MCDI

MAESTRÍA EN CIENCIA DE DATOS E INFORMACIÓN

INFOTEC CENTRO DE INVESTIGACIÓN E INNOVACIÓN EN TECNOLOGÍAS DE LA INFORMACIÓN Y COMUNICACIÓN

DIRECCIÓN ADJUNTA DE INNOVACIÓN Y CONOCIMIENTO
GERENCIA DE CAPITAL HUMANO
POSGRADOS

"SISTEMA DE RECOMENDACIÓN BASADO EN LA DETECCIÓN DE TÓPICO Y ASIGNACIÓN DE TÉRMINOS CLAVES DE LOS DOCUMENTOS ASOCIADOS A REPOSITORIOS INSTITUCIONALES"

PROPUESTA DE INTERVENCIÓN,
Que para obtener el grado de MAESTRO EN CIENCIA DE DATOS E INFORMACIÓN

Presenta:

Ing. René Gerardo Jara Sixtos

Asesor:

Dr. Dagoberto Armenta Medina

Ciudad de México, Marzo, 2023.

CONACYT

INFOTEC POSGRADOS

Autorización de impresión

INFOTEC
POSGRADOS

AUTORIZACIÓN DE IMPRESIÓN Y NO ADEUDO EN BIBLIOTECA

Maestría en Ciencia de Datos e Información, MCDI

Ciudad de México, 1 de marzo de 2023.

La Gerencia de Capital Humano / Gerencia de Investigación hacen constar que el trabajo de titulación

"Sistema de recomendación basado en la detección de tópico y asignación de términos claves de los documentos asociados a repositorios institucionales"

Desarrollado por el alumno: **René Gerardo Jara Sixtos**, y bajo la asesoría del **Dr. Dagoberto Armenta Medina** cumple con el formato de Biblioteca. Por lo cual, se expide la presente autorización para impresión del proyecto terminal al que se ha hecho mención.

Asimismo, se hace constar que no debe material de la biblioteca de INFOTEC.

Vo. Bo.

Mtro. Felipe Alfonso Delgado Castillo
Gerente de Capital Humano

Anexar a la presente autorización al inicio de la versión impresa del trabajo referido que ampara la misma.

C.p.p Servicios Escolares

Tabla de contenido

Índice de figuras

Índice de gráficos

Índice de cuadros

Capítulo 1

Introducción

Capítulo 1. Introducción

Actualmente, acceder a los documentos que concentran los buscadores generales de información a partir de ocurrencias simples de palabras a texto abierto genera grandes listados, con resultados muchos de los cuales no se encuentran relacionados unos con otros, además si no se utiliza la palabra correcta en estas búsquedas se dificulta aún más esta labor. Para mejorar esta experiencia es importante ofrecer a los usuarios alternativas a sus búsquedas y localización. Una de estas maneras es el uso de sistemas de recomendación que sugieren otros documentos relacionados con los que se están consultando. Eso se puede hacer analizando los documentos y explotando los metadatos relacionados con estos.

A partir de los metadatos que acompañan a un documento es posible conocer más de ellos, ya que contienen información semiestructurada que describen sus características intrínsecas. Aplicando técnicas de detección automática de tópicos en textos no estructurados y la identificación y asignación de palabras clave a los documentos depositados, es posible identificar los temas principales de los que trata, evaluando la ocurrencia de ciertas palabras y los patrones que estas ocurrencias forman, permitiendo agrupar los documentos por temáticas y detectando interrelaciones entre los mismos.

1.1 Motivación

Derivado de la gran cantidad de información que se genera actualmente en todos los ámbitos cada vez más se dificulta la búsqueda y localización de documentos de valor para los usuarios, por lo cual es preciso que estos sean clasificados y organizados. Debido a la velocidad y los volúmenes con que se genera esta información, su clasificación y organización se vuelve un proceso humanamente imposible por lo que es cada vez más necesario aplicar técnicas de procesamiento computacional que permitan analizar de manera automática estos grandes cúmulos de información para poder etiquetarlos y agruparlos con sus similares. De esta manera la información puede ser ofrecida y consumida por los usuarios eficientemente, siendo muy conveniente que su clasificación y organización pueda realizarse de manera no supervisada.

1.2 Problemática

Los recursos de información que se alojan en los Repositorios Institucionales de Ciencia Abierta pueden ser consultados mediante búsquedas a texto abierto y consultas de relaciones simples como lo es la búsqueda por autor o materia, lo que conlleva a obtener largos listados de resultados que si bien están ponderados solo se limitan al peso de las palabras que se usan en la cadena de búsqueda. Una vez que se consulta un recurso de información solo se visualiza información relacionada con el mismo autor, sin ofrecer otras alternativas de consulta, como pudiera ser recomendaciones de otros recursos de información relacionados que facilite a los usuarios el acceso a información relevante para su búsqueda.

1.3 Objetivos

1.3.1 Objetivo General

El objetivo es determinar de manera no supervisada los tópicos y palabras clave relacionados con los recursos de información de un Repositorio Institucional de Ciencia Abierta y proponer un sistema de recomendación de recursos de información basado en tópicos para finalmente explorar los tópicos obtenidos de un Repositorio Institucional.

1.3.2 Objetivos Específicos

- Extraer tópicos principales de un recurso de información a partir del preprocesamiento por técnicas de reducción morfológica de los títulos y resúmenes disponibles para consulta en un Repositorio Institucional.
- Evaluar métricas de selección de términos relevantes, se propone TF-IDF, Punto de transición de Goffman y Entropía.
- Evaluar si existe algún otro metadato que permita mejorar la extracción.
- Evaluar las combinaciones de estos metadatos.
- Evaluar la métrica obtenida con un grupo de documentos clasificados.
- Clasificar los recursos de acuerdo con los tópicos extraídos.

- Generar un prototipo de un sistema de recomendación de recursos de información de acuerdo con los tópicos extraídos.
- Obtener y explorar los tópicos de una muestra de Repositorios Institucionales.

1.4 Contribución

El presente trabajo se enfoca en una propuesta de mejora a las consultas de recursos de información de los Repositorios Institucionales de Ciencia Abierta, los cuales son plataformas digitales que contienen los recursos de información académica, científica, tecnológica y de innovación, siendo de gran valor por ser un punto de difusión del conocimiento generado en el país. Para lograr las mejoras en las consultas este proyecto contempla la implementación de enfoques computacionales derivados de técnicas de procesamiento de lenguaje natural y modelado de tópicos.

La aplicación de técnicas de modelado de tópico para la extracción de palabras claves permite relacionar de manera no supervisada los recursos de información de un Repositorio Institucional de Ciencia Abierta.

Con la aplicación de técnicas automáticas de Procesamiento de Lenguaje Natural a los metadatos y el modelado de tópicos de los recursos de información, es posible proponer elementos de catalogación acordes a cada recurso, como son los temas y palabras clave que permita a partir de esta información, recomendar otros recursos de información asociados.

Capítulo 2

Marco Teórico

Capítulo 2. Marco Teórico

2.1 Repositorio institucional

Como parte de las políticas de ciencia abierta en el país (Guajardo, 2020) y con el objetivo de permitir el acceso libre y gratuito a los materiales y recursos de información, que resultan de los procesos de investigación que se producen en México con fondos públicos, en 2017 CONACYT establece el Programa de Repositorios (CONACYT, 2017a) con el objetivo de impulsar la creación de los Repositorios Institucionales de Ciencia Abierta como plataformas digitales e interoperables, para resguardar y ofrecer los recursos de información académica, científica, tecnológica y de innovación generados por instituciones de educación superior y aquellas que realizan investigación científica y tecnológica. Con el objetivo de coordinar la interoperabilidad de estos repositorios se estableció (CONACYT, 2017b) el uso de un esquema de metadatos alineados a OpenAIRE que enumera la información mínima para describir un recurso de información y del protocolo OAI-PMH para el intercambio de esta información.

2.2 Sistema de recomendación

Los Sistemas de recomendación (Recommendation System)(Adomavicius & Tuzhilin, 2005; Bobadilla et al., 2013) aparecen como técnicas para acercar documentos a los usuarios acordes a la información que se está consultando, estos pueden clasificarse en basados en contenido, colaborativos e híbridos (Adomavicius & Tuzhilin, 2005), en el primero las recomendaciones se hacen a partir de documentos similares a los que el usuario está viendo o vio en el pasado, en el caso de los colaborativos se recomiendan documentos que otros usuarios similares relacionaron p.ej. calificándolos, en el último se combina los dos anteriores con la idea de mejorar los resultados.

En el caso del repositorio de información de INFOTEC no se requiere un proceso de registros y login para su consulta, ni tiene procesos de calificación o comentarios sobres los recursos de información lo que dificulta dar seguimiento a las preferencias de los usuarios, por lo que se optó por un enfoque puramente basado en contenidos (Pazzani & Billsus, 2007).

En este enfoque se utilizará la información relacionada con los documentos (metadatos) para extraer sus características, estas se compararán entre los demás documentos y a partir de estos se hará una recomendación de los similares, por ejemplo haciendo un símil en el caso de películas o libros podrían ser otros del mismo género o autor, para que esto sea posible estos metadatos deben estar estructurados, aunque esto no garantiza la calidad de la información, y deben pasar por un proceso de análisis y normalización, con la idea de desechar los datos irrelevantes y extraer esas características que permitan clasificarlos y relacionarlo con sus similares, con el objetivo de hacer los cruces de manera eficiente y filtrar la información que se le ofrece a los usuarios

2.3 Modelado de tópico

El Modelado de Tópicos (Topic Modeling) (Blei, 2012) surge como un enfoque probabilístico a la idea de que los documentos que están relacionados con uno o más temas (tópico) reflejan esta relación en un grupo de palabras las cuales están presentes en mayor o menor medida dependiendo de esta relación con los temas, así que cada tema tendrá un grupo de palabras que aparecerán en los documentos relacionados con él, pero si lo está a varios temas aparecerán también otros conjuntos de palabras y en el caso de temas emergentes sé irá creando su propio grupo de palabras conforme vaya madurando.

Esta técnica permite procesar grandes cantidades de información sin intervención humana y no requiere que la información esté previamente etiquetada, ya que parten de procesar todas las palabras y calcular su distribución dentro del documento y en el conjunto de documentos.

Para reducir la complejidad al representar un documento(Baeza-Yates & Ribeiro-Neto, 2011) es recomendable realizar ciertos pasos previos como la eliminación de palabras vacías (Stopwords) las cuales no aportan información relevante en la mayoría de los casos (como el

nuestro) y la reducción morfológica (como el Stemming) el cual reduce las variantes de una misma palabra, además con esto reducimos las dimensiones de los datos y con ello simplificamos su procesamiento. A partir de estos datos se genera una representación simplificada de cada documento agregando la frecuencia en que aparece cada uno de estos términos simplificados sin importar su orden y también su gramática, esta representación ahora debe ser procesada mediante técnicas de pesado de términos (TF-IDF, Goffman, Entropía) con la intención de ponderar las palabras de mayor aporte y descartar palabras de uso común, que será la entrada para el modelado. Finalmente, con el objetivo de encontrar los temas a partir del análisis de las palabras, esto se puede hacer mediante técnicas como el LDA y LSA.

2.3.1 TF-IDF

Una de las técnicas de pesado de términos más populares es TF-IDF (Term Frequency - Inverse Document Frequency)(Baeza-Yates & Ribeiro-Neto, 2011) la cual permite cuantificar la relevancia de ciertas palabras en un documento comparándolo dentro de un corpus de ellos, mediante la determinación de la frecuencia de aparición de un término en un documento por la proporción inversa de ese mismo término en todos los documentos del corpus, con esto obtendremos pesos altos cuando el término ocurre en un grupo pequeño de documentos y bajos cuando lo hacen en un solo documento y aún más bajos cuando aparece en todos los documentos, permitiendo con esto usarlo como un marcador de relevancia.

El peso se calcula mediante la siguiente fórmula (Ramos, 2003):

Para: un conjunto de documentos D, un término t y un documento d

$$TF-IDF_{t,d}=f_{t,d}*\log\left(\frac{n}{f_{t,D}}\right)$$

Donde $f_{t,d}$ es el número de ocurrencias del término t en el documento d, n es el número total de documentos en D, y $f_{t,D}$ es el número de ocurrencias del término t en el conjunto de documentos D.

2.3.2 Punto de transición de Goffman

La ley de Zipf plantea (Urbizagástegui Alvarado & Restrepo Arango, 2011) que al escribir preferimos usar más palabras comunes con respecto a las menos conocidas, proponiendo 2 ecuaciones las cuales describen el comportamiento de la distribución en frecuencia de palabras en un documento, una para palabras de alta frecuencia y otra para palabras de baja frecuencia, Goffman propone que en la región donde estas dos ecuaciones se encuentran (Punto de transición) es donde se localizan las palabras de mayor relevancia de un texto. Entonces si tenemos un listado de ocurrencia de palabras ordenados por frecuencia las que se encuentren más cerca de este punto se consideraran de mayor relevancia.

Este punto de transición (Boyce & Lockard, 1975) se encuentra donde la distribución de baja frecuencia termina y comienza la distribución de alta frecuencia y está dado por:

$$PTG = \frac{-1 \pm \sqrt{1+8I_1}}{2}$$

Donde PTG es la frecuencia donde se va a encontrar estos dos curvas (Punto de transición), así como I_1 es el número de términos que tienen frecuencia 1.

2.3.3 Entropía

El uso de la entropía para el pesado de términos propone el cálculo de la LogEntropy (Dumais, 1991) o Entropía inversa (Quesada, 2007) como métrica la cual plantea que mientras más entropía tenga un término menos información transmite acerca de los documentos en los que aparece, por lo tanto, tiene menos relevancia en su uso. Por ejemplo, si una palabra aparece en todos los documentos esta no aporta gran significado para el entendimiento del mismo y obtendrá un peso de esta métrica bajo, por el otro lado una palabra que tiene menos apariciones obtendrá un peso alto.

La fórmula para el cálculo de este peso está dada por (Pincombe, 2004):

Para: un conjunto de documentos D, un término t y un documento d:

$$E_t = 1 + \frac{\sum_{t=1}^{n} (P_{t,d} * \log(P_{t,d}))}{\log(n)}$$

Donde n es el número total de documentos en D y la probabilidad $P_{t,d}$ está dada por:

$$P_{t,d} = \frac{f_{t,d}}{f_{t,D}}$$

Donde $f_{t,d}$ es el número de ocurrencias del término t en el documento d y $f_{t,D}$ es el número de ocurrencias del término t en el conjunto de documentos D.

2.3.4 LSA

El Análisis Semántico Latente (Latente Semantic Analysis) (Dumais, 1991) propone que las palabras tienen una estructura oculta en su uso y plantea que esta puede ser estimada aplicando técnicas estadísticas para el pesado de términos para construir la representación de los términos contenidos en los documentos para formar una Matriz Término-Documento (Term-Document Matrix o TDM) y aplicando la Descomposición en valores singulares (Singular Vector Decomposition o SVD) para la reducción de dimensión de esta matriz y obtener así la estructura de las asociaciones latentes.

El procedimiento de SVD propone que, si se tiene una matriz X de dimensiones m x n, esta puede ser descompuesta en el producto de 3 matrices

$$\underset{mxn}{X} = \underset{mxr}{T} \quad \underset{rxr}{S} \quad \underset{rxn}{O^T}$$

Donde T y O son ortogonales y S es diagonal y r es el rango de X

Peros si tomamos solo los primeros k valores de S con sus correspondientes columnas en T y O obtendremos una aproximación de la matriz X, de rango k y tan cercana también como k

$$\underset{mxn}{X} \approx \underset{mxn}{\hat{X}} = \underset{mxk}{T} \quad \underset{kxk}{S} \quad \underset{kxn}{O^T}$$

Aplicando esta idea al procesamiento de la información se puede construir una matriz TDM que caracterice nuestros documentos y los pesos de los términos que estos contienen y descomponerla mediante SVD y tomar solo los primeros k componentes que contienen las estructuras de relaciones de los términos y los documentos eliminando con esto el “ruido”, hay que considerar que si se usa una k muy pequeña se puede perder información sobre las relaciones.

Capítulo 3

Metodología

Capítulo 3. Metodología

3.1 Descripción General

Para el desarrollo de esta propuesta de aplicación se utilizó Python como lenguaje de trabajo por su facilidad de uso en aplicaciones de Ciencia de Datos, así como por variedad de librerías que tiene para este mismo fin, además se usó como set de datos el del Repositorio Institucional de INFOTEC.

Para la realización se seguirán los siguientes pasos:

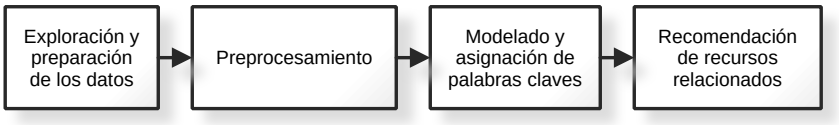

Figura 1: Pasos a seguir
Fuente: Elaboración propia.

3.2 Exploración y preparación de los datos

Como ya se mencionó anteriormente los Repositorios Institucionales (RI) exponen los datos de sus recursos de información según los Lineamientos Específicos para Repositorios (CONACYT, 2017b) en un EndPoint basado en el protocolo OAI-PMH los cuales una vez cosechados deben ser limpiados y preparados según las características de la fuente de información para su procesamiento.

3.2.1 Conjunto de Datos

Los lineamientos específicos para repositorios definen que es posible obtener hasta 26 metadatos (embebidos en 16 elementos) de los recursos de información, ya que no todos ellos son de uso obligatorio, entre los que podemos encontrar: Título, Autor, Nivel de acceso, Condición de licencia, Materia, Descripción, Editor, Colaboradores, Tipo de resultado científico, Idioma, Audiencia y sus respectivos identificadores y referencias. Estos metadatos describen algunas características de los recursos alojados, se propuso trabajar y explotar los siguientes cuatro:

- Título (Title)
- Descripción (Description)
- Materia (Subject)
- Idioma (Language)

3.2.2 Cosecha

Para el proceso de extracción de los metadatos se planteó la utilización de la librería Sickle que permite realizar la cosecha de metadatos expuestos bajo el protocolo OAI-PMH, el cual se parametrizó con la URL de consulta de este protocolo en el repositorio de INFOTEC (https://infotec.repositorioinstitucional.mx).

3.2.3 Análisis exploratorio

Para conocer las características propias del set de datos perteneciente al repositorio se debe realizar una revisión exploratoria de los datos con el objetivo de definir las características del tratamiento óptimo que se debe aplicar a la información antes de ser procesada.

3.3 Preprocesamiento

Con el fin de mejorar los resultados es necesario aplicar un proceso de limpieza y normalización de los datos, este paso debe considerar los datos de entrada, estas técnicas tienen como finalidad eliminar la mayor cantidad de datos que no son relevantes y entregar la información de manera homogénea y simplificada, estas pueden iniciar con la eliminación de patrones de cadenas de texto irrelevante como las direcciones de internet, correos electrónicos, números de teléfono, identificadores alfanuméricos, etc., y la normalización del texto para reducir las palabras iguales escritas de diferentes maneras que se puede lograr transformando todas las letras a minúsculas y eliminando los signos de puntuación y caracteres especiales, incluso pueden aplicarse técnicas como Stemming para una mayor reducción de variantes.

También se incluyen procesos de separación en unidades más simples (Tokens) para su procesamiento, estas pueden ser párrafos, sentencias, palabras o incluso silabas dependiendo de las necesidades.

3.4 Modelado y asignación de palabras clave

Para el pesado de términos se consideraron tres técnicas con el fin de evaluar su eficiencia: TF-IDF, Punto de Transición de Goffman, y Entropía, estas técnicas se aplicaron al listado de Tokens extraídos en el preprocesamiento para los metadatos Title, Description y en el resultado de la concatenación del listado de Tokens de los dos anteriores.

