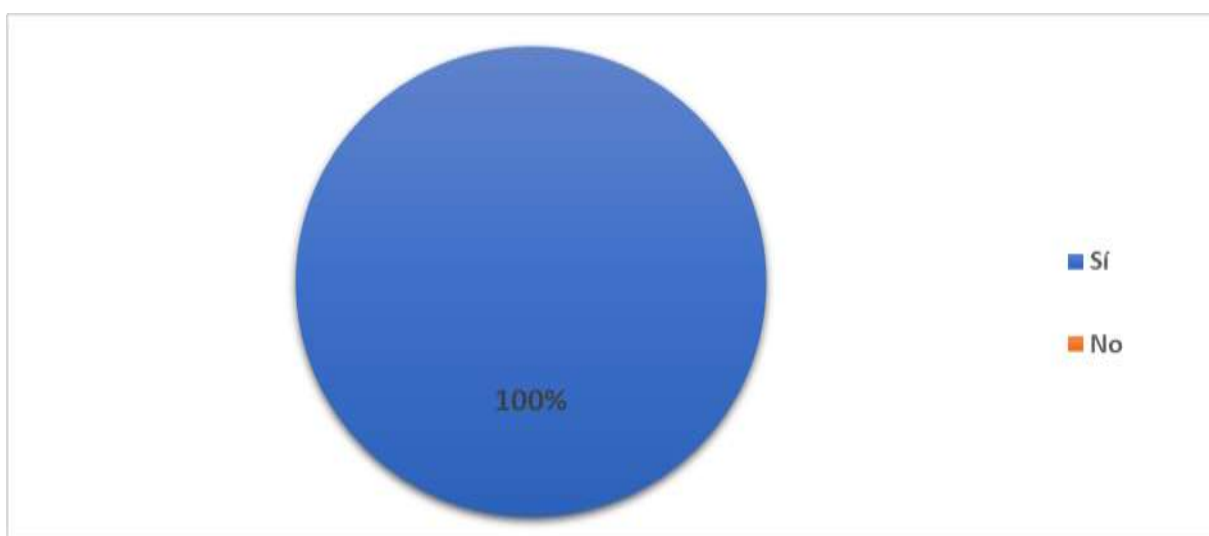


Figura 13. Opinión de expertos sobre el aporte del dashboard para los microemprendedores. Información tomada de la entrevista realizada en Google Forms. Elaborado por Ruth Ruiz.

Análisis: De acuerdo con los resultados que se reflejan en la figura 13. se observa que el 100% de los entrevistados consideran que construir un Dashboard con información pública de emprendimientos aportaría con información de utilidad para los microemprendedores.

7. ¿Qué herramienta web gratuita considera apropiada para desarrollar un dashboard?

Tabla 14.

Herramienta gratuita para Dashboard.

Descripción	Frecuencia de respuesta	Porcentaje
Power Bi	2	50%
Tableau Public	2	50%
Inetsoft	0	0%

Syncfusion	0	0%
------------	---	----

Información tomada de la entrevista realizada en Google Forms. Elaborado por Ruth Ruiz.

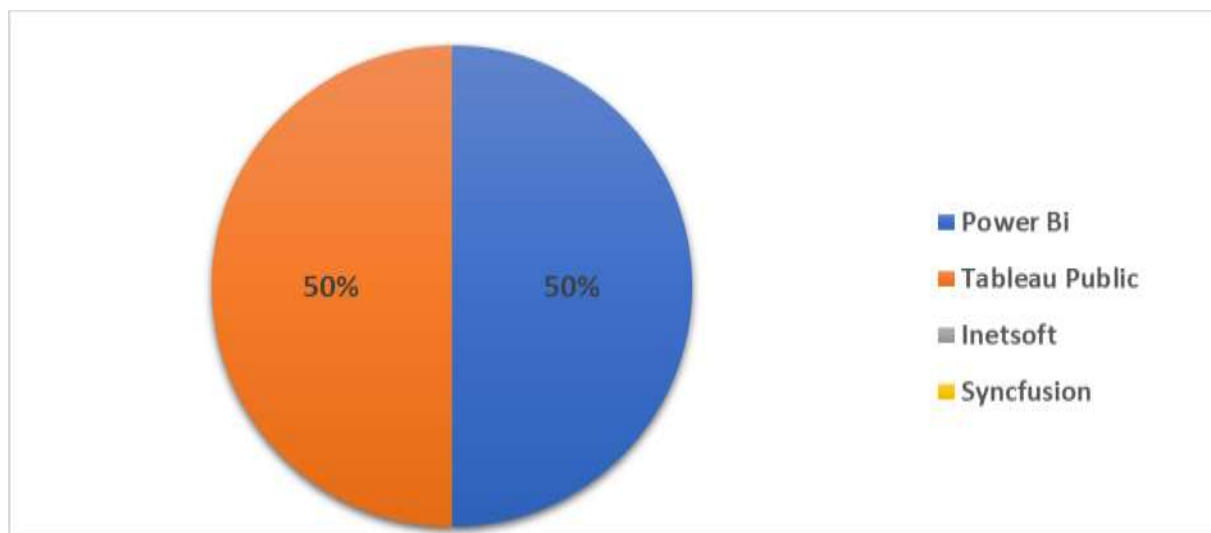


Figura 14. Herramienta gratuita para dashboard. Información tomada de la entrevista realizada en Google Forms. Elaborado por Ruth Ruiz.

Análisis: De acuerdo con los resultados que se reflejan en la figura 14. se observa que el 50% de los entrevistados se inclinan por recomendar Power Bi como herramienta web gratuita para el desarrollo de un Dashboard mientras que el otro 50% recomienda Tableau Public.

8. ¿Cuáles son los tipos de gráficos estadísticos que considera estarían mejor ambientados para presentar datos extraídos de la web de un directorio de emprendimientos?

Tabla 15.

Tipos de gráficos estadísticos para dashboard.

Descripción	Frecuencia de respuesta	Porcentaje
Gráfico de barras	2	50%
Gráfico circular	2	50%
Gráfico de línea	3	75%
Cartograma	2	50%

Información tomada de la entrevista realiza en Google Forms. Elaborado por Ruth Ruiz.

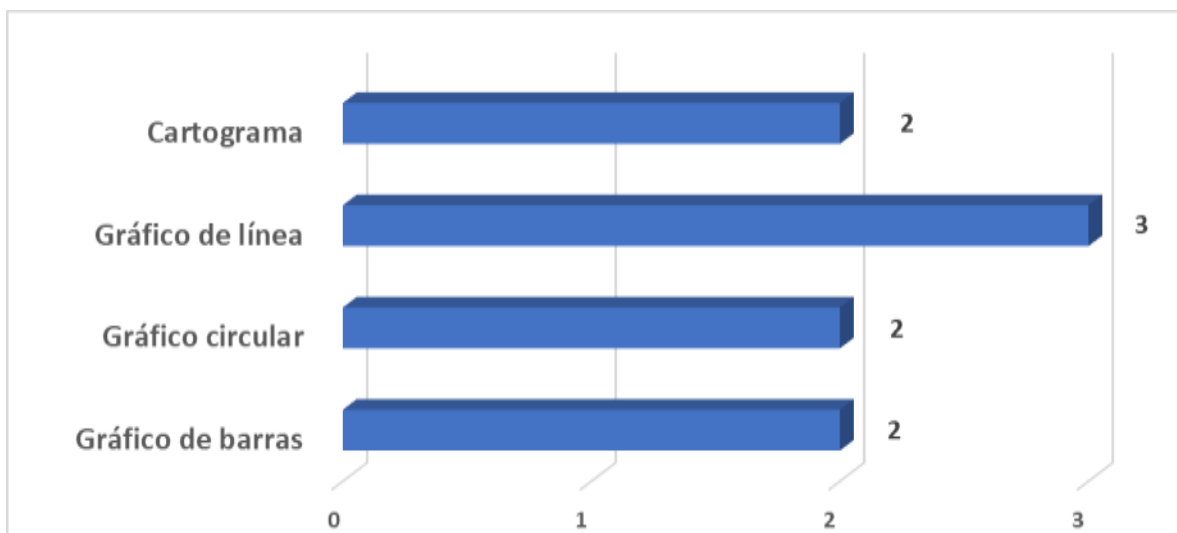


Figura 15. Tipos de gráficos estadísticos para Dashboard. Información tomada de la entrevista realizada en Google Forms. Elaborado por Ruth Ruiz.

Análisis: De acuerdo con los resultados que se reflejan en la figura 15. se observa que el gráfico de línea, según el 75% de los entrevistados, es el mejor ambientado para presentar datos extraídos de un directorio web de emprendimientos, mientras que el cartograma, gráfico circular y gráfico de barras cuentan cada uno de ellos con el respaldo del 50% de los entrevistados.

3.2. Metodología a desarrollar

Con relación a la investigación realizada acerca de las metodologías de minería de datos, en un principio se consideró utilizar la metodología KDD, que es uno de los primeros ejemplares para el desarrollo de proyectos relacionados con minería de datos que tiene como propósito dar sentido a los datos a través de técnicas y métodos. Para completar la metodología hay que cumplir con 9 fases, que van desde la comprensión del dominio de aplicación, determinación de técnica y algoritmos de minería hasta la interpretación y utilización del nuevo conocimiento, etapas en particular que de acuerdo con los objetivos planteados carecen de relevancia, sin embargo, la investigación de las metodologías sirvió como referencia, al igual que el marco de trabajo de web scraping mencionado en el marco teórico, para proponer una metodología propia que se acople adecuadamente a la integración de los objetivos y a las necesidades del proyecto de investigación, por lo tanto, las etapas quedarían de la siguiente manera:

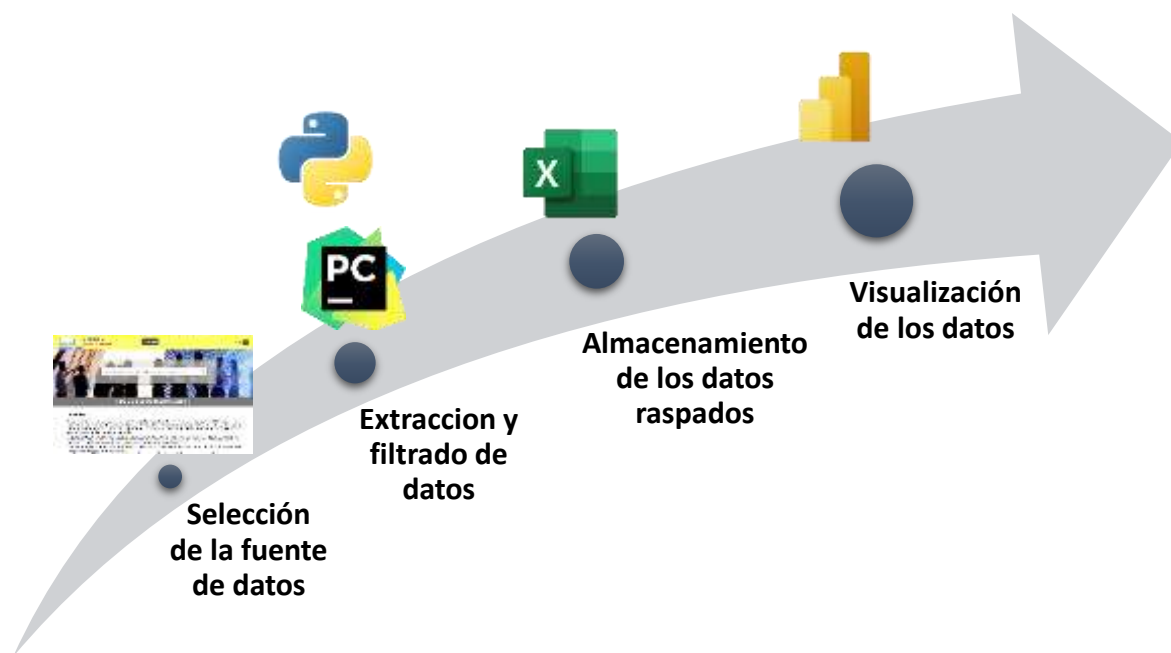


Figura 16. Esquema general de la metodología. Elaborado por Ruth Ruiz.

3.2.1. Factibilidad técnica

Para llevar a cabo la metodología propuesta se escogió un conjunto de herramientas necesarias en el desarrollo del proyecto que van desde la fase Extracción y filtrado de datos hasta la Visualización de los datos. El lenguaje de programación Python mediante el IDE Pycharm permitirá recopilar los datos de la web, los archivos .CSV leídos mediante Excel constituirán la herramienta de almacenamiento de datos y Power Bi constituirá la herramienta de visualización.

A continuación, se especifican los factores técnicos a considerar de las herramientas y el recurso que se utilizará para la ejecución del proyecto.

Hardware	Características
Notebook Asus Q405UA	Windows 10 Home
	Procesador Intel® core™ i5-8350u @
	1.60Ghz (8 CPUs).
	8 GB RAM
	500 GB SDD

Software	Requisitos	Versión
Python	SO: Linux y Windows 7, 8, 8.1 y 10. Mayor a 4GB RAM. Versiones de 64 bits de Microsoft Windows 10,8. Min: 2 GB RAM; Recomendado: 8 GB RAM.	3.9.6
Pycharm	2,5 GB de espacio en disco duro, SDD recomendado. 1024x768 mínimo de resolución de pantalla Python 2.7, Python 3.5 o más	Community 2021.2
Excel	Windows, paquete de Office Versión de 32 o 64 bits en Windows 7/10/ Server 2008 R2 o posterior.	2016
Power Bi Desktop	1 GB RAM. Procesador 1Ghz or faster Resolución de pantalla 1440*900	2.96.1061.0

3.2.2. Selección de la fuente de datos

En la actualidad, existe un número limitado de sitios web con información relacionada al emprendimiento o microemprendimiento, muchos de los sitios identificados a pesar de ser entidades públicas (leer Art. 18 de la Constitución de la República del Ecuador) entorpecen el proceso de obtención de información con la imposición de condicionales de acceso como: una consulta a la vez con RUC a la mano, medidas de seguridad con antibot para el monitoreo y detección de tráfico inusual y bloqueo de IP. Para la primera fase del esquema general de la metodología que consiste en la selección de fuentes de datos, tomando como referencia la regla a considerar para ejecutar web scraping de (Mitchell, 2013) en fundamentación legal, solo se efectuará en aquellos sitios en donde las políticas de navegación, términos y condiciones lo permitan.

- **SRI en línea**



Figura 17. Medidas de seguridad en SRI para consultar de RUC. Información tomada del SRI en línea. Elaborado por el SRI.

El sitio web SRI en línea, contiene servicios para contribuyentes y para todo tipo de usuarios en general, en el cual, se puede gestionar por medio del RUC información detallada y actualizada del emprendimiento.

Las desventajas del sitio web del SRI en línea es que para realizar un raspado web se debe contar con los RUC de los emprendimientos, además, cuenta con captcha para impedir que un bot realice varias consultas.

- **Superintendencia de Compañías, Valores y Seguros**

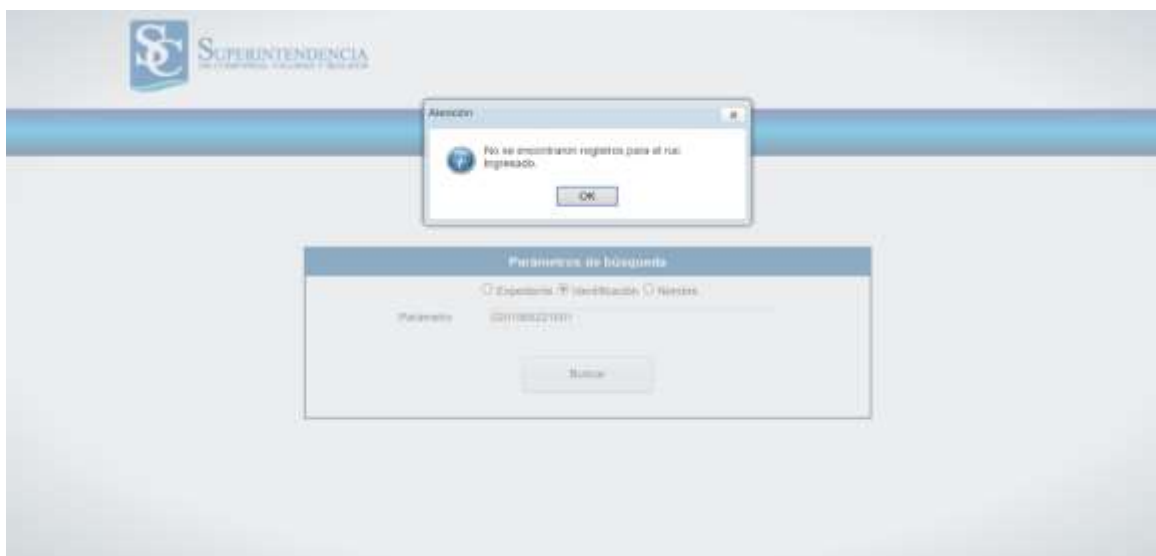


Figura 18. RUC consultado no encontrado. Información tomada de la Superintendencia de compañías, valores y seguros. Elaborado por la SUPERCIA.

