Vol. 1, Año 5, No. 11, 2025
ISSN 2954-4211

Revista Internacional Socio-Innova-Tec del Altiplano

Ontology proposal to categorize economic activities in cities of Mexico
Propuesta de una ontología para categorizar actividades económicas en municipios de México

Mariana TORRES-HERRERA[1*], Elías RUIZ-HERNÁNDEZ[1], German CUAYA-SIMBRO[1]

[1]*TecNM- ITS del Oriente del Estado de Hidalgo. Carretera Apan-Tepeapulco Km 3.5, Colonia Las Peñitas, Apan Hidalgo, México.*
*(*0009-0009-5442-4945, 0000-0002-6659-3780, 0000-0001-6303-154X*)*

Sent date: 08/October/2024 Acceptance date: 13/June/2025

Abstract:

In this work, a data-based methodology is proposed to obtain an ontology that describes the predominant economic activities of the municipalities of Mexico. The learned ontology is based on Word embedding techniques and unsupervised learning with data from the National Statistical Directory of Economic Units (DENUE). The proposed model differs from some other ontology proposals that are based on subjective estimates of experts in the smart city domain. In contrast, this proposal determines economic activities that already exist in the municipalities of Mexico and proposes an ontology that describes the municipalities in terms of their development in concepts or dimensions. The main concepts found in this proposal are industry, economy, health, food, communications, culture, environment, mobility, among others. Thus, the ontology measures levels of development (similar or different) in each of these concepts for the municipalities of Mexico. In municipalities with similar levels of development, joint development opportunities can be discovered: what economic units a municipality is missing that another already has. This methodology can be extrapolated to other municipalities or cities in other countries if similar information on economic units is available.

***Keywords*:** Ontology, cities, mining, clustering, development.

Resumen:

En este trabajo se propone una metodología basada en datos para obtener una ontología que describa las actividades económicas preponderantes de los municipios de México. La ontología aprendida está basada en técnicas de *Word embedding* y aprendizaje no supervisado con datos del Directorio Estadístico Nacional de Unidades Económicas (DENUE). El modelo propuesto se diferencia de algunas otras propuestas de ontologías que están basadas en estimaciones subjetivas de expertos en el dominio de ciudades inteligentes. Por contraste, en esta propuesta se determinan actividades económicas que ya existen en los municipios de México y se propone una ontología que describe a los municipios en términos de su desarrollo en conceptos o dimensiones. Los conceptos principales hallados en esta propuesta son: industria, economía, salud, alimentación, comunicaciones, cultura, ambiente, movilidad, entre otros. Así, la ontología mide niveles de desarrollo (similar o diferente) en

cada uno de estos conceptos para los municipios de México. En municipios con niveles de desarrollo similar se pueden descubrir oportunidades conjuntas de desarrollo: qué unidades económicas le faltan a un municipio que otro ya tiene. La presente metodología es extrapolable a otros municipios o ciudades de otros países si se cuenta con información similar de unidades económicas.

***Palabras clave*:** Ontología, ciudades, minería, agrupamiento, desarrollo.

* *Corresponding author. E-mail: 22030625m@itesa.edu.mx*
Tel. +52 771-228-66-65

1. Introducción

Actualmente, el tema de las métricas para determinar grados desarrollo entre ciudades, así como descubrir los desafíos que muchas de ellas enfrentan en el mundo, está adquiriendo mayor relevancia en Asia, Europa y Estados Unidos debido al aumento demográfico y la posible manifestación de dilemas en el desarrollo urbano, como la contaminación, la sobrepoblación, la indigencia y el crecimiento económico desigual. En este contexto, resulta de creciente interés examinar la información presente en los censos y bases de datos similares para simplificar la comprensión de esta situación y mejorar la toma de decisiones de entidades gubernamentales, sociales y privadas. Estas decisiones pueden conducir a una mejora en la calidad de los servicios públicos, un desarrollo social más amplio y equilibrado, e incluso aumentar el desempeño económico del sector privado en varias ciudades (Chourabi *et al*., 2012; De Nicola & Villani, 2021). En un horizonte más amplio, estas decisiones pueden llevar al incremento de ciudades que se cataloguen como inteligentes al reducir las brechas y diferencias entre ellas (Ageed *et al*., 2021).

