INSTITUTO TECNOLÓGICO Y DE ESTUDIOS SUPERIORES DE OCCIDENTE

CIFOVIS

Identidades e Inclusión Social

PROYECTO DE APLICACIÓN PROFESIONAL (PAP)

Programa de Desarrollo Local y Fortalecimiento del Tejido Social I

ITESO, Universidad Jesuita de Guadalajara

2E11A BOSQUE ESCUELA

Ciencia de Datos en el Bosque La Primavera

PRESENTAN

Ingeniería en biotecnología – Kairos Noemi Castañeda Andrade

Ingeniería y Ciencia de Datos – Diego Canales Morales

Profesor(es) PAP: Juan Fernando Escobar Ibáñez, Luis Rául Martínez García

Tlaquepaque, Jalisco, mayo de 2026

ÍNDICE

REPORTE PAP

Presentación Institucional de los Proyectos de Aplicación Profesional

Los Proyectos de Aplicación Profesional (PAP) son experiencias socio-profesionales de los alumnos que desde el currículo de su formación universitaria- enfrentan retos, resuelven problemas o innovan una necesidad sociotécnica del entorno, en vinculación (colaboración) (co-participación) con grupos, instituciones, organizaciones o comunidades, en escenarios reales donde comparten saberes.

El PAP, como espacio curricular de formación vinculada, ha logrado integrar el Servicio Social (acorde con las Orientaciones Fundamentales del ITESO), los requisitos de dar cuenta de los saberes y del saber aplicar los mismos al culminar la formación profesional (Opción Terminal), mediante la realización de proyectos profesionales de cara a las necesidades y retos del entorno (Aplicación Profesional).

El PAP es un proceso acotado en el tiempo en que estudiantes, actores sociales y profesores se asocian colaborativamente y en red, en un proyecto, e incursionan en un mundo social, como actores que enfrentan verdaderos problemas y desafíos traducibles en demandas pertinentes y socialmente relevantes. Frente a éstas transfieren experiencia de sus saberes profesionales y demuestran que saben hacer, innovar, co-crear o transformar en distintos campos sociales.

El PAP trata de sembrar en los estudiantes una disposición permanente de encargarse de la realidad con una actitud comprometida y ética frente a las disimetrías sociales. En otras palabras, se trata del reto de "saber y aprender a transformar".

El Reporte PAP consta de tres componentes:

El primer componente refiere al ciclo participativo del PAP, en donde se documentan las diferentes fases del proyecto y las actividades que tuvieron lugar durante el desarrollo de éste y la valoración de las incidencias en el entorno.

El segundo componente presenta los productos elaborados de acuerdo con su tipología.

El tercer componente es la reflexión crítica y ética de la experiencia, el reconocimiento de las competencias y los aprendizajes profesionales que el estudiante desarrolló en el transcurso de su labor.

Resumen

Durante el semestre de primavera 2026 en el proyecto de aplicación profesional Bosque Escuela (PAP2E11A), en el Programa de desarrollo local y fortalecimiento del tejido social I, tuvimos como propósito realizar una revisión de literatura en torno al bosque La Primavera para identificar los temas predominantes y tendencias de investigación con el objetivo de generar una base de datos que permitiera analizar rápidamente los temas centrales y vacíos de información sobre el BLP. Para ello se utilizó un pipeline automatizado de extracción de datos con la API de Google Scholar y Procesamiento de Lenguaje Natural (NLP) con el modelo "es_core_news_lg" de spaCy para la minería de textos en documentos PDF. Encontramos 54 artículos científicos que contenían las palabras clave programadas, pero después de analizarlos solo 19 hablaban específicamente del BLP. Los temas más abordados son: 1) flora, 2) conservación del bosque. Como productos obtuvimos una base de datos estructura, es decir, la información extraída de los 54 artículos se organizó automáticamente en columnas de la siguiente manera: Título, Autores, Año, Publicado en, Link, Resumen, Citas, Ruta Local PDF, Tema_Central, Localización_Geografica y Resultados_Sintesis. Además, desarrollamos un programa en Python para automatizar dicho proceso, y una carpeta en OneDrive que contiene los 54 artículos estudiados.

1. Ciclo participativo del Proyecto de Aplicación Profesional

El PAP es una experiencia de aprendizaje y de contribución social integrada por estudiantes, profesores, actores sociales y responsables de las organizaciones, que de manera colaborativa aplican sus conocimientos y generan alternativas para dar respuesta a problemáticas de un contexto específico y en un tiempo delimitado. Por tanto, la experiencia PAP supone un proceso en lógica de proyecto, así como de un estilo de trabajo participativo y recíproco entre los involucrados.

1.1 Entendimiento del ámbito y del contexto

Estamos en medio de una crisis socioambiental muy importante derivada del sistema socioeconómico capitalista, así como de la pérdida de hábitat, la urbanización y

contaminación. Entre los factores más alarmantes se encuentra la gran cantidad de especies extintas y de especies amenazadas, que han llevado a considerar el inicio de la sexta extinción masiva de especies en la historia del planeta. Las afectaciones a los ecosistemas y la pérdida de biodiversidad representa graves impactos negativos en el bienestar humano por la disminución de los servicios ecosistémicos, es decir, la pérdida de biodiversidad afecta la producción de alimentos, disponibilidad de agua y regulación del clima, lo que genera riesgos en la salud, la economía y la estabilidad social (Ceballos et al., 2020).

Ante esta pérdida de biodiversidad y los impactos en el correcto funcionamiento de los ecosistemas, las áreas protegidas surgen como estrategia para conservar la biodiversidad, ya que protegen a la naturaleza y los recursos que hay en ellas. De esta forma, las diferentes modalidades de áreas protegidas (p.ej. parques nacionales, reservas naturales, zonas salvajes, áreas de conservación comunitaria) y su biodiversidad proporcionan alimentos, agua potable, medicamentos y protección frente a los impactos de los desastres naturales. Además tienen un papel muy importante en la mitigación del cambio climático, pues las áreas protegidas a nivel mundial almacenan al menos el 15% del carbono. Sin embargo, existen limitaciones y retos como, la falta de evaluación de su efectividad, problemas en la planeación y en el monitoreo, además de que se necesita mejorar la gobernanza y gestión, en otras palabras, las áreas protegidas son fundamentales, pero su éxito depende de cómo se gestionan y evalúan (IUCN, 2026).

Según la Comisión Nacional de Áreas Naturales Protegidas en México (2025), las Áreas Naturales Protegidas (ANP) son las herramientas más efectivas para conservar los ecosistemas, permitir la adaptación de la biodiversidad y enfrentar los efectos del cambio climático. En México actualmente se administran 232 ANP de carácter federal 29 entidades federativas que cubren 98, 000, 719 hectáreas y existen 602 Áreas Destinadas Voluntariamente a la Conservación con una superficie de 1, 301, 037. 94 hectáreas, lo que en conjunto representan el 11.76% del territorio terrestre y el 23.78% de territorio marino (Fig. 1).

Figura 1. Áreas Naturales Protegidas en México (Comisión Nacional de Áreas Naturales Protegidas, 2025).

En el estado de Jalisco se encuentra La Primavera, un área de Protección de Flora y Fauna que fue decretada como tal en 199X y que cubre una superficie de 30, 500 has en los municipios de Zapopan, Tlajomulco de Zúñiga, Tala y El Arenal. Según el Gobierno de México (2026) cuenta con 742 especies de flora, 7 especies de peces, 49 especies de reptiles, 20 especies de anfibios, 200 especies de aves y 59 especies de mamíferos. Este espacio tiene gran importancia ambiental, debido a que alberga una importante cantidad de especies de flora y fauna, además predominan ecosistemas como bosques de encinos y pinos, selvas y vegetación hidrófila. Por otro lado, regula y amortigua el clima y es una zona de captación hídrica por sus manifestaciones termales, como el río caliente, balneario de los volcanes, cerritos colorados, planillas y algunas fumarolas dispersas, cuenta con dos yacimientos geotérmicos del subsuelo, uno a 600 metros y otro a 2,000 mil metros (CONANP, s. f.).

Esta ANP se ve amenazada constantemente por el crecimiento del Área Metropolitana de Guadalajara, pues invaden zonas rurales y forestales cercanas al bosque, lo que provoca el cambio de uso de suelo, fragmentación del ecosistema y presión sobre el bosque, también

tierras agrícolas están siendo desplazas por la urbanización y transformadas en prácticas más intensivas generando el deterioro del suelo y del agua. Otras amenazas son el aumento de actividades turísticas dentro del bosque, modificación del entorno natural y presión sobre los recursos ecológicos (Rios-Llamas & Hernández-Vázquez, 2023).