Este sitio web corresponde a la Superintendencia de Compañías, Valores y Seguros, a través del portal web portal.supercias.gob.ec ofrece varios servicios como tramites en línea, portal de documentos, publicaciones y resoluciones, investigaciones y estudios, portal de información.

En este último servicio, se puede conseguir información únicamente de sociedades privadas que se definen como personas jurídicas de derecho privado, entre ellas se encuentran compañías anónimas, de responsabilidad limitada, de economía mixta, entre otras (Facturas Rápidas Ecuador, 2020). Además, el sitio web contiene medidas de seguridad como CAPTCHA y bloqueo de IP.

- **Directorio de emprendimientos**

Figura 19. Página de emprendimientos. Información tomada del sitio web Directorio de emprendimiento. Elaborado por Directorio de emprendimientos

Este portal web facilita la búsqueda de emprendimientos tanto de personas naturales como jurídicas con capacidad legal para realizar transacciones comerciales, permite agregar a los usuarios registrados información referente a productos y/o servicios que ofertan, posibilitando a los emprendedores contar con un espacio virtual para exponer su negocio. La página determina con claridad que toda información adicional registrada a través del portal será de acceso público, además de actualizarse constantemente.

En la raíz del dominio del sitio web, se encuentra un archivo robots.txt de estado /disable para un listado específico de bots pertenecientes a compañías como SEMrush, Blexbot, Ahrefs, Dotmic DotBot, Majestic-12, la mayoría de los bots que diseñan estas empresas son de rastreo con fines de Marketing o campañas publicitarias y bots de indexación de sitios web de comercio electrónico.

El Directorio de emprendimientos cuenta con certificación digital acreditado por la autoridad certificante Let's Encryptes que ofrece certificados gratuitos con la finalidad de crear una web más segura, por lo tanto, el sitio web es óptimo para realizar web scraping, por establecer una conexión segura, permitir la ejecución de bots y declarar al portal web de uso público.

3.2.3. Extracción y filtrado de datos

3.2.3.1. Inspeccionar los elementos del lenguaje de marcado

El primer paso consiste en inspeccionar y comprender la estructura del portal web Directorio de Emprendimiento para identificar los elementos a extraer y encontrar patrones dentro del lenguaje de marcado. Los tipos de elementos básicos de un archivo .html son: Etiquetas; Atributos y NavigableString que sirven para encerrar diferentes partes del contenido para que se perciban y comporten de distintas maneras. En la figura 20. está el html de la página de inicio de Directorio de emprendimientos sri-en-linea.com/emprendimientos/

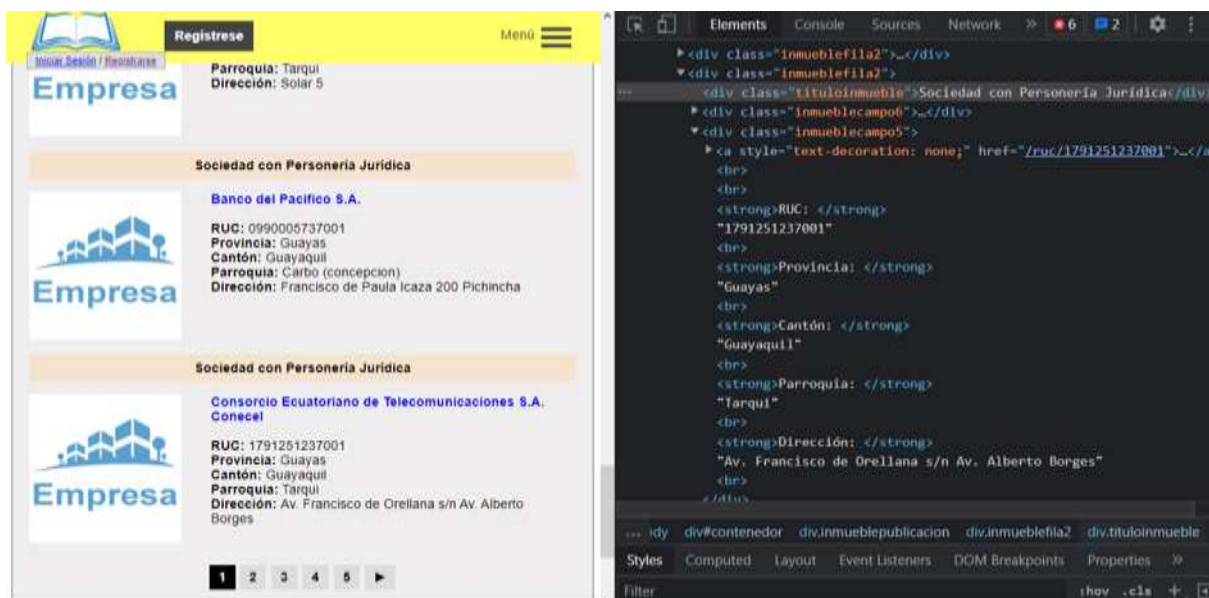


Figura 20. Parte del HTML de la página de la sección de Emprendimientos. Información tomada de Directorio de Emprendimientos. Elaborado por Directorio de Emprendimiento.

Dentro del sitio web Directorio de emprendimiento sección /emprendimiento existe un total de 357191 páginas habilitadas, cada una de las páginas presenta información resumida como RUC, provincia, cantón, parroquia y dirección de 15 emprendimientos. Estos pequeños recuadros que se ven en la figura 20. comparten en común una misma clase llamada `class="inmueblecampo5"` contenida dentro de una etiqueta `<div>`, por lo tanto, este será el primer elemento de interés que se incluirá dentro del código de scraping.

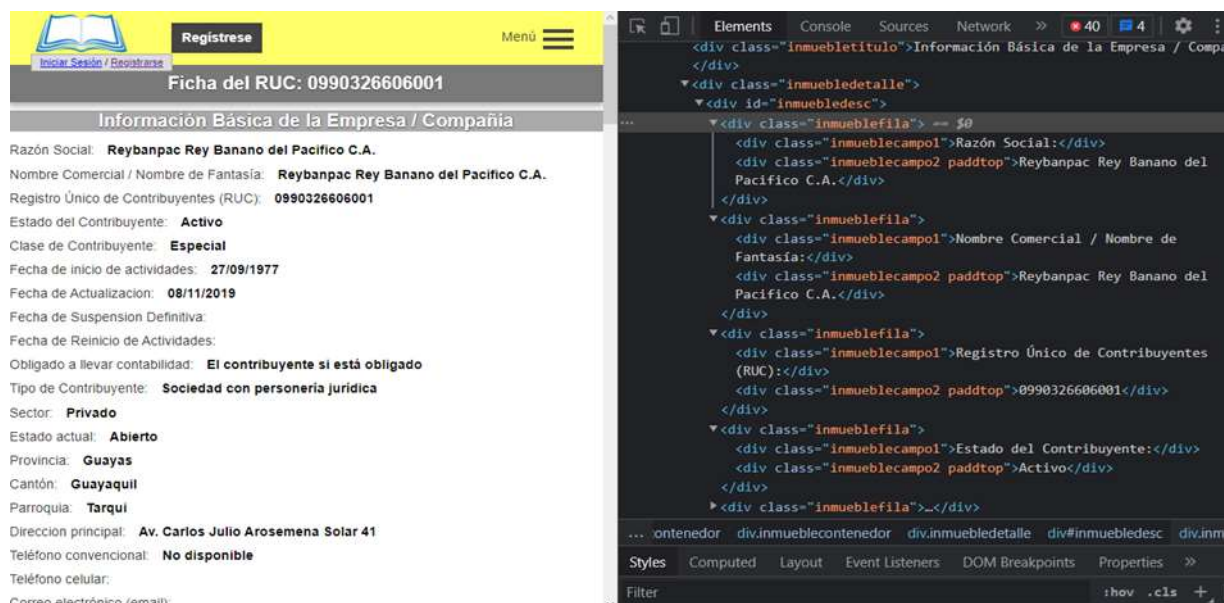


Figura 21. Parte del HTML de la página de un emprendimiento. Información tomada del Directorio de Emprendimientos. Elaborado por el Directorio de Emprendimientos.

La figura 21. muestra como luce la página de un emprendimiento. En esta página el elemento de interés es `class="inmueblecampo2 paddtop"` que se encuentra encerrada en la etiqueta `<div>`, esta es la clase en común entre cada línea de la página y contiene los datos del emprendimiento.

3.2.3.2. Extracción de los datos

En este punto se utiliza el lenguaje de programación Python para rastrear el contenido del portal web, en un principio se consideró utilizar alguna aplicación de escritorio o extensión web como Octoparse o Webscraper.io pero debido a las limitaciones del plan gratuito como: número de páginas limitada para raspar datos, funcionalidades orientadas más hacia el contexto de scraper (raspado en una sola página) y no de crawler (recolección de enlaces del sitio web y raspado en cada enlace) y una comunidad de soporte pobre, se optó por el lenguaje mencionado sugerido por los profesionales entrevistados.

Python se caracteriza por ser una herramienta poderosa para utilizar la técnica Web Scraping, las librerías BeautifulSoup y Request son la dupla perfecta para recopilar datos eficazmente, estas y otras más como Pandas que sirve para almacenar los datos raspados serán las protagonistas del script a desarrollar.

Una vez inspeccionado la estructura del lenguaje de marcado del sitio web para encontrar los elementos de interés y los patrones dentro de la página, se procede a desarrollar el código de scraping, para ello, se utiliza Pycharm como entorno de desarrollo integrado de código abierto que ofrece compatibilidad de primer nivel con el lenguaje de programación Python e integra funcionalidades de editor de código inteligente para detectar y corregir código sobre la marcha.

Primero se debe instalar Python, se descargó la última versión del programa desde la página oficial www.python.org/downloads/, no fue necesario instalar el gestor de librerías pip porque en la versión que se descargó ya viene instalado. Luego se instaló Pycharm, también se puede descargar desde la página oficial www.jetbrains.com/es-es/pycharm/ y cuenta con dos versiones, Professional y Community. Para la creación y ejecución del código la versión Community es suficiente.

Antes de proceder a desarrollar el código, se debe descargar las librerías BeautifulSoup, Requests y Pandas. En la terminal de Pycharm se insertaron los siguientes comandos:

- `pip install beautifulsoup4`
- `pip install requests`
- `pip install pandas`

Beautifulsoup4 es una potente librería de Python capaz de extraer información típicamente de archivos XML o HTML mediante el análisis del árbol de elementos con métodos como `html.parser`, `lxml.parser` y `html5lib.parser`.

Requests es la librería HTTP de Python con cualidades simples y elegantes que permite hacer peticiones HTTP a un servidor web a través de métodos como `GET()` o `POST()`.

Pandas es una librería de Python para el análisis de datos que proporciona una estructura flexible a los datos para que se puedan manipular eficientemente.

3.2.3.3. Código

En esta parte se presenta el script desarrollado que permite extraer los datos deseados. Para extraer los datos que se necesitan, el código se dividirá en dos secciones, en la primera sección, se encargará de conseguir desde el portal web Directorio de Emprendimiento las url's

que pertenecen solamente a emprendimientos de Guayaquil (lo que se conoce como crawler) y, la otra sección, se encargará de conseguir los datos de cada emprendimiento de Guayaquil (lo que se conoce como scraper).

En la figura 22. se muestran las librerías importadas que se utilizaron para la primera sección del código.

```
1 import pandas as pd
2 from bs4 import BeautifulSoup
3 import requests
```

Figura 22. Librerías importadas para extracción de url. Información tomada de Pycharm. Elaborado por Ruth Ruiz.

Como se mencionó en el apartado **Inspeccionar los elementos del lenguaje de marcado**, dentro del portal web Directorio de emprendimientos sri-en-linea.com/emprendimientos existe un total de 357191 páginas, cada una de estas páginas muestra información de 15 emprendimientos. Según el análisis html (figura 20) la clase `class="inmueblecampo5"` que encierra la información resumida de cada emprendimiento, contiene un elemento en particular identificado como `<a>`, a es una etiqueta html que permite manejar hiperenlaces dentro de una página web, al mismo tiempo, la clase `class="inmueblecampo5"` contiene un dato importante que permite filtrar la selección de url's, ese campo se encuentra contenido en una de las etiquetas `` de la clase y se llama cantón. A continuación, se adjunta la captura del código.

```
def links_gye (pagina):
    url_raiz = "https://sri-en-linea.com/emprendimientos/"
    respuesta = requests.get(url_raiz+str(pagina))
    if respuesta.status_code == 200:
        extractor = BeautifulSoup(respuesta.content, "html.parser")
        datosclase = extractor.find_all("div", class_="inmueblecampo5")
        saveurl = list()
        for i in datosclase:
            busqueda = i.find_all("strong")
            canton = busqueda[i].nextSibling
            if canton == "Guayaquil":
                busqueda2 = i.find("a", href=True)
                url_gye = "https://sri-en-linea.com"+busqueda2['href']
                saveurl.append(url_gye)
        return saveurl
```

Figura 23. Función para extraer url's de varias páginas. Información tomada de Python. Elaborado por Ruth Ruiz.

Se crea una función llamada "links_gye" para agrupar el conjunto de instrucciones que constituyen la lógica del web crawler y así encapsular esta parte del código que será reutilizada varias veces, ya que tendrá que repetirse conforme a la cantidad de páginas (en este caso son 357191 páginas).

Se tomó como url a <https://sri-en-linea.com/emprendimientos/> puesto que, al navegar entre paginas mediante los botones inferiores de esta, se logró identificar lo poco cambiante que es; en la url solo se agrega al final el número de página cliqueado, es decir, si quiero ir a la página 2 desde los botones CSS de la página, el url lucirá así <https://sri-en-linea.com/emprendimientos/2>.

La parte clave de esta sección de código está en la línea 7 y 9, en donde Requests mediante la petición HTTP .GET() solicita el recurso web y BeautifulSoup mediante el analizador html.parser interactúa con el lenguaje de marcado (solo si el recurso está disponible, estado OK) y extrae la estructura HTML de la página para convertirla en un árbol de objeto Python que se pueda manipular libremente. Con el uso de métodos como find() o find_all() se consigue filtrar de manera sencilla el archivo HTML a través de los elementos del código para encontrar una etiqueta o el grupo de etiquetas que se necesiten.

La propiedad NextSibling de BeautifulSoup se usa para retornar el nodo siguiente, este valor se almacena en la variable *cantón*, el filtro para obtener las url's que contenga información de emprendimientos de Guayaquil se realiza a partir de la línea 15. Mediante el "if" se permite seleccionar que etiquetas <a> pueden agregar el valor href a la lista saveurl() con el método append(), para añadir ítems a la lista; esta lista es la palabra reservada de retorno de la función.

```

21 enlacesgye=list()
22 p_min = 1
23 p_max = 89280
24 for e in range(p_min,p_max):
25     enlacesgye.extend(links_gye(e))
26 print(enlacesgye)

```

Figura 24. Presentación de URL filtradas. Información tomada de Pycharm. Elaborado por Ruth Ruiz.

Por otra parte, en la figura 24. se incluye el ciclo for que iterará en un rango determinado (que representa a las páginas por recorrer) utilizando el contador (e) como argumento de la función links_gye para conseguir simular la acción de cambiar de página en página y escoger las url's que pertenecen a los emprendimientos del cantón Guayaquil. Debido a la gran cantidad de páginas que se tenía que recorrer y al tener como recurso de procesamiento un equipo que no está diseñado para trabajar de forma continua, como lo es una computadora personal, se dividió la carga de trabajo en cuatro partes del total de páginas, es decir, se tomaron 4 días para recolectar los datos; los 3 primeros días se recorrieron 89280 y el ultimo día 89351, lo que da un total de 357191 páginas, la lista enlacesgye() utiliza el método extend() que une una lista con


```

with open('guayana99279.csv') as File:
    reader = csv.DictReader(File)
    razon_social = list()
    nombre_comercial = list()
    ruc = list()
    edc = list()
    cdc = list()
    finicio = list()
    factualizaciones = list()
    feusdefinitiva = list()
    freinicio = list()
    obligadocant = list()
    tipocontribuyente = list()
    estactual = list()
    parroquia = list()
    direccion = list()
    codigociu = list()
    for row in reader:
        valores = list(row.values())
        cambiartipo = str(valores)
        url=cambiartipo[2:44]
        page = requests.get(url)
        sopa = BeautifulSoup(page.content, 'html.parser')
        empregyes = sopa.find_all('div', class_='inmueblecampo? paddtop')
        razon_social.append(empregye[0].text)
        nombre_comercial.append(empregye[1].text)
        ruc.append(empregye[2].text)
        edc.append(empregye[3].text)
        cdc.append(empregye[4].text)
        finicio.append(empregye[5].text)
        factuizacion.append(empregye[6].text)
        feusdefinitiva.append(empregye[7].text)
        freinicio.append(empregye[8].text)
        obligadocant.append(empregye[9].text)
        tipocontribuyente.append(empregye[10].text)
        estactual.append(empregye[12].text)
        parroquia.append(empregye[15].text)
        direccion.append(empregye[16].text)
        codigociu.append(empregye[20].text)

```

Figura 27. Extracción de información de emprendimientos de Guayaquil. Información tomada de Pycharm. Elaborado por Ruth Ruiz.