Como ya se mencionó la técnica TF-IDF permite inferir la importancia o el peso de una palabra a partir de su ocurrencia dentro de un documento contra la ocurrencia de la misma en el corpus. Este proceso consiste en contabilizar el número de apariciones de cada término en todos los documentos generando una Matriz Documento-Término (Document-Term Matrix o DTM) y así obtener la Frecuencia del Término (TF) o sea cuantas veces aparece un término en cada documento, así como la Frecuencia Inversa del Documento (IDF) a partir de generar una Bolsa de Palabras (Bag of Words) que se forma con el listado de todos los términos y el número de apariciones de este en todo el corpus, para a continuación aplicar la fórmula completa de TF-IDF para cada término, dando como resultado una Matriz Término-Documento (TDM) con los términos, los documentos y los pesos TF-IDF calculados. Para aplicar esta técnica se usó la implementación de TF-IDF de la librería de Sklearn.

La siguiente técnica que se aplicó fue Punto de Transición de Goffman, con la cual se limita el número de palabras relevantes a partir de la frecuencia con que aparecen en un corpus de documentos considerando que las más relevantes se encuentran al centro de esta distribución. Para el cálculo de esta se implementó el algoritmo Python contando el número de apariciones de cada término para generar la DTM, estos datos se agruparon para generar la Bolsa de Palabras del corpus, a partir de esto se contabilizó la cantidad de términos con aparición única y aplicando este valor a la fórmula de Goffman se calculó el punto de corte PTG.

Finalmente, se aplicó el cálculo de la Entropía el cual también contrasta la aparición de una palabra en un documento contra la misma en el corpus. Para este cálculo se implementó el algoritmo en Python, contabilizando las palabras para generar la DTM, estos datos se agruparon en un listado para generar una Bolsa de Palabras del corpus, a continuación se calcula la probabilidad de cada término y se aplica en la fórmula de Entropía para calcular el peso de cada palabra, con esto se generó una TDM con los términos, los documentos y los pesos de LogEntropy calculados.

3.5 Recomendación de recursos relacionados

Para el sistema de recomendaciones se propuso utilizar los términos obtenidos por el modelado de tópicos como evidencia de relación entre los recursos de información, bajo la premisa de que si 2 documentos obtuvieron el mismo término en la extracción del modelado de tópicos, estos podrían tener alguna relación, a partir de este principio se procesaron los listados de términos extraídos en el procesamiento anterior y se generó para cada recurso de información una relación de otros documentos donde también se había extraído el mismo término.

3.5.1 Sistema de recomendaciones

El repositorio de información de INFOTEC está implementado en un desarrollo propio basado en una versión del Dspace, la cual según la documentación del mismo está implementado en Java, y PostgreSQL, sin embargo, otros repositorios están desarrollados en esta y otras plataformas diferentes, debido a esto si se desea hacer la integración en algún repositorio de una funcionalidad de recomendación, una alternativa es agregar los listados generados por este proceso de extracción de tópicos directamente a las bases de datos e implementar la consulta y despliegue en Java y adecuar las vistas del recurso

Una alternativa para no modificar la BD de la aplicación es generar un grupo de servicios de BackEnd donde sea posible consultar la información generada por el proceso de extracción de tópicos y mediante llamados del repositorio realizar peticiones a este BackEnd para agregar estos recursos relacionados como información extra en el despliegue.

3.6 Tecnologías y librerías utilizadas.

3.6.1 Sickle

La librería Sickle (Loesch, 2020) implementa un cliente en Python del protocolo OAI-PMH que permite hacer consultas a repositorios que expongan datos en este protocolo.

3.6.2 Langdetect

La librería Langdetect (Danilák, 2021) implementa la librería “language-detection” de Nakatani Shuyo en Python y permite determinar el idioma de un texto de manera muy simple.

3.6.3 Sklearn

La librería Scikit-learn (Pedregosa et al., 2011) es una librería de Aprendizaje automático (Machine Learning o ML) escrita en Python la cual implementa una gran variedad de algoritmos de Clasificación regresión y agrupamiento entre otros, es ampliamente usado por su arquitectura rapidez y facilidad de uso.

Capítulo 4

Análisis y Evaluación

Capítulo 4. Análisis y Evaluación

4.1 Exploración y preparación de los datos

4.1.1 Cosecha

Se parametrizó la librería Sickle para listar los registros alojados en el repositorio de INFOTEC en la dirección https://infotec.repositorioinstitucional.mx/oai/request, se limitó la consulta para obtener solo los recursos activos y con esto se logró obtener 361 registros en el repositorio (al 30 de septiembre de 2021), de los cuales se pudo obtener una visión general de los metadatos y su uso dentro del repositorio.

4.1.2 Análisis exploratorio

Una de las primeras revisiones que se hizo fue determinar la cantidad de datos nulos, al buscarse este tipo de información se obtuvo el siguiente listado:

	Nulos
title	0
creator	0
contributor	151
publisher	36
date	0
type	0
description	11
audience	121
subject	0
identifier	0
relation	105
rights	0
language	11
format	1
source	325
coverage	356

Cuadro 1: Metadatos nulos
Fuente: Elaboración propia.

Se puede observar qué hay una gran cantidad de metadatos definidos nulos, sin embargo, dentro de los que se seleccionaron para ser explotados solo 11 están en este caso en los metadatos descripción y lenguaje. Lo que nos indica que es un set de datos bastante completo para nuestras necesidades.

Al explorar la serie del metadato Title se encontró que existen algunos casos donde hay más de un elemento asignado

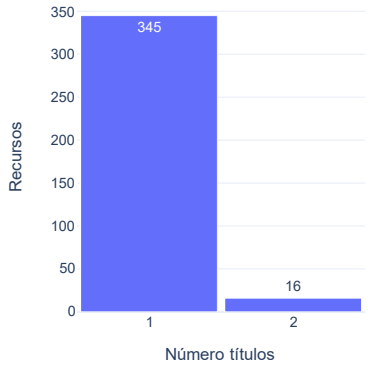

Gráfico 1: Recursos con múltiples títulos
Fuente: Elaboración propia.

Continuando se observó que aunque los datos en su mayoría están en español, en algunos casos se encuentra en inglés, en los registros con más de un título se observa un caso en el que está en varios idiomas (inglés y español) y en otros se agrega información que parecería describir características del recurso de información.

Revisando la longitud de las palabras se observó que los títulos van de 2 a 52 palabras.

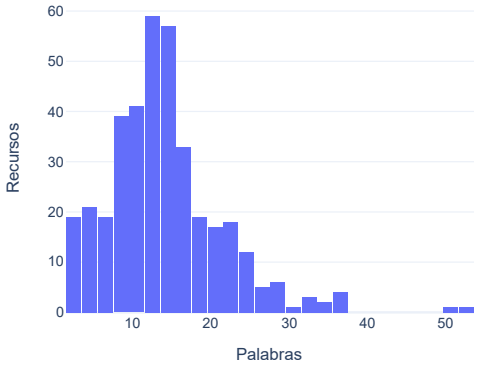

Gráfico 2: Palabras en el título
Fuente: Elaboración propia.

Explorando la serie del metadato Description se pudo observar qué, aunque es obligatorio según los lineamientos no todos los registros contenían esta información además existen algunos casos donde hay más de un elemento asignado.

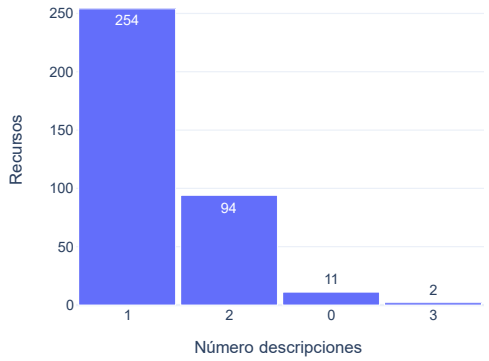

Gráfico 3: Recursos con múltiples descripciones
Fuente: Elaboración propia.

Siguiendo con la revisión de los datos se observó también que en su mayoría están en español y en otros en inglés además en los registros con más de un elemento algunos contienen información en varios idiomas, pero en otros se agrega otra que parecería describir características del recurso de información e incluso solo contienen URLs.

id	title	description
1027/425	Análisis documental sobre el tema del big data y su impacto en los derechos humanos	El presente artículo tiene como objetivo brindar al lector una aproximación sobre los estudios publicados en los últimos años que dan cuenta del manejo de la infraestructura tecnológica y la gestión del conocimiento que se genera a través de los análisis de grandes cúmulos de datos o macrodatos (conocidos también como big data analytics) relacionados al tema de los derechos humanos. Actualmente, los individuos alrededor del mundo pueden ver vulnerados sus derechos humanos a través del manejo indiscriminado de la herramienta big data, ya que la información que se genera día a día y segundo a segundo por medio de los dispositivos tecnológicos —como los teléfonos inteligentes— abarca desde los hábitos de consumo de las personas hasta aspectos de su vida privada, como pueden ser sus creencias religiosas o sus datos biométricos. Así, la vulneración de derechos humanos se puede dar desde la manera en que se generan, almacenan y, en general, se tratan los datos de las personas, quienes en ocasiones desconocen cómo es que se están obteniendo y utilizando sus datos. Como resultado, se encontró que la mayoría de las investigaciones bajo estos parámetros centran el análisis en las distintas normas jurídicas en materia de privacidad y protección de datos, tendientes a regular la manera en que se realiza la minería de datos. Sin embargo, se debe considerar que no solo el derecho a la privacidad se pone en riesgo, sino que existen otros derechos humanos que pueden ser vulnerados al hacer un mal uso de estas tecnologías; por ejemplo, al generar discriminación a partir de la elaboración de listas negras que segreguen a las personas o promuevan el racismo, o al constituir un obstáculo a la libertad de expresión, por mencionar solo algunos casos. Véase reporte "1.1.1.1 Número de publicaciones arbitradas. Enero - Junio de 2020": http://infotec.repositorioinstitucional.mx/jspui/handle/1027/421
1027/427	A zero-knowledge proof system with algebraic geometry techniques	Current requirements for ensuring data exchange over the internet to fight against security breaches have to consider new cryptographic attacks. The most recent advances in cryptanalysis are boosted by quantum computers, which are able to break common cryptographic primitives. This makes evident the need for developing further communication protocols to secure sensitive data. Zero-knowledge proof systems have been around for a while and have been considered for providing authentication and identification services, but it has only been in recent times that its popularity has risen due to novel applications in blockchain technology, Internet of Things, and cloud storage, among others. A new zero-knowledge proof system is presented, which bases its security in two main problems, known to be resistant, up to now, against quantum attacks: the graph isomorphism problem and the isomorphism of polynomials problem. Véase reporte "1.1.1.1 Número de publicaciones arbitradas. Enero - Junio de 2020": http://infotec.repositorioinstitucional.mx/jspui/handle/1027/421

Cuadro 2: Ejemplo de recursos con varias descripciones
Fuente: Elaboración propia.

En cuanto la longitud de los datos se observó que las descripciones van de 1 a 425 palabras

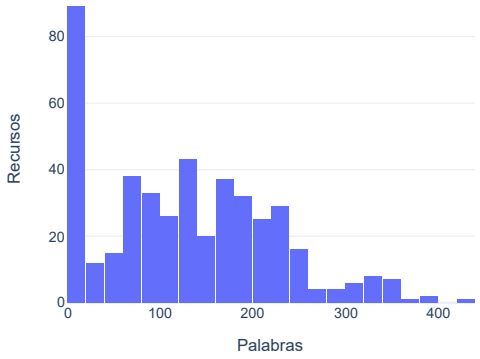

Gráfico 4: Palabras en la descripción
Fuente: Elaboración propia.

Al explorar la serie del metadato Language se pudo observar que no todos los registros contenían esta información a pesar qué es un elemento obligatorio según los lineamientos (CONACYT, 2017b), de la información se vio que: 308 están etiquetados como recursos en español, 42 en inglés y 11 no contenían información.

Revisando los lineamientos estos mencionan que este elemento hace referencia al idioma del recurso de información por lo que podría o no concordar con el título y/o la descripción. No se encontró ninguna manera explícita de identificar el idioma de los metadatos Title y Description.

id	title	description	language
1027/203	Latent Dirichlet Allocation complement in the vector space model for Multi-Label Text Classification	En la tarea de clasificación de texto uno de los principales problemas es elegir qué características dan los mejores resultados. Se pueden utilizar diversas características como palabras, n-gramos, n-gramos sintácticos de varios tipos (etiquetas POS, relaciones de dependencia, mezclas, etc.) o se pueden considerar combinaciones de estas características. Además, se pueden aplicar algoritmos para la reducción de la dimensionalidad de estos conjuntos de características, como la Asignación de Dirichlet Latente (LDA). En este artículo, consideramos la tarea de clasificación de texto de varias etiquetas y aplicamos varios conjuntos de características. Consideramos un subconjunto de archivos multi-etiquetados del corpus de Reuters-21578. Utilizamos valores tf-IDF tradicionales de las características e intentamos considerar e ignorar las palabras de parada. También probamos varias combinaciones de características, como bigrams y unigrams. También experimentamos con la adición de los resultados LDA en Vector Space Models como nuevas características. Estos últimos experimentos tuvieron los mejores resultados.	eng
1027/261	Mexico: insertion of ICT services in global value chains, capabilities and public policy	This study aims to identify the process by which firms venture into the information, technology and communication (ICT) services global value chain and the obstacles they face. We analyse the prevailing mode of governance of the global value chain (GVC) taking into consideration the degree of complexity of transactions, the ability to codify information, customer-supplier interrelations, and suppliers' capabilities to meet buyers' demands. Our enquiry on the mode of governance could not be as precise as we might have expected. We found a group of firms in transition between captive and relational modes of governance. We identified three clusters with different strategies. We show that complexity of services is associated with the mode of governance. Low added value is associated with captive governance. Outsourcing is especially relevant in firms with captive or hierarchical governance, and a substantial part of exports are from firms with relational governance, characterised by their high capabilities, with more complex services.	spa

Cuadro 3: Ejemplo de recursos con títulos y descripciones en diferentes idiomas
Fuente: Elaboración propia.

Explorando la serie del metadato Subject que tiene como propósito determinar el área de conocimiento, esto acotado dentro de un vocabulario controlado definido por Conacyt y que permite agregar palabras claves y códigos de clasificación se observó que en todos los casos la información estaba relacionada con vocabularios controlados o Tesauros (p.ej. Tesauro de la UNESCO, LEM, Tesauro de la Suprema corte de Justicia de la Nación, Etc.) incluso algunos referenciados con múltiples nombres.

	Recursos
cti	361
Tesauro de la UNESCO	119
LEM	103
UNESCO	45
Tesauro Jurídico de la Suprema Corte de Justicia de la Nación	34
Tesauro de la Suprema Corte de Justicia de la Nación	34
Tesauro UNESCO	19
MDPI	6
LEMB	5
LMA	5
Tesauro ISOC de economía	3
LC	2
Tesuaro de la Suprema Corte de Justicia de la Nación	2
Tesauro ITESO	2
Tesuaro de la UNESCO	2
Library of Congress	1
TESAURO DE LA UNESCO	1
IEEE Access	1
Tesauro Jurídico Suprema Corte de Justicia de la Nación	1
Tesauro spines	1
IJSRM	1
Tesauro de la UNSECO	1
MPDI	1
Tesauro del ITESO	1
Tesauro de laa Suprema Corte de Justicia de la Nación	1

Cuadro 4: Ejemplo de orígenes de clasificación
Fuente: Elaboración propia.

Del análisis exploratorio y de las características de los textos que se encontraron y al no haber una identificación explícita del idioma se determina la necesidad de implementar una herramienta de detección de idioma. Con el fin de etiquetar con este los textos tanto del título como de la descripción y poder tratar los datos de manera independiente según el idioma en el que se encuentra, acotándolo para este caso solo a dos idiomas español e inglés.

Se realizó una detección de idioma mediante la librería Langdetect, al aplicarla en el título se determinó que el 15% de los registros estaban en inglés y en tanto que en la descripción el 10% de los registros estaban en esta misma condición.

Además, al encontrarse múltiples elementos en título y descripción se determinó que en caso de que estuvieran en el mismo idioma estos se concatenaran para formar una sola cadena.

4.2 Preprocesamiento

Se aplicó un preprocesamiento de limpieza y normalización, considerando los datos observados durante la exploración, en el cual además de transformar los textos a minúsculas, eliminar números y signos de puntuación, se eliminaron los patrones compatibles con URL.

Después de esta normalización y debido a que la información se encontró en algunos casos separada en múltiples partes y en diferentes idiomas (español e inglés) de manera no etiquetada se agregó un proceso para la identificación de estos idiomas dentro de los datos a procesar y se tomó la decisión de que en caso de que existieran dos elementos del mismo campo en el mismo idioma, estos fueran considerados de manera concatenada como uno solo, y en caso de que no existiese ningún elemento que cumpliera con la condición se procesaba un listado sin Tokens.

Finalmente, se aplicó un proceso de reducción morfológica de palabras (Stemmer) y de eliminación de palabras vacías (Stopwords), finalmente se realizó la separación de Tokens por palabras.

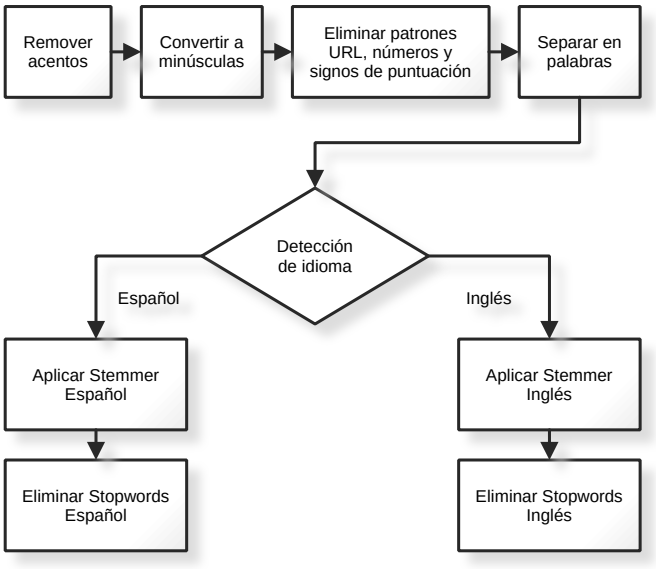

Figura 2: Preprocesamiento
Fuente: Elaboración propia.

Esto se estableció, ya que de la identificación de idioma se observó que había recursos de información que contenían datos en español, inglés y en ambos idiomas

Idioma	title	description
Español	304	306
Ingles	56	36
Ambos	1	8

Cuadro 5: Determinación de Idioma
Fuente: Elaboración propia.

Por lo que fue necesario separar el procesamiento de esta información según el idioma determinado, para después separar los textos en Tokens, de lo cual se obtuvo:

Idioma	Título	Descripcion	Titulo + Descripcion
Español	1150	5552	5711
Ingles	327	1697	1772

Cuadro 6: Tokens del corpus inicial
Fuente: Elaboración propia.

Con el fin de reducir la complejidad de procesamiento se determinó aplicar algunas técnicas para disminuir el tamaño de este corpus. Estos Tokens obtenidos fueron optimizados eliminando las palabras vacías (Stopwords) tanto en español como en inglés con la librería de NLTK, además a los Tokens originalmente obtenidos se les aplicó la técnica de Stemmer con el fin de reducir el número de variantes de la misma palabra, en ambas técnicas lograron reducir el número de Tokens:

	Sin Stopwords			Después de Stemmer		
Idioma	Título	Descripción	Título + Descripción	Título	Descripción	Título + Descripción
Español	1109	5440	5599	964	3229	3330
Ingles	303	1608	1683	310	1252	1305

Cuadro 7: Tokens sin Stopwords y después del Stemmer
Fuente: Elaboración propia.