Marco Conceptual

Para el desarrollo del presente trabajo se consideran algunos conceptos clave que permiten comprender el enfoque técnico y metodológico utilizado, iniciando con el concepto de ciencia de datos, este es un campo interdisciplinario donde la estadística, la programación y el análisis computacional se combinan para extraer conocimiento útil a partir de grandes volúmenes de información (IBM, 2026). La ciencia de datos en este proyecto se aplica para transformar información científica dispersa sobre el BLP en un dataset estructurado, para lograr analizar patrones, organizar el conocimiento existente y facilitar la toma de decisiones basada en datos.

Otro concepto importante por mencionar es el Procesamiento de Lenguaje Natural (NLP), es una rama de la inteligencia artificial que permite a los sistemas informáticos interpretar, analizar y procesar lenguaje humano en formato textual (Stryker & Holdsworth, 2026). En este proyecto se utiliza el NLP para identificar entidades relevantes dentro de los artículos científicos, como ubicaciones geográficas, especies o temas de investigación, permitiendo estructurar automáticamente la información contenida en los documentos.

Finalmente, tenemos el concepto de datos estructurados, estos consisten en información organizada en formatos definidos, como tablas o bases de datos, lo que facilita su análisis, consulta y procesamiento automático (Jonker & Gomstyn, 2026).

1.2 Caracterización de la organización

El escenario de intervención para este proyecto es Bosque Escuela, una iniciativa estratégica dentro del PAP Programa de Desarrollo Local y Fortalecimiento del Tejido Social I. Esta organización funciona como un espacio de gestión del conocimiento y educación ambiental, cuyo objetivo primordial es la protección y restauración del Bosque La Primavera (BLP) mediante la vinculación académica y comunitaria.

En este contexto, la organización requiere de herramientas tecnológicas avanzadas para procesar la vasta, pero dispersa, información científica generada sobre el área protegida. Mi participación como estudiante de la ingeniería en Ingeniería y Ciencia de Datos se alinea con la necesidad de la organización de modernizar sus capacidades de análisis de información. El equipo de trabajo está conformado por:

- **Diego Canales Morales:** Responsable del área de Ciencia de Datos. Mi rol principal consistió en el diseño y ejecución de la arquitectura técnica para la recuperación y estructuración de evidencia científica. Esto incluyó el desarrollo de algoritmos de *web scraping* y modelos de Procesamiento de Lenguaje Natural (NLP) para transformar documentos PDF en activos de información consultables.
- **Kairos Noemi Castañeda Andrade:** Responsable de la creación del documento RPAP, redacción del resumen y puntos 1.1, 1.3, 1.4, 1.6, 3, y 3.1.
- **Asesores del PAP:** Los profesores Juan Fernando Escobar Ibáñez y Luis Raúl Martínez García, quienes supervisaron la pertinencia técnica y la orientación social del proyecto hacia el fortalecimiento del tejido social y el cuidado del territorio.

1.3 Identificación de la(s) problemática(s)

Existe una falta de estructuración en la información científica disponible sobre el BLP. A pesar de la gran producción científica, los datos sobre localización de especies, resultados de investigaciones y periodos de estudio se encuentran en formatos PDF no estructurados, lo que impide un análisis territorial rápido y una toma de decisiones informada basada en evidencia científica integrada.

1.4. Planeación de alternativa(s), objetivo(s) particular(es) y formalización de acuerdos de colaboración

De tal manera que, nuestro objetivo es desarrollar e implementar un pipeline automatizado de minería de datos y procesamiento de lenguaje natural (NLP) para extraer, limpiar y

estructurar información técnica de artículos científicos sobre el BLP, transformando documentos PDF en un dataset analítico. Mientras que nuestros objetivos específicos son: automatizar la recuperación de documentos, estructurar información mediante NLP, sintetizar contenidos técnicos y, finalmente, analizar el dataset.

La estrategia metodológica para alcanzar los objetivos del proyecto se diseñó como un proceso evolutivo de ingeniería de datos, comenzando con una fase de configuración técnica donde se establecieron las herramientas de Python y los controladores de navegación necesarios para interactuar con la web de forma automatizada. Una vez listo el entorno, la fase operativa inició con la ejecución de búsquedas especializadas mediante una API conectada a Google Scholar, utilizando términos específicos como "Bosque La Primavera" Jalisco para garantizar que el sistema recuperara la mayor cantidad de metadatos y documentos PDF posibles.

Tras obtener este volumen inicial de información, el trabajo se centró en un proceso riguroso de curaduría y limpieza, donde validamos la integridad de cada archivo y su relevancia temática; esto fue fundamental para reducir el ruido informativo, logrando filtrar el universo inicial de 54 documentos hasta obtener una muestra final de 19 artículos que hablaban exclusivamente del bosque. Con este corpus depurado, implementamos el procesamiento de lenguaje natural (NLP) utilizando el modelo es_core_news_lg de spaCy, lo que permitió que el sistema realizara un análisis semántico profundo para reconocer automáticamente lugares geográficos y términos técnicos de flora y fauna sin depender de listas cerradas.

Finalmente, el proceso cerró con el uso de expresiones regulares y ventanas de contexto para segmentar las secciones de resultados y conclusiones de cada obra. Esto nos permitió sintetizar los hallazgos de cada autor en bloques de texto coherentes, consolidando toda la información en un dataset analítico estructurado que facilita la identificación de tendencias y problemáticas actuales en el territorio del Bosque La Primavera.

1.5. Desarrollo de la propuesta de mejora

El desarrollo de la propuesta se estructuró mediante la implementación de un pipeline de procesamiento de datos de extremo a extremo, diseñado para transformar literatura científica no estructurada en un activo de información analítico. El proceso se dividió en las siguientes fases operativas:

A Adquisición y Recuperación Automatizada

La fase inicial se enfocó en la recolección masiva de registros bibliográficos mediante la API de Google Scholar proporcionada por la página SerpApi. Para garantizar la precisión y exhaustividad de la búsqueda, se programaron scripts en Python utilizando Selenium WebDriver, empleando los siguientes criterios de búsqueda (queries):

- "Bosque La Primavera" Jalisco
- "Bosque de la Primavera" Guadalajara
- "Bosque La Primavera" Zapopan

Esta automatización permitió navegar de forma autónoma por portales institucionales, identificando enlaces permanentes (handles) y ejecutando descargas físicas de archivos PDF, superando las limitaciones de metadatos incompletos en la web.

B Depuración y Curaduría del Corpus

Una vez obtenidos los archivos, se procedió a una validación de integridad técnica para confirmar la legibilidad de cada documento. Se aplicó un filtro de relevancia temática estricto, en el cual, de un universo inicial de 54 documentos, se identificó que solo 19 artículos constituían investigaciones centradas específicamente en el BLP. Esta depuración fue esencial para eliminar el ruido informativo de proyectos ajenos al contexto territorial de interés.

C Minería de Texto y Procesamiento de Lenguaje Natural (NLP)

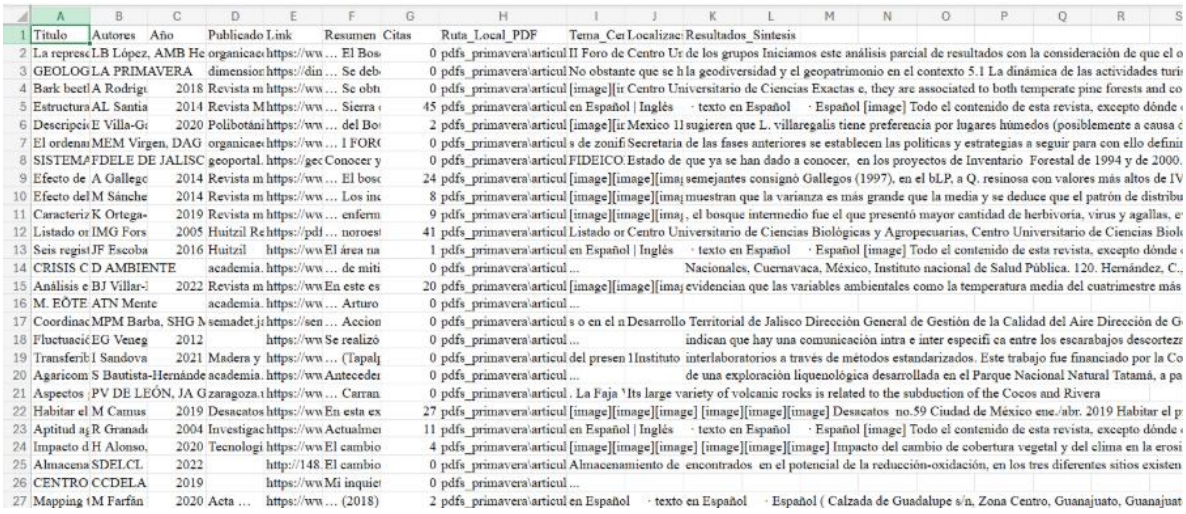
Para el análisis profundo, se implementó el modelo de lenguaje "es_core_news_lg" de spaCy. Al ser un modelo robusto (Large), permitió realizar una minería de texto abierta y semántica sobre el cuerpo completo de los documentos. El sistema reconoció automáticamente entidades geográficas y términos técnicos de flora y fauna, utilizando contadores de frecuencia para determinar los temas predominantes basándose en la repetición textual.