Por otro lado, para entender los datos que se encuentran en censos estadísticos, se necesitan herramientas de Minería de Textos y otras que ayuden a comprender de manera automática y por tanto, rápida, la información que una determinada ciudad puede tener (Karani, 2018). En este tenor, los modelos semánticos, en su mayoría, describen relaciones entre palabras basadas en jerarquías o redes, lo que les permite ser útiles para la descripción de conceptos (Djenouri et al., 2022). Sin embargo, su implementación puede resultar costosa, degradando la eficiencia al desarrollar un modelo de este tipo. Estos modelos requieren entonces de un marco adecuado para su implementación en contextos con entornos dinámicos, datos en crecimiento y/o entornos con recursos computacionales limitados (Malkawi et al., 2022).

En este artículo se propone una ontología reducida generada a partir de la clasificación de conceptos textuales obtenidos de datos urbanos de diversas ciudades. Esta ontología busca agrupar estos conceptos en función de sus similitudes basándose en un esquema de incrustación léxica, también conocido como *word embedding* (Karani, 2018). La ontología propuesta surge de aplicar algoritmos de aprendizaje no supervisado para que parte de la estructura emerja de una manera más cercana a la realidad de los sectores económicos y sociales provenientes de las ciudades. Los datos recopilados provienen de ciudades de México y se compara la estructura de la ontología con otras ontologías desarrolladas subjetivamente (De Nicola & Villani, 2021).

El presente artículo presenta un enfoque innovador para el modelado ontológico de actividades comerciales en municipios mexicanos utilizando técnicas de aprendizaje automático y minería de textos. Se destaca la importancia de este enfoque debido a los desafíos en manejar grandes volúmenes de información demográfica. Lo anterior podría ser intratable de procesar por una persona (Ageed *et al.,* 2021). La ontología propuesta se basa en técnicas de aprendizaje no supervisado y *word embedding* para agrupar conceptos similares. El objetivo es facilitar la comprensión de la situación urbana y mejorar la toma de decisiones. Se compara esta ontología con otras desarrolladas subjetivamente.

Se implementan algoritmos de clustering de palabras con *embeddings* (incrustación léxica), lo que arroja resultados automáticos en la distribución de clústeres y la detección de *outliers* (Na *et al.,* 2010). Se discuten las implicaciones de estos hallazgos y se proponen áreas de investigación futura, incluida la sectorización de actividades comerciales y la mejora de la visualización de datos mediante *dashboards* o mapas de calor a escala municipal. En resumen, este estudio ofrece una base para describir y comparar el desarrollo urbano y económico en diferentes ciudades mexicanas, con el potencial de ser extrapolado a otros países y regiones.

El proyecto se centró en la recolección de datos del Directorio Estadístico Nacional de Unidades Económicas (DENUES), que proporciona una importante fuente de información sobre diversas actividades económicas en los municipios (INEGI, 2023). Posteriormente, se llevó a cabo la limpieza y preparación de los datos para asegurar su calidad y utilidad en los análisis subsecuentes, así como la propuesta de una ontología de dimensiones, construida a partir de los datos recolectados, que permite una mejor estructuración y análisis de los indicadores urbanos.

Con esto, se tiene el objetivo de describir una ontología para ayude a tener una especie de indicador de desarrollo urbano para ciudades en el contexto de municipios de México que también puede servir en el caso de ayudar a descubrir potencialmente ciudades inteligentes (o ciudades próximas a ello) que permita de esta manera determinar cuáles son las áreas claves que permiten a una ciudad, determinar su grado de maduración, ya sea a nivel económico, en desarrollo o en el contexto de una ciudad inteligente (Chourabi *et al.,* 2012; Djenouri *et al.,* 2022). Esta ontología busca aprenderse a partir de datos urbanos y demográficos y usando técnicas de aprendizaje no supervisado para poder ser extrapolable a otros países.