Con la sentencia **with ... as ...** que contiene como expresión la función básica de Python, Open(), se abre el archivo .csv que almacena los datos impresos de la figura 25. (en la etapa de **Almacenamiento de datos raspados** se explicará cómo se realiza este paso) a través del objeto File devuelto por la expresión que será utilizado por la clase csv.DictReader encargada de leer la información en un formato de diccionario (compuesto por una clave y un valor, 'key' : 'valor') cuya clave será la primera fila que se encuentra en el .csv. Hasta este punto los datos contenidos en la variable **reader** son de tipo 'csv.DictReader', para realizar peticiones HTTP, el dato a insertar debe ser de tipo string, por lo tanto, con el uso del método de diccionario .values() se retorna el valor del diccionario, sin embargo, con esto no basta para obtener el string.


```

C:\Users\Ruth_Roxana\PycharmProjects\pythonProject\venv\Scripts\python.exe "C:/Users/Ruth_Roxana/PycharmProjects/pythonProject/scrapper de emprendimiento.py"
<class 'dict_values'>
dict_values(['https://api-mg-limon.com/ruc/0997244835001'])
<class 'dict_values'>
dict_values(['https://api-mg-limon.com/ruc/1798978648001'])
<class 'dict_values'>
dict_values(['https://api-mg-limon.com/ruc/09989806419001'])

```

Figura 28. Error de la clase dictReader. Información tomada de Pycharm. Elaborado por Ruth Ruiz.

Como se aprecia en la figura 28. la salida del contenido de la variable valores no contiene puramente datos de tipo string. Para solucionar esto, y tomando en cuenta que la cantidad de caracteres de cada enlace no cambia de 42, se decidió aprovechar la característica particular de las cadenas de texto que permite tratarlas como lista de solo lectura, por lo tanto, se transformó el tipo de dato de la variable valores a string y luego mediante la creación de otra variable que se le asigna el nombre de url se obtiene el substring utilizando la notación de porciones cadena_de_caracteres[posición_inicial:posición_final].

Con este último paso realizado, es posible solicitar el recurso web del enlace de entrada con el método .GET(). Posteriormente la variable sopa extraerá y analizará mediante el parser.html el lenguaje de marcado para obtener todas las etiquetas <div> que encierren la clase class='inmueblecampo 2 paddtop' por contener los datos del emprendimiento. Existen 21 elementos dentro del HTML de la página compartiendo la misma clase, para obtener el texto de las líneas de interés se especifica la posición de la línea seguida del .text con la finalidad de conseguir el valor y almacenarlas en listas diferentes como se muestra en la figura 26. Dentro del entorno de Pycharm los datos se visualizan como se muestra en la figura 29. (la presentación de los datos viene acotada por defecto).

```

C:\Users\Ruth_Roxana\PycharmProjects\pythonProject\venv\Scripts\python.exe "C:/Users/Ruth_Roxana/PycharmProjects/pythonProject/scrapper de emprendimiento.py"

```

	RAZON_SOCIAL	...	COUIGO_CIIU
0	Tiendas Industriales Asociadas Tia S. A.	...	0471102
1	Corporacion El Rosado S.A.	...	0471102
2	Keybanpac Key Banana del Pacifico C.A.	...	A032101
3	Industrial Pesquera Santa Priscila S.A.	...	A032102
4	Operadora y Procesadora de Productos Marinos O...	...	A032102
5	Negocios Industriales Real N.I.R.S.A. S.A.	...	C102002
6	Junta de Beneficencia de Guayaquil	...	0061803
7	Distribuidora Farmaceutica Ecuatoriana Difare	...	0464922
8	Banco del Pacifico S.A.	...	0641901
9	Consorcio Ecuatoriano de Telecomunicaciones S...	...	3612001
10	Agroazucar Ecuador S.A.	...	C107202
11	Corporacion Agricola San Juan S.a Casjoca	...	A011900

Figura 29. Imprimir datos de emprendimientos. Información tomada de Pycharm. Elaborado por Ruth Ruiz.

Este proceso se repite 4 veces y corresponde a la cantidad de partes en la que se dividió el raspado de datos.

3.2.4. Almacenamiento de los datos raspados

Como se mencionó en el apartado **Código**, se realizó 2 scripts por separado, uno bajo el contexto de crawler que permite recopilar las url's de emprendimientos de Guayaquil y el otro bajo el contexto de scraper que utiliza las url's recopiladas por el primer script para extraer los datos de cada emprendimiento.

Los datos raspados deben ser almacenados en algún lugar para usos posteriores, con ese fin, se utilizarán archivos .CSV capaces de representar de manera sencilla datos en formato de tabla. Mediante separadores como por ejemplo ',' ó ';' se crean columnas y con saltos de línea se crean filas.

La figura 30. muestra el fragmento de código utilizado para crear el .CSV que contendrá los hiperenlaces del cantón Guayaquil.

```

37 df = pd.DataFrame({'Enlaces de Gye': enlacesgye})
38 nombre = 'pagina'+str(p_max-1)+'.csv'
39 df.to_csv(nombre, index=False)

```

Figura 30. Código para almacenar las url's en .CSV. Información tomada de Pycharm. Elaborado por Ruth Ruiz.

Con la ayuda de Pandas (pd es el sobrenombre de la librería para trabajar dentro del script) los datos raspados pueden obtener un formato de datos tabular y con la función DataFrame es posible estructurar estos datos en 2 dimensiones (filas y columnas) y guardarlos sin importar el tipo de dato ni el tamaño de la variable que lo contenga.

Para crear el DataFrame se debe especificar el nombre del campo (columna) y la variable que incorpora el registro (fila), en este caso se le asignó al nombre del campo 'Enlace de Gye' y como registro al listado enlacegye() (mirar figura 24). Con la función .to_csv se crea el archivo .CSV que almacenará los valores del DataFrame y se deshabilita en este caso la generación de índices con index=False por ser innecesarios.

```

37 df = pd.DataFrame({'RAZON_SOCIAL':razon_social, 'NOMBRE_COMERCIAL':nombre_comercial, 'NUMERO_RUC':ruc,
38                  'ESTADO_CONTRIBUYENTE':edc, 'CLASE_CONTRIBUYENTE':cdc, 'FECHA_INICIO_ACTIVIDADES':finicio,
39                  'FECHA_ACTUALIZACION':factualizacion, 'FECHA_SUSPENSION_DEFINITIVA':fsusdefinitiva,
40                  'FECHA_REINICIO_ACTIVIDADES':freinicio, 'OBLIGADO':obligadocont,
41                  'TIPO_CONTRIBUYENTE':tipocontribuyente, 'ESTADO_ACTUAL':estactual,
42                  'DESCRIPCION_PARRQUIA': parroquia, 'DIRECCION': direccion, 'CODIGO_CIUD':codigociu})
43 df.to_csv('Directorio de emprendimiento 1.csv', index=False, sep=';', encoding='utf-8-sig')

```

Figura 31. Código para almacenar los datos de cada emprendimiento. Información tomada de Pycharm. Elaborado por Ruth Ruiz.

La figura 31. utiliza la misma lógica de la figura 30. para crear el .CSV que almacenará los datos de cada emprendimiento del cantón Guayaquil, pero, con más campos y registros, y con nuevos parámetros dentro de la función `.to_csv` como `sep` y `encoding`. El parámetro `sep` es utilizado para indicar un tipo particular de separador de columnas, en este caso, se insertó el “;” para prevenir errores en la separación de columnas al momento de leer los archivos .CSV en Excel, debido a que, cuando se trabaja con valores que pueden contener distintos tipos de datos la probabilidad de que estos valores contengan “,” es alta, por ello, se cambia el separador por uno que no es común entre los datos que se van a guardar.

El otro parámetro es `encoding`, este parámetro permite especificar la codificación con la que se almacenará el archivo creado, en este caso se codifican los datos en UTF-8-sig (sig es la abreviatura de firma que sirve para indicar que el texto del archivo es Unicode) lo que permite codificar caracteres acentuados como las tildes, virgulillas o diéresis comunes en el lenguaje español.

Este proceso se repite 4 veces que corresponde a la cantidad de partes en la que se dividió el raspado de datos.

Una vez completado el proceso, se unen los 4 archivos .CSV que contienen la información de cada emprendimiento para luego filtrarlos desde el año 2020, dentro de la carpeta de proyecto se muestran como aparece en la figura 32. Los archivos con extensión `.csv` pueden leerse en Microsoft Excel y ser editados, la convergencia de estos archivos en Excel queda como se ve en la figura 33.

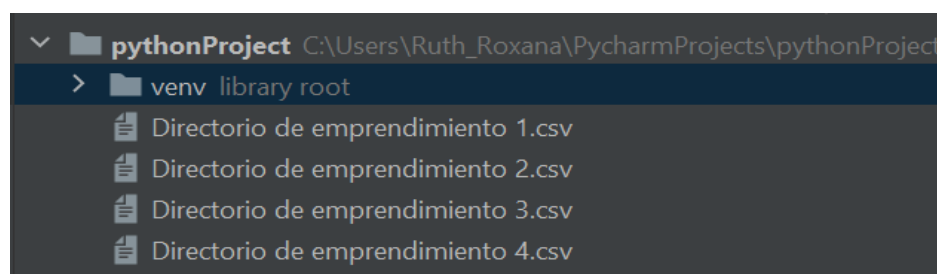


Figura 32. Carpeta *PycharmPrjects* con archivos CSV. Información tomada de *Pycharm*. Elaborado por Ruth Ruiz.

The screenshot shows an Excel spreadsheet with a table containing personal and company data. The columns include: RAZON SOCIAL, NOMBRE, COMPLENOMBRE, ESTADO, CONTI, CLASE, CORREO, FECHA, FECHA ACTUAL, FECHA SUSP, FECHA REINICIO, TIPO, CONTRIBUYENTE, ESTADO, DESCRIPCION, DIRECCION, and CODIGO CIIU. The CODIGO CIIU column contains various 6-digit codes representing economic activities.

Figura 33. Archivos CSV en Excel. Información tomada de Excel. Elaborado por Ruth Ruiz.

La última columna del Excel que se muestra en la figura 33. contiene el código de Clasificación Industrial Internacional Uniforme que permite clasificar uniformemente las actividades de carácter económico en Sección, División, Grupo, Clases, Subclases y Actividad. El propósito principal de este código es reunir y presentar datos estadísticos acorde a las actividades económicas (Instituto Nacional de Estadísticas y Censos, 2012).

Tomando como ejemplo el primer código en la tabla, el significado de G476102 sería el siguiente.



Sección: Comercio al por mayor y al por menor, reparación de vehículos automotores y motocicletas;

División: Comercio al por menor, excepto el de vehículos automotores y motocicletas;

Grupo: Venta al por menor de productos culturales y recreativos en comercios especializados;

Clase: Venta al por menor de libros, periódicos y artículos de papelería en comercios especializados

Subclase: Venta al por menor de libros, periódicos y artículos de papelería en comercios especializados.

Actividad: Venta al por menor de periódicos en establecimientos especializados.

Antes de pasar a la siguiente etapa, en el archivo Directorio de Emprendimientos se añadirán 6 columnas más, estas columnas contendrán el significado de los caracteres del código CIIU de cada emprendimiento debido a que el código como tal no sirve de mucho para poder clasificar las actividades económicas. Por esta razón, desde la página <https://aplicaciones2.ecuadorencifras.gob.ec/SIN/> se descargó el archivo CIIU REV 4.0 EXCEL. Con la función BUSCARV(valor a buscar, lugar y rango en donde buscar, columna que devuelve el valor, aprobar coincidencia o no) se cruzaron las bases de datos de Excel. Para el argumento de la función (valor a buscar) se utiliza la función IZQUIERDA(texto; cantidad de caracteres a extraer acorde al nivel).

	CODIGO CIIU	SECTOR	DIVISION	GRUPO	CLASE	SUBCLASE	ACTIVIDAD
1							
2	562901	ACTIVIDADES DE A	SERVICIO DE SUMINISTRO DE C	OTRAS ACTIV	OTRAS ACTIV	OTRAS ACTIV	Actividades de contratistas de servicio de comidas (por ejemplo, para compañías de transpor
3	5960907	OTRAS ACTIVIDAD	OTRAS ACTIV	OTRAS ACTIVIDAD	OTRAS ACTIV	OTRAS ACTIV	Actividades de limpiabotas (betuneros), porteadores de maletas, personas encargadas de est
4	M691001	ACTIVIDADES PRO	ACTIVIDADES	ACTIVIDADES	ACTIVIDADES	ACTIVIDADES	Actividades de representación jurídica de los intereses de una parte contra otra, sea o no ant
5	5960907	OTRAS ACTIVIDAD	OTRAS ACTIV	OTRAS ACTIVIDAD	OTRAS ACTIV	OTRAS ACTIV	Actividades de limpiabotas (betuneros), porteadores de maletas, personas encargadas de est
6	G479902	COMERCIO AL POR	COMERCIO A VENTA AL POR ME	OTRAS ACTIV	OTRAS ACTIV	OTRAS ACTIV	Venta al por menor por comisionistas (no dependientes de comercios); Incluye actividades de
7	R900001	ARTES, ENTRETE	ACTIVIDADES	ACTIVIDADES	ACTIVIDADES	ACTIVIDADES	Producción de obras de teatro, conciertos, óperas, espectáculos de danza y otras actividades
8	5960907	OTRAS ACTIVIDAD	OTRAS ACTIV	OTRAS ACTIVIDAD	OTRAS ACTIV	OTRAS ACTIV	Actividades de limpiabotas (betuneros), porteadores de maletas, personas encargadas de est
9	5960907	OTRAS ACTIVIDAD	OTRAS ACTIV	OTRAS ACTIVIDAD	OTRAS ACTIV	OTRAS ACTIV	Actividades de limpiabotas (betuneros), porteadores de maletas, personas encargadas de est
10	5960907	OTRAS ACTIVIDAD	OTRAS ACTIV	OTRAS ACTIVIDAD	OTRAS ACTIV	OTRAS ACTIV	Actividades de limpiabotas (betuneros), porteadores de maletas, personas encargadas de est
11	G477111	COMERCIO AL POR	COMERCIO A VENTA AL POR ME	VENTA AL PO	VENTA AL PO	VENTA AL PO	Venta al por menor de prendas de vestir y peletería en establecimientos especializados;
12	5960907	OTRAS ACTIVIDAD	OTRAS ACTIV	OTRAS ACTIVIDAD	OTRAS ACTIV	OTRAS ACTIV	Actividades de limpiabotas (betuneros), porteadores de maletas, personas encargadas de est
13	563002	ACTIVIDADES DE A	SERVICIO DE	ACTIVIDADES DE	ACTIVIDADES	ACTIVIDADES	Actividades de preparación y servicio de bebidas para su consumo inmediato en: cafés, tiend
14	U682002	ACTIVIDADES INM	ACTIVIDADES	ACTIVIDADES	ACTIVIDADES	ACTIVIDADES	Administración de bienes inmuebles a cambio de una retribución o por contrato.
15	M711011	ACTIVIDADES PRO	ACTIVIDADES DE	ACTIVIDADES	ACTIVIDADES	ACTIVIDADES	Actividades de asesoramiento técnico de arquitectura en diseño de edificios y dibujo de plan
16	O841101	ADMINISTRACIÓN	ADMINISTRACIÓN	ADMINISTRACIÓN	ADMINISTRACIÓN	ADMINISTRACIÓN	Desempeño de las funciones ejecutivas y legislativas de los órganos y organismos centrales, r
17	O841101	ADMINISTRACIÓN	ADMINISTRACIÓN	ADMINISTRACIÓN	ADMINISTRACIÓN	ADMINISTRACIÓN	Desempeño de las funciones ejecutivas y legislativas de los órganos y organismos centrales, r

Figura 34. Columnas agregadas con significado de código CIIU. Información tomada de Excel. Elaborado por Ruth Ruiz.

3.2.5. Visualización de los datos

La visualización de los datos es la última etapa de la metodología propuesta, para este último paso se consideraron las 2 herramientas recomendadas por los entrevistados, Power Bi y Tableau Public. Ambas herramientas permiten crear tableros de mando visualmente llamativos y comprensibles, sin embargo, se inclinó por la primera opción al poseer una versión gratuita bastante generosa y brindar todos los servicios Pro en un tiempo de duración de 2 meses si el registro se realiza con una cuenta institucional u organizacional.

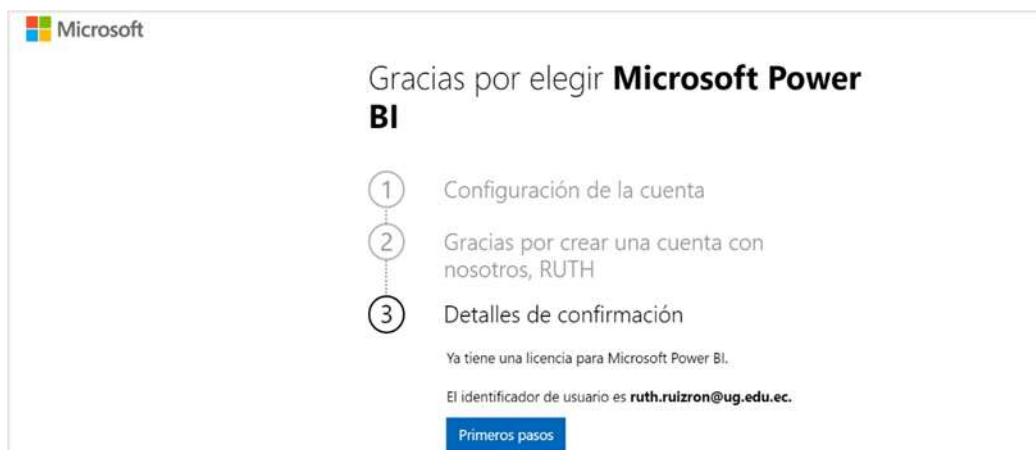


Figura 35. Registro con cuenta institucional en Power Bi. Información tomada de Power Bi. Elaborado por Power Bi.