D **Extracción de Contexto mediante Expresiones Regulares (Regex)**

Finalmente, se emplearon expresiones regulares combinadas con ventanas de contexto para segmentar los documentos y localizar secciones críticas como resúmenes, objetivos y conclusiones. Se configuró el script para capturar bloques de hasta 1,000 caracteres, permitiendo obtener una síntesis coherente de los resultados y tendencias reportadas (positivas o negativas), así como identificar las especies que reciben mayor atención académica actualmente.

1.6. Valoración de productos, resultados e impactos

A continuación, se muestra el dataset (base de datos) obtenido por medio del pipeline automatizado capaz de recuperar, procesar y estructurar información proveniente de artículos en formato PDF. Este dataset organizó los datos de cada artículo en columnas de la siguiente manera: Título, Autor(es), Año, Publicado en, Enlace a la publicación, Resumen, Citas, Ruta Local PDF, Tema_Central, Localización_Geografica y Resultados_Sintesis (Ilustración 2), en la parte de anexos se puede acceder a esta información (Anexo 1).



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	Titulo	Autores	Año	Publicado	Link	Resumen	Citas	Ruta_Local_PDF	Tema_Cer	Localizac	Resultados_Síntesis								
2	La represc	LB López, AMB He		organicae	https://ww	… El Bos	0	pdfs_primavera\articul	II Foro de Centro Ur		de los grupos Iniciamos este análisis parcial de resultados con la consideración de que el o								
3	GEOLOG	LA PRIMAVERA		dimensior	https://din	… Se deb	0	pdfs_primavera\articul	No obstante que se h		la geodiversidad y el geopatrimonio en el contexto 5.1 La dinámica de las actividades turí								
4	Bark beetl	A Rodrigu	2018	Revista m	https://ww	… Se obt	0	pdfs_primavera\articul	[image][ir	Centro Universitario de Ciencias Exactas e, they are associated to both temperate pine forests and co									
5	Estructura	AL Santia	2014	Revista M	https://ww	… Sierra	45	pdfs_primavera\articul	en Español | Inglés	· texto en Español	· Español [image] Todo el contenido de esta revista, excepto dónde								
6	Descripció	E Villa-G	2020	Polibotáni	https://ww	… del Bo	2	pdfs_primavera\articul	[image][ir	Mexico 11	sugieren que L. villaregalis tiene preferencia por lugares húmedos (posiblemente a causa d								
7	El ordena	MEM Virgen, DAG		organicae	https://ww	… 1 FOR	0	pdfs_primavera\articul	s de zonifi	Secretaría de las fases anteriores se establecen las políticas y estrategias a seguir para con ello defini									
8	SISTEMA	FDELE DE JALISC		geoportal.	https://gec	Conocer y	0	pdfs_primavera\articul	FIDEICO	Estado de que ya se han dado a conocer, en los proyectos de Inventario Forestal de 1994 y de 2000.									
9	Efecto de	A Gallego	2014	Revista m	https://ww	… El bosc	24	pdfs_primavera\articul	[image][image][ima	semejantes consignó Gallegos (1997), en el bLP, a Q. resinosa con valores más altos de IV									
10	Efecto del	M Sánche	2014	Revista m	https://ww	… Los inc	8	pdfs_primavera\articul	[image][image][ima	muestran que la varianza es más grande que la media y se deduce que el patrón de distribu									
11	Caracteriz	K Ortega-	2019	Revista m	https://ww	… enferm	9	pdfs_primavera\articul	[image][image][ima	, el bosque intermedio fue el que presentó mayor cantidad de herbivoría, virus y agallas, e									
12	Listado or	IMG Fors	2005	Huitzil Re	https://pdf	… noroes	41	pdfs_primavera\articul	Listado or	Centro Universitario de Ciencias Biológicas y Agropecuarias, Centro Universitario de Ciencias Biol									
13	Seis regist	JF Escoba	2016	Huitzil	https://ww	El área na	1	pdfs_primavera\articul	en Español | Inglés	· texto en Español	· Español [image] Todo el contenido de esta revista, excepto dónde								
14	CRISIS C	D AMBIENTE		academia.	https://ww	… de miti	0	pdfs_primavera\articul	…		Nacionales, Cuernavaca, México, Instituto nacional de Salud Pública. 120. Hernández, C.,								
15	Análisis e	BJ Villar-I	2022	Revista m	https://ww	En este es	20	pdfs_primavera\articul	[image][image][ima	evidencian que las variables ambientales como la temperatura media del cuatrimestre más									
16	M. EÖTE	ATN Mente		academia.	https://ww	… Arturo	0	pdfs_primavera\articul	…										
17	Coordinac	MPM Barba, SHG M		semadet.j	https://sen	… Accion	0	pdfs_primavera\articul	s o en el n	Desarrollo Territorial de Jalisco Dirección General de Gestión de la Calidad del Aire Dirección de G									
18	Fluctuació	EG Veneg	2012		https://ww	Se realizó	0	pdfs_primavera\articul	…		indican que hay una comunicación intra e inter específi ca entre los escarabajos descortez								
19	Transferib	I Sandova	2021	Madera y	https://ww	… (Tapalp	0	pdfs_primavera\articul	del presen	Instituto	interlaboratorios a través de métodos estandarizados. Este trabajo fue financiado por la Co								
20	Agaricom	S Bautista-Hernánde		academia.	https://ww	Anteceder	0	pdfs_primavera\articul	…		de una exploración liquenológica desarrollada en el Parque Nacional Natural Tatamá, a pa								
21	Aspectos	PV DE LEÓN, JA G		zaragoza.u	https://ww	… Carran	0	pdfs_primavera\articul	. La Faja V	Its large variety of volcanic rocks is related to the subduction of the Cocos and Rivera									
22	Habitar el	M Camus	2019	Desacatos	https://ww	En esta ex	27	pdfs_primavera\articul	[image][image][image]	[image][image][image] Desacatos no.59 Ciudad de México ene./abr. 2019 Habitar el p									
23	Aptitud ag	R Granad	2004	Investigac	https://ww	Actualmer	11	pdfs_primavera\articul	en Español | Inglés	· texto en Español	· Español [image] Todo el contenido de esta revista, excepto dónde								
24	Impacto d	H Alonso,	2020	Tecnologí	https://ww	El cambio	4	pdfs_primavera\articul	[image][image][image]	[image][image][image] Impacto del cambio de cobertura vegetal y del clima en la erosi									
25	Almacena	SDELCL	2022		http://148.	El cambio	0	pdfs_primavera\articul	Almacenamiento de	encontrados en el potencial de la reducción-oxidación, en los tres diferentes sitios existen									
26	CENTRO	CCDELA	2019		https://ww	Mi inquie	0	pdfs_primavera\articul	…										
27	Mapping t	M Farfán	2020	Acta …	https://ww	… (2018)	2	pdfs_primavera\articul	en Español	· texto en Español	· Español (Calzada de Guadalupe s/n, Zona Centro, Guanajuato, Guanajuat								

Ilustración 2 Pipeline automatizado.

Al analizar la base de datos obtenida encontramos que de los 54 artículos descargados, 19 (35% del total) están enfocados en el BLP, mientras que el resto son de otras regiones de México y del mundo (Fig. 2).