Al aplicar ambas técnicas en conjunto finalmente se obtuvo un corpus de:

Idioma	Título	Descripcion	Título + Descripcion
Español	924	3126	3227
Ingles	286	1163	1216

Cuadro 8: Tokens sin Stopwords y con Stemmer
Fuente: Elaboración propia.

Obteniendo resultados como los siguientes:

id	title	language	title_lang	title_tk_es	title_tk_en
1027/72	Mapas: representación de los SAD de acuerdo con la naturaleza de su información	spa	es	map; representacion; sad; acuerd; naturalez; informacion	
1027/357	De la teoría del crecimiento económico hacia un cambio de paradigma tecnológico sustentable; From the economic growth theory towards a sustainable technological paradigm shift	spa	es; en	teori; crecimient; econom; haci; cambi; paradigm; tecnolog; sustent	econom; growth; theori; toward; sustain; technolog; paradigm; shift

Cuadro 9: Ejemplo de Tokens extraídos del título
(language: Idioma declarado, title_lang: Idioma determinado, title_tk_es: Tokens en español, title_tk_en: Tokens en inglés)
Fuente: Elaboración propia.

Aquí se puede ver el resultado del procesamiento del título, el idioma determinado de manera no supervisada (title_lang), así como los Tokens identificados tanto en español (title_tk_es), como en inglés (title_tk_en), en el registro 1027/72 se puede ver que en caso de que solo exista el título en español la determinación de Tokens en inglés muestra un conjunto vacío ([ ]) así como el registro 1027/357 en el cual se determinó que los elementos del título corresponden a los dos idiomas, por lo tanto, la determinación de los Tokens quedo separada por idioma.

id	description	language	description_lang	description_tk_es	description_tk_en
1027/167	El presente trabajo es resultado de la investigación realizada como parte de los requisitos señalados para obtener el grado de Maestría en Dirección Estratégica de las Tecnologías de la Información y Comunicación en el Fondo de Información y Documentación para la Industria INFOTEC. La investigación es una implementación de un proyecto, que es uno de los tipos de trabajos finales establecidos en el Reglamento de Estudios de Posgrado de INFOTEC.	spa	es	present; trabaj; result; investigacion; realiz; part; requisit; senal; obten; grad; maestri; direccion; estrateg; tecnologi; informacion; comunicacion; fond; informacion; documentacion; industri; infotec; investigacion; implementacion; proyect; tip; trabaj; final; establec; reglament; estudi; posgr; infotec	
1027/427	Current requirements for ensuring data exchange over the internet to fight against security breaches have to consider new cryptographic attacks. The most recent advances in cryptanalysis are boosted by quantum computers, which are able to break common cryptographic primitives. This makes evident the need for developing further communication protocols to secure sensitive data. Zero-knowledge proof systems have been around for a while and have been considered for providing authentication and identification services, but it has only been in recent times that its popularity has risen due to novel applications in blockchain technology, Internet of Things, and cloud storage, among others. A new zero-knowledge proof system is presented, which bases its security in two main problems, known to be resistant, up to now, against quantum attacks: the graph isomorphism problem and the isomorphism of polynomials problem. Véase reporte "1.1.1.1 Número de publicaciones arbitradas. Enero - Junio de 2020": http://infotec.repositorioinstitucional.mx/jspui/handle/1027/421	eng	en; es	veas; report; numer; public; arbitr; ener; juni	current; requir; ensur; data; exchang; internet; fight; secur; breach; consid; new; cryptograph; attack; recent; advanc; cryptanalysi; boost; quantum; comput; abl; break; common; cryptograph; primit; make; evid; need; develop; communic; protocol; secur; sensit; data; zero; knowledg; proof; system; around; consid; provid; authent; identif; servic; recent; time; popular; risen; due; novel; applic; blockchain; technolog; internet; thing; cloud; storag; among; new; zero; knowledg; proof; system; present; base; secur; two; main; problem; known; resist; quantum; attack; graph; isomorph; problem; isomorph; polynomi; problem

Cuadro 10: Ejemplo de Tokens extraídos de la descripción,
(language: Idioma declarado, description_lang: Idioma determinado, description_tk_es: Tokens en español, description_tk_en: Tokens en inglés)
Fuente: Elaboración propia.

En este ejemplo del procesamiento de la descripción se puede ver, la determinación del idioma de manera no supervisada (description_lang) así como los Tokens identificados tanto en español (description_tk_es), como en inglés (description_tk_en), en el registro 1027/167 se puede ver que en caso de que solo exista la descripción en español la determinación de Tokens en inglés muestra un conjunto vacío ([ ]) así como el registro 1027/427 en el cual se determinó que los elementos de la descripción corresponden a los dos idiomas y, por lo tanto, la determinación de los Tokens quedo separada por idioma.

id	title	description	language	title_lang	description_lang	title_description_tk_es	title_description_tk_en
1027/159	Guía Metodológica Básica para la Implementación de Telefonía Celular como Medio de Contacto con la Ciudadanía	El gobierno electrónico es considerado como una de las vertientes más relevantes dentro de esta revolución, las relaciones del Estado con sus ciudadanos se están dando de una manera vertiginosa y los Gobiernos lo saben bien, por eso les interesa poder usar este medio para atender las necesidades y requerimientos de sus gobernados. Los trámites electrónicos de Gobierno, son uno de las canales más demandados para atender estos 7 requerimientos, de ahí la importancia de entender como la Administración Pública está planteando el uso de la tecnología, la innovación y del cambio tecnológico.	spa	es	es	gui; metodolog; basic; implementacion; telefoni; celul; medi; contact; ciudadani; gobiern; electron; consider; vertient; relev; dentr; revolucion; relacion; ciudadan; dand; maner; vertigin; gobi; sab; bien; interes; pod; usar; medi; atend; neces; requer; gobern; tramit; electron; gobiern; canal; demand; atend; requer; ahi; import; entend; administracion; public; plant; uso; tecnologi; innovacion; cambi; tecnolog	
1027/427	A zero-knowledge proof system with algebraic geometry techniques	Current requirements for ensuring data exchange over the internet to fight against security breaches have to consider new cryptographic attacks. The most recent advances in cryptanalysis are boosted by quantum computers, which are able to break common cryptographic primitives. This makes evident the need for developing further communication protocols to secure sensitive data. Zero-knowledge proof systems have been around for a while and have been considered for providing authentication and identification services, but it has only been in recent times that its popularity has risen due to novel applications in blockchain technology, Internet of Things, and cloud storage, among others. A new zero-knowledge proof system is presented, which bases its security in two main problems, known to be resistant, up to now, against quantum attacks: the graph isomorphism problem and the isomorphism of polynomials problem. Véase reporte "1.1.1.1 Número de publicaciones arbitradas. Enero - Junio de 2020": http://infotec.repositorioinstitucional.mx/jspui/handle/1027/421	eng	en	en; es	veas; report; numer; public; arbitr; ener; juni	zero; knowledg; proof; system; algebra; geometri; techniqu; current; requir; ensur; data; exchang; internet; fight; secur; breach; consid; new; cryptograph; attack; recent; advanc; cryptanalysi; boost; quantum; comput; abl; break; common; cryptograph; primit; make; evid; need; develop; communic; protocol; secur; sensit; data; zero; knowledg; proof; system; around; consid; provid; authent; identif; servic; recent; time; popular; risen; due; novel; applic; blockchain; technolog; internet; thing; cloud; storag; among; new; zero; knowledg; proof; system; present; base; secur; two; main; problem; known; resist; quantum; attack; graph; isomorph; problem; isomorph; polynomi; problem

Cuadro 11: Ejemplo de Tokens extraídos de la concatenación del título y la descripción,
(title_lang: Idioma detectado en el título, description_lang: Idioma detectado en la descripción, title_description_tk_es: Tokens en español, title_description_tk_en: Tokens en inglés)
Fuente: Elaboración propia.

Este sería el resultado de la concatenación del título y la descripción en el registro 1027/159, en este caso la determinación del idioma de manera no supervisada (title_lang, description_lang) así como la concatenación de los Tokens identificados tanto en español (title_description_tk_es), como en inglés (title_description_tk_en) donde también se muestra el caso de que solo exista el elemento en español entonces la determinación de Tokens en el otro idioma contiene un conjunto vacío ([ ]). Igualmente en el registro 1027/427 se determinó que los elementos están en ambos idiomas y también la determinación de los Tokens quedó separada por idioma.

4.3 Modelado y Asignación de palabras clave

Para el modelado y asignación de palabras claves se aplicó el siguiente proceso realizando el cálculo de pesos con las tres técnicas propuestas y luego se compararon los resultados para elegir la técnica que obtenía los mejores resultados para nuestras necesidades.

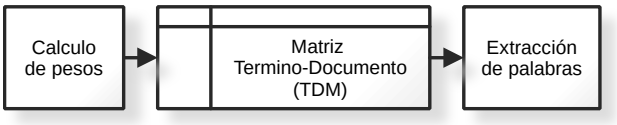

Figura 3: Modelado y asignación de palabras
Fuente: Elaboración propia.

Como ya se mencionó se propusieron tres técnicas que se aplicaron a nuestros datos, inicialmente se usó la técnica TF-IDF implementada en la librería de Sklearn y a partir de la TDM obtenida sé ordenaron y seleccionaron los Tokens de mayor valor hasta obtener al menos 5 para cada documento.

Para el caso de la implementación que se hizo del cálculo del Punto de Transición de Goffman, con el PTG calculado se evaluó la distancia entre la frecuencia de cada Token y este punto y se seleccionaron los términos que más se acercaran dentro del listado de Tokens de cada documento para obtener al menos 5 para cada uno.

Finalmente, para la métrica de Entropía implementada a partir de la TDM calculada para cada documento se ordenaron y extrajeron los términos con valores mayores dentro de los Tokens para obtener al menos 5 para cada uno.

4.3.1 Title

A continuación, se muestra tres ejemplos del resultado de la extracción de palabras claves para el elemento título en español (Los resultados del procesamiento en inglés se pueden consultar en el Anexo I) con cada técnica se puede apreciar que los términos parecen ser relevantes en el título de los recursos de información

id	title	title_idf_es	title_goffman_es	title_etp_es
1027/142	Ciudades inteligentes en Iberoamérica; ejemplos de iniciativas desde el sector privado, la sociedad civil, el gobierno y la academia	iberoamer; priv; socied; ejempl; inici; academi; civil	ciudad; iberoamer; priv; socied; inteligent; sector; gobiern	ciudad; priv; inteligent; sector; gobiern
1027/477	IMPLEMENTACIÓN DE UN MÉTODO DE AJUSTE ESTACIONAL COMBINADO UTILIZANDO CLUSTERING JERÁRQUICO SOBRE PATRONES ESTACIONALES DE SERIES DE TIEMPO	utiliz; tiemp; clustering; seri; combin; jerarqu; estacional	utiliz; patron; implementacion; seri; metod; estacional; ajust	utiliz; patron; implementacion; metod; ajust
1027/107	Recomendaciones Jurídicas para las instituciones financieras en el tratamiento de las huellas dactilares de sus usuarios en el proceso de verificación de la identificación en las sucursales de México	usuari; tratamient; huell; dactilar; sucursal; verificacion; identificacion	institu; financ; tratamient; jurid; recomend; mexic; proces	institu; jurid; recomend; mexic; proces

Cuadro 12: Ejemplo de palabras claves extraídas del título en español
(title_idf_es: TF-IDF, title_goffman: Punto de transición de Goffman, title_etp_es: Entropía)
Fuente: Elaboración propia.

Se puede ver que aparecen los mismos términos cuando comparamos la salida entre cada técnica, por lo que se revisó que tanto sucede este fenómeno y se observa que Goffman y Entropía tienen 5 términos comunes en dos terceras partes de los registros, así como TF-IDF y Goffman tiene de 5 a 7 términos comunes en la mitad de los recursos.

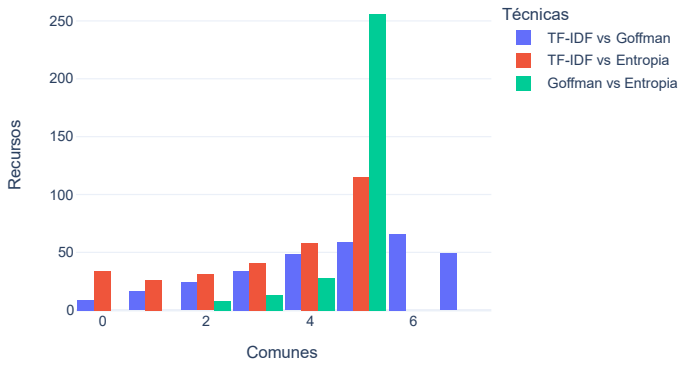

Gráfico 5: Términos comunes del título entre técnicas
Fuente: Elaboración propia.

4.3.2 Description

Para el elemento descripción en español a continuación se muestran dos ejemplos del resultado de la extracción para cada técnica.

id	description	description_idf_es	description_goffman_es	description_etp_es
1027/354	La obra está elaborada a partir del análisis y las experiencias en el ámbito de los soportes digitales; las problemáticas y los medios de intercambio de información que se manejan en la actualidad. Plantea, desde un inicio, otorgar diversos panoramas a situaciones presentadas en varias instituciones de la comunidad científica, con respecto a los formatos y contenidos elaborados gracias a las Tecnologías de la Información y Comunicación. El gran apogeo que ha tenido la aplicación de las TIC en el desarro-llo y creación de obras literarias y de investigación científica y académica trajo con-sigo diversos análisis para el uso, arropamiento y organización de estos materiales.	arrop; elabor; traj; divers; llo; desarr; cientif	digital; divers; aplicacion; experient; comun; organizacion; medi	tecnologi; present; part; uso; informacion
1027/454	El presente estudio expone los lineamientos sobre la implementación de un Ciberejército en el Estado Mexicano, brindando la información necesaria desde el punto de vista jurídico sobre los elementos necesarios para una debida funcionalidad y operación del mismo. Se exponen cuáles son los métodos de creación de los ciberejércitos, los sistemas de prevención, sistemas de ataques y sistemas de resiliencia, además de los métodos utilizados para la ciberinteligencia y el ciberespionaje dentro de los parámetros para ejercer estas actividades. También se exponen los lineamientos jurídicos que justifican el uso de la tecnología aplicada en la materia, además de los estatutos internacionales que regulan y promueven la creación u la actividad de un ciberejército. Solución estratégica	expon; creacion; ciberejercit; metod; lineamient; sistem; adem	activ; element; jurid; mism; tambi; deb; materi	tecnologi; estudi; present; uso; informacion

Cuadro 13: Ejemplo de palabras claves extraídas de la descripción en español
(description_idf_es: TF-IDF, description_goffman: Punto de transición de Goffman, description_etp_es: Entropía)
Fuente: Elaboración propia.

Ahora se revisó si para la descripción también aparecen los mismos términos entre las diferentes técnicas de extracción, para lo que se evaluó el número de términos comunes extraídos por las diferentes técnicas y se observó que a diferencia del título para la mayoría de los recursos cada técnica extrajo diferentes términos.

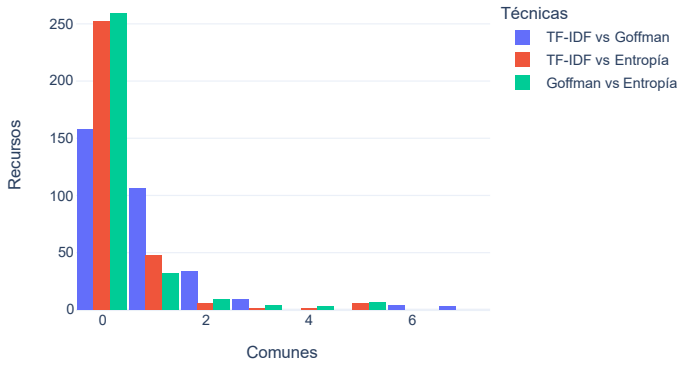

Gráfico 6: Términos comunes de la descripción entre técnicas
Fuente: Elaboración propia.

4.3.3 Title+Description

Finalmente, para los términos extraídos de la concatenación de los elementos título y descripción se puede observar que al igual que la descripción en general cada técnica determinó identificadores diferentes para la mayoría de los recursos.

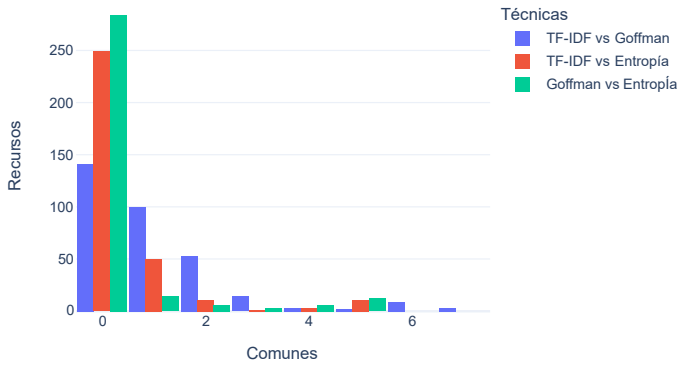

Gráfico 7: Términos comunes del título + descripción entre técnicas
Fuente: Elaboración propia.

Aunque parecería que los resultados del Gráfico 5 validarían las técnicas de Goffman y Entropía al generar los mismos elementos al ver los resultados se pudo identificar (Cuadros 12 y 13) los resultados de la técnica de TF-IDF generaba términos más consistentes para palabras claves que Goffman y Entropía, ya que estos últimos mostraban términos más generales.

4.3.4 Subject

Ya que los lineamientos permiten que dentro del metadato Subject además de la materia también se agreguen palabras claves que categoricen el recurso de información pudiendo venir estas de los autores, se evaluaron los datos contenidos dentro de este elemento con la idea de utilizarlas como etiquetas de referencia para la validación de las palabras claves que se están extrayendo, sin embargo, en el caso de este repositorio no es utilizada esta opción y todos los datos pertenecen a vocabularios controlados como se puede ver en el Cuadro 4. Estos metadatos fueron procesados y en los que fue posible se aislaron estas palabras, y a los datos extraídos se le aplicaron los mismos procesos de limpieza y normalización que al resto de los textos para luego comprobar la cantidad de términos comunes entre estos y los extraídos por cada técnica.

Para el elemento título a continuación vemos el resultado del registro 1027/231

id	title	subject_ext	subject_tk	title_idf_es	title_goffman_es	title_etp_es
1027/231	Propuesta de intervención sobre el uso de Tecnologías de Información y Comunicaciones para la mejora de la gestión educativa en el Distrito Federal	Tecnologías de la información y comunicación; Gestión educacional; Sistema educativo	tecnologi; informacion; comunicacion; gestion; educacional; sistem; educ	mejor; federal; educ; comun; intervencion; gestion; distrit	mejor; propuest; tecnologi; federal; uso; gestion; informacion	propuest; tecnologi; uso; gestion; informacion

Cuadro 14: Ejemplo de Tokens encontrados en el Subject y los extraídos del título con cada técnica
(subject_ext: textos extraídos del Subject, subject_tk: Tokens del Subject, title_idf_es: Tokens con TF-IDF, title_goffman_es: Tokens con Goffman, title_etp_es: Tokens con Entropía)
Fuente: Elaboración propia.

Se pueden identificar algunas palabras comunes entre los conjuntos de términos aislados y los extraídos por las diferentes técnicas, por lo que se revisó en cuantos recursos ocurría.