Luego de haber obtenido la licencia para Microsoft Power Bi, se descargó desde la Microsoft Store la aplicación Power BI Desktop para instalarla de manera local y mejorar la experiencia de diseño del tablero de mando, la versión de escritorio posee una interfaz muy parecida a las otras herramientas de Office.

3.2.5.1. Conexión con origen de datos

Power BI Desktop permite conectar cientos de orígenes de datos de la Suite Office, entre ellas, Excel, en este proceso se importan los datos recabados y almacenados en Directorio de emprendimientos.xlsx con el formato de tabla que permite administrar y analizar datos relacionados.

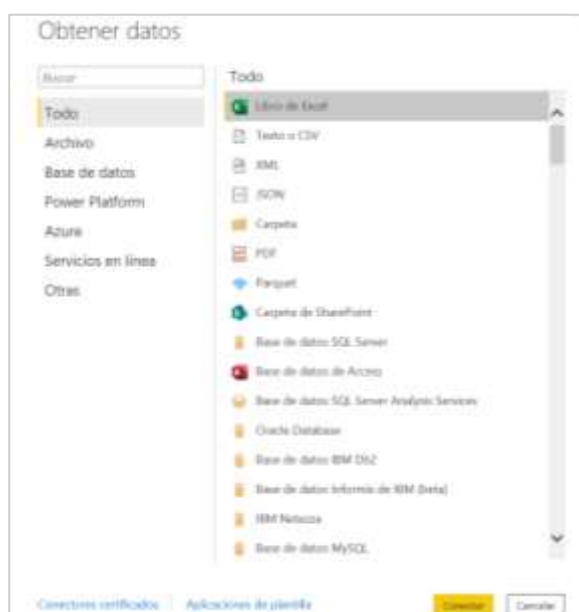


Figura 36. Origen de datos. Información tomada de Power Bi. Elaborado por Power Bi.

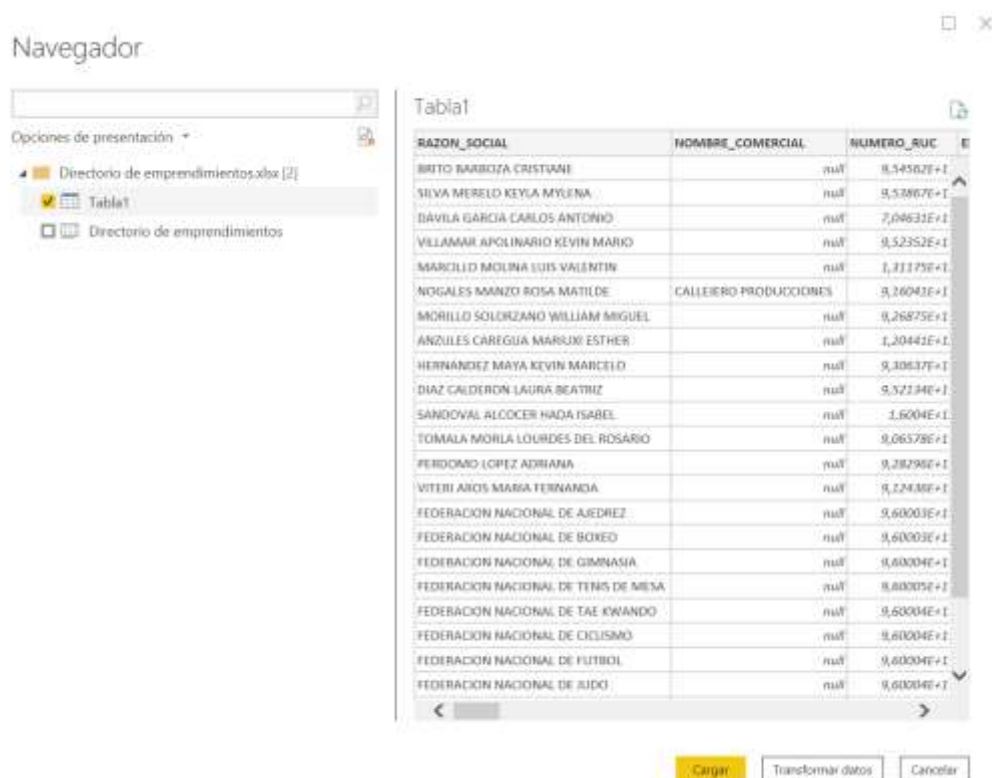


Figura 37. Selección de tabla dentro del archivo *Directorio de emprendimientos*. Información tomada de Power BI. Elaborado por Power BI.

3.2.5.2. Diseño del Dashboard

Una vez importado el conjunto de datos en Power BI se procede a diseñar el dashboard que permitirá explorar visualmente los datos, mediante objetos visuales creados y organizados sobre un área del lienzo de la vista **Informe**.

Cada uno de estos objetos visuales puede contener uno o varios campos de la tabla importada mediante la *acción arrastrar y soltar*. La figura 38. muestra el resultado de adherir desde el panel **Visualización**, el objeto “Segmentación de datos” que permite crear un filtro en cascada con el objetivo de agrupar la Clasificación Industrial Internacional Uniforme (CIIU), este filtro se sincronizará con un conjunto de elementos visuales para crear informes centralizados de las distintas actividades comerciales.

El conjunto de elementos visuales que se utiliza son Gráficos de líneas, Gráficos de barras y Gráficos circulares, el primero fue el más recomendado entre los entrevistados. Los gráficos de líneas permiten demostrar la tendencia de las actividades económicas con el paso del tiempo; se insertan 2 gráficos de este tipo, uno para visualizar la cantidad de emprendimiento que

iniciaron en una actividad económica por año y el otro para visualizar la cantidad de emprendimientos que se suspendieron definitivamente en una actividad económica por año.



Figura 38. Ejemplo de objeto visual Segmentación de datos con varios campos. Información tomada de Power BI. Elaborado por Ruth Ruiz.

Los gráficos de barras se caracterizan por representar la longitud de un conjunto de datos por categorías, este tipo de gráfico fue recomendado por el 50% de los entrevistados, en este caso se utilizan 2 gráficos de barras para representar el total de emprendimientos por actividad económica y el total de emprendimientos distribuidos por parroquia.

El último gráfico utilizado es el circular, este tipo de gráfico también fue recomendado por el 50% de los entrevistados. El uso más común del gráfico circular es el de visualizar porciones o sectores de la totalidad de un conjunto de datos, dentro del tablero de mando se destinó 4 de estos gráficos para presentar las 5 actividades más comunes en el mercado, 5 actividades con mayor duración en el mercado, 5 actividades con menor duración en el mercado y el promedio de duración de una actividad tomando como referencia el inicio y suspensión de la actividad por año.

Otro de los gráficos recomendados por el 50% de los entrevistados fue el cartograma, este tipo de gráfico no se incluye dentro del dashboard final debido a los errores de ubicación que se generan constantemente. El resultado final se muestra en la figura 39.

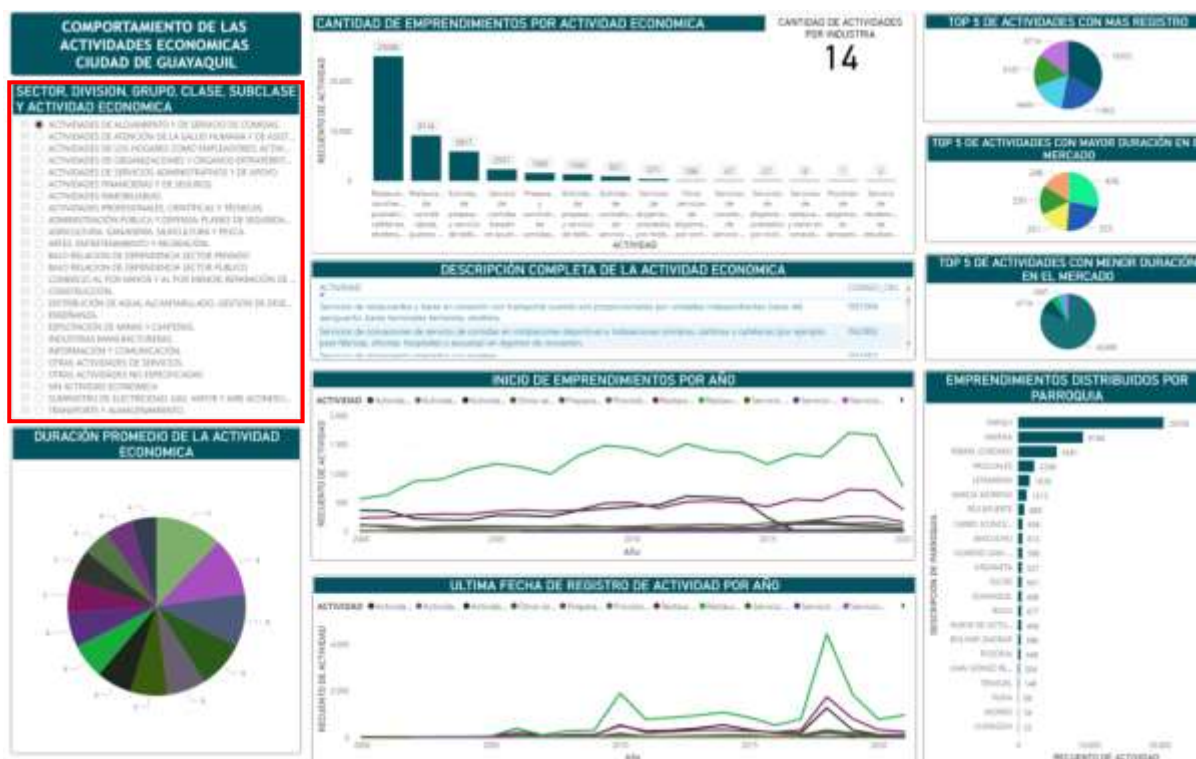


Figura 39. Dashboard final. Información tomada de Power Bi. Elaborado por Ruth Ruiz.

El objeto visual que está encerrado en el recuadro rojo es el filtro principal que interactúa con 6 objetos del lienzo, estos son: El grafico circular (Duración promedio de la actividad económica), los gráficos de línea (Inicio de emprendimiento y última fecha de registro de actividad económica por año), los gráficos de barras (Emprendimientos distribuidos por parroquia y Cantidad de emprendimientos por actividad económica) y por último la Tarjeta (Cantidad de actividades por industria). Los 3 gráficos circulares restantes son independientes, no comparten filtros con los otros objetos.

Existen distintas maneras de llegar a conocer el comportamiento de una actividad económica en el mercado, la primera es buscando a través del objeto “Segmentación de datos” que representa un filtro en cascada que va desde Sección, División, Grupo, Clase, Subclase hasta llegar a la actividad, figura 40.

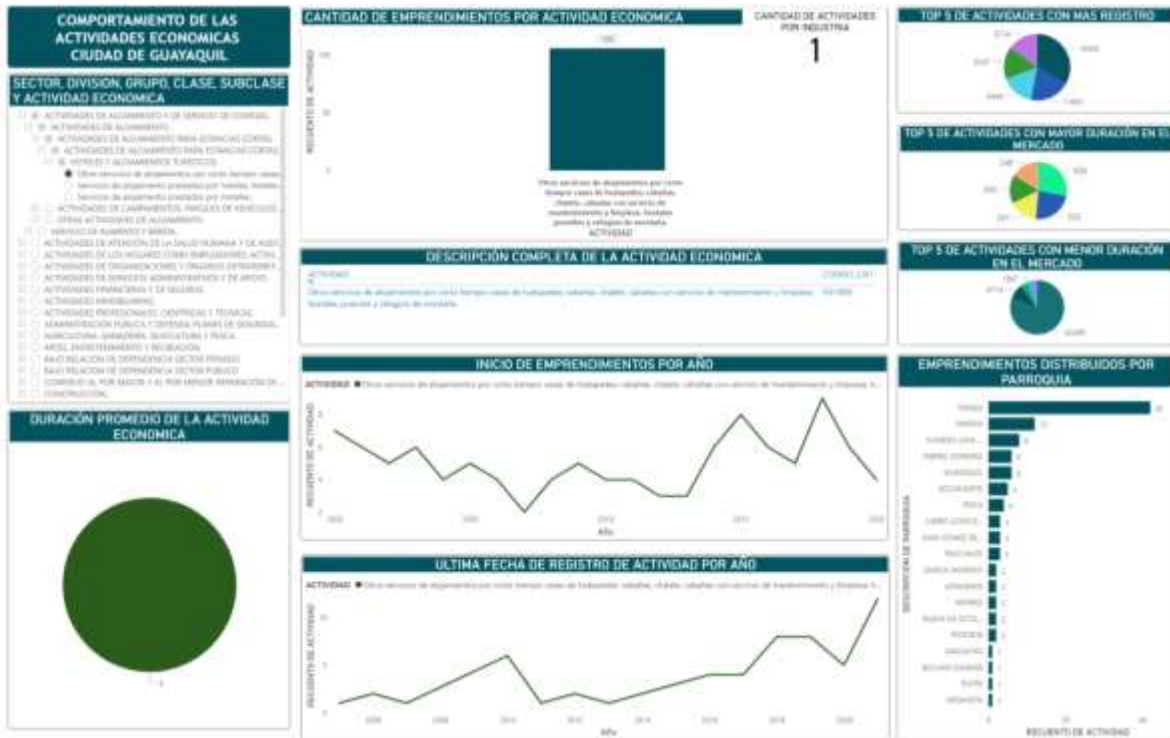


Figura 40. Filtro 1 del dashboard. Información tomada de Power Bi. Elaborado por Ruth Ruiz.

La segunda manera es marcar tan solo el primer nivel del objeto “Segmentación de datos” y utilizar como filtro al grafico de barras que muestra el recuento de emprendimientos por actividad económica, figura 41.

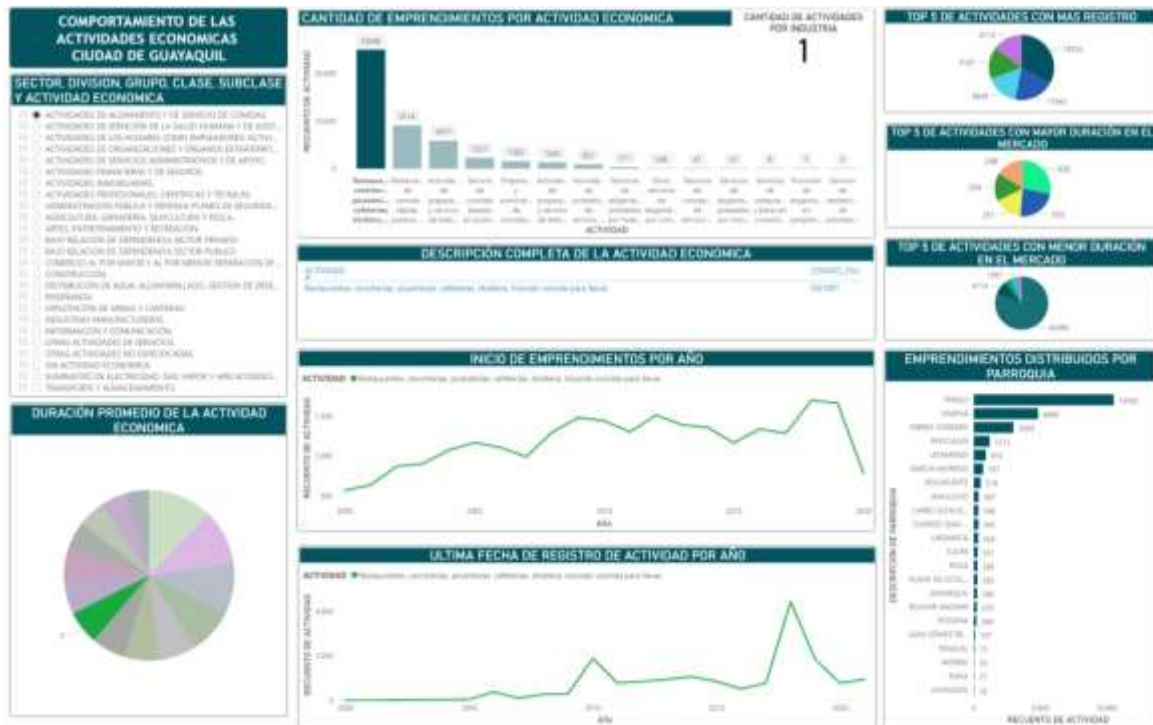


Figura 41. Filtro 2 del dashboard. Información tomada de Power Bi. Elaborado por Ruth Ruiz.