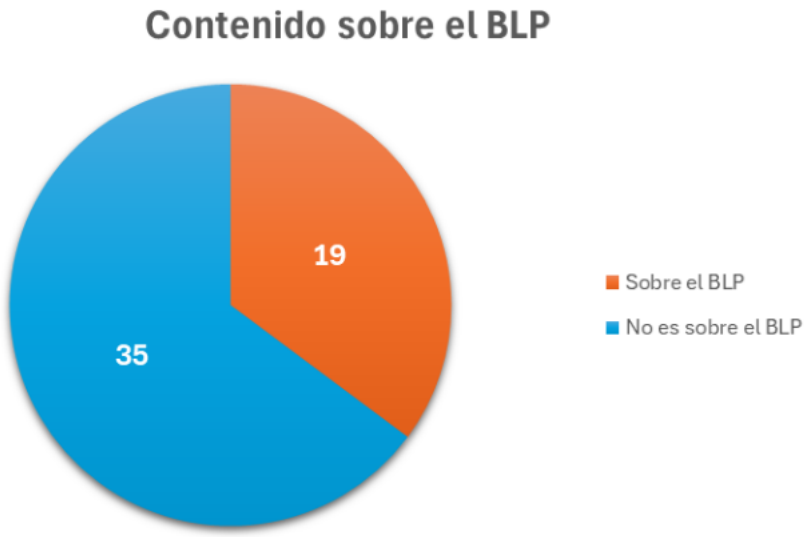


Figura 2. Contenido principal en cada uno de los artículos descargados.

Los temas centrales que fueron abordados en las publicaciones son: Incendios, Actividad humana, Flora, Fauna, Conservación, Manejo, Coberturas del suelo y Gestión del área protegida, siendo Conservación y Flora los más frecuentes (9 y 8, respectivamente; Fig. 3).

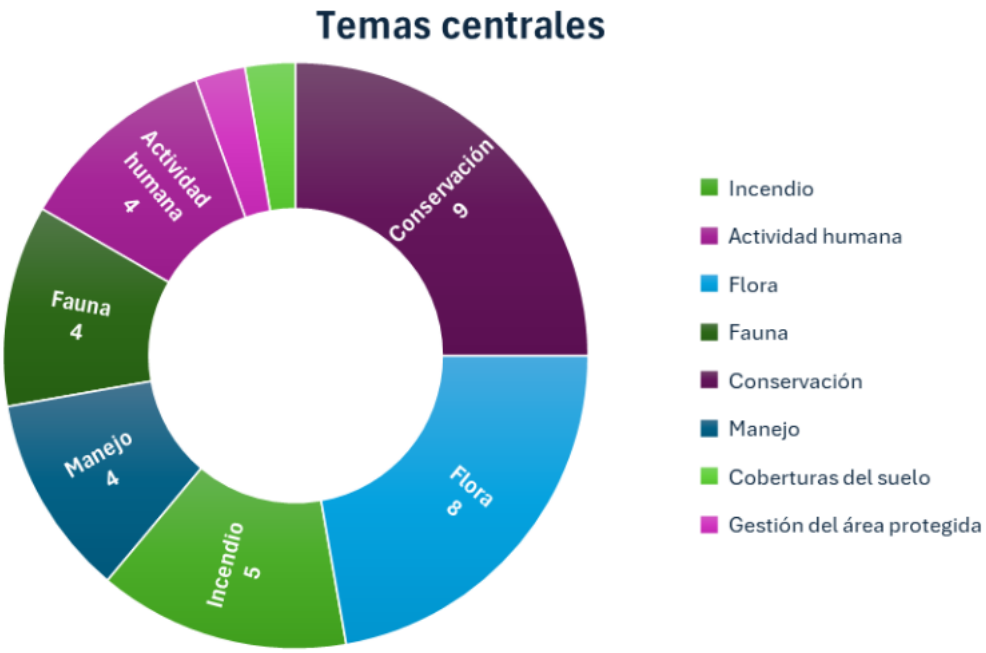


Figura 3. Temas centrales de los artículos que hablan sobre el BLP.

Al revisar con detalle el tema central de Actividad humana, se encontraron 3 registros sobre Actividades antropogénicas en el BLP y un registro sobre Urbanización (Fig. 4).

Figura 4. Subtemas del tema central de la Actividad humana.

Para el tema central de Flora, encontramos que los artículos hablaban de varios tipos de flora siendo el pino el más mencionado con 4 registros, mientras que plantas, árboles, distribución de especies vegetales y encino solo se mencionaban una vez (Fig. 5).

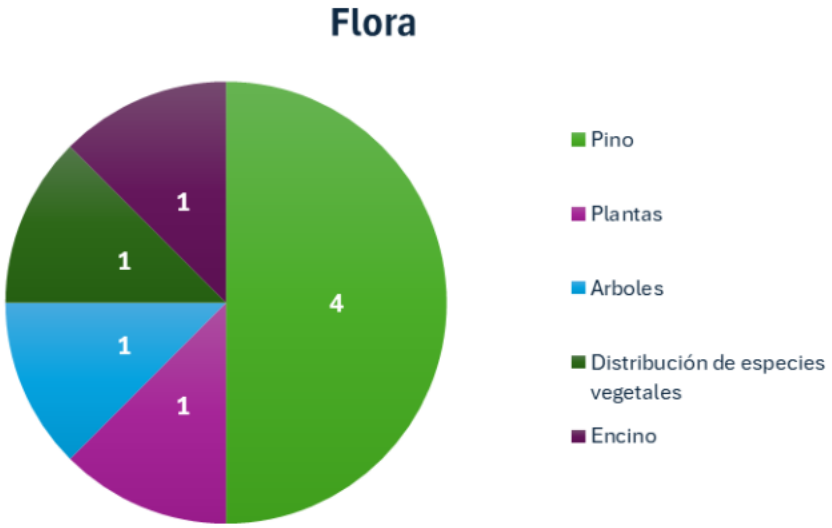


Figura 5. Subtemas en el tema central de Flora.

En el caso del tema central de Fauna, encontramos que se mencionaron diferentes especies por artículo, como aves, escarabajos, murciélagos y guajolote norteño (Fig. 6).

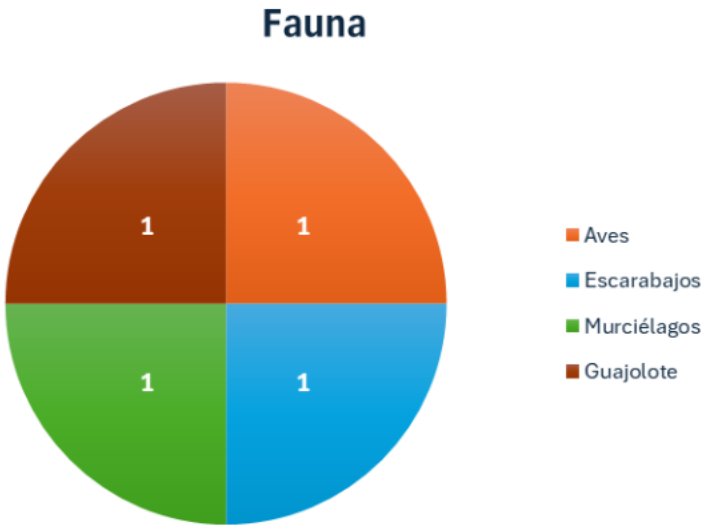


Figura 6. Subtemas en el tema central de Fauna.

Mientras que para el tema central de Conservación se dividía entre la conservación del bosque y la conservación del agua, notándose que la mayoría de los artículos hablaban sobre la conservación del bosque (Fig. 7).

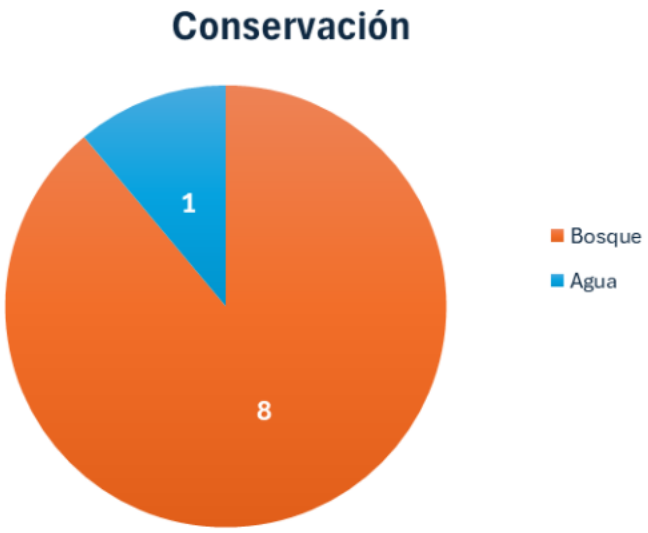


Figura 7. Subtemas en el tema central de Conservación.