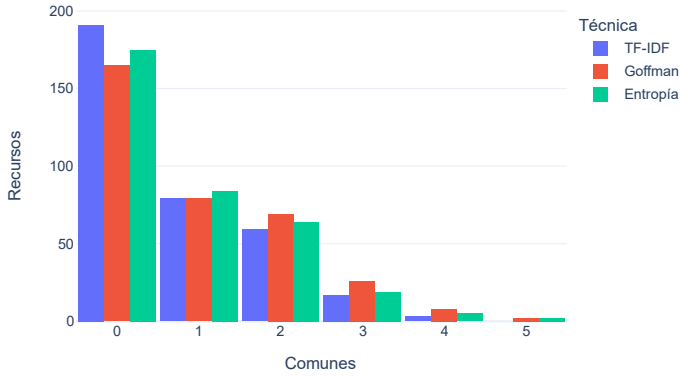

Gráfico 8: Términos comunes extraídos del título y el Subject
Fuente: Elaboración propia.

Y se observó que sin importar la técnica en dos terceras partes de los recursos no había ninguna concordancia entre los términos extraídos del título y los términos aislados del Subject.

Se repitió el ejercicio para el elemento descripción y a continuación vemos el resultado del registro 1027/231.

id	description	subject_ext	subject_tk	description_idf_es	description_goffman_es	description_etp_es
1027/231	Esta propuesta de intervención plantea un modelo basado en estándares para el uso de las Tecnologías de la Información y Comunicación (TIC), que permita establecer un modelo de conectividad viable para los planteles educativos del Distrito Federal, así como la definición de la especificación funcional de un sistema de información para mejorar la gestión y el control escolar, considerando su articulación con los instrumentos de planeación, estadística e indicadores de desempeño en todos los ámbitos del sistema educativo, desde las escuelas del D.F. hasta las instancias de coordinación en la Secretaría de Educación Pública.; Caso de estudio	Tecnologías de la información y comunicación; Gestión educacional; Sistema educativo	tecnologi; informacion; comunicacion; gestion; educacional; sistem; educ	viabl; educ; instanci; especificacion; plantel; escuel; f	mejor; establec; federal; consider; intervencion; gestion; educacion	propuest; tecnologi; asi; uso; informacion

Cuadro 15: Ejemplo de Tokens encontrados en el Subject y los extraídos de la descripción con cada técnica (subject_ext: textos extraídos del Subject, subject_tk: Tokens del Subject, description_idf_es: Tokens con TF-IDF, description_goffman_es: Tokens con Goffman, description_etp_es: Tokens con Entropía)
Fuente: Elaboración propia.

Se puede observar que en este caso también se encuentran algunos términos comunes, por lo que también se procedió a comprobar en cuantos recursos sucedía.

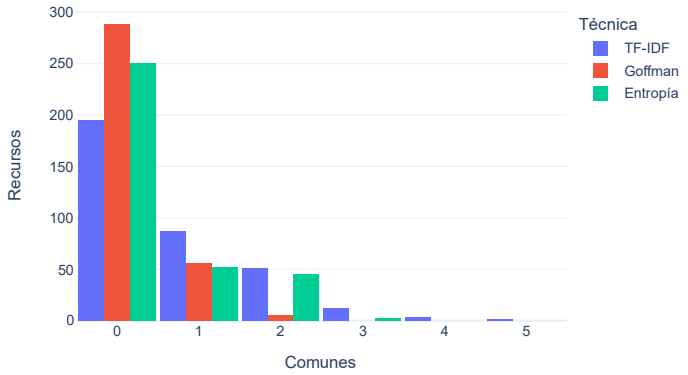

Gráfico 9: Términos comunes extraídos de la descripción y el Subject
Fuente: Elaboración propia.

También se observó que en este caso el número de términos comunes era aún más bajo sobre todo en las técnicas de Goffman y Entropía.

Finalmente, se realizó el ejercicio con la concatenación del título con la descripción y a continuación vemos el resultado del registro 1027/231.

id	title	description	subject_ext	subject_tk	title_description_idf_es	title_description_goffman_es	title_description_etp_es
1027/231	Propuesta de intervención sobre el uso de Tecnologías de Información y Comunicaciones para la mejora de la gestión educativa en el Distrito Federal	Esta propuesta de intervención plantea un modelo basado en estándares para el uso de las Tecnologías de la Información y Comunicación (TIC), que permita establecer un modelo de conectividad viable para los planteles educativos del Distrito Federal, así como la definición de la especificación funcional de un sistema de información para mejorar la gestión y el control escolar, considerando su articulación con los instrumentos de planeación, estadística e indicadores de desempeño en todos los ámbitos del sistema educativo, desde las escuelas del D.F. hasta las instancias de coordinación en la Secretaría de Educación Pública.; Caso de estudio	Tecnologías de la información y comunicación; Gestión educacional; Sistema educativo	tecnologi; informacion; comunicacion; gestion; educacional; sistem; educ	viabl; federal; educ; intervencion; gestion; plantel; distrit	mejor; federal; consider; comun; intervencion; educacion; control	propuest; tecnologi; asi; uso; informacion

Cuadro 16: Ejemplo de Tokens encontrados en el Subject y los extraídos de la concatenación de título y descripción con cada técnica
(subject_ext: textos extraídos del Subject, subject_tk: Tokens del Subject, title_description_idf_es: Tokens con TF-IDF, title_description_goffman_es: Tokens con Goffman, title_description_etp_es: Tokens con Entropía)
Fuente: Elaboración propia.

Se observó el mismo comportamiento mostrando términos generales, por lo que se efectuó la evaluación de que tanto ocurría este fenómeno.

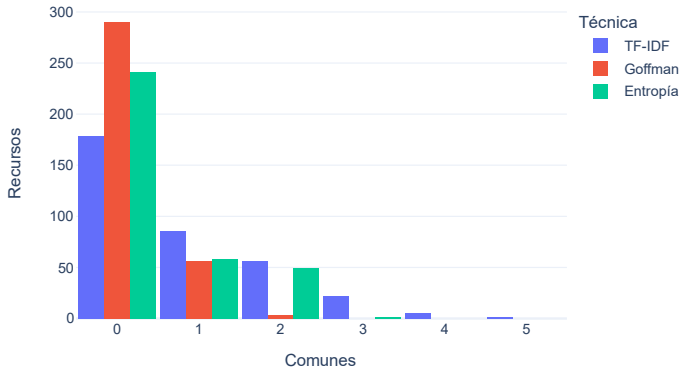

Gráfico 10: Términos comunes extraídos del título+descripción y el Subject
Fuente: Elaboración propia.

El resultado no fue muy diferente al presentado con solo la descripción, igualmente en el caso de las técnicas Goffman y Entropía tuvieron menor concordancia.

4.4 Recomendación de recursos

Como se planteó se analizaron los términos extraídos de cada recurso y se buscó su aparición en otros recursos como evidencia de relación, un ejemplo del resultado de la relación obtenida usando TF-IDF para el elemento título en español del registro 1027/387, usando esta técnica se extrajeron 7 términos y se pudo relacionar con 5 registros, se puede apreciar que en todos los casos los registros tienen cierta relación con el principal.

id	title	title_idf_es	title_related
1027/387	Blockchain en la optimización de la comprobación del gasto en actividades de vigilancia del espectro radioeléctrico	comprobacion; radioelectr; vigil; espectr; blockchain; gast; optimizacion	
1027/380	Entorno regulatorio del precio del espectro radioeléctrico y sus efectos en la introducción de tecnologías móviles de quinta generación en México	entorn; radioelectr; espectr; quint; generacion; preci; regulatori	radioelectr; espectr
1027/473	Estrategias de sensibilización y confianza para el uso de Blockchain en transferencias bancarias en México	bancari; transferent; blockchain; sensibilizacion; uso; confianz; estrategi	blockchain
1027/444	Optimización de la distribución de vacunas para prevenir epidemias	epidemi; preven; vacun; distribucion; optimizacion	optimizacion
1027/126	Diagnóstico y propuesta de mejora en procesos de un sistema de gestión gubernamental: Caso sistema para la evaluación de resultados de los organos de vigilancia y control; Propuesta de intervención	propuest; diagnost; vigil; intervencion; result; sistem; organ	vigil
1027/160	Caso: Secretaría de Hacienda y Crédito Público; La eficiencia del gasto en TIC en la administración Pública Federal	eficient; secretari; haciend; federal; public; credit; gast	gast

Cuadro 17: Ejemplo de recursos relacionados con un registro por el título usando TF-IDF
(title_idf_es: Tokens por TF-IFD, title_related: Tokens comunes)
Fuente: Elaboración propia.

En el caso de la técnica de Goffman se extrajeron 7 términos y se pudo relacionar con otros 4 registros de los cuales se aprecia cierta relación a excepción del 1027/428.

id	title	title_goffman_es	title_related
1027/387	Blockchain en la optimización de la comprobación del gasto en actividades de vigilancia del espectro radioeléctrico	activ; vigil; radioelectr; espectr; blockchain; gast; optimizacion	
1027/380	Entorno regulatorio del precio del espectro radioeléctrico y sus efectos en la introducción de tecnologías móviles de quinta generación en México	efect; tecnologi; movil; introduccion; radioelectr; espectr; mexic	radioelectr; espectr
1027/473	Estrategias de sensibilización y confianza para el uso de Blockchain en transferencias bancarias en México	transferent; blockchain; sensibilizacion; uso; confianz; mexic; estrategi	blockchain
1027/444	Optimización de la distribución de vacunas para prevenir epidemias	epidemi; preven; vacun; distribucion; optimizacion	optimizacion
1027/428	1.5.1.1 Número de Actividades de Divulgación Científica y Tecnológica. Enero - Junio 2020	activ; tecnolog; juni; numer; divulgacion; ener; cientif	activ

Cuadro 18: Ejemplo de recursos relacionados con un registro por el título usando Goffman
(title_goffman_es: Tokens por Goffman, title_related: Tokens comunes)
Fuente: Elaboración propia.

El resultado para la entropía se extrajeron 5 términos y se logró relacionar con otros 3 registros los cuales muestran cierta relaciona a excepción del 304.

id	title	title_etp_es	title_related
1027/387	Blockchain en la optimización de la comprobación del gasto en actividades de vigilancia del espectro radioeléctrico	activ; vigil; blockchain; gast; optimizacion	gast; vigil; blockchain; optimizacion; activ
1027/473	Estrategias de sensibilización y confianza para el uso de Blockchain en transferencias bancarias en México	transferent; blockchain; uso; mexic; estrategi	blockchain
1027/444	Optimización de la distribución de vacunas para prevenir epidemias	epidemi; preven; vacun; distribucion; optimizacion	optimizacion
1027/428	1.5.1.1 Número de Actividades de Divulgación Científica y Tecnológica. Enero - Junio 2020	activ; tecnolog; juni; numer; ener	activ

Cuadro 19: Ejemplo de recursos relacionados con un registro por el título usando Entropía
(title_etp_es: Tokens por Entropía, title_related: Tokens comunes)
Fuente: Elaboración propia.

Sí revisamos la cantidad de registros que relaciono cada técnica de los términos extraídos del título, se puede observar que TF-IDF relaciono no más de 20 registros en la mayoría de sus registros mientras que Goffman y Entropía la media está aproximadamente en 50 además que relacionaron más recursos que TF-IDF.

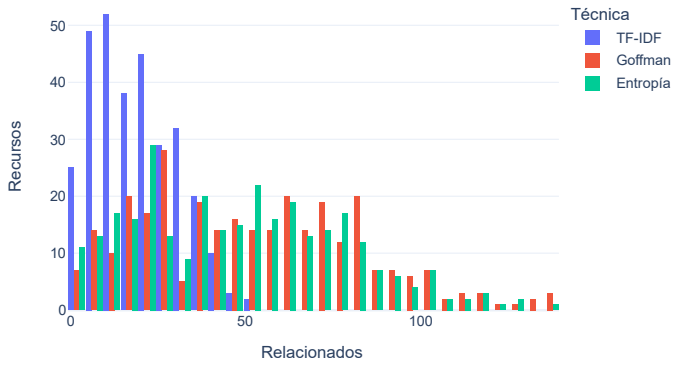

Gráfico 11: Recursos relacionados por título por técnica
Fuente: Elaboración propia.

Además, si calculamos la distancia de Jaccard entre los conjuntos de recursos relacionados por cada una de las técnicas se observa que Goffman y Entropía se encuentran por encima del 0.6 lo que se interpreta como que relacionan los mismos registros

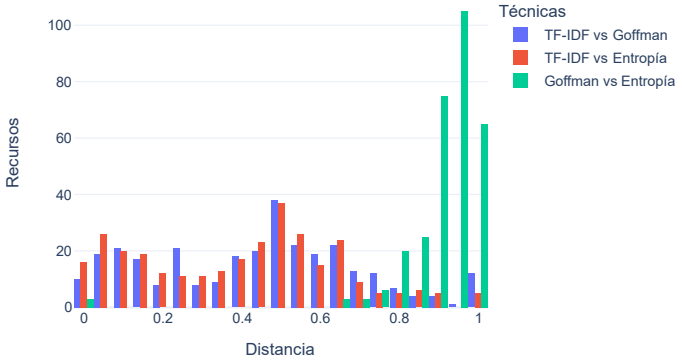

Gráfico 12: Distancia Jaccard de recursos relacionados por el título entre técnicas
Fuente: Elaboración propia.

Para el elemento descripción en español a continuación se muestra el resultado de la relación para cada técnica para el registro 1027/106, usando el TF-IDF se extrajeron 7 términos y se pudo relacionar con 5 registros.

id	description	description_idf_es	description_related	title
1027/106	El Manual Administrativo de Aplicación General en Materia de Tecnologías de Información y Comunicaciones (MAAGTIC) es una norma orientada a extender las buenas prácticas en gestión de servicios tecnológicos a todas las dependencias de la administración pública federal. Esta guía constituye una herramienta para acelerar los trabajos de implantación, acortar las brechas que prevalezcan en las instituciones y prevenirlas con oportunidad frente a errores comunes y prácticas inadecuadas. Asimismo, contribuye a su adopción ágil y efectiva, a través de recomendaciones prácticas, sugerencias útiles y ejemplos reales.	inadecu; practic; comun; acort; sugerent; maagtic; prevalezc	comun; acort; prevalezc; maagtic; practic; sugerent; inadecu	Guía práctica para acelerar la adopción del MAAGTIC en la Administración Pública Federal
1027/175	El Gobierno Federal, por conducto de la Secretaría de la Función Pública (SFP), implementó y coordinó un programa conocido como “Reforma Regulatoria”, cuya intención fue reducir significativamente la normatividad interna de las dependencias y entidades de la Administración Pública Federal. Resultado de esta iniciativa, surgió la imposición de un conjunto de normas de observancia obligatoria para todas las entidades y dependencias que, en el caso de las Tecnologías de Información y Comunicaciones, se trató de un “Modelo de Gobierno” estandarizado y aplicable a través de la implementación de un “Manual Administrativo de Aplicación General en Materia de Tecnologías de Información y Comunicaciones” (MAAGTIC), elemento fundamental para el desarrollo del modelo propuesto.	federal; surgi; entidad; dependent; comun; imposicion; maagtic	comun; maagtic	Desarrollo de un modelo simplificado de gestión de TIC para agencias gubernamentales de estructuras reducidas.
1027/254	El constante avance en materia de tecnologías de la información, seguridad y comunicaciones ha llevado no solo a las empresas del sector privado si no a las instituciones públicas a mejorar sus procesos para incluir las Tecnologías de Información y Comunicaciones (TIC) como herramienta de apoyo fundamental en la automatización de estos. En este trabajo se pretende plasmar la solución a un problema mediante el uso de tecnologías de la información en una delegación del Gobierno del Distrito Federal, automatizando el proceso en materia de gestión y control documental.; Reporte analítico de experiencia laboral	tecnologi; automatizacion; materi; comun; delegacion; distrit; informacion	comun	Desarrollo e implementación de sistema automatizado de gestión de correspondencia en la Delegación Miguel Hidalgo
1027/174	El patrimonio cultural inmaterial entre los pueblos originarios y entre los individuos se transmite en gran medida de forma oral, mediante la repetición de los contenidos. La particularidad de esa forma de transferencia de la memoria de una comunidad es que las experiencias culturales se modifican, se enriquecen e inclusive se pierden para siempre. Las tradiciones y las costumbres en México nos dotan de identidad y son el reflejo de miles de años de cultura de nuestro pueblo, por lo que no podemos permitir que se pierdan. La hipótesis de éste trabajo es la siguiente: “si a una organización social de orden local, se le dota de un repositorio con herramientas tecnológicas de gestión de contenidos fáciles de usar; al estar cerca de la comunidad, puede ser motor de recopilación y sistematización de ésta información en formato de texto, audio o video, mediante el apoyo de miembros de la propia comunidad que gustan de la tecnología y que se organizan en redes sociales para potenciar el conocimiento”. Dicha organización social puede ser gubernamental, no gubernamental o una combinación de ambos. Para los fines de éste trabajo, se considerará la inserción del proyecto en un gobierno municipal.	puebl; dot; pierd; cultural; social; comun; conten	comun	Propuesta de Repositorio Digital Colaborativo para Salvaguardar el Patrimonio Cultural Inmaterial

Cuadro 20: Ejemplo de recursos relacionados con un registro por la descripción usando TF-IDF
(description_idf_es: Tokens por TF-IFD, description_related: Tokens comunes)
Fuente: Elaboración propia.

Aquí con la técnica Goffman también se obtuvieron 7 términos, pero se lograron relacionar 1027/106 recursos incluidos los 5 determinados por TF-IDF, aquí ya se puede observar que los términos detectados son más generales lo que podría explicar el gran número de recursos relacionados.

id	description	description_goffman_es	description_related	title
1027/106	El Manual Administrativo de Aplicación General en Materia de Tecnologías de Información y Comunicaciones (MAAGTIC) es una norma orientada a extender las buenas prácticas en gestión de servicios tecnológicos a todas las dependencias de la administración pública federal. Esta guía constituye una herramienta para acelerar los trabajos de implantación, acortar las brechas que prevalezcan en las instituciones y prevenirlas con oportunidad frente a errores comunes y prácticas inadecuadas. Asimismo, contribuye a su adopción ágil y efectiva, a través de recomendaciones prácticas, sugerencias útiles y ejemplos reales.	federal; general; aplicacion; comun; gestion; administracion; materi	comun; administracion; materi; federal; aplicacion; general; gestion	Guía práctica para acelerar la adopción del MAAGTIC en la Administración Pública Federal
1027/175	El Gobierno Federal, por conducto de la Secretaría de la Función Pública (SFP), implementó y coordinó un programa conocido como “Reforma Regulatoria”, cuya intención fue reducir significativamente la normatividad interna de las dependencias y entidades de la Administración Pública Federal. Resultado de esta iniciativa, surgió la imposición de un conjunto de normas de observancia obligatoria para todas las entidades y dependencias que, en el caso de las Tecnologías de Información y Comunicaciones, se trató de un “Modelo de Gobierno” estandarizado y aplicable a través de la implementación de un “Manual Administrativo de Aplicación General en Materia de Tecnologías de Información y Comunicaciones” (MAAGTIC), elemento fundamental para el desarrollo del modelo propuesto.	element; federal; aplicacion; implement; comun; administracion; materi	comun; administracion; materi; federal; aplicacion	Desarrollo de un modelo simplificado de gestión de TIC para agencias gubernamentales de estructuras reducidas.
1027/227	Las instituciones públicas pueden utilizar las Tecnologías de Información y Comunicación (TIC) con tres objetivos: mejorar sus procesos internos, ofrecer servicios a la ciudadanía o mejorar trámites con la iniciativa privada. El término comúnmente aceptado para el uso de las TIC en el gobierno es e-gobierno. El gobierno mexicano ha hecho uso de las TIC por varias décadas. Inicialmente, conforme al avance en materia de cómputo, se utilizaban equipos especializados en el tratamiento y almacenamiento de la información, con sistemas de información hechos a la medida y muy orientados a los procesos internos. El presente proyecto considera que tres especificaciones pueden dar respuesta a estas interrogantes: la utilización de arquitecturas empresariales, ITIL y COBIT. La primera en el sentido de planeación estratégica tecnológica, dónde estamos, a dónde hay que llegar y qué hacer para llegar; la segunda en el sentido de un modelo de servicios en materia de TIC institucional; y el último en el sentido del control y verificación del modelo completo. La integración de un modelo de administración de TIC general, empleando el método analítico-sintético y usando como ejemplo la Comisión de Derechos Humanos del Distrito Federal (CDHDF).; Proyecto de caso de estudio	mejor; federal; especif; comun; human; administracion; materi	administracion; comun; materi; federal	Modelo de administración de TIC para la Administración Pública. El caso de la CDHDF
1027/126	La Secretaría de la Función Pública (SFP), como dependencia del Poder Ejecutivo Federal, en el ámbito de sus atribuciones y facultades que le encomienda la Ley Orgánica de la Administración Pública Federal (LOAPF, 2015), así como otras leyes, reglamentos, y demás disposiciones aplicables, vigila que los servidores públicos federales se apeguen a la legalidad durante el ejercicio de sus funciones, sanciona a los que no lo hacen así; promueve el cumplimiento de los procesos de control y fiscalización del gobierno federal, de disposiciones legales en diversas materias, dirige y determina la política de compras públicas de la Federación, coordina y realiza auditorías sobre el gasto de recursos federales, coordina procesos de desarrollo administrativo, gobierno digital, opera y encabeza el Servicio Profesional de Carrera (SPC), coordina la labor de los Órganos Internos de Control (OIC) en cada dependencia del gobierno federal y evalúa la gestión de las entidades, también a nivel federal.	federal; materi; tambi; cad; gestion; administracion; polit	administracion; gestion; materi; federal	Diagnóstico y propuesta de mejora en procesos de un sistema de gestión gubernamental: Caso sistema para la evaluación de resultados de los organos de vigilancia y control; Propuesta de intervención

Cuadro 21: Ejemplo de recursos relacionados con un registro por la descripción usando Goffman
(description_goffman_es: Tokens por Goffman, description_related: Tokens comunes)
Fuente: Elaboración propia.