Y la última, es parecida a la segunda solo que en lugar de utilizar el grafico de barra se utiliza la tabla de texto ubicada en la parte inferior del grafico de barras (Recuento de emprendimientos). Esto es posible debido a la reciprocidad automática de filtrado entre objetos, al ingresar un objeto automáticamente los campos ingresados se convierten en filtros para toda la página, para especificar lo contrario se selecciona el objeto que no se desea utilizar como filtro y luego se marca el icono **Ninguno** (encerrado en recuadro rojo) del objeto que no se debe filtrar, figura 42.

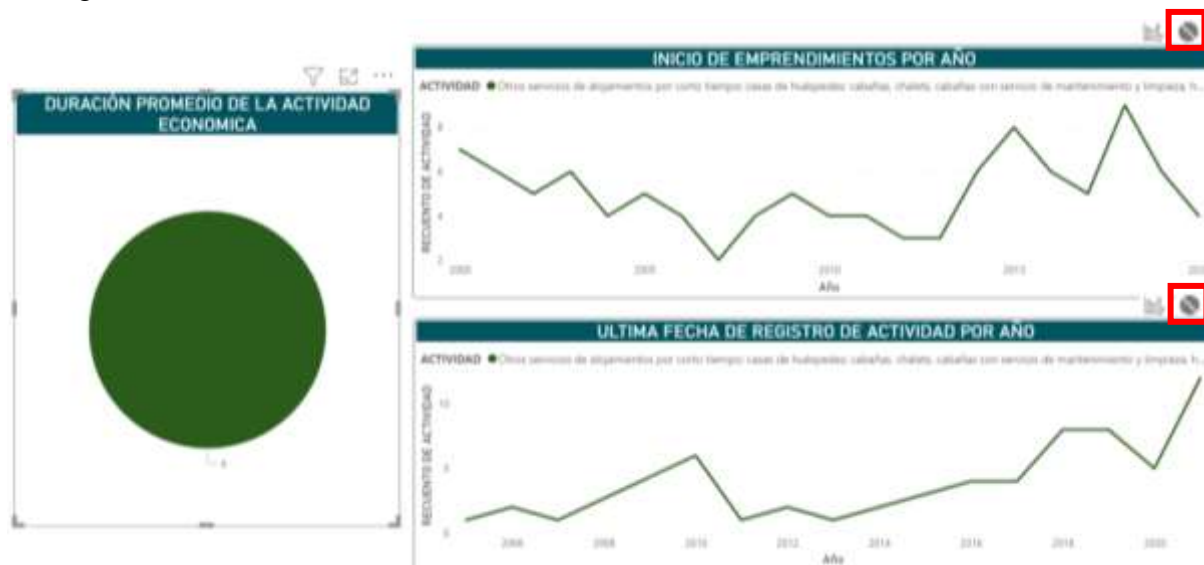


Figura 42. Negar reciprocidad de filtrado entre objetos. Información tomada de Power Bi. Elaborado por Ruth Ruiz.

3.3. Conclusiones y recomendaciones

3.3.1. Conclusiones

- Con base en la investigación documental, se concluye que es posible ejecutar técnicas de web scraping en medios digitales debido a que no está definido un marco legal para dicha técnica en Ecuador, sin embargo, es importante revisar detenidamente las políticas de privacidad y uso que la página impone para no violar los derechos del autor.
- Se concluye que la técnica mas optima para realizar raspado web son los analizadores HTML y las peticiones HTTP ya que permiten identificar patrones en el contenido del lenguaje de marcado de todo un sitio web y recolectar datos de múltiples páginas del sitio automáticamente.

- Se concluye que la herramienta más útil y adecuada para desarrollar web scraping es Python, por ser fácil de comprender, flexible y por manejar un entorno enriquecido de herramientas para la minería de datos como librerías que permiten construir un código funcional de extracción totalmente personalizado y almacenar datos en tiempos periódicos de ejecución.
- Utilizando la técnica de web scraping en el portal “Directorio de emprendimientos” y la visualización del tablero de mando, se concluye que el presente trabajo de investigación aporta con conocimiento técnico del uso de las herramientas y a los microemprendedores la facilidad de obtener información relacionada al comportamiento de las actividades económicas en la ciudad de Guayaquil.

3.3.2. Recomendaciones

- Comprobar que el gestor de paquetes PIP de Python esté actualizado en la última versión para evitar conflictos en la instalación de librerías y en la ejecución del código.
- Evitar el uso de cartogramas como objetos visuales para la construcción de un dashboard debido a la falta de desarrollo de mapas en países sudamericanos en herramientas que integren Bing Maps.
- Desarrollar un dashboard a futuro donde los datos puedan ser actualizados constantemente y en tiempo real para conocer en los próximos años el comportamiento de las actividades económicas en el mercado.

ANEXOS

Anexo 1.

Decreto Ejecutivo 101

UTILIZACION DE SOFTWARE LIBRE EN LA ADMINISTRACION PUBLICA

Decreto Ejecutivo 1014

Registro Oficial 322 de 23-abr.-2008

Ultima modificación: 25-abr.-2011

Estado: Reformado

Rafael Correa Delgado

PRESIDENTE CONSTITUCIONAL DE LA REPUBLICA

Considerando:

Que en el apartado g) del numeral 6 de la Carta Iberoamericana de Gobierno Electrónico, aprobada por la IX Conferencia Iberoamericana de Ministros de Administración Pública y Reforma del Estado, realizada en Chile el 1 de junio del 2007, se recomienda el uso de estándares abiertos y software libre, como herramientas informáticas;

Que es el interés del Gobierno alcanzar soberanía y autonomía tecnológica, así como un significativo ahorro de recursos públicos y que el software libre es en muchas instancias un instrumento para alcanzar estos objetivos;

Que el 18 de julio del 2007 se creó e incorporó a la estructura orgánica de la Presidencia de la República la Subsecretaría de Informática, dependiente de la Secretaría General de la Administración, mediante Acuerdo No. 119, publicado en el Registro Oficial No. 139 de 1 de agosto del 2007 ;

Que el numeral 1 del artículo 6 del Acuerdo No. 119, faculta a la Subsecretaría de Informática a elaborar y ejecutar planes, programas, proyectos, estrategias, políticas, proyectos de leyes y reglamentos para el uso de software libre en las dependencias del Gobierno Central; y,

En ejercicio de la atribución que le confiere el numeral 9 del artículo 171 de la Constitución Política de la República.

Decreta:

Art. 1.- Establecer como política pública para las entidades de la Administración Pública Central la utilización de software libre en sus sistemas y equipamientos informáticos.

Art. 2.- Se entiende por software libre, a los programas de computación que se pueden utilizar y distribuir sin restricción alguna, que permitan su acceso a los códigos fuentes y que sus aplicaciones puedan ser mejoradas.

Estos programas de computación tienen las siguientes libertades:

- a) Utilización del programa con cualquier propósito de uso común;
- b) Distribución de copias sin restricción alguna;
- c) Estudio y modificación del programa (Requisito: código fuente disponible); y,
- d) Publicación del programa mejorado (Requisito: código fuente disponible).

Art. 3.- Las entidades de la Administración Pública Central previa a la instalación del software libre en sus equipos, deberán verificar la existencia de capacidad técnica que brinde el soporte necesario para el uso de este tipo de software.

Anexo 2.

Código Orgánico de la Economía Social



Disposiciones especiales sobre ciertas obras

Parágrafo Primero

Del software y bases de datos

Apartado Primero

Del software de código cerrado y bases de datos

Art. 131.- Protección de software.- El software se protege como obra literaria. Dicha protección se otorga independientemente de que hayan sido incorporados en un ordenador y cualquiera sea la forma en que estén expresados, ya sea como código fuente; es decir, en forma legible por el ser humano; o como código objeto; es decir, en forma legible por máquina, ya sea sistemas operativos o sistemas aplicativos, incluyendo diagramas de flujo, planos, manuales de uso, y en general, aquellos elementos que conformen la estructura, secuencia y organización del programa.

Se excluye de esta protección las formas estándar de desarrollo de software.

Art. 132.- Adaptaciones necesarias para la utilización de software.- Sin perjuicio de los derechos morales del autor, el titular de los derechos sobre el software, o el propietario u otro usuario legítimo de un ejemplar del software, podrá realizar las adaptaciones necesarias para la utilización del mismo, de acuerdo con sus necesidades, siempre que ello no implique su utilización con fines comerciales.

Art. 133.- Titulares de derechos.- Es titular de los derechos sobre un software el productor, esto es, la persona natural o jurídica que toma la iniciativa y responsabilidad de la realización de la obra. Se presumirá titular, salvo prueba en contrario, a la persona cuyo nombre conste en la obra o sus copias de la forma usual.

Dicho titular está además autorizado para ejercer en nombre propio los derechos morales sobre la obra, incluyendo la facultad para decidir sobre su divulgación.

El productor tiene el derecho exclusivo de impedir que terceras personas realicen sin su consentimiento versiones sucesivas del software y software derivado del mismo.

Las disposiciones del presente artículo podrán ser modificadas mediante acuerdo entre los autores y el productor.

Concordancias:

- CÓDIGO CIVIL (TÍTULO PRELIMINAR), Arts. 32



Art. 134.- Actividades permitidas sin autorización.- Se permite las actividades relativas a un software de lícita circulación, sin que se requiera autorización del autor o titular, ni pago de valor alguno, en los siguientes casos:

1. La copia, transformación o adaptación del software que sea necesaria para la utilización del software por parte del propietario u otro usuario legítimo de un ejemplar del mismo;
2. La copia del software por parte del propietario u otro usuario legítimo de un ejemplar del mismo que sea con fines de seguridad y archivo, es decir, destinada exclusivamente a sustituir la copia legítimamente obtenida, cuando esta ya no pueda utilizarse por daño o pérdida;
3. Las actividades de ingeniería inversa sobre una copia legítimamente obtenida de un software que se realicen con el único propósito de lograr la compatibilidad operativa entre programas o para fines de investigación y educativos;
4. Las actividades que se realicen sobre una copia legítimamente obtenida de un software con el único propósito de probar, investigar o corregir su funcionamiento o la seguridad del mismo u otros programas, de la red o del computador sobre el que se aplica;
- y, 5. La utilización de software con fines de demostración a la clientela en los establecimientos comerciales en que se expongan o vendan o reparen equipos o programas computacionales, siempre que se realice en el propio local o de la sección del establecimiento destinadas a dichos objetos y en condiciones que eviten su difusión al exterior.

Art. 135.- Excepción a la reproducción.- No constituye reproducción de un software, a los efectos previstos en el presente Título, la introducción del mismo en la memoria interna del respectivo aparato, para efectos de su exclusivo uso personal.

Art. 136.- Uso lícito del software.- Salvo pacto en contrario, será lícito el aprovechamiento del software para su uso en varias estaciones de trabajo mediante la instalación de redes, estaciones de trabajo u otros procedimientos similares.

Art. 137.- Excepción a la transformación.- No constituye transformación, a los efectos previstos en el presente Título, la adaptación de un software realizada por el propietario u otro usuario legítimo para la utilización exclusiva del software.

Art. 138.- Prohibición de transferencia a las modificaciones efectuadas a un software.- Las adaptaciones o modificaciones permitidas en este Parágrafo no podrán ser transferidas bajo ningún título, sin que medie autorización previa del titular del derecho respectivo. Asimismo, los ejemplares obtenidos en la forma indicada no podrán ser transferidos bajo ningún título, salvo que lo sean conjuntamente con el programa que les sirvió de matriz y con la autorización del titular.

Anexo 3.

Ley Orgánica de Transparencia y Acceso a la Información Pública

LEY ORGANICA DE TRANSPARENCIA Y ACCESO A LA INFORMACION PUBLICA, Arts. 9, 23

CODIGO CIVIL (TITULO PRELIMINAR), Arts. 31

CODIGO CIVIL (LIBRO IV), Arts. 1572

Art. 5.- Publicidad.- El Estado, de conformidad con la Ley, pondrá en conocimiento de las ciudadanas o ciudadanos, la existencia de registros o bases de datos de personas y bienes y en lo aplicable, la celebración de actos sobre los mismos, con la finalidad de que las interesadas o interesados y terceras o terceros conozcan de dicha existencia y los impugnen en caso de afectar a sus derechos.

Concordancias:

LEY ORGANICA DE TRANSPARENCIA Y ACCESO A LA INFORMACION PUBLICA, Arts. 1

LEY DE COMPAÑÍAS, Arts. 439

Art. 6.- Accesibilidad y confidencialidad.- Son confidenciales los datos de carácter personal, tales como: ideología, afiliación política o sindical, etnia, estado de salud, orientación sexual, religión, condición migratoria y los demás atinentes a la intimidad personal y en especial aquella información cuyo uso público atente contra los derechos humanos consagrados en la Constitución e instrumentos internacionales.

El acceso a estos datos sólo será posible con autorización expresa del titular de la información, por mandato de la ley o por orden judicial.

También son confidenciales los datos cuya reserva haya sido declarada por la autoridad competente, los que estén amparados bajo sigilo bancario o bursátil, y los que pudieren afectar la seguridad interna o externa del Estado.

La autoridad o funcionario que por la naturaleza de sus funciones custodie datos de carácter personal, deberá adoptar las medidas de seguridad necesarias para proteger y garantizar la reserva de la información que reposa en sus archivos.

Para acceder a la información sobre el patrimonio de las personas el solicitante deberá justificar y motivar su requerimiento, declarar el uso que hará de la misma y consignar sus datos básicos de identidad, tales como: nombres y apellidos completos, número del documento de identidad o ciudadanía, dirección domiciliaria y los demás datos que mediante el respectivo reglamento se determinen. Un uso distinto al declarado dará lugar a la determinación de responsabilidades, sin perjuicio de las acciones legales que el/la titular de la información pueda ejercer.

La Directora o Director Nacional de Registro de Datos Públicos, definirá los demás datos que integrarán el sistema nacional y el tipo de reserva y accesibilidad.

Concordancias:

CONSTITUCION DE LA REPUBLICA DEL ECUADOR, Arts. 18, 66, 77, 91

CODIGO DE LA NIÑEZ Y ADOLESCENCIA, Arts. 54, 251, 317

LEY ORGANICA DE LA CONTRALORIA GENERAL DEL ESTADO, Arts. 79

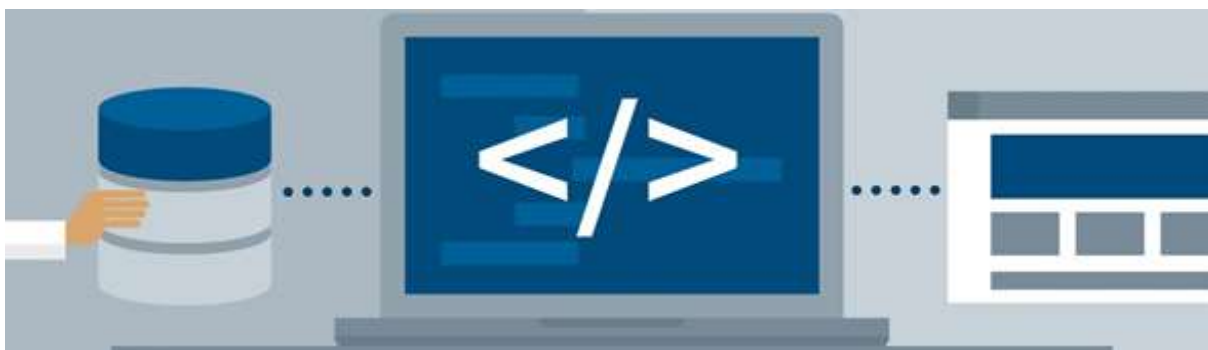
LEY ORGANICA DE EMPRESAS PUBLICAS, LOEP, Arts. 46

CODIGO DEL TRABAJO, Arts. 45, 310

LEY ORGANICA DE TRANSPARENCIA Y ACCESO A LA INFORMACION PUBLICA, Arts. 6, 9, 19

Anexo 4.

La Entrevista



SOPORTE AL MICROEMPRENDIMIENTO POR MEDIO DE PROTOTIPO DE DASHBOARD DE DATOS OBTENIDOS MEDIANTE WEB SCRAPING EN MEDIOS DIGITALES

Objetivo: Recopilar información con relación a la parte técnica (herramientas y software) que involucra la creación de un tablero de mando y un raspador web como soporte para los microemprendedores desde el punto de vista de profesionales en las diferentes áreas para el sustento del trabajo de titulación.

Apellidos y Nombre del experto consultado:

Escriba el título de tercer nivel:

Escriba el título de cuarto nivel:

Área de estudio en la que desempeña actualmente:

1. ¿Cuál es su nivel de experiencia con trabajos relacionados a web scraping?
 - Alto
 - Media
 - Bajo
 - Nada
2. ¿Qué lenguaje de programación cree que es adecuado para realizar web scraping?
 - Python
 - PHP
 - Ruby
 - C++

3. ¿Considera ético realizar web scraping en páginas que contengan medidas de seguridad como captcha o robots.txt * / disable?

- Sí
- No

4. En base a la pregunta anterior, justifique su respuesta.

5. ¿Cuál es su nivel de experiencia con trabajos relacionados a Dashboard?

- Alto
- Media
- Bajo
- Nada

6. Según su opinión, una base de datos construida con información pública de emprendimientos integrada en un dashboard aportaría con información de utilidad para los microempreendedores.