Finalmente, para el tema central de Manejo, este tenía como subtemas el manejo forestal, del fuego y de especies, donde el subtema con mayores registros fue sobre el manejo forestal (Fig. 8)

Figura 8. Subtemas en el tema central de Manejo.

Descripción técnica del pipeline implementado

El pipeline automatizado se desarrolló íntegramente en Python y se estructuró en tres módulos secuenciales: (A) recuperación bibliográfica mediante la API de Google Scholar, (B) descarga física de documentos PDF filtrados por dominio académico, y (C) análisis profundo de texto mediante Procesamiento de Lenguaje Natural (NLP) con spaCy. A continuación, se describe el funcionamiento de cada módulo y se presenta el código fuente correspondiente.

Módulo A - Recuperación de artículos vía SerpApi (Google Scholar)

El primer módulo se conecta a la API de SerpApi para realizar consultas automatizadas en Google Scholar. Se definen tres queries de búsqueda que combinan distintos nombres del bosque con municipios clave, garantizando cobertura geográfica en los resultados. Por cada query se iteran páginas de 20 resultados hasta alcanzar un máximo de 300 registros, extrayendo metadatos como título, autores, año, revista, enlace, resumen y número de citas. Al finalizar, se eliminan duplicados con pandas (ya que un mismo artículo puede aparecer en múltiples búsquedas) y se exporta el dataset a un archivo CSV codificado en UTF-8 (Tabla 1).

Tabla 1. Script de recuperación automatizada de artículos desde Google Scholar mediante SerpApi

```
from serpapi import GoogleSearch
import pandas as pd
import time
import re
import requests
import os
import fitz
import spacy
from collections import Counter

API_KEY = "8b556fc5ca59..."  # Clave de autenticación SerpApi

# Términos de búsqueda especializados para el Bosque La Primavera
QUERIES = [
    '"Bosque La Primavera" Jalisco',
    '"Bosque de la Primavera" Guadalajara',
```

```
    '"Bosque La Primavera" Zapopan'
]

def limpiar_publicacion(pub_string):
    autores, revista, anio = "N/A", "N/A", "N/A"
    if pub_string:
        match_anio = re.search(r'\b(19|20)\d{2}\b', pub_string)
        if match_anio:
            anio = match_anio.group(0)
        partes = pub_string.split(' - ')
        if len(partes) >= 1:
            autores = partes[0]
        if len(partes) >= 2:
            revista = partes[1].replace(anio, '').strip(', ')
    return autores, revista, anio

def extraer_datos_profesional():
    resultados_totales = []

    for q in QUERIES:
        start = 0
        print(f'--- Iniciando búsqueda: {q} ---')

        while start < 300:  # Hasta 300 resultados por query
            params = {
                "engine": "google_scholar",
                "q": q,
                "api_key": API_KEY,
                "start": start,
                "num": 20,
                "hl": "es"
            }
            search = GoogleSearch(params)
            results = search.get_dict()
            organic = results.get("organic_results", [])

```

```
            if not organic:
                break

            for art in organic:
                pub_info = art.get("publication_info", {}).get("summary", "")
                autores, revista, anio = limpiar_publicacion(pub_info)
                resultados_totales.append({
                    "Título":      art.get("title"),
                    "Autores":     autores,
                    "Año":         anio,
                    "Publicado en": revista,
                    "Link":        art.get("link"),
                    "Resumen":     art.get("snippet"),
                    "Citas":       art.get("inline_links", {})
                                       .get("cited_by", {}).get("total", 0)
                })

            start += 20
            print(f'Progreso: {len(resultados_totales)} artículos
acumulados...')
            time.sleep(0.5)

    df = pd.DataFrame(resultados_totales)
    # Eliminar duplicados: mismo artículo puede aparecer en varias queries
    df = df.drop_duplicates(subset=['Título', 'Autores'])

    df.to_csv("investigaciones_bosque_primavera_final.csv",
              index=False, encoding="utf-8-sig")
    print(f'¡Éxito! {len(df)} artículos únicos exportados.')

if __name__ == "__main__":
    extraer_datos_profesional()
```

Módulo B – Descarga automatizada de PDFs por dominio académico

El segundo módulo lee el CSV generado en el paso anterior e intenta descargar el PDF de cada artículo. Para evitar descargar contenido irrelevante o bloqueado por derechos de autor,

se aplica un filtro de dominio que solo acepta URLs provenientes de repositorios académicos confiables (SciELO, Redalyc y ResearchGate) o enlaces que terminen directamente en .pdf. Cada descarga incluye un encabezado de User-Agent para simular un navegador real y evitar bloqueos por detección de bots. La ruta local de cada archivo descargado se registra en una nueva columna del DataFrame, y el dataset actualizado se exporta a un segundo CSV llamado investigaciones_con_rutas.csv, que sirve como entrada para el módulo de análisis NLP (Tabla 2).

Tabla 2. Script de descarga selectiva de documentos PDF desde repositorios académicos confiables.

```
def descargar_pdfs(csv_path, carpeta_destino='pdfs_primavera'):
    if not os.path.exists(carpeta_destino):
        os.makedirs(carpeta_destino)

    df = pd.read_csv(csv_path)
    df['Ruta_Local_PDF'] = os.path.join(os.getcwd(), carpeta_destino)

    for index, row in df.iterrows():
        url = row['Link']

        # Validar que el URL sea texto y no un valor NaN
        if not isinstance(url, str):
            print(f'Saltando índice {index}: Link vacío o inválido.')
            continue

        # Filtrar solo fuentes académicas confiables
        dominios_validos = ['scielo.org', 'redalyc.org', 'researchgate.net']
        if any(d in url for d in dominios_validos) or url.endswith('.pdf'):
            try:
                nombre_archivo = f'articulo_{index}.pdf'
                ruta_completa  = os.path.join(carpeta_destino, nombre_archivo)

                # Headers para evitar bloqueos por detección de bots
                headers  = {'User-Agent': 'Mozilla/5.0'}
                response = requests.get(url, timeout=15, headers=headers)

```

```
                if response.status_code == 200:
                    with open(ruta_completa, 'wb') as f:
                        f.write(response.content)
                    df.at[index, 'Ruta_Local_PDF'] = ruta_completa
                    print(f'Descargado {index}: {str(row["Título"])[:30]}...')

                time.sleep(1.5)  # Pausa para no saturar los servidores

            except Exception as e:
                print(f'Error en link {index}: {e}')

    nuevo_csv = 'investigaciones_con_rutas.csv'
    df.to_csv(nuevo_csv, index=False, encoding='utf-8-sig')
    return nuevo_csv

if __name__ == "__main__":
    archivo = "investigaciones_bosque_primavera_final.csv"
    print('Iniciando descarga de documentos del Bosque La Primavera...')
    ruta_actualizada = descargar_pdfs(archivo)
    print(f'Proceso terminado. Archivo generado: {ruta_actualizada}')
```

Módulo C - Minería de texto y análisis NLP con spaCy

El tercer módulo es el más complejo, ya que ejecuta el análisis lingüístico y semántico sobre el texto completo de cada PDF. El proceso inicia cargando el modelo de lenguaje es_core_news_lg de spaCy, que al ser un modelo Large ofrece mayor precisión en el reconocimiento de entidades nombradas (NER) y análisis morfosintáctico en español. Se amplía el límite de caracteres del modelo a 2,000,000 para poder procesar documentos extensos sin truncar el texto.

Para cada artículo, el texto se extrae página por página con la librería PyMuPDF (fitz) y se somete a una función de limpieza que elimina artefactos tipográficos (saltos de línea irregulares, fragmentos de metadatos editoriales como ISSN, URLs, números de volumen y referencias a derechos de autor). Una vez limpio, el texto se procesa con el modelo NLP, que permite identificar de manera automática cinco tipos de información:

- Localización geográfica (etiqueta LOC): el modelo detecta topónimos y nombres de lugares mencionados en el documento. Se aplica un filtro adicional para excluir falsos

positivos frecuentes en artículos científicos (palabras como "Universidad", "Figura" o "Tabla" que spaCy puede clasificar erróneamente como entidades geográficas). Los seis lugares con mayor frecuencia de aparición se registran como la localización del estudio.