Para la técnica de Entropía se obtuvieron 5 términos, pero en este caso se relacionaron 245 recursos, también se puede observar que los términos extraídos son generales lo que también podría explicar la gran cantidad de recursos relacionados

id	description	description_etp_es	description_related	title
1027/106	El Manual Administrativo de Aplicación General en Materia de Tecnologías de Información y Comunicaciones (MAAGTIC) es una norma orientada a extender las buenas prácticas en gestión de servicios tecnológicos a todas las dependencias de la administración pública federal. Esta guía constituye una herramienta para acelerar los trabajos de implantación, acortar las brechas que prevalezcan en las instituciones y prevenirlas con oportunidad frente a errores comunes y prácticas inadecuadas. Asimismo, contribuye a su adopción ágil y efectiva, a través de recomendaciones prácticas, sugerencias útiles y ejemplos reales.	trabaj; tecnologi; public; general; informacion	public; tecnologi; informacion; trabaj; general	Guía práctica para acelerar la adopción del MAAGTIC en la Administración Pública Federal
1027/422	Este reporte analítico de experiencia laboral tiene como finalidad identificar y facilitar, al interior de las dependencias y entidades de la Administración Pública Federal, los requisitos que deben observarse para proteger los datos personales en las contrataciones que éstas celebren con los particulares en servicios relacionados con cómputo en la nube. El reporte se adentra en el estudio de los procedimientos de contratación que lleva a cabo la administración pública para cumplir con las obligaciones que tiene para con sus gobernados. Después se abordan las áreas de oportunidad en la contratación de servicios de cómputo en la nube, la regulación existente y las consideraciones de la administración pública para proteger los datos personales para contratar servicios de cómputo en la nube. Posteriormente se mencionan los procedimientos de contratación que realiza INFOTEC que como ente público es uno de los principales proveedores especializado en tecnologías de la información y comunicación. Finalmente se aborda el compromiso de la administración pública en el cumplimiento que mandata la constitución para la protección de los datos personales de sus gobernados. Reporte analítico de experiencia laboral	estudi; tecnologi; public; realiz; informacion	public; tecnologi; informacion	Requisitos que deben cumplirse para proteger los datos personales en la contratación de servicios de cómputo en la nube en la administración pública federal
1027/167	El presente trabajo es resultado de la investigación realizada como parte de los requisitos señalados para obtener el grado de Maestría en Dirección Estratégica de las Tecnologías de la Información y Comunicación en el Fondo de Información y Documentación para la Industria INFOTEC. La investigación es una implementación de un proyecto, que es uno de los tipos de trabajos finales establecidos en el Reglamento de Estudios de Posgrado de INFOTEC.	trabaj; tecnologi; present; part; informacion	informacion; tecnologi; trabaj	Metodología de implementación de un modelo de procesos para PyMES de software: norma mexicana NMX-I-059.NYCE-2005
1027/259	"El empleo de las tecnologías de información y comunicaciones (TIC) en las tareas de la administración pública data de varias décadas. Sus promotores iniciales las aprovecharon para aligerar la carga de funciones administrativas, agilizar la realización de censos y estadísticas nacionales, así como facilitar actividades de monitoreo y control. Aquellos pioneros —profesores y servidores públicos— incursionaron con éxito en el uso de las TIC en labores de gobierno, aun cuando en esos años no era posible dimensionar su potencial en la oferta de servicios para la población." (SIC)	tecnologi; asi; public; uso; informacion	public; tecnologi; informacion	Las nuevas tecnologías de información y el cambio necesario en la administración pública en México

Cuadro 22: Ejemplo de recursos relacionados con un registro por la descripción usando Entropía
(description_etp_es: Tokens por Entropía, description_related: Tokens comunes)
Fuente: Elaboración propia.

Si evaluamos en número de recursos que relacionan estos términos es muy diferente entre cada una de las técnicas mientras que TF-IDF la mayoría relacionan hasta 20 Goffman relaciona 150 y entropía relaciona más de 250.

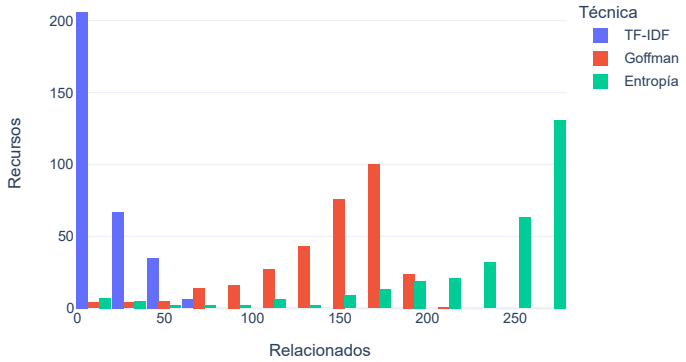

Gráfico 13: Recursos relacionados por descripción por técnicas
Fuente: Elaboración propia.

En este caso la distancia de Jaccard entre los conjuntos de recursos relacionados por cada técnica en su mayoría está por debajo del 0.5, sin embargo, esto es predecible, tan solo porque los conjuntos de recursos relacionados son de tamaños muy dispares entre cada técnica.

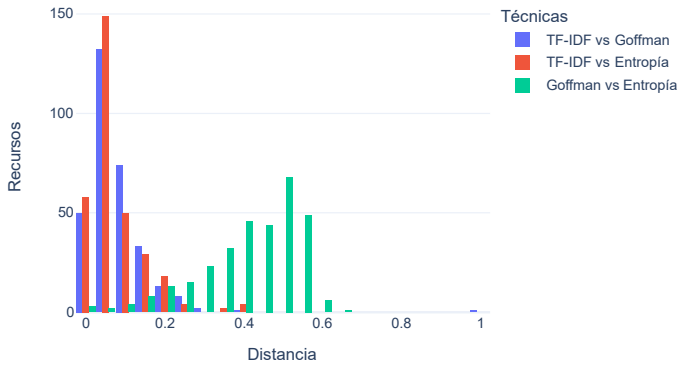

Gráfico 14: Distancia Jaccard de recursos relacionados por la descripción entre técnicas
Fuente: Elaboración propia.

Finalmente, para la concatenación de los elementos título y descripción se puede observar que se repite el comportamiento observado en el procesamiento de solo la descripción donde el TF-IDF relacionó pocos elementos, pero Goffman y Entropía relacionaron varias veces más recursos.

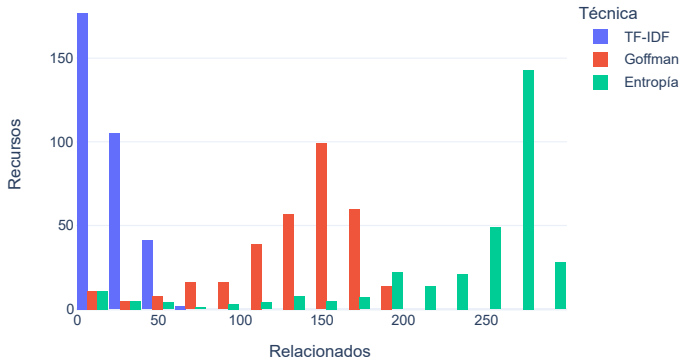

Gráfico 15: Recursos relacionados por título + descripción entre técnicas
Fuente: Elaboración propia.

De igual forma la distancia de Jaccard entre los diferentes conjuntos tan solo por sus diferencias de dimensiones las distancias no superan los 0.5.

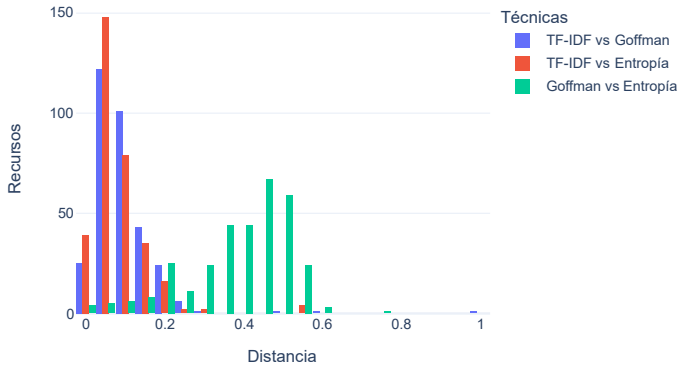

Gráfico 16: Distancia Jaccard de recursos relacionados por el título + descripción entre técnicas
Fuente: Elaboración propia.

Al comparar los resultados y se pudo observar en una muestra de recursos de información que la técnica de TF-IDF generaba un grupo de relaciones más acotadas y que Goffman y Entropía relacionaban muchos más recursos lo que podría significar que se están seleccionando términos generales.

Estos resultados difieren con los reportados para el indexado Latent Semantic Indexing (LSI) entre la TF-IDF y la Entropia (Dumais, 1991), ya que lo esperado seria que esta última entregara mejores resultados, sin embargo, al ser una aplicación diferente de la misma métrica para nuestras necesidades TF-IDF generan un número más acotado de recursos relacionados.

4.5 Agrupación de recursos

Ya que materia también caracterizan la información se evaluó si a partir de los términos extraídos a los títulos en español por TF-IDF se podía encontrar una relación con las áreas del conocimiento, dado que en los lineamientos este nivel de catalogación es obligatoria y debe estar relacionados con algunas de las 7 áreas del conocimiento del catálogo de Conacyt (CONACYT, s/f), fueron procesados y se encontró que en el caso de este repositorio solo está relacionado con 4 de las 7:

- Ingeniería y Tecnología
- Ciencias Sociales
- Ciencias Físico Matemáticas y Ciencias de la Tierra
- Humanidades y Ciencias de la Conducta

Las tres áreas restantes: Biología y Química, Ciencias Agropecuarias y Biotecnología y Medicina y Ciencias de la Salud, no aparecen muy probablemente porque no están relacionados con los campos de interés de INFOTEC.

Para identificar si era posible relacionar de manera no supervisada estas áreas y los documentos se usó la matriz de pesos calculados mediante TF-IDF, la cual se redujo dimencionalmente mediante la técnica LSA (Latent Semantic Analysis) para obtener los 5 vectores principales de los cuales se obtuvieron los 20 términos de mayor peso para generar una nube de palabras de cada uno de estos y se observó que no tienen correspondencia visible a las 5 áreas de conocimiento

Gráfico 17: Nubes de palabras de los principales vectores encontrados
Fuente: Elaboración propia.

En un vistazo rápido se puede ver que hay términos que se repiten entre cada una de las nubes generadas, pero las palabras principales si están relacionadas con las aéreas de interés del centro. También se utilizó estos vectores para obtener los documentos de mayor peso para estos 5 vectores principales.

1	2	3	4	5
Entrevistas a líderes mexicanos en Tecnologías de la Información y Comunicación (TIC). Hacia un modelo de e-Gobierno municipal para México	Modelo de Transferencia de Conocimiento	Protección de datos personales a través de herramientas de procesamiento automatizado de datos: desafíos y recomendaciones	Tecnologías de información y comunicación en la administración pública: conceptos, enfoques, aplicaciones y resultados	Casos de Estudio
Datos abiertos México 2017 - 2018	Transferencia de conocimiento organizacional: Modelo y solución	La protección de datos personales y la privacidad en México: concepto y regulación	Uso y apropiación de las tecnologías de la información y comunicación (TIC) en las Pymes de Aguascalientes	Modelo de administración de TIC para la Administración Pública. El caso de la CDHDF
Propuesta de un modelo de auditoría integral que garantice el cumplimiento de disposiciones en materia de datos personales y seguridad de la información en México	Hacia un modelo de transferencia de conocimiento para las organizaciones mexicanas	La protección de datos personales y la privacidad en México: concepto y regulación II	Las nuevas tecnologías de información y el cambio necesario en la administración pública en México	Caso: Secretaría de Hacienda y Crédito Público; La eficiencia del gasto en TIC en la administración Pública Federal
La protección de datos personales y la privacidad en México: concepto y regulación	La transferencia de conocimiento en las organizaciones	La protección de datos personales de menores en redes sociales: desafíos y recomendaciones	La cadena de las tecnologías de información y comunicación: Política pública y estrategias empresariales	Casos de estudio Volumen
Implementación de un modelo de gestión de conocimiento y tecnología para una PyME de TIC en México, D.F.	La gestión del conocimiento en la Oficina de Transferencia de Conocimiento de INFOTEC	La transferencia de conocimiento en las organizaciones	Estado, apropiación social de las tecnologías de la información y comunicación y pobreza	Implentación de servicios de correo electrónico y portales web, para la administración pública federal
Sistematización de obligaciones en materia de protección de datos personales para el sector público en el Estado de México	El conocimiento organizacional	La gestión del conocimiento en la Oficina de Transferencia de Conocimiento de INFOTEC	Desafíos y oportunidades para los Derechos Humanos frente a las Tecnologías de la Información y la Comunicación (TIC)	Tres municipios mexicanos exitosos: estudios de caso sobre e-Gob

Cuadro 23: Ejemplo de documentos pertenecientes a cada uno de los 5 grupos
Fuente: Elaboración propia.

Y se puede observar que hay documentos (verde) que se encuentran en más de un grupo. Finalmente, fueron graficados estos vectores tratando de visualizar de manera global la información como otra manera de identificar si estas agrupaciones correspondían a la distribución de las áreas del conocimiento.

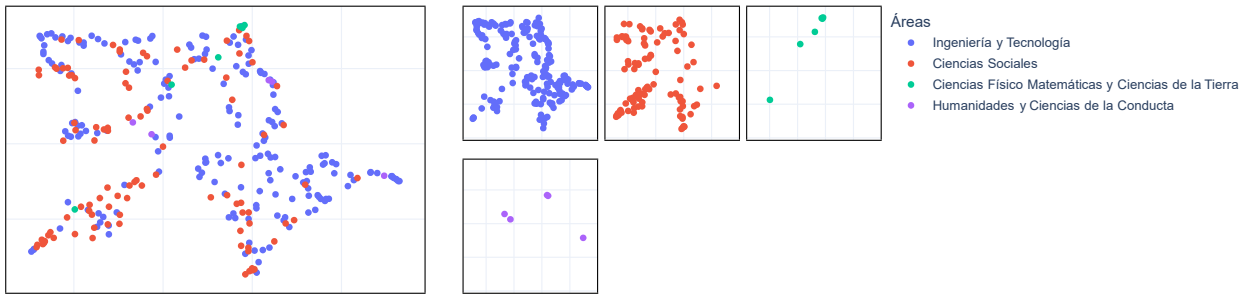

Gráfico 18: Vectores extraídos por LSA agrupados por Área del conocimiento
Fuente: Elaboración propia.

Como se puede ver no se logró visualizar ninguna agrupación relacionada con las áreas del conocimiento, sin embargo, se aprecian completamente integrados los grupos de Ingeniería y tecnología y Ciencias sociales que es precisamente la vocación del INFOTEC, otra hipótesis seria que los grupos que se formaron podrían estar relacionados con las diversas líneas de trabajo de los posgrados que ofrece el centro, sin embargo, no es posible corroborarlo porque no se encuentran etiquetados.

Se sabe que LSA puede relacionar documentos de manera semejante hasta en un 60% como lo haría el juicio humano (Pincombe, 2004), pero en nuestro caso no fue posible apreciar esas agrupaciones con alguna de las categorías etiquetadas.

4.6 Comparación con otros repositorios

Como se había propuesto para evaluar el desempeño de las técnicas utilizadas se aplicaron los mismos procesos para la extracción de términos y de documentos relacionados en otros repositorios, el del Centro de Investigación y Docencia Económicas, A.C. (CIDE) y del Centro de Investigaciones Biológicas del Noroeste S.C. (CIBNOR).

Los resultados se detallan en los Anexos II y III, en términos generales los resultados fueron consistentes con los obtenidos en el repositorio de INFOTEC donde la técnica TF-IDF logra mejores resultados para las necesidades de esta propuesta, sin embargo, un hallazgo destacable es que los usuarios de estos repositorios si hacen uso de las palabras claves de autor en el metadato de Subject, lo que permitió probar la idea descrita en la sección 4.3.4, para usar estos datos como etiquetas de referencia para validar las técnicas de extracción de términos.

En el caso del repositorio del CIDE fueron filtrados los recursos de información que contenían palabras claves del autor la cuales fueron preprocesadas con las mismas reglas definidas en la sección 3.3 y se compararon los términos extraídos por cada técnica y de los cuales se obtuvieron los siguientes gráficos:

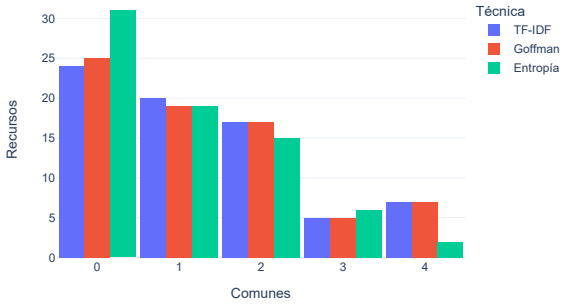

Gráfico 19: Términos comunes extraídos del título y el Subject del Autor CIDE (Gráfico 8)
Fuente: Elaboración propia.

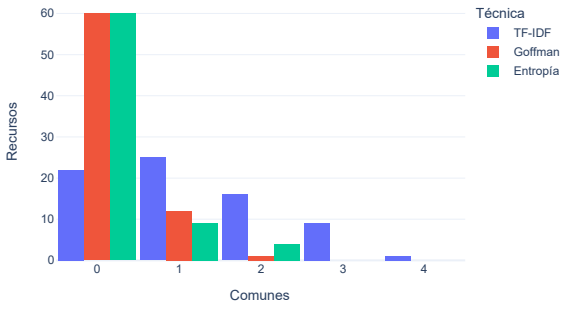

Gráfico 20: Términos comunes extraídos de la descripción y el Subject del Autor CIDE (Gráfico 9)
Fuente: Elaboración propia.