En la actualidad, es posible abordar una serie de cuestiones cruciales que van más allá de la mera recopilación de datos y requieren un enfoque analítico y estratégico. ¿Cuáles son los criterios más pertinentes para identificar patrones de desarrollo urbano comparables entre diferentes ciudades mexicanas? ¿Cómo podemos determinar qué áreas o sectores, ya sean industriales, comerciales, sanitarios u otros, ofrecen una imagen más precisa del nivel de desarrollo urbano o social de una ciudad en el contexto de la inteligencia urbana?

Una posible aproximación para abordar estas cuestiones podría ser la creación de una estructura ontológica que permita representar cuantitativamente las similitudes y diferencias en términos de desarrollo sanitario, educativo y económico entre las distintas ciudades. Esta

estructura ontológica podría basarse en indicadores clave como la tasa de alfabetización, la accesibilidad a los servicios de salud, la infraestructura educativa y de salud, el producto interno bruto (PIB) per cápita, la diversidad económica, entre otros (Rahayu *et al.,* 2022).

Al considerar las actividades económicas, es esencial no sólo observar su presencia o ausencia en una ciudad, sino también evaluar su importancia relativa en el contexto urbano (Alti *et al.,* 2016). Por ejemplo, determinadas ciudades pueden tener una presencia destacada en sectores como la manufactura, la tecnología, el turismo o la agricultura, lo que influye significativamente en su dinámica económica y su desarrollo general.

Además, al examinar áreas de oportunidad en términos de la presencia o ausencia de ciertas actividades económicas, es crucial identificar sectores que podrían ser prometedores para el crecimiento económico y la diversificación de una ciudad. Esto podría implicar identificar industrias emergentes con potencial de crecimiento, promover políticas que fomenten la inversión en sectores clave y mejorar la infraestructura necesaria para apoyar el desarrollo de estas actividades económicas.

2. Marco teórico

El análisis del desarrollo urbano y social de las ciudades mexicanas requiere un enfoque integral que considere una amplia gama de factores, desde indicadores demográficos y de salud hasta la diversidad económica y las oportunidades de crecimiento. Aplicando herramientas analíticas y estratégicas apropiadas, es posible identificar patrones, tendencias y áreas de oportunidad que pueden contribuir a promover un desarrollo urbano más equitativo y sostenible en todo el país.

El concepto de ciudades inteligentes ha ganado considerable atención en la última década, impulsado por la necesidad de mejorar la calidad de vida urbana a través del uso eficiente de la tecnología y los datos. Numerosos estudios han investigado algunos enfoques para optimizar la coincidencia de ontologías, abordando así la eficiencia de este proceso. Estas soluciones se pueden clasificar en dos grupos principales:

En el trabajo "Understanding Smart Cities: An Integrative Framework," los autores proponen un marco integral que identifica los componentes clave de las ciudades inteligentes. Este estudio se basa en la recopilación y análisis de datos urbanos para definir y evaluar las dimensiones de una ciudad inteligente (Chourabi *et al.,* 2012).

Por otro lado en Smart City Ontologies and Their Applications: A Systematic Literature Review, se revisa sistemáticamente el uso de ontologías en el contexto de las ciudades inteligentes, destacando su papel en la interoperabilidad de datos y procesos, la gestión de big data y el razonamiento automatizado. La revisión clasifica las ontologías según los subdominios urbanos que abordan, como energía inteligente, salud inteligente y gestión de crisis, y discute los desafíos y las oportunidades que presentan las tecnologías semánticas para mejorar los servicios de las ciudades inteligentes. El estudio realiza una revisión sistemática de la literatura para investigar cómo las ontologías apoyan los servicios de ciudades inteligentes, identificando los problemas abordados y los logros alcanzados hasta la