- Especies y temas predominantes: se combinan entidades de tipo MISC y ORG con los sustantivos con mayor frecuencia de aparición en el documento. Esto permite capturar tanto nombres de especies y géneros taxonómicos como términos técnicos recurrentes del campo de estudio, sin requerir listas predefinidas.
- Tema central (Resumen/Abstract): una expresión regular busca el bloque de texto que sigue a encabezados como "Resumen", "Abstract" o "Introducción", capturando entre 150 y 800 caracteres. Si no se encuentra la sección, se toman los primeros 400 caracteres del documento como aproximación.
- Síntesis de resultados (Conclusiones/Discusión): otra expresión regular localiza la sección de cierre del artículo y extrae hasta 1,000 caracteres representativos, permitiendo identificar los hallazgos y tendencias reportadas por los autores.
- Año de estudio: se extraen todos los años (formato 19XX o 20XX) mencionados en el texto y se registra el más frecuente como el año de referencia del estudio, técnica que resulta robusta incluso cuando el año no aparece explícitamente en los metadatos.

Toda la información extraída se consolida en nuevas columnas del DataFrame original y se exporta al archivo analisis_exhaustivo_primavera.csv, que constituye el dataset analítico final del proyecto (Tabla 3).

Tabla 3. Script de análisis NLP con spaCy para extracción de entidades, temas y síntesis de resultados de cada documento.

```
import spacy, fitz, re, os, pandas as pd
from collections import Counter


# Cargar modelo Large para máxima precisión semántica
try:
    nlp = spacy.load("es_core_news_lg")
    nlp.max_length = 2_000_000  # Ampliar límite para artículos largos
    print("Modelo Large cargado correctamente.")
except:
    print(" Modelo Large no encontrado. Usando modelo base.")
```

```
    nlp = spacy.load("es_core_news_sm")

def limpiar_texto_profundo(texto):
    """Elimina ruido tipográfico manteniendo la estructura semántica."""
    texto = re.sub(r'\s+', ' ', texto)
    basura = ['download', 'free pdf', 'issn', 'http', 'www.',
              'vol.', 'página', 'copyright']
    for b in basura:
        texto = re.sub(rf'(?i).{{0,20}}{b}.{{0,20}}', '', texto)
    return texto.strip()

def mineria_calidad_total(ruta_csv):
    df = pd.read_csv(ruta_csv)
    # Filtrar solo archivos que existen físicamente en disco
    df = df[df['Ruta_Local_PDF'].apply(
        lambda x: os.path.isfile(str(x)) if pd.notna(x) else False],
    ).copy()

    columnas = ['Tema_Central', 'Resultados_Sintesis',
                'Anio_Estudio', 'Localizacion_Geografica',
                'Especies_y_Temas_Detectados']
    for col in columnas:
        df[col] = ''

    for idx, row in df.iterrows():
        try:
            doc_pdf = fitz.open(row['Ruta_Local_PDF'])

            # Extraer todo el texto del documento
            texto_completo = ''.join(p.get_text() for p in doc_pdf)
            texto_limpio   = limpiar_texto_profundo(texto_completo)

            # Procesar con spaCy (NLP completo sobre el documento)
            doc_nlp = nlp(texto_limpio)

            # ── 1. LOCALIZACIÓN GEOGRÁFICA ──────────────────────────
```

```
            lugares = [e.text for e in doc_nlp.ents
                       if e.label_ == 'LOC' and len(e.text) > 4]
            # Excluir falsos positivos comunes en artículos científicos
            excluir = ['universidad', 'figura', 'tabla', 'cuadro']
            lugares = [l for l in lugares
                       if not any(x in l.lower() for x in excluir)]
            df.at[idx, 'Localizacion_Geografica'] = ', '.join(
                loc for loc, _ in Counter(lugares).most_common(6))

            # ── 2. ESPECIES Y TEMAS ──────────────────────────────
            entidades  = [e.text.lower() for e in doc_nlp.ents
                          if e.label_ in ['MISC', 'ORG']]
            sustantivos = [t.text.lower() for t in doc_nlp
                           if t.pos_ == 'NOUN'
                           and not t.is_stop and len(t.text) > 4]
            conteo = Counter(entidades + sustantivos)
            df.at[idx, 'Especies_y_Temas_Detectados'] = ', '.join(
                t for t, _ in conteo.most_common(15))

            # ── 3. TEMA CENTRAL (abstract / resumen) ─────────────
            tema = re.search(
                r'(?i)(resumen|abstract|introducción)[\s:]+(.{150,800}?\.)',
                texto_limpio)
            df.at[idx, 'Tema_Central'] = (
                tema.group(2).strip() if tema else texto_limpio[:400])

            # ── 4. RESULTADOS / CONCLUSIONES ─────────────────────
            res = re.search(
r'(?i)(conclusiones|resultados|discusión)[\s:]+(.{200,1000}?\.)',
                texto_limpio)
            df.at[idx, 'Resultados_Sintesis'] = (
                res.group(2).strip() if res else 'Revisar sección final del
PDF')

            # ── 5. AÑO DE ESTUDIO (por frecuencia) ───────────────
            fechas = re.findall(r'\b(19|20)\d{2}\b', texto_limpio)
```

```
            if fechas:
                df.at[idx, 'Anio_Estudio'] =
Counter(fechas).most_common(1)[0][0]

            print(f'☑ Completado: {idx} — {row["Título"][:40]}')
            doc_pdf.close()

        except Exception as e:
            print(f'  Error en {idx}: {e}')

    df.to_csv("analisis_exhaustivo_primavera.csv",
              index=False, encoding="utf-8-sig")
    print("¡Dataset de alta calidad generado!")

if __name__ == "__main__":
    mineria_calidad_total("investigaciones_con_rutas.csv")
```

Discusión de resultados

Las síntesis de literatura son ejercicios muy relevantes para identificar tendencias y vacíos de información para ser atendidos en futuros trabajos de investigación. Con la metodología implementada en este trabajo identificamos 19 artículos que se han desarrollado en el Bosque La Primavera. Aunque existen artículos científicos realizados en el BLP que no fueron detectados en este trabajo, este resultado puede estar relacionado con el Proceso de Lenguaje Natural utilizado. Si bien el NLP permite transformar grandes volúmenes de texto no estructurado en información organizada y útil, puede tener limitaciones para acertar en la interpretación de palabras que cuentan con múltiples significados o frases con interpretaciones distintas (Luca, 2025). En este sentido, se registraron 35 artículos que no tenían que ver con el BLP sino con contaminación en cuerpos de agua en otros estados de México, amenazas en otros bosques, problemas de género respecto a la mujer, microorganismos, entre otros, lo que puede estar relacionado con un bajo nivel de exigencia en la cantidad de palabras que tenían que coincidir con las palabras de búsqueda para ser incluidos. De manera similar, Delgado y Uzcátegui (2022) utilizaron NLP, pero además aplicaron spaCy y un modelo llamado BETO, el cual fue entrenado por la técnica de Dense Passage Retrieval (DRP) para mejorar la recuperación de información relevante e interpretar

la intención y el significado del texto. Con esto, los autores lograron una precisión del 90%, demostrando que el procesamiento de lenguaje natural puede permitir darle sentido al lenguaje humano, logrando búsquedas naturales, clasificaciones automáticas y resultados más acertados. Por ello, se recomienda que futuros trabajos busquen mejorar el proceso del NLP para que realice búsquedas más precisas, así como evaluar la posibilidad de integrar otras metodologías y herramientas

A pesar de tener una muestra de solo 19 artículos sobre el BLP, observamos una tendencia de enfoques en la investigación en torno al BLP hacia el tema de Conservación y Flora. Esto es un resultado importante, ya que permite identificar temas que no están bien representados en la literatura como la Fauna, la Gestión del área protegida y/o sobre cómo la Actividad humana está amenazando al BLP. El análisis más detallado de estas temáticas, así como de los alcances y limitaciones de cada trabajo, permitirá plantear futuros esfuerzos de investigación que favorezcan la toma de decisiones con base en evidencia científica.