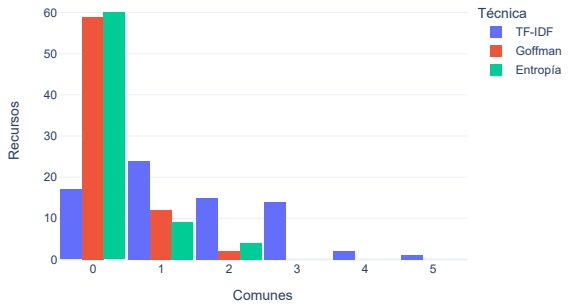

Gráfico 21: Términos comunes extraídos del título+descripción y el Subject del Autor CIDE (Gráfico 10)
Fuente: Elaboración propia.

Si comparamos los gráficos 19 al 21 con los obtenidos del repositorio de referencia (Gráficos 8 al 10) se pueden identificar que se obtienen más palabras comunes entre las extraídas y las asignadas por el autor (CIDE), que entre las extraídas y las provenientes de vocabularios controlados (INFOTEC), además se observa que TF-IDF logra extraer más términos que se encuentran entre los definidos por los autores como palabras claves que las otras dos técnicas.

En cuanto al repositorio del CIBNOR la gran mayoría de los recursos contienen palabras claves asignadas por el autor, así que al comparar los términos extraídos por cada técnica y los encontrados en el metadato Subject asignados por los autores se obtuvieron los siguientes gráficos:

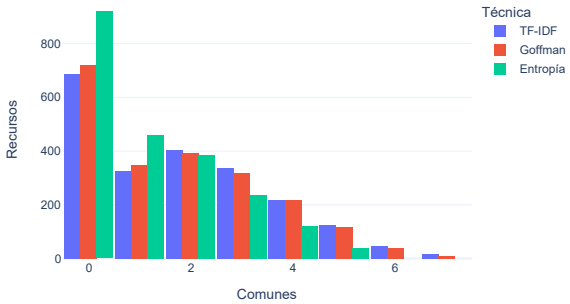

Gráfico 23: Términos comunes extraídos del título y el Subject CIBNOR (Gráfico 8)
Fuente: Elaboración propia.

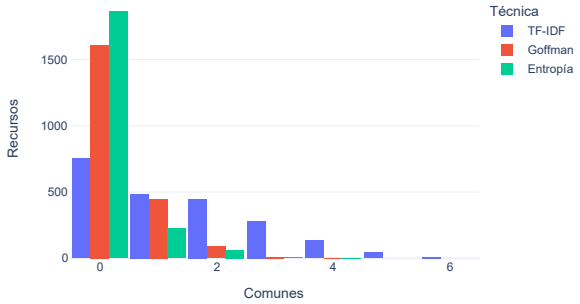

Gráfico 22: Términos comunes extraídos de la descripción y el Subject CIBNOR (Gráfico 9)
Fuente: Elaboración propia.

Donde si comparamos los gráficos 22 al 24 con los obtenidos del CIDE y se observa el

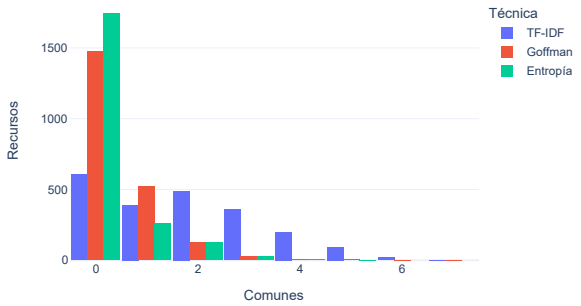

Gráfico 24: Términos comunes extraídos del título+descripción y el Subject CIBNOR (Gráfico 10)
Fuente: Elaboración propia.

mismo comportamiento, donde TF-IDF obtiene mejores resultados.

Como ya mencionamos anteriormente para el LSA (Pincombe, 2004) se esperaría que utilizando la Entropía se obtengan mejores resultados en los documentos relacionados, sin embargo, para nuestra aplicación se puede ver en los gráficos 21 y 24 que la técnica TF-IDF genera mayor número de conciencias entre los términos determinados y las palabras definidas por los autores en el campo Subject.

Con el uso de la información del metadato Subject como fuente de datos etiquetados, podemos pensar que la técnica TF-IDF aplicada a la concatenación título + descripción permite extraer palabras claves de mejor calidad.

4.7 Prueba de concepto

Para la evaluación de los resultados de las diferentes métricas se desarrolló una prueba de concepto en Python con el siguiente modelo.

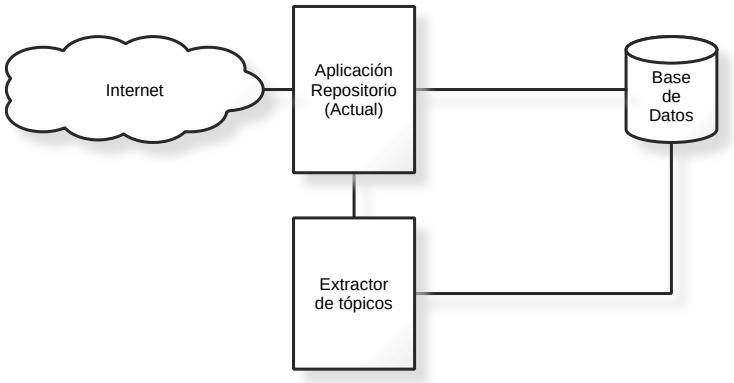

Figura 4: Modelo propuesto
Fuente: Elaboración propia.

En esta prueba de concepto, que implementa el Extractor de tópicos en un Script en Python que a demanda extrae los tópicos y almacenan los resultados en la Base de Datos implementada en MongoDB, la cual es consultada por la Aplicación Repositorio la cual fue implementada en Flask.

4.7.1 Extractor de tópicos

Este es un Script de Python que a demanda a partir del URL del servicio OAI-PMH del repositorio de INFOTEC consulta los datos de los recursos de información alojados en el mismo, para aplicarles los procesos de limpieza y normalización descritos con anterioridad, para a continuación aplicarle las tres técnicas en las condiciones evaluadas, el resultado es almacenado en BD.

4.7.2 Base de datos

Los datos obtenidos por el paso anterior se almacenan en una BD MongoDB en una colección para el repositorio de información, debido a que este manejador está orientado a documentos no fue necesario crear una estructura de datos relacionales.

4.7.3 Aplicación (Repositorio)

La capa de consulta se implementó en Flask con Pymongo para las consultas a BD, para simplificar la implementación de la visualización se utilizó Bootstrap.

La interfaz de consulta consta de un buscador de coincidencias simple de palabras la cual realiza las búsquedas de palabras completas a texto abierto (que está implementada de manera más avanzada en cualquiera de los repositorios de información) y una página de detalle del recurso de información con un listado de recursos relacionados con su título, descripción y palabras términos extraídos adicionalmente existe la posibilidad de realizar esta visualización para cada campo y métrica que sé probó.

Capítulo 5

Conclusiones y Trabajo Futuro

Capítulo 5. Conclusiones y Trabajo Futuro

Los sistemas de aprendizaje automático imponen retos para ser evaluados, ya que es necesario tener datos previamente etiquetados o la validación de un ojo experto para poder evaluar su desempeño, en el caso de esta propuesta realizada a sobre los datos del repositorio de INFOTEC no fue posible tener información previamente etiquetada para cumplir este objetivo para corroborar que nuestras suposiciones, sin embargo, al aplicar las mismas técnicas en otros repositorios similares, y además explotar algunos datos previamente etiquetados que contienen, al contrastar los resultados fue posible tomar por válidas las técnicas seleccionadas.

Para el modelado de tópicos como ya se mencionó anteriormente no se obtuvo para esta aplicación resultados claros para la recomendación de recursos de información si se abre el camino para evaluar otros tipos de relaciones que se encuentran entre los documentos del repositorio que pudieran ser determinadas con el apoyo de un ojo experto.

En cuanto a las técnicas evaluadas los resultados reflejan porque TF-IDF y Entropía son las técnicas más ampliamente usadas en la ciencia de datos en el caso de punto de corte de Goffman se ha reportado como una buena alternativa por el pequeño grupo de investigadores que han evaluado su comportamiento, pero no se pudo ver su efectividad en nuestra aplicación propuesta.

La selección de metadatos propuestos inicialmente cumplieron su objetivo al ser útiles para la extracción de relaciones, pero al revisar más a fondo los datos tanto en el repositorio de referencia como en otros similares es evidente que el esquema de interoperabilidad propuesto por CONACYT es lo bastante robusto para permitir que cada organización que lo implemente sea la qué decida que tan rico de información desea poblar sus repositorios permitiendo a los que explotamos estos datos poder realizar mayor o menor aprovechamiento de su información.

Para la extracción y explotación de los resultados de esta solución de ciencia de datos se propuso un modelo muy ligero que permite extraer de manera independiente al repositorio los recursos relacionados y almacenarlos en BD y desplegarlos con una consulta básica a la misma, la cual fue probada de manera independiente al repositorio, sin embargo, es posible

implementar estas consultas dentro de cada repositorio agregando esta consulta al despliegue de las páginas.

5.1 Conclusiones

A partir de los resultados obtenidos en esta propuesta podemos pensar que posible aplicar técnicas de aprendizaje automático no supervisado para procesar los datos expuestos por el Repositorio Institucional de Ciencia Abierta de INFOTEC y lograr relacionar los recursos de información que este contiene, según lo encontrado en este trabajo la técnica de TF-IDF aplicada a los textos incluidos en los metadatos Title y Description de cada contienen términos que permiten relacionar a los recursos por sus contenidos a manera de recomendación.

Al aplicar las mismas técnicas para contrastar los resultados con otros repositorios como el del CIDE y CIBNOR fue posible observar comportamientos similares e incluso validar suposiciones al tener estos una mayor cantidad de información etiquetada, sin embargo, si se requiere aplicarlo de manera amplia en los demás repositorios como en toda solución de ciencia de datos es necesario realizar un análisis exploratorio previo para robustecer o personalizar principalmente el preprocesamiento de los datos para adaptarlo a las características de cada repositorio.

5.2 Trabajo Futuro

Este trabajo permitió identificar algunas líneas que se podrían seguir para optimizar la aplicación de estas las técnicas, así como mejoras que podrían requerirse para aplicar el sistema de manera productiva, dentro de los que se pueden mencionar:

Procesar los repositorios para obtener además del título y descripción también la totalidad del contenido de cada recurso de información, ya que estos comparten el documento completo en PDF aunque no de manera estructurada u homologada.

Evaluar la posibilidad de determinar ngramas para identificar términos de más de una palabra.

Evaluar los puntos del corte por ejemplo en el caso del TF-IDF en este ejercicio al trabajar con relativamente pocos registros se utilizaron todos los datos, pero en el caso de que se requiera implementarlo de manera más amplia es necesario evaluar puntos de corte para reducir el tamaño de las matrices a procesar.

Referencias

Adomavicius, G., & Tuzhilin, A. (2005). Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering*, *17*(6), 734–749. https://doi.org/10.1109/TKDE.2005.99

Baeza-Yates, R., & Ribeiro-Neto, B. (2011). *Modern information retrieval: The concepts and technology behind search* (Second edition). Addison Wesley.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77–84. https://doi.org/10.1145/2133806.2133826

Bobadilla, J., Ortega, F., Hernando, A., & Gutiérrez, A. (2013). Recommender systems survey. *Knowledge-Based Systems*, *46*, 109–132. https://doi.org/10.1016/j.knosys.2013.03.012

Boyce, B., & Lockard, M. (1975). Automatic and manual indexing performance in a small file of medical literature. *Bulletin of the Medical Library Association*, *63*(4), 378–385.

CONACYT. (s/f). *Repositorio Nacional Catálogos*. Recuperado el 30 de noviembre de 2021, de http://catalogs.repositorionacionalcti.mx/

CONACYT. (2017a). *Lineamientos Generales de Ciencia Abierta*. http://www.siicyt.gob.mx/index.php/normatividad/conacyt-normatividad/programas-vigentes-normatividad/lineamientos/lineamientos-generales-de-ciencia-abierta

CONACYT. (2017b). *Lineamientos Específicos para Repositorios*. http://www.siicyt.gob.mx/index.php/normatividad/conacyt-normatividad/programas-vigentes-normatividad/lineamientos/lineamientos-especificos-para-repositorios

Danilák, M. (2021). *Langdetect* (1.0.8) [Python]. https://github.com/Mimino666/langdetect (Original work published 2014)

Dumais, S. T. (1991). Improving the retrieval of information from external sources. *Behavior Research Methods, Instruments, & Computers*, *23*(2), 229–236. https://doi.org/10.3758/BF03203370

Gefen, D., Endicott, J. E., Fresneda, J. E., Miller, J., & Larsen, K. R. (2017). A Guide to Text Analysis with Latent Semantic Analysis in R with Annotated Code Studying Online Reviews and the Stack Exchange Community. *Communications of the Association for Information Systems*, *41*, 450–496. https://doi.org/10.17705/1CAIS.04121

Guajardo, M. (2020). *Factores determinantes para la implementación del esquema de metadatos para repositorios de datos de investigación de la Política de Ciencia Abierta en México* (pp. 143–160).

Landauer, T. K. (Ed.). (2014). *Handbook of latent semantic analysis* (1. iss. in paperback). Routledge.

Lee, J. (2020). *Benchmarking Language Detection for NLP*. Medium. https://towardsdatascience.com/benchmarking-language-detection-for-nlp-8250ea8b67c

Loesch, M. (2020). *Sickle: OAI-PMH for Humans* (0.7.0) [Python]. https://github.com/mloesch/sickle

Pazzani, M. J., & Billsus, D. (2007). Content-Based Recommendation Systems. En P. Brusilovsky, A. Kobsa, & W. Nejdl (Eds.), *The Adaptive Web* (Vol. 4321, pp. 325–341). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-72079-9_10

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*(85), 2825–2830.

Pincombe, B. (2004). *Comparison of Human and Latent Semantic Analysis (LSA) Judgements of Pairwise Document Similarities for a News Corpus*. 44.

Quesada, J. (2007). *Creating your own LSA space*. 71–85.

Ramos, J. (2003). *Using TF-IDF to Determine Word Relevance in Document Queries*.

Urbizagástegui Alvarado, R., & Restrepo Arango, C. (2011). La ley de Zipf y el punto de transición de Goffman en la indización automática. *Investigación Bibliotecológica. Archivonomía, Bibliotecología e Información*, *25*(54), 71. https://doi.org/10.22201/iibi.0187358xp.2011.54.27482

Anexos

Anexos

ANEXO I: Resultados con el repositorio de INFOTEC en Inglés

En este anexo veremos los resultados de los mismos procesos descritos en la sección 4.3, pero aplicados a los no más de 57 elementos que fueron determinados en idioma inglés (Cuadro 5).

De los conjuntos de términos extraídos igualmente se compararon para identificar que tantos elementos tenían en común para los metadatos procesados.

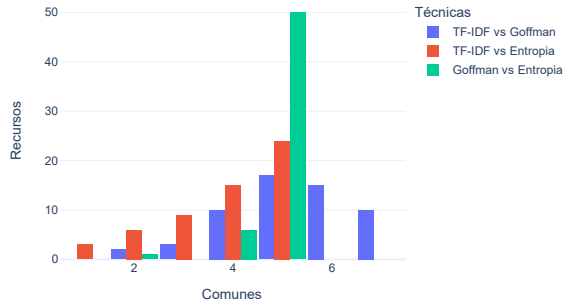

Gráfico 25: Términos comunes del título en Inglés (Gráfico 5)
Fuente: Elaboración propia.

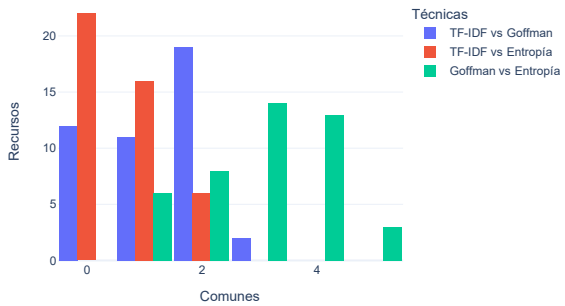

Gráfico 26: Términos comunes de la descripción en Inglés entre técnicas (Gráfico 6)
Fuente: Elaboración propia.

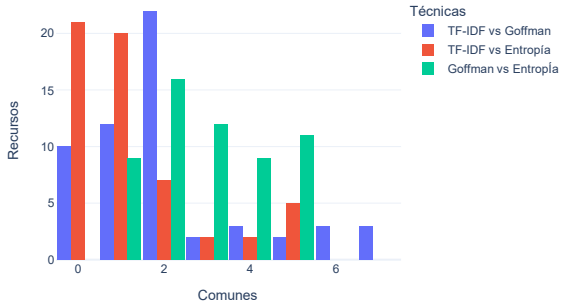

Gráfico 27: Términos comunes del título + descripción en Inglés entre técnicas (Gráfico 7)
Fuente: Elaboración propia.

El resultado para el título (Grafico 25) es muy similar al obtenido en español en cuanto los obtenidos de la descripción y la concatenación de ambos aunque si muestran una tendencia a no tener términos comunes como en español si parecería haber más, esto podría deberse a que el tamaño del corpus es muy pequeño.

Igualmente, se identificaron los recursos relacionados a través de los términos obtenidos por cada técnica para los metadatos analizados.

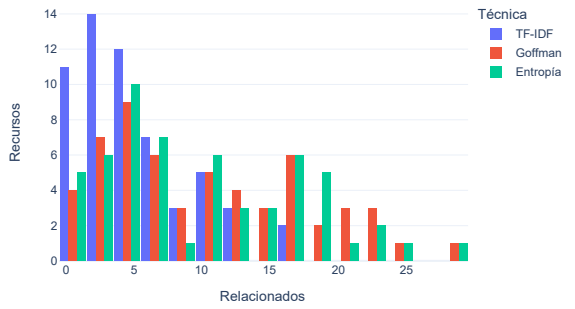

Gráfico 28: Recursos relacionados por título en Inglés por técnica (Gráfico 11)
Fuente: Elaboración propia.

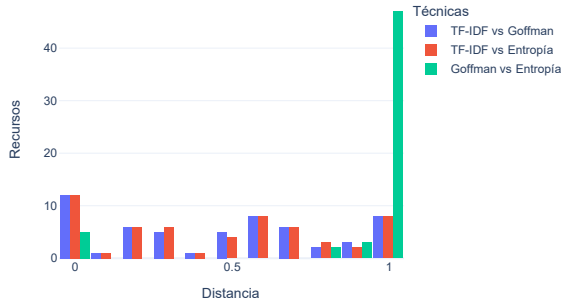

Gráfico 29: Distancia Jaccard de recursos relacionados por el título en Inglés entre técnicas (Gráfico 12)
Fuente: Elaboración propia.

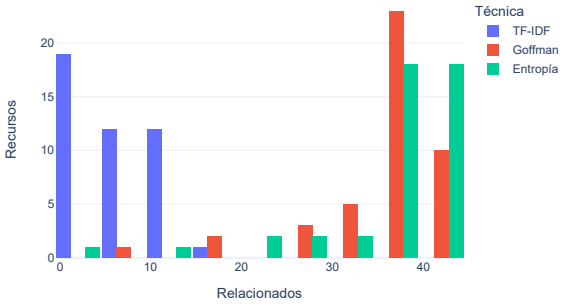

Gráfico 30: Recursos relacionados por descripción en Inglés por técnicas (Gráfico 13)
Fuente: Elaboración propia.