- Sí
- No

7. ¿Qué herramienta web gratuita considera apropiada para desarrollar un dashboard?

- Power Bi
- Tableau Public
- Inetsoft
- Syncfusion

8. ¿Cuáles son los tipos de gráficos estadísticos que considera estarían mejor ambientados para presentar datos extraídos de la web de un directorio de emprendimientos?

- Gráficos de barras
- Gráfico circular
- Gráfico de línea
- Cartograma

Anexo 5.

Instalación de Python

1. Descargar Python desde la página oficial <https://www.python.org/downloads/>



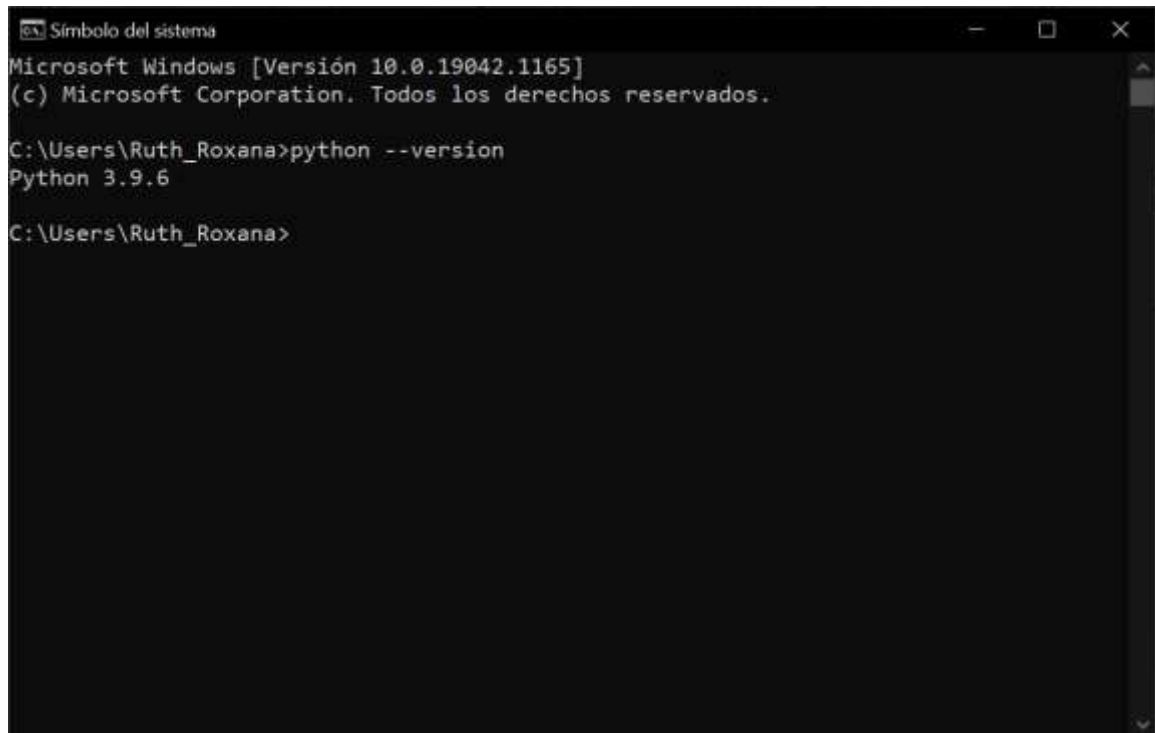
2. Instalación de Python

Una vez finalizada la descarga se debe ejecutar el archivo descargado y marcar las casillas *Instalar lanzador para todos los usuarios* y *Agregar Python (versión descargada) al Path*, esta segunda casilla permitirá que Python sea reconocido por el CMD (Símbolo del sistema) para que pueda ser administrado. Luego, se escoge la opción de instalación *Instalar ahora*, con esta opción, Python se instalará en el directorio de usuario, además de proporcionar un conjunto de bibliotecas estándar, el lanzador y el sistema de gestión de paquetes PIP.



Una vez terminada la instalación, se da clic en *Close*.

Para comprobar la versión de Python desde CMD se debe ingresar el comando **python --version** o simplemente **py --version**.

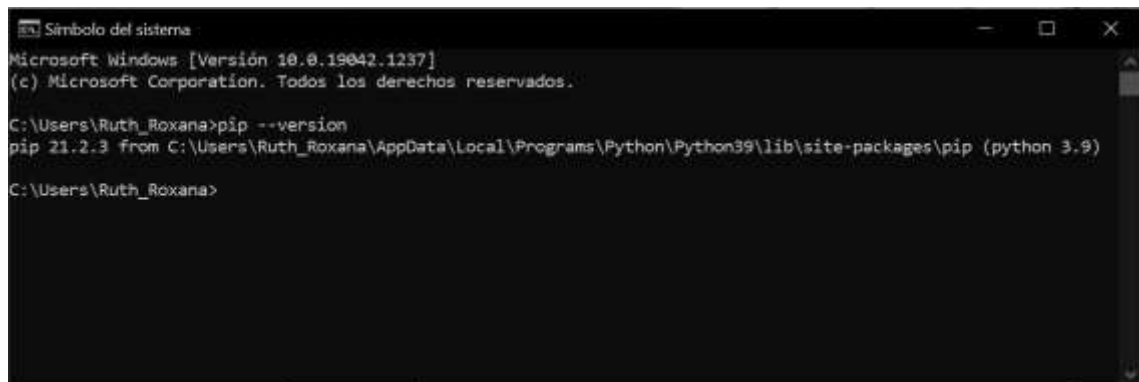


```
Símbolo del sistema
Microsoft Windows [Versión 10.0.19042.1165]
(c) Microsoft Corporation. Todos los derechos reservados.

C:\Users\Ruth_Roxana>python --version
Python 3.9.6

C:\Users\Ruth_Roxana>
```

Si se desea consultar la versión del sistema de gestión de librerías PIP, se debe ingresar el comando **pip --version**



```
Símbolo del sistema
Microsoft Windows [Versión 10.0.19042.1237]
(c) Microsoft Corporation. Todos los derechos reservados.

C:\Users\Ruth_Roxana>pip --version
pip 21.2.3 from C:\Users\Ruth_Roxana\AppData\Local\Programs\Python\Python39\lib\site-packages\pip (python 3.9)

C:\Users\Ruth_Roxana>
```

A pesar de que la ruta indique que la versión pip corresponde con la última versión Python instalada, PIP se actualiza constantemente. Por lo tanto, es necesario ejecutar el comando **python -m pip install -U pip**, este comando se encargará de buscar una versión actual y si existe, la instala.

```

Símbolo del sistema
Microsoft Windows [Versión 10.0.19042.1237]
(c) Microsoft Corporation. Todos los derechos reservados.

C:\Users\Ruth_Roxana>pip --version
pip 21.2.3 from C:\Users\Ruth_Roxana\AppData\Local\Programs\Python\Python39\lib\site-packages\pip (python 3.9)

C:\Users\Ruth_Roxana>python -m pip install -U pip
Requirement already satisfied: pip in c:\users\ruth_roxana\appdata\local\programs\python\python39\lib\site-pack
ages (21.2.3)
Collecting pip
  Using cached pip-21.2.4-py3-none-any.whl (1.6 MB)
Installing collected packages: pip
  Attempting uninstall: pip
    Found existing installation: pip 21.2.3
    Uninstalling pip-21.2.3:
      Successfully uninstalled pip-21.2.3
Successfully installed pip-21.2.4

C:\Users\Ruth_Roxana>

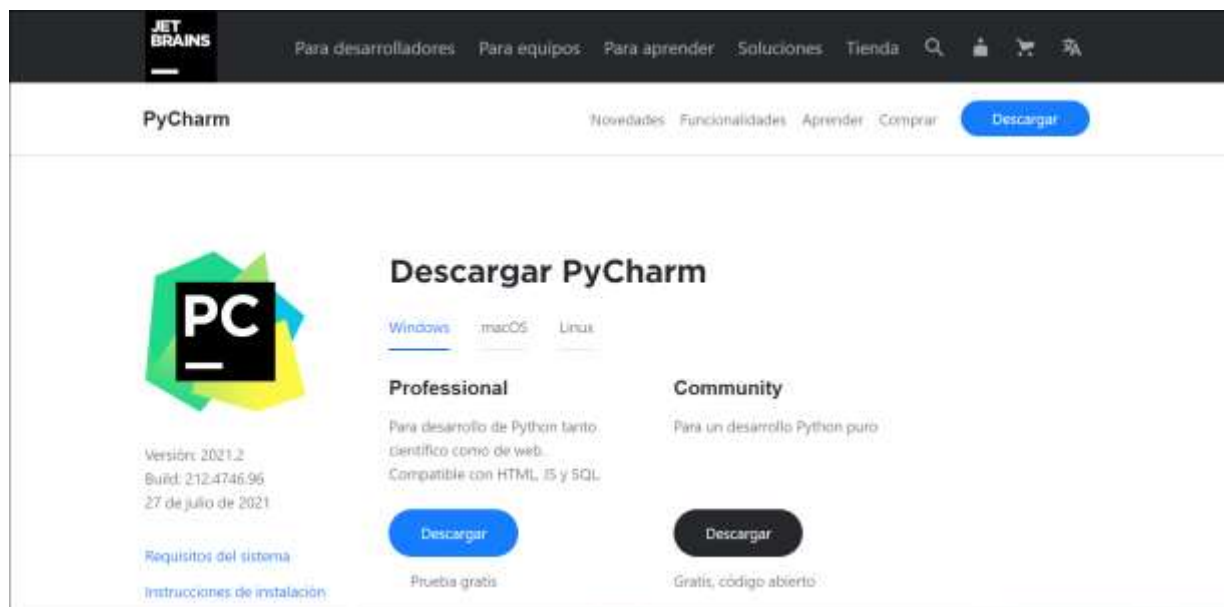
```

Con el primer comando la versión era **pip 21.2.3** después de ejecutar el siguiente comando se encontró una nueva versión **pip 21.2.4**. Este paso es importante para evitar conflictos al momento de instalar las librerías a utilizar para construir el raspador web.

Anexo 6.

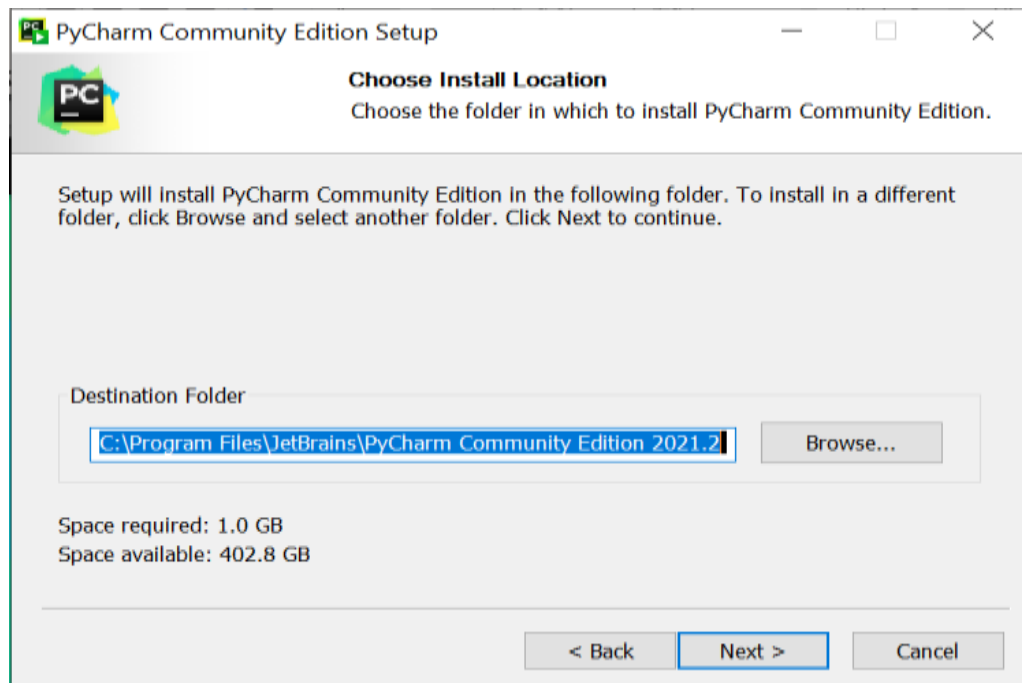
Instalación de Pycharm

1. Descargar Pycharm Community desde la página oficial <https://www.jetbrains.com/es-es/pycharm/download>.

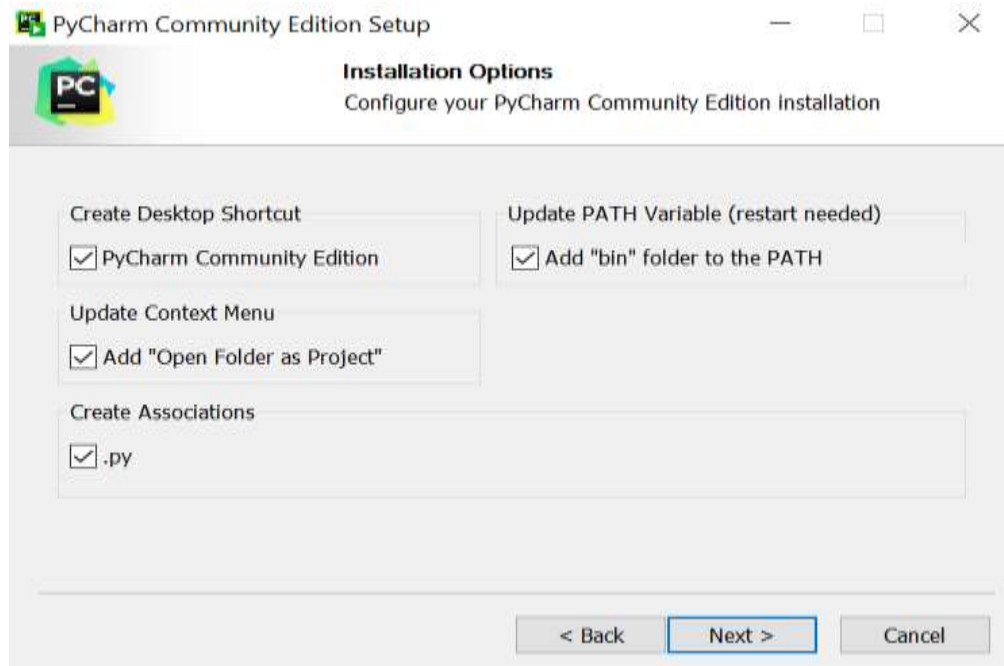


2. Instalación de Pycharm

Una vez terminada la descarga se debe ejecutar el archivo descargado y seleccionar la ubicación de la instalación.



En este punto, Pycharm proporciona opciones de instalación, “Crear asociación” es importante marcarla para que Pycharm se convierta en el editor de archivos .py.

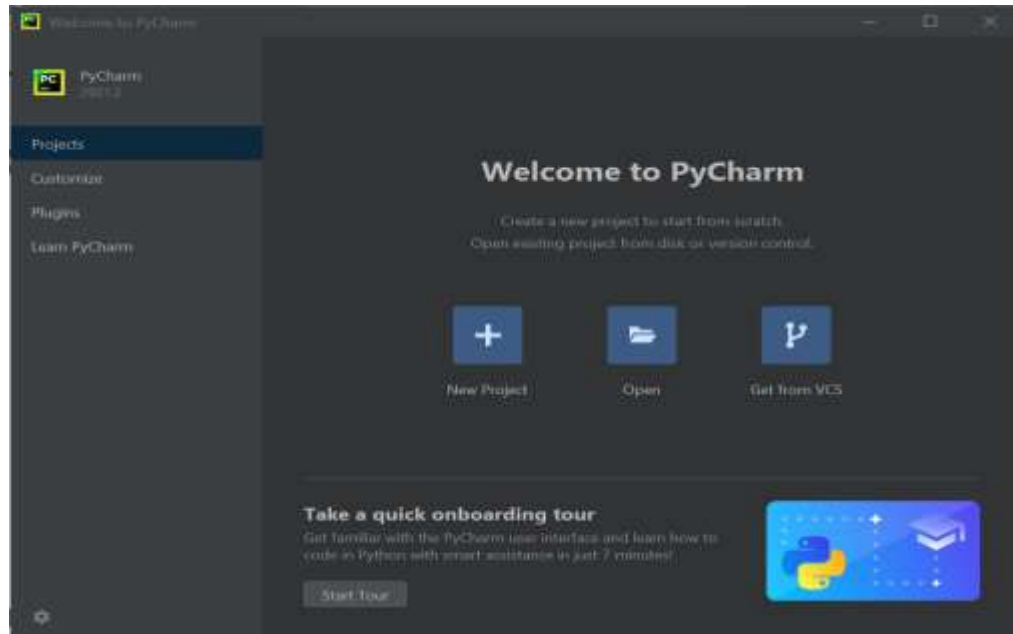


Después de este paso se da clic en Next / Instalar / Finalizar.