1.7. Bibliografía y otros recursos

Ceballos, G., Ehrlich, P. R., & Raven, P. H. (2020). Vertebrates on the brink as indicators of biological annihilation and the sixth mass extinction. Proceedings Of The National Academy Of Sciences, 117(24), 13596-13602. https://doi.org/10.1073/pnas.1922686117

Comisión Nacional de Áreas Naturales Protegidas. (2025, 4 junio). Áreas Naturales Protegidas. Gobierno de México. https://www.gob.mx/conanp/documentos/areas-naturales-protegidas-278226

CONANP. (s. f.). Áreas Naturales Protegidas. https://descubreanp.conanp.gob.mx/es/conanp/ANP?suri=88

Delgado, H., & Uzcátegui, F. (2022). Procesamiento de Lenguaje Natural para la Búsqueda Semántica en Repositorios Académicos. ResearchGate. https://www.researchgate.net/publication/365476862_Procesamiento_de_Lenguaje_Natural_para_la_Busqueda_Semantica_en_Repositorios_Academicos

Gobierno de México. (2026). Bosque La primavera. Bosque la Primavera. Recuperado 22 de febrero de 2026, de https://bosquelaprimavera.jalisco.gob.mx/#

IBM. (2026). Data Science. IBM. https://www.ibm.com/think/topics/data-science

IUCN. (2026, 23 abril). Effective protected areas. https://iucn.org/our-work/topic/effective-protected-areas

Jonker, A., & Gomstyn, A. (2026). Structured vs Unstructured Data. IBM. https://www.ibm.com/think/topics/structured-vs-unstructured-data

Luca, C. (2025, abril). Natural Language Processing (NLP) for document analysis. ResearchGate. https://www.researchgate.net/publication/390941525_Natural_Language_Processing_NLP_for_Document_Analysis

Rios-Llamas, C., & Hernández-Vázquez, S. (2023). Endangered rural economies in the periurban area of Bosque La Primavera, Guadalajara, Mexico. Scielo, 19(1). https://www.scielo.cl/article_plus.php?lng=es&pid=S0718-235X2023000100049&tlng=en

Stryker, C., & Holdsworth, J. (2026). Natural language processing. IBM. https://www.ibm.com/think/topics/natural-language-processing

1.8. Anexos

Anexo 1. Pipeline automatizado (base de datos), script y artículos recuperados: 3-CienciaDatos

2. Productos

A continuación se muestra el código utilizado para la búsqueda de información.

```
from serpapi import GoogleSearch
import pandas as pd
import time
import re
import requests
import os
import fitz
import spacy
from collections import Counter

API_KEY = "8b556fc5ca59..."  # Clave de autenticación SerpApi

# Términos de búsqueda especializados para el Bosque La Primavera
```

```
QUERIES = [
    '"Bosque La Primavera" Jalisco',
    '"Bosque de la Primavera" Guadalajara',
    '"Bosque La Primavera" Zapopan'
]

def limpiar_publicacion(pub_string):
    autores, revista, anio = "N/A", "N/A", "N/A"
    if pub_string:
        match_anio = re.search(r'\b(19|20)\d{2}\b', pub_string)
        if match_anio:
            anio = match_anio.group(0)
        partes = pub_string.split(' - ')
        if len(partes) >= 1:
            autores = partes[0]
        if len(partes) >= 2:
            revista = partes[1].replace(anio, '').strip(', ')
    return autores, revista, anio

def extraer_datos_profesional():
    resultados_totales = []

    for q in QUERIES:
        start = 0
        print(f'--- Iniciando búsqueda: {q} ---')

        while start < 300:  # Hasta 300 resultados por query
            params = {
                "engine": "google_scholar",
                "q": q,
                "api_key": API_KEY,
                "start": start,
                "num": 20,
                "hl": "es"
            }
            search = GoogleSearch(params)
```

```
            results = search.get_dict()
            organic = results.get("organic_results", [])

            if not organic:
                break

            for art in organic:
                pub_info = art.get("publication_info", {}).get("summary", "")
                autores, revista, anio = limpiar_publicacion(pub_info)
                resultados_totales.append({
                    "Título":      art.get("title"),
                    "Autores":     autores,
                    "Año":         anio,
                    "Publicado en": revista,
                    "Link":        art.get("link"),
                    "Resumen":     art.get("snippet"),
                    "Citas":       art.get("inline_links", {})
                                       .get("cited_by", {}).get("total", 0)
                })

            start += 20
            print(f'Progreso: {len(resultados_totales)} artículos
acumulados...')
            time.sleep(0.5)

    df = pd.DataFrame(resultados_totales)
    # Eliminar duplicados: mismo artículo puede aparecer en varias queries
    df = df.drop_duplicates(subset=['Título', 'Autores'])

    df.to_csv("investigaciones_bosque_primavera_final.csv",
              index=False, encoding="utf-8-sig")
    print(f'¡Éxito! {len(df)} artículos únicos exportados.')

if __name__ == "__main__":
    extraer_datos_profesional()
```

```
def descargar_pdfs(csv_path, carpeta_destino='pdfs_primavera'):
```

```
    if not os.path.exists(carpeta_destino):
        os.makedirs(carpeta_destino)

    df = pd.read_csv(csv_path)
    df['Ruta_Local_PDF'] = os.path.join(os.getcwd(), carpeta_destino)

    for index, row in df.iterrows():
        url = row['Link']

        # Validar que el URL sea texto y no un valor NaN
        if not isinstance(url, str):
            print(f'Saltando índice {index}: Link vacío o inválido.')
            continue

        # Filtrar solo fuentes académicas confiables
        dominios_validos = ['scielo.org', 'redalyc.org', 'researchgate.net']
        if any(d in url for d in dominios_validos) or url.endswith('.pdf'):
            try:
                nombre_archivo = f'articulo_{index}.pdf'
                ruta_completa  = os.path.join(carpeta_destino, nombre_archivo)

                # Headers para evitar bloqueos por detección de bots
                headers  = {'User-Agent': 'Mozilla/5.0'}
                response = requests.get(url, timeout=15, headers=headers)

                if response.status_code == 200:
                    with open(ruta_completa, 'wb') as f:
                        f.write(response.content)
                    df.at[index, 'Ruta_Local_PDF'] = ruta_completa
                    print(f'Descargado {index}: {str(row["Título"])[:30]}...')

                time.sleep(1.5)  # Pausa para no saturar los servidores

            except Exception as e:
                print(f'Error en link {index}: {e}')

```

```
    nuevo_csv = 'investigaciones_con_rutas.csv'
    df.to_csv(nuevo_csv, index=False, encoding='utf-8-sig')
    return nuevo_csv

if __name__ == "__main__":
    archivo = "investigaciones_bosque_primavera_final.csv"
    print('Iniciando descarga de documentos del Bosque La Primavera...')
    ruta_actualizada = descargar_pdfs(archivo)
    print(f'Proceso terminado. Archivo generado: {ruta_actualizada}')
```

```
import spacy, fitz, re, os, pandas as pd
from collections import Counter

# Cargar modelo Large para máxima precisión semántica
try:
    nlp = spacy.load("es_core_news_lg")
    nlp.max_length = 2_000_000  # Ampliar límite para artículos largos
    print("Modelo Large cargado correctamente.")
except:
    print(" Modelo Large no encontrado. Usando modelo base.")
    nlp = spacy.load("es_core_news_sm")

def limpiar_texto_profundo(texto):
    """Elimina ruido tipográfico manteniendo la estructura semántica."""
    texto = re.sub(r'\s+', ' ', texto)
    basura = ['download', 'free pdf', 'issn', 'http', 'www.',
              'vol.', 'página', 'copyright']
    for b in basura:
        texto = re.sub(rf'(?i).{{0,20}}{b}.{{0,20}}', '', texto)
    return texto.strip()

def mineria_calidad_total(ruta_csv):
    df = pd.read_csv(ruta_csv)
    # Filtrar solo archivos que existen físicamente en disco
    df = df[df['Ruta_Local_PDF'].apply(
        lambda x: os.path.isfile(str(x)) if pd.notna(x) else False],
```

```
    ).copy()

    columnas = ['Tema_Central', 'Resultados_Sintesis',
                'Anio_Estudio', 'Localizacion_Geografica',
                'Especies_y_Temas_Detectados']
    for col in columnas:
        df[col] = ''

    for idx, row in df.iterrows():
        try:
            doc_pdf = fitz.open(row['Ruta_Local_PDF'])

            # Extraer todo el texto del documento
            texto_completo = ''.join(p.get_text() for p in doc_pdf)
            texto_limpio   = limpiar_texto_profundo(texto_completo)

            # Procesar con spaCy (NLP completo sobre el documento)
            doc_nlp = nlp(texto_limpio)

            # ── 1. LOCALIZACIÓN GEOGRÁFICA ──────────────────────
            lugares = [e.text for e in doc_nlp.ents
                       if e.label_ == 'LOC' and len(e.text) > 4]
            # Excluir falsos positivos comunes en artículos científicos
            excluir = ['universidad', 'figura', 'tabla', 'cuadro']
            lugares = [l for l in lugares
                       if not any(x in l.lower() for x in excluir)]
            df.at[idx, 'Localizacion_Geografica'] = ', '.join(
                loc for loc, _ in Counter(lugares).most_common(6))

            # ── 2. ESPECIES Y TEMAS ─────────────────────────────
            entidades  = [e.text.lower() for e in doc_nlp.ents
                          if e.label_ in ['MISC', 'ORG']]
            sustantivos = [t.text.lower() for t in doc_nlp
                           if t.pos_ == 'NOUN'
                           and not t.is_stop and len(t.text) > 4]
            conteo = Counter(entidades + sustantivos)
```

```
            df.at[idx, 'Especies_y_Temas_Detectados'] = ', '.join(
                t for t, _ in conteo.most_common(15))

            # ── 3. TEMA CENTRAL (abstract / resumen) ──────────────
            tema = re.search(
                r'(?i)(resumen|abstract|introducción)[\s:]+(.{150,800}?\.)',
                texto_limpio)
            df.at[idx, 'Tema_Central'] = (
                tema.group(2).strip() if tema else texto_limpio[:400])

            # ── 4. RESULTADOS / CONCLUSIONES ──────────────────────
            res = re.search(
r'(?i)(conclusiones|resultados|discusión)[\s:]+(.{200,1000}?\.)',
                texto_limpio)
            df.at[idx, 'Resultados_Sintesis'] = (
                res.group(2).strip() if res else 'Revisar sección final del
PDF')

            # ── 5. AÑO DE ESTUDIO (por frecuencia) ────────────────
            fechas = re.findall(r'\b(19|20)\d{2}\b', texto_limpio)
            if fechas:
                df.at[idx, 'Anio_Estudio'] =
Counter(fechas).most_common(1)[0][0]

            print(f'☑ Completado: {idx} — {row["Título"][:40]}')
            doc_pdf.close()

        except Exception as e:
            print(f'  Error en {idx}: {e}')

    df.to_csv("analisis_exhaustivo_primavera.csv",
              index=False, encoding="utf-8-sig")
    print("¡Dataset de alta calidad generado!")

if __name__ == "__main__":
    mineria_calidad_total("investigaciones_con_rutas.csv")
```