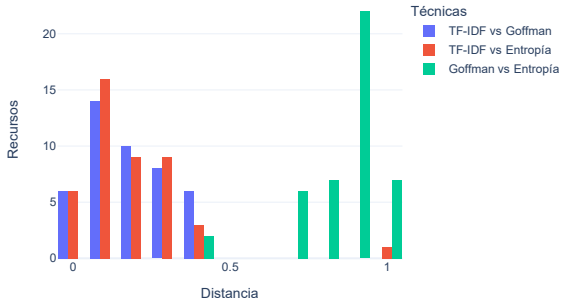

Gráfico 31: Distancia Jaccard de recursos relacionados por la descripción en Inglés entre técnicas (Gráfico 14)
Fuente: Elaboración propia.

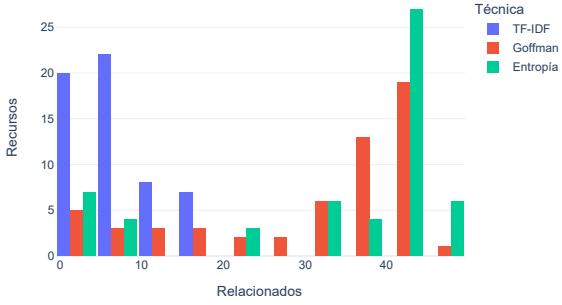

Gráfico 32: Recursos relacionados por título + descripción en Inglés entre técnicas (Gráfico 15)
Fuente: Elaboración propia.

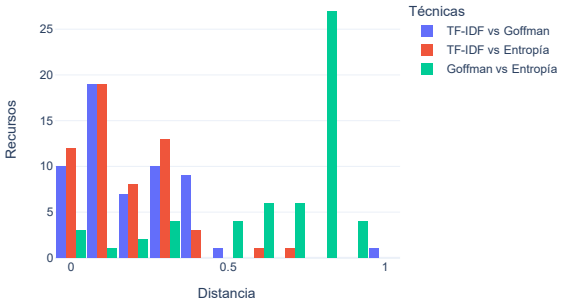

Gráfico 33: Distancia Jaccard de recursos relacionados por el título + descripción entre técnicas (Gráfico 16)
Fuente: Elaboración propia.

Para el título, descripción y la concatenación de ambos se obtuvieron gráficas (Grafico 28 - 33) consistentes con lo observado al procesar sus contrapartes en español.

Finalmente, también se realizó la agrupación de recursos y se obtuvieron los 5 vectores más representativos con los que se construyeron las siguientes nubes de palabras.

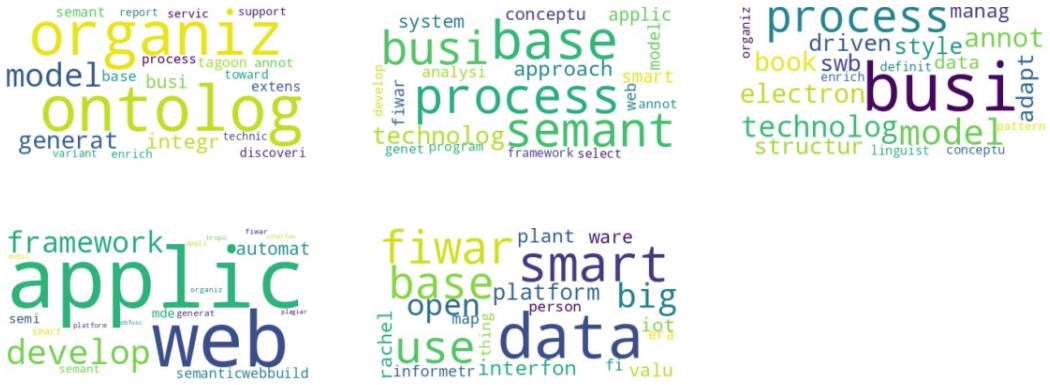

Gráfico 34: Nubes de palabras en Inglés de los principales vectores encontrados (Gráfico 17)
Fuente: Elaboración propia.

Al igual que su contraparte en español también se aprecian términos repetidos entre cada nube, y los términos extraídos corresponden a las áreas TIC del centro, con estos vectores también se obtuvieron los documentos de mayor pero para cada uno.

1	2	3	4	5
Generating Ontologies through Organizational Modeling	An i* based approach for conceptual modeling of business process technology	SWB Process: A Business Process Management System driven by Semantic Technologies	A MDE Framework for semi-automatic development of Web applications	Using Transparency Models to evaluate Open Data Systems: a case study using a Mexican Open Data Platform
Generating Ontologies through Organizational Modeling	SWB Process: A Business Process Management System driven by Semantic Technologies	Definition of Linguistic Syntactic Patterns for Semantic Annotation of Business Process Models (Technical report)	SemanticWebBuilder: A Framework for Semantic Web Applications Development	The value of our personal data in the Big Data and the Internet of all Things Era
TAGOOn+: Generation and Integration of Organizational Ontologies	Definition of Linguistic Syntactic Patterns for Semantic Annotation of Business Process Models (Technical report)	A business model for electronic books	Applying Tropos modeling for Smart mobility applications based on the FIWARE platform	Informetric Mapping of “Big Data” in FI-WARE
Towards supporting business services discovery through the integration of organizational models with ontologies	RESyS: Towards a rule-based recommender system based on semantic reasoning	Adaptation of business models to the Structure-in-5 organizational style	Automatic Generation of Summary Obfuscation Corpus for Plagiarism Detection	Rachel: An IoT smart plant based on FIWARE
Extension and integration of i* models with ontologies	Applying Tropos modeling for Smart mobility applications based on the FIWARE platform	An i* based approach for conceptual modeling of business process technology	Web Service SWePT: A Hybrid Opinion Mining Approach	Applying Tropos modeling for Smart mobility applications based on the FIWARE platform
Enriching Organizational Models through Semantic Annotation	SemanticWebBuilder: A Framework for Semantic Web Applications Development	Enriching Organizational Models through Semantic Annotation	Enriching Organizational Models through Semantic Annotation	Strategy for the automated diagnostic of the openess degree in government data

Cuadro 24: Ejemplo de documentos pertenecientes a cada uno de los 5 grupos (Cuadro24)
Fuente: Elaboración propia.

Se observan varios documentos repetidos entre grupos esto también está influenciado por la poca cantidad de documentos en este corpus.

ANEXO II: Comparación de resultados con el repositorio CIDE

Con el fin de comparar los resultados obtenidos de las técnicas propuestas se realizó el procesamiento de los datos expuestos por el Centro de Investigación y Docencia Económicas, A.C. (CIDE), para lo cual se realizaron los mismos procesos descritos en la sección 4.3, para lo cual se cosechó del repositorio del centro que expone sus datos en la URL: https://cide.repositorioinstitucional.mx/oai/request, igualmente se limitó la consulta para obtener los recursos activos y se obtuvieron 718 registros (al 30 de noviembre de 2021).

Se llevó a cabo una revisión exploratoria para identificar la consistencia y similitud de los datos con respecto a lo ya encontrado en el repositorio de referencia INFOTEC.

	Nulos
title	0
creator	0
contributor	656
publisher	435
date	0
type	0
description	12
audience	0
subject	0
identifier	0
relation	566
rights	0
language	114
format	1
source	697
coverage	---

Cuadro 25: Metadatos nulos CIDE (Cuadro 1)
Fuente: Elaboración propia.

Se encontró que aunque existe elementos nulos son un número bajo en las series de metadatos de interés, además se analizó la serie del metadato Subject para identificar el origen de las clasificaciones asignadas por este centro a los recursos de información.

También se observó que en este caso el origen es un vocabulario controlado (LCSH) y en muy pocos casos (AUTHOR y Author) se puede asumir que se refieren a palabras claves determinadas por el autor.

En la identificación del idioma se encontró que igualmente contenía información en español, inglés y en ambos idiomas.

Idioma	title	description
Español	435	183
Ingles	242	11
Ambos	41	512

Cuadro 26: Determinación de Idioma CIDE (Cuadro 6)
Fuente: Elaboración propia.

A continuación se realizó la extracción de palabras claves para los tres metadatos: Title, Description y ambos concatenados en idioma español y se compararon sus salidas.

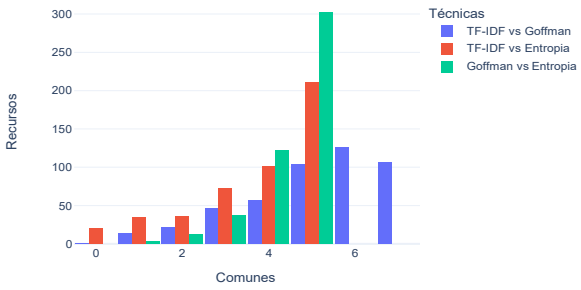

Gráfico 35: Términos comunes del título entre técnicas CIDE (Gráfico 5)
Fuente: Elaboración propia.

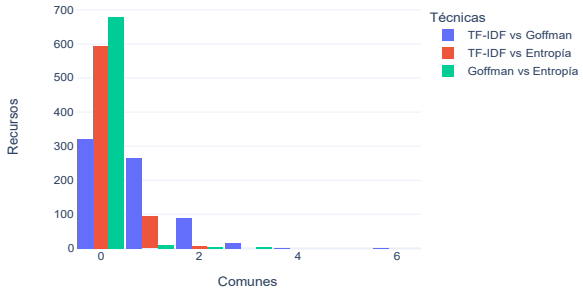

Gráfico 36: Términos comunes de la descripción entre técnicas CIDE (Gráfico 6)
Fuente: Elaboración propia.

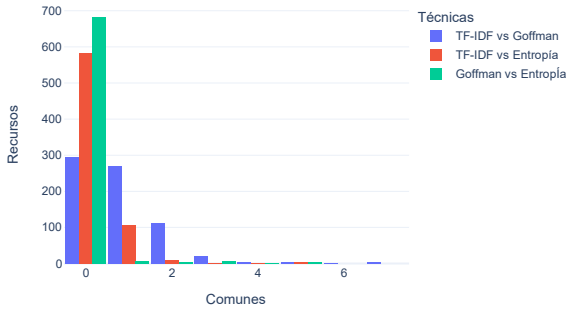

Gráfico 37: Términos comunes del título + descripción entre técnicas CIDE (Gráfico 7)
Fuente: Elaboración propia.

Se puede observar un comportamiento muy similar al encontrado en el repositorio de referencia, donde para el metadato Title, Goffman y Entropía determinan los mismos términos y para Description y la concatenación de ambos generan términos diferentes en su mayoría.

En este repositorio también se exploró utilizar el metadato Subject como etiquetas de referencia a las palabras claves extradidas, ya que como se puede ver en el Cuadro 26 además de los vocabularios controlados CTI y LCSH se usa en algunos casos palabras asignadas por el autor, al igual que en repositorio de referencia fueron extraídos los textos y procesados con las mismas reglas y se buscaron términos comunes.

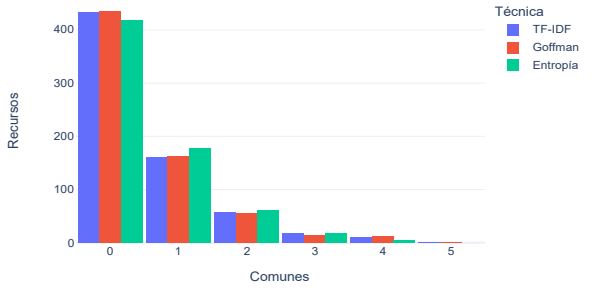

Gráfico 38: Términos comunes extraídos del título y el Subject CIDE (Gráfico 8)
Fuente: Elaboración propia.

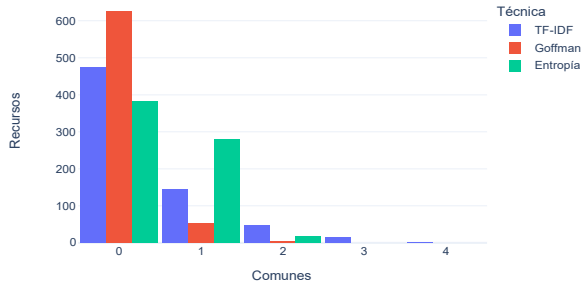

Gráfico 39: Términos comunes extraídos de la descripción y el Subject CIDE (Gráfico 9)
Fuente: Elaboración propia.

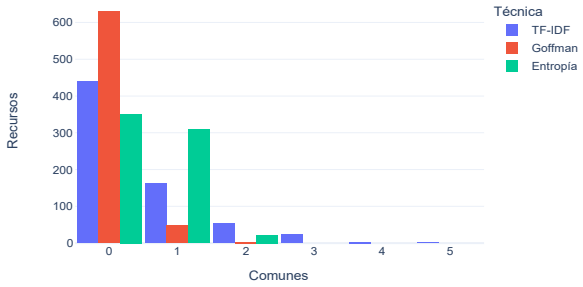

Gráfico 40: Términos comunes extraídos del título+descripción y el Subject CIDE (Gráfico 10)
Fuente: Elaboración propia.

Se puede ver en los gráficos 38 a 40 comportamientos similares a los del repositorio de referencia donde la mayoría de los recursos no tienen términos comunes entre los extraídos y los asignados en el campo Subject. Se decidió acotar estas gráficas solo los recursos etiquetados con palabras claves procedentes del autor con resultados positivos que se explican en la sección 4.6.

Igualmente con los términos extraídos se obtuvieron los recursos relacionados por cada una de las técnicas, y entre ellas, para los conjuntos de recursos obtenidos del título se obtuvieron las siguientes gráficas.

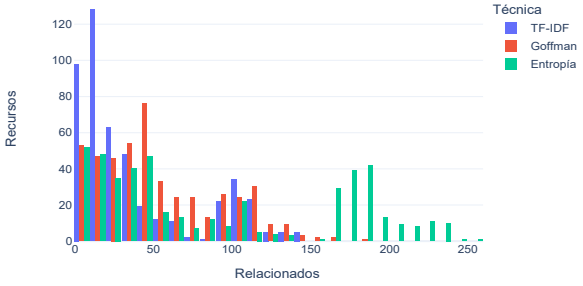

Gráfico 41: Recursos relacionados por título por técnica CIDE (Gráfico 11)
Fuente: Elaboración propia.

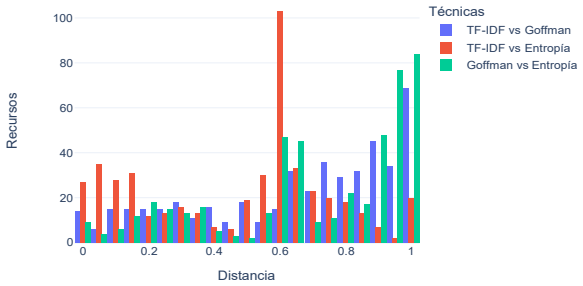

Gráfico 42: Distancia Jaccard de recursos relacionados por el título entre técnicas CIDE (Gráfico 12)
Fuente: Elaboración propia.

Si se compara con los gráficos correspondientes del repositorio de referencia para el elemento título se observa que igualmente TF-IDF relaciono como máximo un número proporcional a la cantidad de recursos en el repositorio, pero en su mayoría se mantuvieron por debajo de los 30 recursos, en cuanto a la comparativa de recursos relacionados por cada técnica las distancias de Jaccard de estos conjuntos es alta para en parte por el tamaño de los conjuntos y también se pueden entender que cada técnica relaciono diferentes recursos de información, aunque de manera global la mitad de los recursos generando conjuntos similares con las tres técnicas.

Igualmente, se generaron los metadatos de descripción y para la concatenación de este con el título.

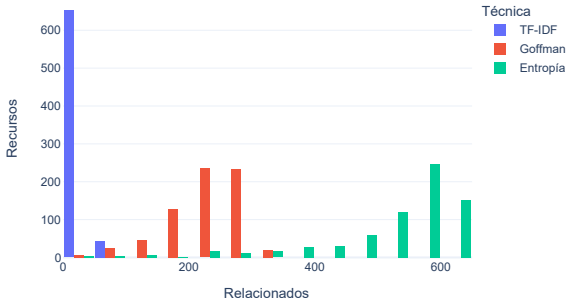

Gráfico 43: Recursos relacionados por descripción por técnicas CIDE (Gráfico13)
Fuente: Elaboración propia.

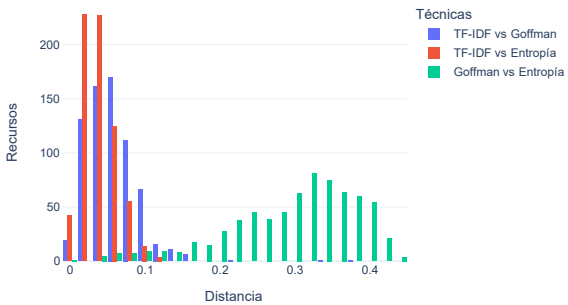

Gráfico 44: Distancia Jaccard de recursos relacionados por la descripción entre técnicas CIDE (Gráfico 14)
Fuente: Elaboración propia.

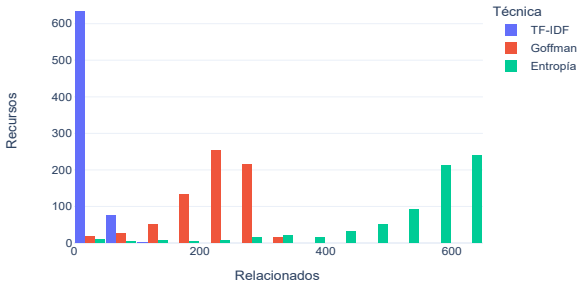

Gráfico 45: Recursos relacionados por título + descripción entre técnicas CIDE (Gráfico 15)
Fuente: Elaboración propia.

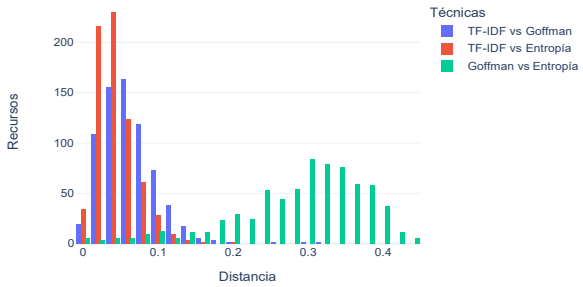

Gráfico 46: Distancia Jaccard de recursos relacionados por el título + descripción entre técnicas CIDE (Gráfico 16)
Fuente: Elaboración propia.

En este caso los resultados de las gráficas 43 a 46 si resultaron muy similares a sus contrapartes en el repositorio de referencia.

Finalmente, se corrieron los procesos de agrupación de recursos y se construyo una nube palabras con los de mayor peso para cada uno de los 5 vectores principales.

Gráfico 47: Nubes de palabras de los principales vectores encontrados CIDE (Gráfico 17)
Fuente: Elaboración propia.

Se observa que también se repiten términos entre las últimas dos nubes generadas. Esto se ve reflejado en los archivos representativos de estos vectores como se pueden ver los primeros grupos están claramente diferenciados por sus nombres, pero los últimos dos tienen documentos comunes, y que los documentos pertenecen claramente a grupos de temática similares entre ellos.