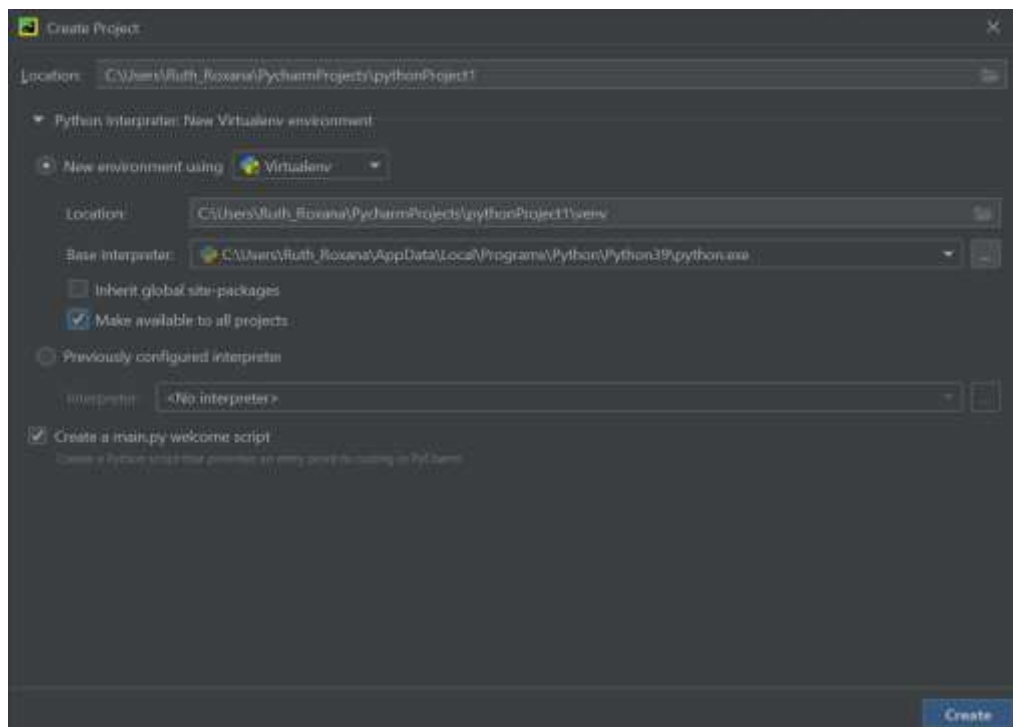
Anexo 7.

Crear un nuevo proyecto.

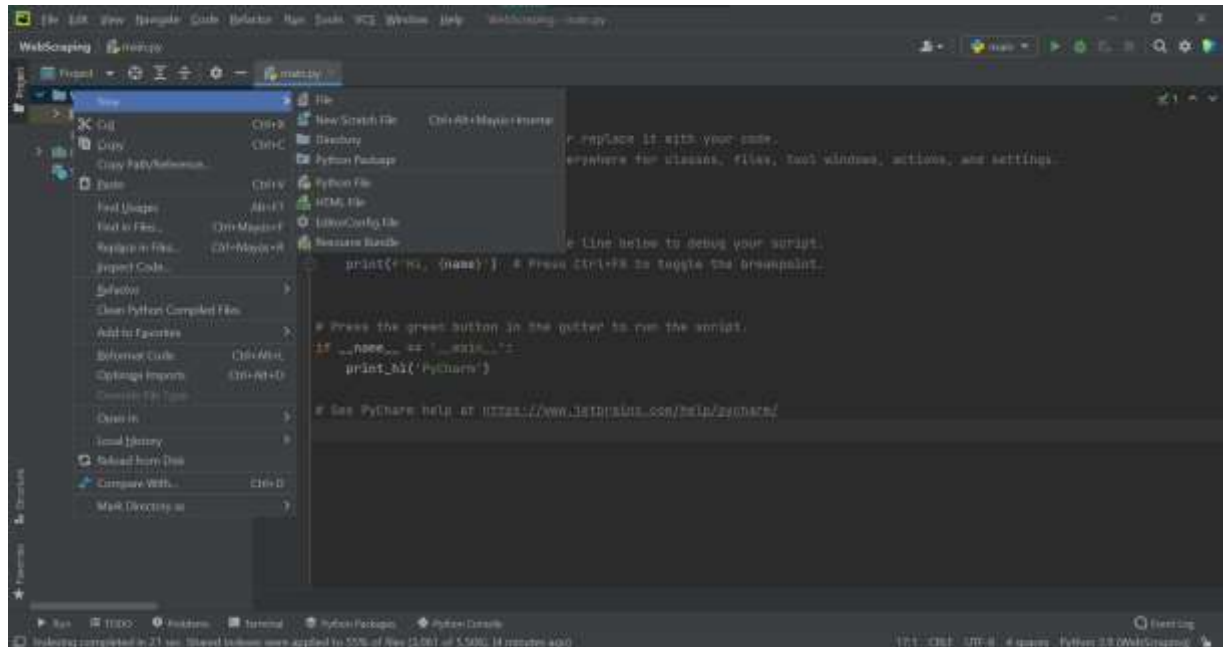
1. Escoger crear nuevo proyecto.



2. En la siguiente pantalla, dentro de la línea **Ubicación** al final se introduce el nombre de la carpeta que contendrá los scripts a desarrollar, en este caso la carpeta se llama **pythonProject1**, luego dar clic en **Crear**.



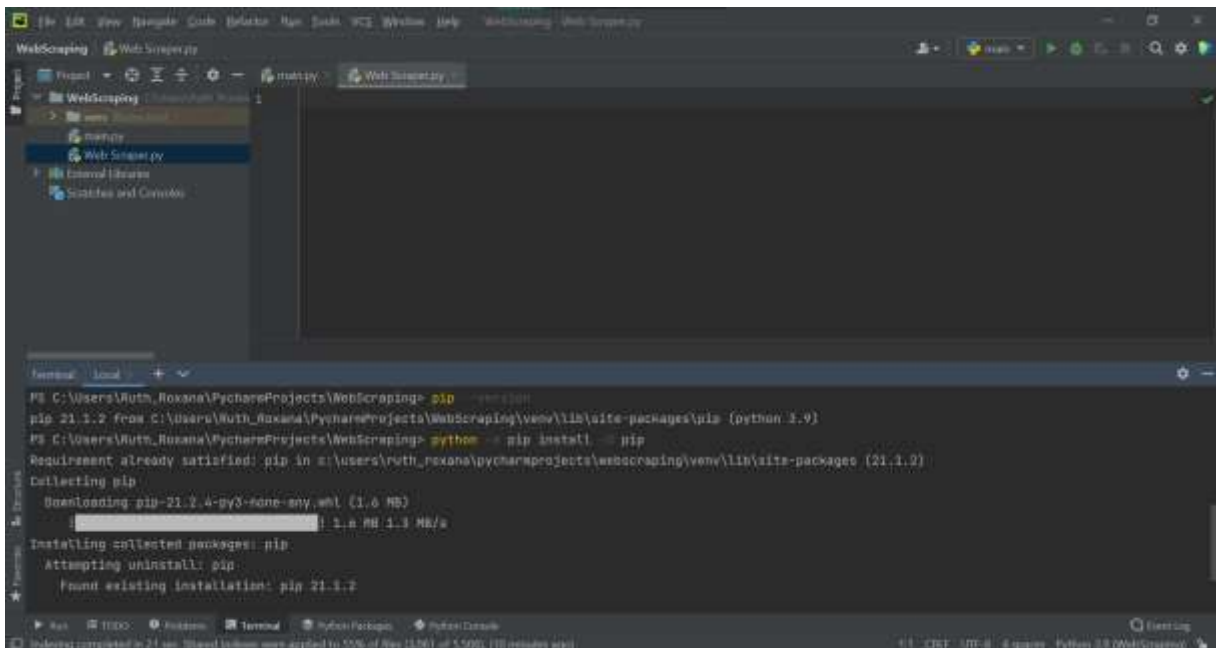
3. Para crear un nuevo archivo .py, se da clic derecho sobre la carpeta del proyecto en Nuevo / Archivo / Ingresar nombre del script y listo.



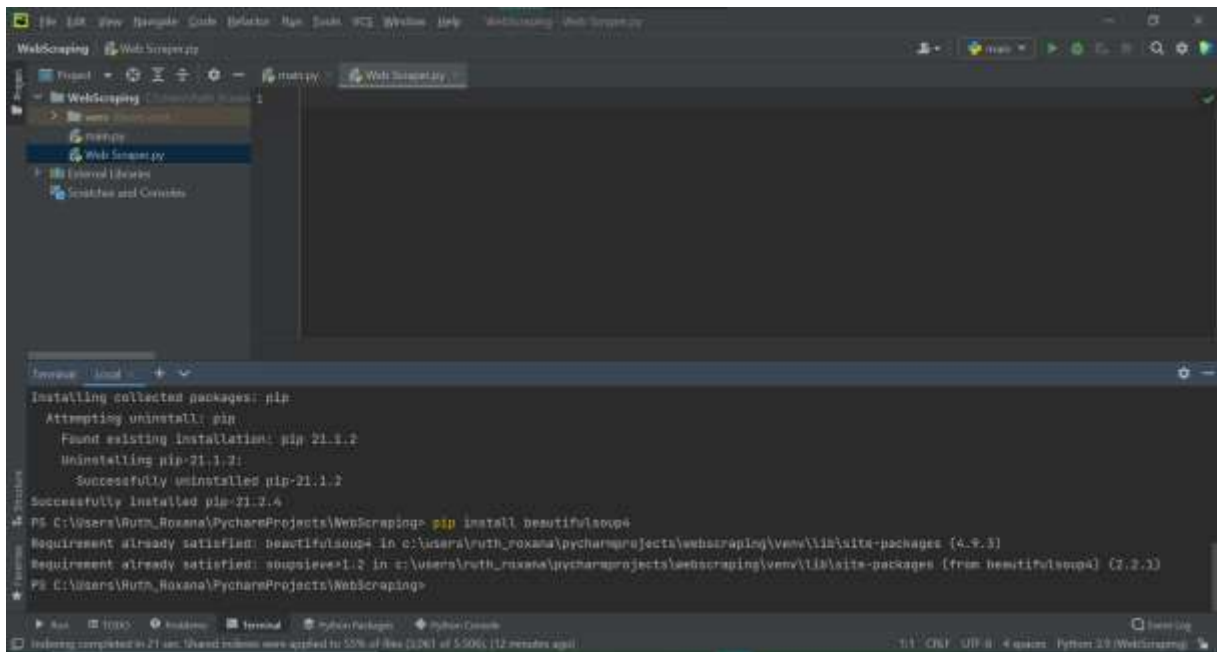
Anexo 8.

Instalar librerías desde la terminal de Pycharm

Desde Pycharm también se puede consultar la versión del gestor de librerías PIP, solo basta con insertar el mismo comando **pip --version** y, para actualizar el gestor, también se ejecuta el mismo comando **python -m pip install -U pip**.



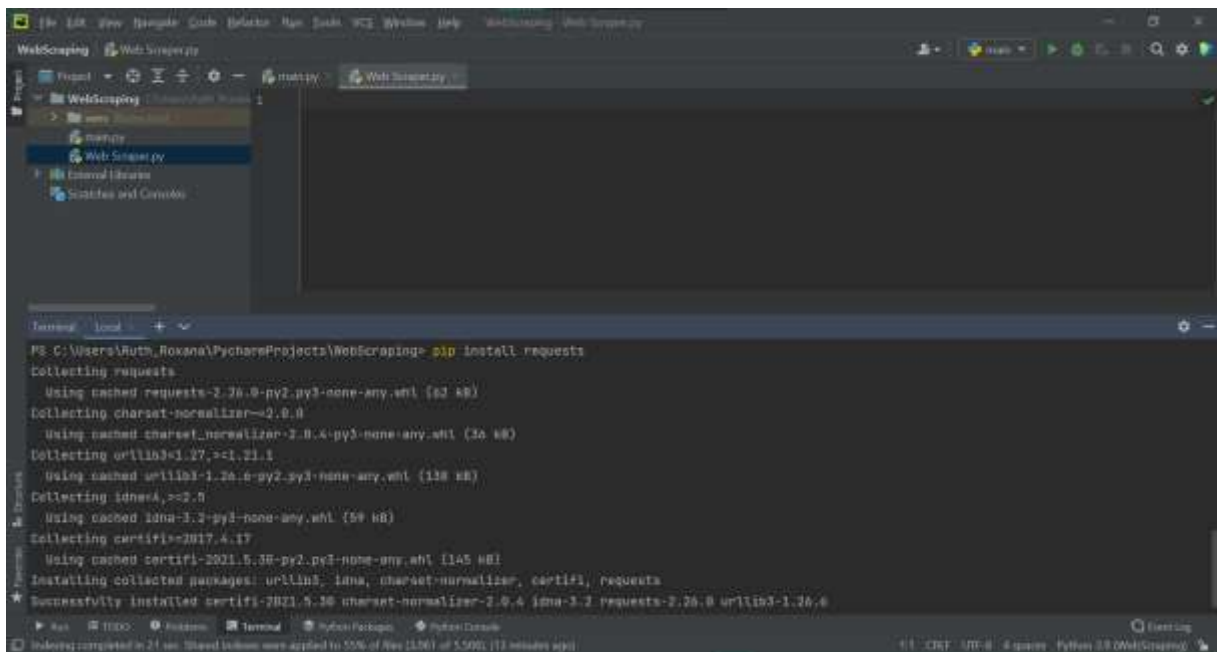
Para instalar la librería de Beautiful Soup se ingresa el comando **pip install beautifulsoup4**.



```

Installing collected packages: pip
  Attempting uninstall: pip
    Found existing installation: pip 21.1.2
    Uninstalling pip-21.1.2:
      Successfully uninstalled pip-21.1.2
    Successfully installed pip-21.2.4
PS C:\Users\Ruth_Roxana\PycharmProjects\WebScraping> pip install beautifulsoup4
Requirement already satisfied: beautifulsoup4 in c:\users\ruth_roxana\pycharmprojects\webcraping\venv\lib\site-packages (4.9.3)
Requirement already satisfied: soupsieve>1.2 in c:\users\ruth_roxana\pycharmprojects\webcraping\venv\lib\site-packages (from beautifulsoup4) (2.2.3)
PS C:\Users\Ruth_Roxana\PycharmProjects\WebScraping>
  
```

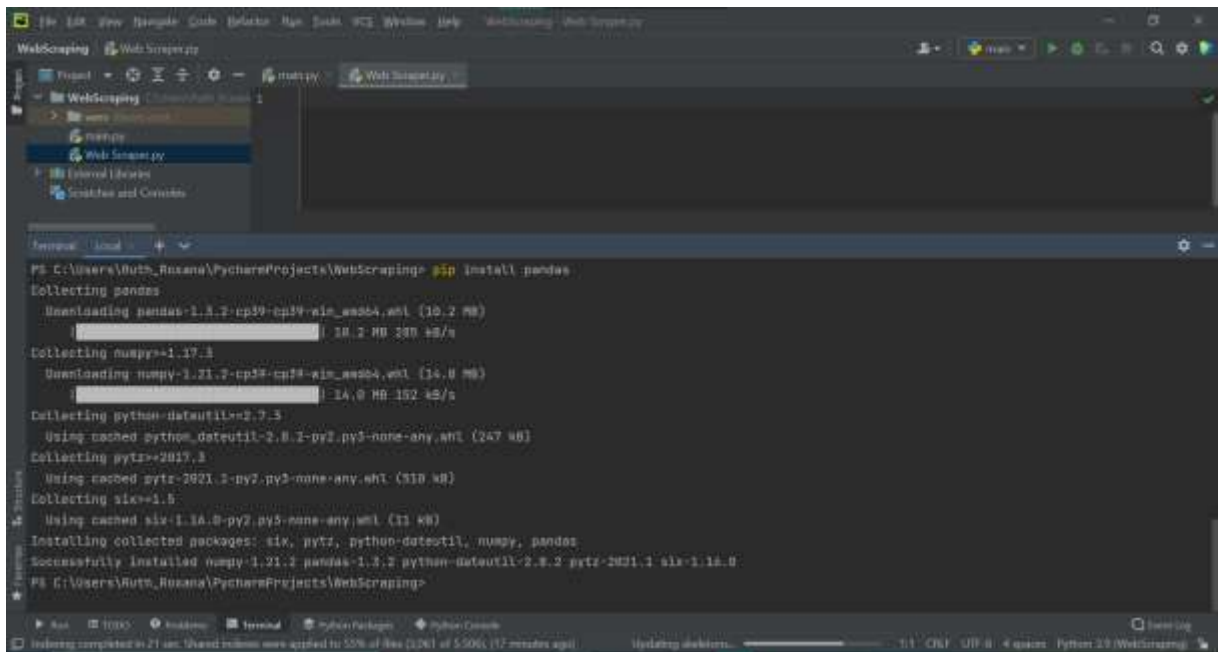
Para instalar la librería Requests se ingresa el comando **pip install requests**.



```

PS C:\Users\Ruth_Roxana\PycharmProjects\WebScraping> pip install requests
Collecting requests
  Using cached requests-2.26.0-py2.py3-none-any.whl (62 kB)
Collecting charset-normalizer<2.0.0
  Using cached charset-normalizer-2.0.4-py3-none-any.whl (36 kB)
Collecting urllib3<1.27,>=1.21.1
  Using cached urllib3-1.26.6-py2.py3-none-any.whl (138 kB)
Collecting idna<=2.0
  Using cached idna-3.2-py3-none-any.whl (59 kB)
Collecting certifi<2017.4.17
  Using cached certifi-2021.5.38-py2.py3-none-any.whl (1145 kB)
Installing collected packages: urllib3, idna, charset-normalizer, certifi, requests
Successfully installed certifi-2021.5.38 charset-normalizer-2.0.4 idna-3.2 requests-2.26.0 urllib3-1.26.6
  
```

Para instalar la librería Pandas se ingresa el comando **pip install requests**.