Código 1, 2 y 3 Juntos. Script Completo utilizado para el desarrollo del pipeline

3. Reflexión crítica y ética de la experiencia

El RPAP tiene también como propósito documentar la reflexión sobre los aprendizajes en sus múltiples dimensiones, las implicaciones éticas y los aportes sociales del proyecto para compartir una comprensión crítica y amplia de las problemáticas en las que se intervino.

El Bosque La Primavera es un área rica en cuanto a la biodiversidad que alberga, además de todos los recursos naturales que contiene, por lo que es de vital importancia estudiarlo, conocerlo y protegerlo de la mejor manera posible para preservar nuestra riqueza geológica. Debemos comprender que no somos la única especie que importa en esta casa común, y que además necesitamos de esta riqueza para preservar nuestra sobrevivencia y salud. Esperamos que con el uso de herramientas de la Ciencia de Datos se logre un estudio exhaustivo e integro para generar un conocimiento en torno al BLP y que de esta manera se genere conciencia de cómo convivir con otras especies.

3.1 Sensibilización ante las realidades

Se realizaron visitas a una parte del BLP donde pudimos conocer la situación que atraviesa nuestra riqueza natural, sistemas completos se ven afectados por nuestras actividades humanas, desde la fauna, flora, agua, tierra y hasta comunidades de personas que conviven a diario con los efectos de la contaminación que todos generamos. Nos hizo preguntarnos si realmente somos conscientes del daño que podemos causar al no gestionar tanto nuestros residuos como nuestros recursos ambientales. Estamos hablando de personas, animales y plantas que no pueden cambiar su realidad de un día a otro, no es fácil simplemente irse del lugar donde viven, por lo que tienen que adaptarse a un lugar tóxico, lo cual es completamente injusto. Todos tenemos derecho a vivir en un ambiente sano y limpio, conforme a esto, esperamos que facilitar el estudio de esta ANP ayude a comprender la situación y actuar adecuadamente para ayudar a todas estas personas, fauna y flora.

3.2 Aprendizajes logrados

Durante el desarrollo de este proyecto enfrentamos retos que nos exigieron salir de nuestra zona de conocimiento técnico habitual. El desafío más significativo al inicio fue la descarga automatizada de documentos PDF desde Google Scholar, ya que nunca habíamos trabajado con web scraping aplicado a repositorios académicos. Sabíamos que era posible hacer scraping en general, pero no tenía claridad sobre cómo estructurarlo sobre una fuente como Google Scholar ni qué herramientas eran las adecuadas para hacerlo de manera confiable. Descubrir que existía una API especializada como SerpApi que permitía interactuar con Google Scholar de forma programática fue un aprendizaje que no anticipamos, y que abrió una forma completamente nueva de pensar en la recuperación automatizada de información académica.

Otro reto importante fue la toma de decisiones a lo largo del pipeline: desde definir qué dominios académicos incluir en el filtro de descarga, hasta decidir qué variables extraer del texto con NLP y cómo manejar los casos donde el modelo no encontraba la información esperada. Cada una de esas decisiones implicaba un balance entre precisión y cobertura que aprendí a evaluar de manera más crítica conforme avanzaba el proyecto.

Este proyecto nos ayudó a desarrollar profesionalmente en el diseño de arquitecturas de datos de extremo a extremo, en el uso de NLP aplicado a texto científico en español, y en la capacidad de tomar decisiones técnicas con información incompleta.

3.3 Congruencia con perfil de egreso

Selecciona de cada una de la(s) carrera(s) que colaboraron en el desarrollo de este PAP la competencia que más se puso en práctica y/o se desarrolló:

Humanidades	☐ Arquitectura	Choose an item.
	☐ Arte y creación	Choose an item.
	☐ Ciencias de la comunicación	Choose an item.
	☐ Ciencias de la educación	Choose an item.
	☐ Comunicación y artes audiovisuales	Choose an item.
	☐ Derecho	Choose an item.
	☐ Desarrollo Inmobiliario Sustentable	Choose an item.
	☐ Diseño	Choose an item.
	☐ Diseño de indumentaria y moda	Choose an item.
	☐ Diseño urbano y arquitectura del paisaje	Choose an item.
	☐ Gestión Cultural	Choose an item.
	☐ Gestión Pública	Choose an item.
	☐ Nutrición	Choose an item.
	☐ Periodismo y comunicación pública	Choose an item.
	☐ Psicología	Choose an item.

	☐ Publicidad y comunicación estratégica	Choose an item.
	☐ Relaciones internacionales	Choose an item.
	☐ Traducción e interpretación	
Ingenierías	☐ Ingeniería ambiental/ y tecnologías sustentables	Choose an item.
	☐ Ingeniería Civil	Choose an item.
	☐ Ingeniería de Alimentos	Choose an item.
	☐ Ingeniería Electrónica	Choose an item.
	☒ Ingeniería en Biotecnología	Innovar y emprender proyectos que generen riqueza a su entorno, promoviendo la creación de empleos de calidad y ofreciendo productos que cubran necesidades sociales mediante el diseño, la ingeniería y el desarrollo de bioproyectos.
	☐ Ingeniería en Ciberseguridad	Choose an item.
	☐ Ingeniería en Desarrollo de Software	Choose an item.
	☐ Ingeniería en Inteligencia Artificial	Choose an item.
	☐ Ingeniería en Mecatrónica	Choose an item.
	☐ Ingeniería en Nanotecnología	Choose an item.
	☐ Ingeniería en Negocios y Servicios Digitales	Choose an item.
	☐ Ingeniería en Sistemas Computacionales	Choose an item.
	☐ Ingeniería en Sistemas Digitales Embebidos	Choose an item.
	☐ Ingeniería Financiera	Choose an item.
	☐ Ingeniería Industrial	Choose an item.
	☐ Ingeniería Mecánica	Choose an item.
	☐ Ingeniería Química	Choose an item.
	☒ Ingeniería y Ciencia de Datos	Desarrollar y ejecutar programas para la gestión de información de recursos naturales y servicios ambientales desde una perspectiva social que comprenda una visión holística y multidisciplinaria.
Negocios	☐ Administración de Empresas y Emprendimiento	Choose an item.
	☐ Comercio y Negocios Globales	Choose an item.
	☐ Contaduría y Gobierno Corporativo	Choose an item.
	☐ Finanzas	Choose an item.
	☐ Hospitalidad y Turismo	Choose an item.
	☐ Mercadotecnia y Dirección Comercial	Choose an item.
	☐ Negocios y Mercados Digitales	Choose an item.
	☐ Recursos Humanos y Talento Organizacional	Choose an item.