1	2	3	4	5
Violencia de género contra las mujeres: modelos de atención, Procuraduría General de Justicia. Guanajuato	Política exterior y opinión pública: México ante el mundo	Calidad de gobierno y rendición de cuentas en las entidades federativas: Oaxaca	Las reformas electorales de 2007	Las reformas electorales de 2007
Violencia de género contra las mujeres: modelos de atención, Procuraduría General de Justicia. Distrito Federal	La incidencia de la opinión pública en la política exterior de México teoría y realidad	Calidad de gobierno y rendición de cuentas en las entidades federativas: Jalisco	Propuesta para una reforma electoral en México	La otra reforma
Violencia de género contra las mujeres: modelos de atención, Procuraduría General de Justicia. Baja California	La incidencia de la opinión pública en la política exterior de México: teoría y realidad	Calidad de gobierno y rendición de cuentas en las entidades federativas: Zacatecas	La otra reforma	La reforma del sector eléctrico mexicano: recomendaciones de política pública
Violencia de género contra las mujeres: modelos de atención, Procuraduría General de Justicia. Tlaxcala	La cultura política de los políticos en el México democrático	La calidad del gobierno y la rendición de cuentas en los estados: una agenda de investigación	Reforma electoral y elecciones presidenciales en Estados Unidos	La reforma electoral pendiente
Violencia de género contra las mujeres: modelos de atención, Procuraduría General de Justicia. Morelos	La estructura de la rendición de cuentas en México: informe sobre la calidad de la información en las cuentas públicas en México	La rendición de cuentas gubernamental: una propuesta para el análisis empírico en las entidades federativas	La reforma electoral pendiente	El Estado mexicano después de su reforma
Violencia de género contra las mujeres: modelos de atención, Procuraduría General de Justicia. Querétaro	El enfoque económico en el estudio de las políticas públicas	La rendición de cuentas del gobierno municipal en México	Una lectura crítica de la reforma electoral en México a raíz de la elección de 2006; A critical reading of the electoral reform in Mexico as a result of the 2006 election	Los años de Salinas: crisis electoral y reformas

Cuadro 27: Ejemplo de documentos pertenecientes a cada uno de los 5 grupos CIDE (Cuadro 24)
Fuente: Elaboración propia.

En cuanto al comparativo para ver si estas agrupaciones están relacionadas con el área del conocimiento no se realizó, ya que todos los recursos de este repositorio pertenecen solo al área de Ciencias Sociales.

ANEXO III: Comparación de resultados con el repositorio CIBNOR

Con el fin de comparar los resultados obtenidos de las técnicas propuestas se realizó el procesamiento de los datos expuestos por el Centro de Investigaciones Biológicas del Noroeste S.C. (CIBNOR), para lo cual se realizaron los mismos procesos descritos en la sección 4.3, para lo cual se cosechó del repositorio del centro que expone sus datos en la URL: https://cibnor.repositorioinstitucional.mx/oai/request, igualmente se limitó la consulta para obtener los recursos activos y se obtuvieron 2264 registros (al 30 de noviembre de 2021).

Se llevó a cabo una revisión exploratoria para identificar la consistencia y similitud de los datos con respecto a lo ya encontrado en el repositorio de referencia INFOTEC.

	Nulos
title	0
creator	0
contributor	151
publisher	36
date	0
type	0
description	11
audience	121
subject	0
identifier	0
relation	105
rights	0
language	11
format	1
source	325
coverage	356

Cuadro 28: Metadatos nulos CIBNOR (Cuadro 1)
Fuente: Elaboración propia.

En este repositorio tampoco se encontraron elementos nulos en las series de metadatos e interés, así que sé continuó el análisis de la serie del metadato Subject con el fin de identificar el origen de las clasificaciones asignadas a los recursos de información.

	Recursos
cti	2264
AUTOR	2105
Autor	52
DIRECTOR	1
EDITORES	1

Cuadro 29: Orígenes de clasificación CIBNOR (Cuadro 5)
Fuente: Elaboración propia.

Se puede ver que para este repositorio solo se asignaron elementos del vocabulario controlado CTI, pero además se agregaron las palabras claves asignadas por el autor en la mayoría de los recursos de información.

Idioma	title	description
Español	1242	184
Ingles	482	429
Ambos	540	1559

Cuadro 30: Cuadro 27: Determinación de Idioma CIBNOR (Cuadro 6)
Fuente: Elaboración propia.

A continuación se realizó la extracción de palabras claves para los tres metadatos: Title, Description y ambos concatenados en idioma español y se compararon sus salidas.

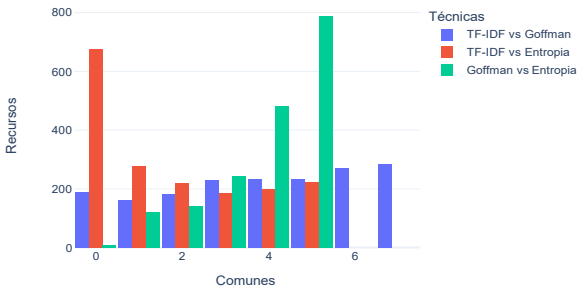

Gráfico 48: Términos comunes del título entre técnicas CIBNOR (Gráfico 5)
Fuente: Elaboración propia.

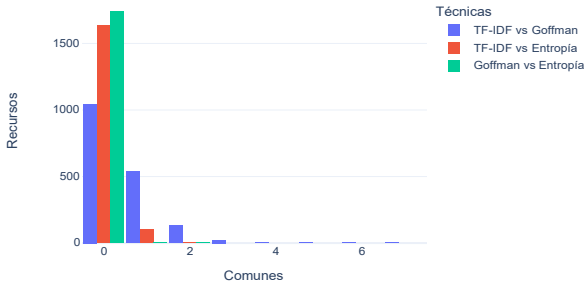

Gráfico 49: Términos comunes de la descripción entre técnicas CIBNOR (Gráfico 6)
Fuente: Elaboración propia.

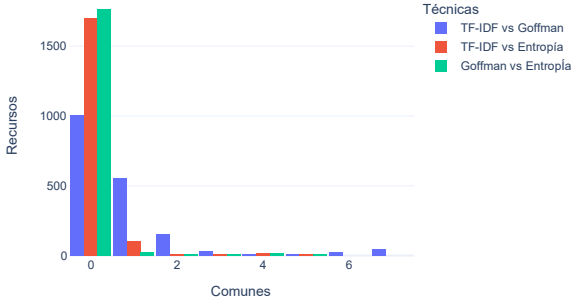

Gráfico 50: Términos comunes del título + descripción entre técnicas CIBNOR (Gráfico 7)
Fuente: Elaboración propia.

En este caso el gráfico 48 su es diferente al del obtenido del repositorio de referencia donde la mayoría de las técnicas obtenían al menos un término comunes entre ellas en este caso TF-IDF y Entropía en una tercera parte de los recursos, sin embargo, Entropía y Goffman si generaron términos comunes, en cuanto a los términos generados para la descripción y para la concatenación de ambos las tres técnicas generaron términos diferentes como en el repositorio de referencia.

Para este repositorio igualmente se exploró la posibilidad de utilizar el metadato Subject como etiquetas de referencia, ya que como se puede ver en el cuadro 29 la mayoría de los recursos están etiquetados con palabras asignadas por el autor, estos textos fueron extraídos y procesados de la misma manera que las del repositorio de referencia y los resultados son los que se muestran en la sección 4.6.

Igualmente, los términos extraídos permitieron obtener los recursos relacionados con estos, para cada una de las técnicas, de estos conjuntos de recursos se obtuvieron los siguientes resultados para el metadato título.

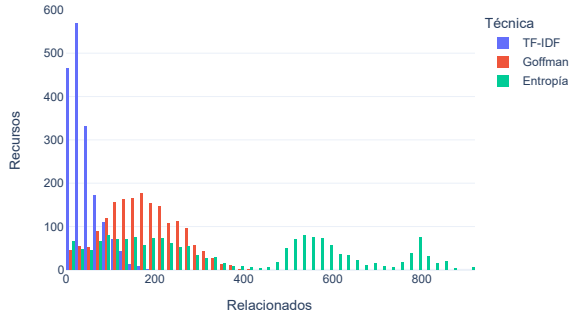

Gráfico 51: Recursos relacionados por título por técnica CIBNOR (Gráfico 11)
Fuente: Elaboración propia.

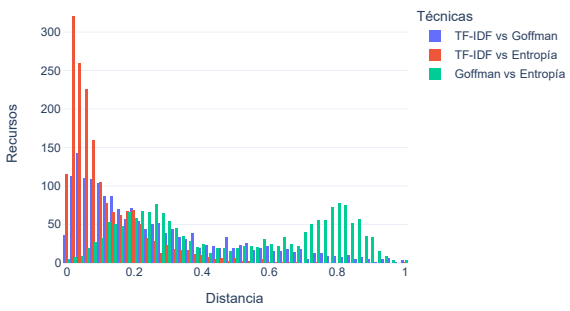

Gráfico 52: Distancia Jaccard de recursos relacionados por el título entre técnicas CIBNOR (Gráfico 12)
Fuente: Elaboración propia.

Sí observamos el número de recursos relacionados por medio de los términos extraídos por técnica TF-IDF en su mayoría son menos de 50 comparado con Goffman que está por los 200 y Entropía que estará por los 500 comparando con los datos obtenidos del repositorio de referencia TF-IDF tiene un comportamiento muy similar. En cuanto a la distancia de Jaccard de estos conjuntos era de esperarse que por los tamaños de conjuntos que cada técnica obtuvo no se observara mucha similitud, sin embargo, si se observa que entre todos Jaccard y Entropía generan en algunos casos conjuntos similares.

En cuanto a los documentos relacionados por medio de los términos extraídos de los metadatos de descripción y su concatenación con el título.

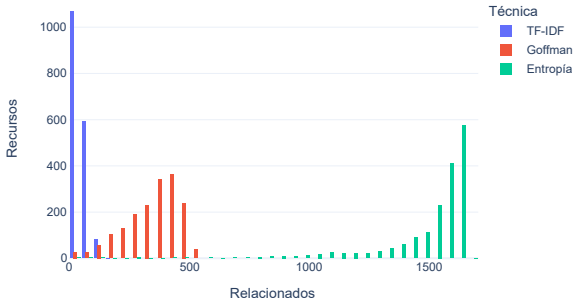

Gráfico 53: Recursos relacionados por descripción por técnicas CIBNOR (Gráfico 13)
Fuente: Elaboración propia.

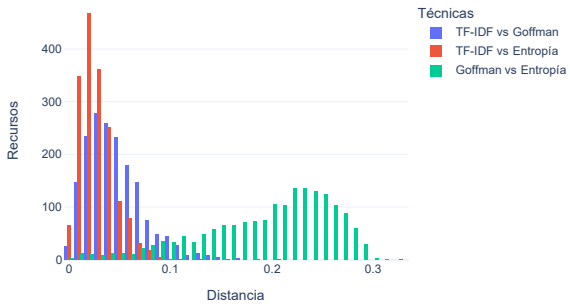

Gráfico 54: Distancia Jaccard de recursos relacionados por la descripción entre técnicas CIBNOR (Gráfico 14)
Fuente: Elaboración propia.

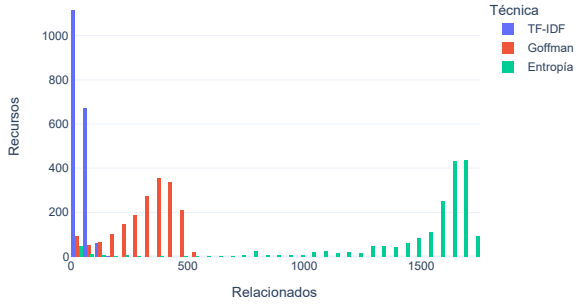

Gráfico 55: Recursos relacionados por título + descripción entre técnicas CIBNOR (Gráfico 15)
Fuente: Elaboración propia.

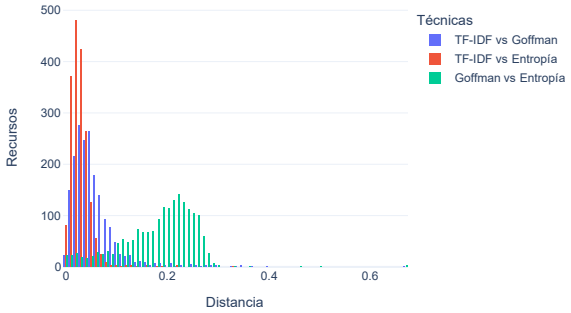

Gráfico 56: Distancia Jaccard de recursos relacionados por el título + descripción entre técnicas CIBNOR (Gráfico 16)
Fuente: Elaboración propia.

Las gráficas 53 a 56 son muy consistente a las obtenidas tanto en el repositorio de referencia como en el de CIDE, donde consistentemente TF-IDF relaciona un número menor de recursos, Goffman de 2 a 3 veces más y que Entropía relaciona la mayor cantidad de recursos.

Por último se ejecutaron los procesos de agrupación de recursos y se construyo una nube de palabras con los términos de mayor peso para los 5 vectores principales.

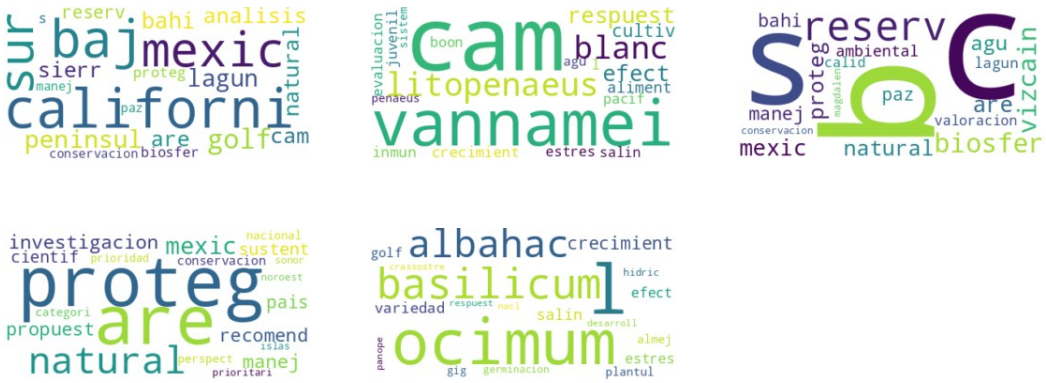

Gráfico 57: Nubes de palabras de los principales vectores encontrados CIBNOR (Gráfico 10)
Fuente: Elaboración propia.

Se puede ver que se generaron nubes aparentemente independientes en el caso de la tercera nube se pueden observar las letras B, C y S como términos independientes que en realidad forman parte del acrónimo del estado de Baja California Sur cede del centro.

1	2	3	4	5
LA SIERRA DE LA LAGUNA DE BAJA CALIFORNIA SUR	Efecto del silicio orgánico sobre la respuesta inmune del camarón blanco Litopenaeus vannamei	ESTRATEGIA PARA EL MANEJO DE LA RESERVA DE LA BIOSFERA EL VIZCAINO, B. C. S., MEXICO	Las áreas naturales protegidas de México	CRECIMIENTO Y DESARROLLO DE VARIEDADES DE ALBAHACA (Ocimum basilicum L.) EN CONDICIONES DE SALINIDAD
Diversidad genética en Sierra de La Laguna, Baja California Sur, México	Efecto de Debaryomyces hansenii en la respuesta antioxidante de juveniles de camarón blanco Litopenaeus vannamei	Hidrología de la Reserva de la Biosfera del Vizcaíno, B. C. S.	Recomendaciones para el manejo sustentable en las áreas naturales protegidas de México; Recommendations for the sustainable management of natural protected areas in Mexic	Emergencia y crecimiento de plántulas de variedades de albahaca (Ocimum basilicum L.) en condiciones salinas
Diversidad y conservación de los peces de la bahía de La Paz, Baja California Sur, México	Genética poblacional de camarón blanco Litopenaeus vannamei en Sinaloa, México	VALORACIÓN HIDROSOCIAL EN LA RESERVA DE LA BIÓSFERA DEL VIZCAÍNO, B.C.S., MÉXICO	LAS ÁREAS NATURALES PROTEGIDAS Y LA INVESTIGACIÓN CIENTÍFICA EN MÉXICO	RESPUESTA DE VARIEDADES DE ALBAHACA (Ocimum basilicum L.) A LA SALINIDAD EN LAS ETAPAS INICIALES DE CRECIMIENTO
LOS MAMÍFEROS DEL ESTADO DE BAJA CALIFORNIA SUR	Análisis fisiológico y genético del desempeño reproductivo del camarón blanco litopenaeus vannamei	Organismo Regulador del Recurso Agua en la Reserva de la Biosfera del Vizcaíno, B.C.S. México.	LA IMPORTANCIA DE LAS ÁREAS NATURALES PROTEGIDAS EN NUESTRO PAÍS	EMERGENCIA Y CRECIMIENTO DE PLÁNTULAS DE VARIEDADES DE ALBAHACA (Ocimum basilicum L.) SOMETIDAS A ESTRÉS HÍDRICO
ÍNDICES DE CAMBIO CLIMÁTICO EN LA RESERVA DE LA BIOSFERA EL VIZCAÍNO, BAJA CALIFORNIA SUR, MÉXICO (1960-2012)	Evaluación del potencial reproductivo de camarón blanco Litopenaeus vannamei en condiciones de domesticación	Calidad de Agua en la Reserva de la Calidad de Agua en la Reserva de la Biosfera del Vizcaíno, B. C. S.	GESTIÓN, MANEJO Y CONSERVACIÓN EN ÁREAS NATURALES PROTEGIDAS	TOLERANCIA A LA SALINIDAD EN VARIEDADES DE ALBAHACA (Ocimum basilicum L.) EN LAS ETAPAS DE GERMINACIÓN, EMERGENCIA Y CRECIMIENTO INICIAL
INDICADORES DE SUSTENTABILIDAD Y PESCA: CASOS EN BAJA CALIFORNIA SUR, MEXICO	El efecto del bacterol – SHRIMP sobre respuesta productiva en juveniles de camarón blanco, Litopenaeus vannamei	Estructura Tarifaria del Recurso Agua en la Reserva de la Biosfera del Vizcaíno, B.C.S. México	Prioridades de Investigación para las Áreas Naturales Protegidas de México	Efecto de un bioestimulante natural como atenuante del estrés salino en variedades de albahaca (Ocimum basilicum L.)

Cuadro 31: Ejemplo de documentos pertenecientes a cada uno de los 5 grupos CIBNOR (Cuadro 24)
Fuente: Elaboración propia.

En el ejemplo de los documentos pertenecientes a cada una de las 5 principales agrupaciones se observa claramente que están relacionadas las temáticas dentro de ellas.

Finalmente, se visualizaron estos vectores extraídos de manera global para identificar si estos correspondían a las áreas del conocimiento.

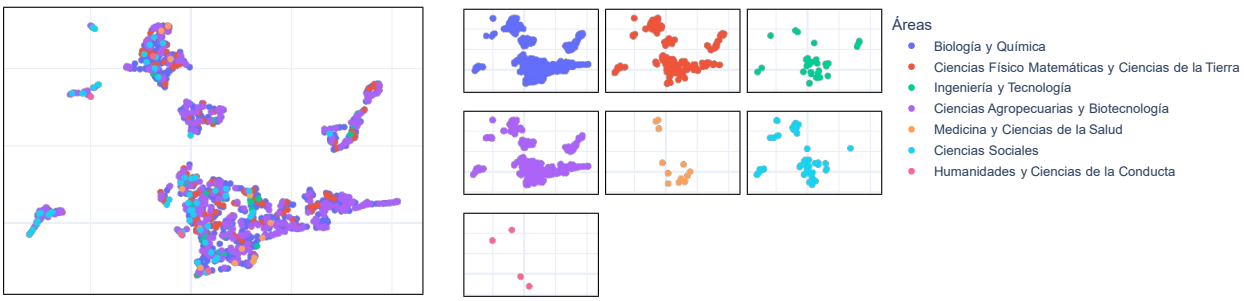

Gráfico 58: Vectores extraídos por LSA agrupados por Área del conocimiento (Gráfico 18)
Fuente: Elaboración propia.

Se observa que si se forman agrupaciones, pero no están relacionadas con las áreas del conocimiento, ya que se puede ver que se encuentran 3 o más áreas en cada agrupación identificable.