The screenshot shows the PyCharm IDE interface. The top toolbar includes icons for Run, Debug, and other development tools. The left sidebar displays the project structure with folders for 'WebScraping' and 'External Libraries'. The main editor area shows a terminal window with the following output:

```
PS C:\Users\Ruth_Rosana\PycharmProjects\WebScraping> pip install pandas
Collecting pandas
  Downloading pandas-1.3.2-cp39-cp39-win_amd64.whl (10.2 MB)
    10.2 MB 280 kB/s
Collecting numpy>=1.17.3
  Downloading numpy-1.21.2-cp39-cp39-win_amd64.whl (14.8 MB)
    14.8 MB 152 kB/s
Collecting python-dateutil>=2.7.3
  Using cached python_dateutil-2.8.2-py2.py3-none-any.whl (247 kB)
Collecting pytz>=2017.3
  Using cached pytz-2021.1-py2.py3-none-any.whl (510 kB)
Collecting six>=1.5
  Using cached six-1.16.0-py2.py3-none-any.whl (11 kB)
Installing collected packages: six, pytz, python-dateutil, numpy, pandas
Successfully installed numpy-1.21.2 pandas-1.3.2 python-dateutil-2.8.2 pytz-2021.1 six-1.16.0
PS C:\Users\Ruth_Rosana\PycharmProjects\WebScraping>
```

The bottom status bar indicates that the indexing is completed in 21 sec, shared indexes were applied to 50% of files (0.261 of 5.000) 177 minutes ago, and the system is updating the dictionary.

Bibliografía

- Alfonzo, I. (1995). *Técnicas de investigación bibliográfica*. Caracas, Venezuela: Contexto Editores.
- Arsalan. (01 de junio de 2021). *Best programming language for web scraping*. Obtenido de Information transformation services: <https://it-s.com/5-best-programming-languages-for-web-scraping/>
- Barría, G. (25 de Agosto de 2020). *Capítulo 12. Minería de datos web*. Recuperado el 05 de Agosto de 2021, de Github.io: <https://arcruz0.github.io/libroadp/web-mining.html#fn49>
- Brath, R., & Peters, M. (15 de octubre de 2004). *Dashboard design: Why design is important*. (DMReview, Ed.) Recuperado el 10 de agosto de 2021, de Furman University: http://cs.furman.edu/~pbatchelor/csc105/articles/TUN_DM_ONLINE.pdf
- Burgos, R., & Villar, L. (Agosto de 2016). Los emprendimientos desde la perspectiva histórica, económica y social, en el escenario mundial y del Ecuador. *Caribeña de Ciencias Sociales*. Obtenido de <https://www.eumed.net/rev/caribe/2016/08/emprendimientos.html>
- Bustamente, G. (2019). Extracción de información para la generación de reportes estructurados a partir de noticias peruanas relacionadas a crímenes. *Trabajo de pregrado*. Pontificia Universidad Católica de Perú, Lima, Perú. Recuperado el 07 de julio de 2021, de <http://hdl.handle.net/20.500.12404/14983>
- Castro, W., & Godino, J. (Enero de 2021). Métodos mixtos de investigación en las contribuciones a los simposios de la SEIEM. *Investigación de Educación Matemática XV*, 99-116. Obtenido de https://www.researchgate.net/publication/277836390_Metodos_mixtos_de_investigacion_en_las_contribuciones_a_los_simposios_de_la_SEIEM_1997-2010
- Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *CRISP-DM 1.0 Guía de minería de datos paso a paso*. Obtenido de <https://www.semanticscholar.org/paper/CRISP-DM-1.0%3A-Step-by-step-data-mining-guide-Chapman-Clinton/54bad20bbc7938991bf34f86dde0babfbd2d5a72>
- Constitución de la República del Ecuador. (21 de Diciembre de 2015). *Constitución de la República del Ecuador*. Obtenido de COSEDE: <https://www.cosedec.gob.ec/wp->

content/uploads/2019/08/CONSTITUCION-DE-LA-REPUBLICA-DEL-ECUADOR.pdf

Decreto Ejecutivo 1014 de software libre Ecuador. (25 de abril de 2011). *Utilización de software libre en la administración pública*. Obtenido de Agencia de Regulacion y Control Hidrocarburifero: <https://www.controlhidrocarburos.gob.ec/wp-content/uploads/MARCO-LEGAL-2016/Registro-Oficial-322-Decreto-Ejecutivo-1014.pdf>

Dexi.io. (05 de may de 2020). *Dexi Digital Commerce Intelligence Suite*. Recuperado el 08 de august de 2021, de Dexi.io: <https://www.dexi.io/software/>

Digital Guide IONOS. (21 de Octubre de 2020). *¿Qué es un web crawler? Cómo las arañas web optimizan Internet*. Recuperado el 29 de Julio de 2021, de IONOS: <https://www.ionos.es/digitalguide/online-marketing/marketing-para-motores-de-busqueda/que-es-un-web-crawler/>

Educ.ar portal. (03 de Agosto de 2017). *Los medios digitales y las nuevas formas de producir contenidos*. Obtenido de educ.ar: <https://www.educ.ar/recursos/132054/los-medios-digitales-y-las-nuevas-formas-de-producir-contenidos>

Espinoza, D. (2020). Diseño y ejecución de arquitectura de descarga, modelamiento y análisis de datos para ampliar servicios en una empresa de tecnologia. *Tesis Pregrado*. Universidad de Chile, Santiago de Chile, Chile. Recuperado el 05 de Julio de 2021, de <http://repositorio.uchile.cl/handle/2250/176838>

Estevez, J. (2017). Ciudades inteligentes y datos abiertos: un dashboard basado en minería de datos. *Trabajo de grado*. Universidad Catolica de Colombia, Bogota, Colombia. Recuperado el 19 de Julio de 2021, de <http://hdl.handle.net/10983/15407>

Facturas Rápidas Ecuador. (21 de diciembre de 2020). *Tipos de contribuyentes*. Obtenido de Facturas Rápidas Facturación Electrónica Ecuador: <https://facturasrapidasec.com/tipos-de-contribuyentes/>

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996, March 15). From Data Mining to Knowledge Discovery in Databases. *AI Magazine*, 17(3). doi:<https://doi.org/10.1609/aimag.v17i3.1230>

Ferrer, J. (2014). *Aplicaciones web*. Madrid, España: RA-Ma Editorial. Recuperado el 01 de Agosto de 2021, de <https://elibro.net/es/lc/uguayaquil/titulos/106407>

- Ferri, C., & Ramírez, M. (2004). *Introducción a la minería de datos*. Madrid, España: PEARSON EDUCACIÓN. Recuperado el Agosto 09, 2021, de <https://elibro.net/es/lc/uguayaquil/titulos/45314>
- García, D. (03 de diciembre de 2020). *Power BI vs Tableau, ¿Qué herramienta de BI elegir?* Obtenido de Bitec: <https://www.bitec.es/noticias-bitec/power-bi-vs-tableau-que-herramienta-de-bi-elegir/>
- Hanretty, C. (2013). *Scraping the web for arts and humanities*. Norwich, Inglaterra: University of east anglia. Recuperado el 26 de July de 2021, de https://chasegoingdigital.files.wordpress.com/2013/02/scraping_book.pdf
- Haro, V. (2018). Diseño e implementacion de un dashboard de soporte academico basado en datos de entornos virtuales de aprendizaje. *Trabajo de master*. Universidad Politecnica de Valencia, Valencia, España. Recuperado el 18 de Julio de 2021, de <http://hdl.handle.net/10251/111761>
- Hernandez Sampieri, R. (2014). *Metodologia de la investigacion* (6 ed.). Ciudad de Mexico, Mexico: MCGRAW-HILL / INTERAMERICANA EDITORES, S.A. DE C.V. Obtenido de <https://www.uca.ac.cr/wp-content/uploads/2017/10/Investigacion.pdf>
- Hernandez, C., & Dueñas, M. (noviembre de 2009). Hacia una metodologia de gestión del conocimiento basada en minería de datos. *Repositorio Institucional Universidad Inca Garcilaso de la vega*, 79-95. Obtenido de <http://hdl.handle.net/20.500.11818/982>
- Imperva. (13 de Julio de 2016). *Web Scraping*. Obtenido de Imperva: <https://www.imperva.com/learn/application-security/web-scraping-attack/?redirect=Distil>
- Import.io. (13 de may de 2020). *Mission-critical web data*. Obtenido de Import.io: <https://www.import.io>
- Instituto La Salle Florida. (2015). *Los Medios Digitales de Comunicación*. Recuperado el 01 de Agosto de 2021, de Calameo: <https://es.calameo.com/books/00388050233cfe31de969>
- Instituto Nacional de Estadísticas y Censos. (junio de 2012). *Clasificacion Nacional de Actividades Economicas*. Obtenido de ecuadorencifras.gob.ec: <https://aplicaciones2.ecuadorencifras.gob.ec/SIN/metodologias/CIU%204.0.pdf>

- Korporate Technologies Group. (27 de septiembre de 2017). *Causas que impiden que su negocio sea rentable*. Obtenido de Grupo Korporate: <https://grupokorporate.com/causas-que-impiden-que-su-negocio-sea-rentable/>
- Lasio, V., Amaya, A., Zambrano, J., & Ordeñana, X. (2020). *Global Entrepreneurship Monitor Ecuador 2019/2020*. Guayaquil: ESPAE, ESPOL. Obtenido de https://www.espae.edu.ec/wp-content/uploads/2021/02/GEM_Ecuador_2019.pdf
- Lasio, V., Ordeñana, X., Caicedo, G., Samaniego, A., & Izquierdo, E. (2018). *Global Entrepreneurship Monitor Ecuador 2017*. Guayaquil: ESPAE - ESPOL. Obtenido de <https://www.gemconsortium.org/file/open?fileId=50078>
- Lema, A. (2016). Implementación de un dashboard para la generación de indicadores de inserción laboral y competencias de graduados de la Carrera de Medicina de la Universidad Central del Ecuador. *Tesis de grado*. Universidad Central del Ecuador, Quito, Ecuador. Obtenido de <http://www.dspace.uce.edu.ec/bitstream/25000/6068/1/T-UCE-0011-259.pdf>
- Libia, R. (obtucre de 2015). *UTILIZACION DE SOFTWARE LIBRE EN LA*. Obtenido de Agencia de regulacion y conrtrol.
- Lopez, J. (2015). Desarrollo de una herramienta software para la extracción de datos sobre el rendimiento de la red electrica. *Trabajo de grado*. Universidad Carlos III de Madrid, Getafe, España. Recuperado el 30 de Julio de 2021
- López, J. (12 de 01 de 2018). Web scraping. *Academia*. Recuperado el 26 de 07 de 2021, de https://www.academia.edu/35895308/Web_scraping
- Lozano Gomez, J. (19 de Mayo de 2020). *Web scraping con Python. Extraer datos de una web. Guía de inicio de Beautiful Soup*. Recuperado el 05 de Agosto de 2021, de J2logo: <https://j2logo.com/python/web-scraping-con-python-guia-inicio-beautifulsoup/>
- Lozano, J. (2020). Implementación de una solución Business Intelligence para apoyar a la toma de decisiones en la empresa Agro Micro Biotech Sac. *Trabajo de grado*. Universidad Privada Antenor Ortega, Trujillo, Peru. Recuperado el 19 de Julio de 2021, de <https://hdl.handle.net/20.500.12759/5591>
- Marres, N., & Weltevrede, E. (2013). Scraping the social? Issues in live social research. *Journal of Cultural Economy*.

- Mitchell, R. (2013). *Instant Web Scraping with Java*. Birmingham, The United Kingdom: Packt Publishing. Retrieved from <http://docplayer.net/24009874-Instant-web-scraping-with-java.html>
- Mousinho, A. (14 de Noviembre de 2019). *Conoce el proceso de Web Scraping y por qué es importante en una estrategia digital*. Recuperado el 03 de Agosto de 2021, de rock content: <https://rockcontent.com/es/blog/web-scraping/>
- Octoparse. (2021, january 25). *Use Octoparse to Download Web Data Easily - User Guide*. Retrieved august 08, 2021, from Octoparse: <https://www.octoparse.com/blog/what-is-octoparse>
- Ordoñez Rivas, L. (17 de Febrero de 2021). *Codigo Organico Integral Penal, COIP*. Obtenido de Ministerio de Defensa Nacional del Ecuador: https://www.defensa.gob.ec/wp-content/uploads/downloads/2021/03/COIP_act_feb-2021.pdf
- Organización Mundial del Comercio. (2016). Informe sobre el comercio mundial 2016. *Organizacion mundial del comercio*, 9. Obtenido de https://www.wto.org/spanish/res_s/booksp_s/world_trade_report16_s.pdf
- Pacheco, F. (2014). *Criptografia desde los sistemas clasicos hasta el futuro de la privacidad*. Buenos Aires, Argentina: Fox Andina. Recuperado el 02 de Agosto de 2021
- ParseHub. (09 de february de 2015). *Extract any data from the web. Easily*. Recuperado el 07 de august de 2021, de Parsehub: <https://www.parsehub.com/intro>
- Raffino, M. (21 de Enero de 2021). *Periodico*. Recuperado el 01 de Agosto de 2021, de concepto.de: <https://concepto.de/periodico/>
- Riquelme, J., Ruiz, R., & Gilbert, K. (2006). Minería de Datos: Conceptos y Tendencias. *Inteligencia Artificial: Revista Iberoamericana de Inteligencia Artificial*, 10(29), 11-18. Obtenido de <http://hdl.handle.net/11441/43290>
- Ron, M. (14 de noviembre de 2019). *Web Scraping*. Obtenido de DerechoECuador.com: https://www.derechoecuador.com/web-scraping#_ftn3
- Sanchez, A., Durán, M., Ballesteros-Ricaurte, J., & Gonzáles-Amarillo, A. (Diciembre de 2020). ScraCOVID-19: Plataforma informativa de contenido digital mediante Scraping y almacenamiento NoSQL. *Revista científica CUC*, 16(2). doi:<https://doi.org/10.17981/ingecuc.16.2.2020.18>

- Sanchez, D. (2020). Plataforma de recomendacion de habilidades tecnologicas segun puesto de trabajo para profesionales de TI, en funcion de la demanda en las bolsas de trabajo digitales. *Repositorio institucional Universidad de Lima*. Universidad de Lima, Lima, Peru. Recuperado el 08 de Julio de 2021, de <https://hdl.handle.net/20.500.12724/12351>
- Saurkar, A., Pathare, K., & Gode, S. (2018, April). An overview on web scraping techniques and tools. *International Journal on Future Revolution in Computer*. Retrieved August 03, 2021, from <http://www.ijfrcsce.org/index.php/ijfrcsce/article/view/1529/1529>
- Scrapy. (2021, april 07). *Scrapy at a glance*. Retrieved from Scrapy: <https://docs.scrapy.org/en/latest/intro/overview.html>
- Sim, L., Ban, K., Tan, T., Sethi, S., & Loh, T. (24 de february de 2017). Development of a clinical decision support system for diabetes care: A pilot study. *PLOS ONE*. doi:<https://doi.org/10.1371/journal.pone.0173021>
- Sumathi, S., & Sivanandam, S. (2006). *Introduction to data mining and its aplications*. (Vol. 29). Heidelberg, Germany: Springer-Verlag. Retrieved from <https://doc.lagout.org/Others/Data%20Mining/Introduction%20to%20Data%20Mining%20and%20its%20Applications%20%5BSumathi%20%26%20Sivanandam%202006-11-14%5D.pdf>
- Tas, O., & Togay, S. (2014). Efectos de la dolarización oficial en una pequeña economía abierta: El caso de Ecuador. *Investigación Económica*, 73, 68. Obtenido de <https://reader.elsevier.com/reader/sd/pii/S0185166715300084?token=E617E520F02A99BEA9D3107B884CAF736F791D2D19D9BA3138CA8EA11C68EE103C92647DEEA2803F486B0AFB854CBF18&originRegion=us-east-1&originCreation=20210909205151>
- Vergara, F. (29 de diciembre de 2017). *Ley Organica del Sistema Nacional de Registro de Datos Pubicos*. Obtenido de Gob.ec: <https://www.gob.ec/sites/default/files/regulations/2018-10/LEY%20SINARDAP.pdf>
- Vicente, N. (2018). *HabScraper: herramienta automatizada para la extracción de datos con web scraping*. Recuperado el 04 de 07 de 2021, de Repositorio Institucional UIB: <http://hdl.handle.net/11201/151095>
- Web Scraper. (09 de December de 2016). *Making web data extraction easy and accessible for everyone*. Obtenido de Web Scraper: <https://webscraper.io>

Zofio, J. (2013). *Aplicaciones web*. Madrid, España: Macmillan Iberia, S.A. Recuperado el 01 de Agosto de 2021, de <https://elibro.net/es/lc/uguayaquil/titulos/43262>