fecha y propone una clasificación de los subdominios urbanos abordados por las ontologías, así como los temas de investigación considerados por la comunidad científica en este ámbito. Además, se destaca la eficacia de las tecnologías semánticas para mejorar el concepto de ciudad inteligente y se discuten los problemas aún sin resolver (De Nicola & Villani, 2021).
El trabajo "Autonomic Semantic-Based Context-Aware Platform for Mobile Applications" introduce la plataforma Kali-Smart, que utiliza tecnologías semánticas y un middleware autónomo para gestionar la adaptación y el razonamiento en aplicaciones móviles en entornos urbanos. La plataforma demuestra cómo las ontologías pueden facilitar la gestión dinámica e inteligente de datos contextuales provenientes de múltiples dispositivos y modalidades, proporcionando una infraestructura flexible para controlar diversas interacciones de los usuarios en tiempo real (Alti et al., 2016).
Como se menciona en (Djenouri et al., 2022), la agrupación en clústeres para instancias basadas en coincidencias de ontologías (COMI) y la minería de patrones para instancias basadas en coincidencias de ontologías (POMI) son dos marcos propuestos en un estudio reciente sobre modelado semántico de ciudades inteligentes.
Estos enfoques buscan abordar el problema de coincidencia de ontologías utilizando técnicas de agrupación y minería de patrones para descubrir conocimiento relevante en datos de ciudades inteligentes.
COMI utiliza el algoritmo *K-medias* para agrupar ontologías altamente correlacionadas, mientras que POMI selecciona propiedades relevantes para el proceso de coincidencia de ontologías. Los experimentos realizados en varias bases de datos, incluidas DBpedia y Ontology Alignment Assessment Initiative, demuestran que COMI y POMI superan a los modelos de coincidencia de ontologías anteriores en términos de costo computacional y calidad de los resultados. Estos marcos muestran efectividad en el manejo de datos heterogéneos a gran escala en entornos de ciudades inteligentes, lo que sugiere su utilidad para aplicaciones prácticas en este campo emergente (Djenouri *et al.,* 2022).

3. Metodología

El enfoque seguido en esta investigación es considerar la información de todos los comercios e instituciones públicas y privadas que contiene un municipio. A continuación, se describe la formulación de la ontología, bajo los supuestos considerados.

3.1. Formulación de la ontología

La ontología propuesta se distingue de otras porque se fundamenta en un método de análisis de datos que permite su generación automática. Por lo anterior. Se tienen algunas consideraciones de esta que se describen a continuación:

- Propósito específico de la ontología: partiendo de la definición de ontología en el campo de Inteligencia Artificial que nos dice: una ontología es una especificación explícita de una conceptualización, es decir proporciona una estructura y contenidos

de forma explícita, definiendo un conjunto de términos básicos y relaciones entre dichos términos a fin de ampliar las definiciones dadas en el vocabulario.

- Representación formal: La ontología se puede describir como una tripleta *O ={D, W, R},* donde:
  - *D*, es un conjunto de dimensiones que son conceptos $d \in D$ que representan conceptos generales aplicables a un municipio y que pueden determinar un grado de desarrollo en un municipio dado. Por ejemplo, *d*=”salud” puede determinar un grado de desarrollo alto o bajo en un municipio dependiendo si este municipio cuenta con muchas o pocas unidades médicas (hospitales, clínicas, farmacias, etc.).
  - *W* representa al conjunto de conceptos o palabras $w \in W$, donde cada una de estas palabras pertenece a una dimensión. Ejemplos de palabras *w* pueden ser “universidad”, “farmacia”, “restaurante”, “mariscos”, etc. Estas palabras se obtienen de un conjunto de datos M_{fxc} que más adelante se describe.
  - *R* que es un conjunto de relaciones que relaciona una palabra *w* hacia una dimensión *d*. La relación que utilizamos en esta ontología es de pertenencia, por abreviar, la relación “in”. Ejemplo: *in(“hospital”, “salud”)*, donde “hospital”∈ *W*, y “salud” ∈ *D*. por simplicidad de la ontología si ocurre *in(w,d)* no ocurre que *in(w, D-{d})*. Igualmente la relación *tiene(w,m)*, representa que el municipio *m* tiene la palabra *w* en la base de datos consultada. Un ejemplo es: *tiene(“monterrey”, “farmacia”)*, indicando que el municipio “monterrey” tiene la palabra “farmacia” al consultar sus unidades económicas en la base de datos.

  De manera más técnica, cada dimensión *D* se obtiene a través de los clústeres del modelo. Las dimensiones obtenidas se describen en la Figura 10, son 12, a saber: *D = {industria, economía, salud, alimentación, comunicaciones, cultura, ambiente, movilidad y transporte, ciudadanía, vivienda, gobernanza, educación}*.
  Cada palabra del conjunto *W* se obtiene de una base de datos de conceptos de unidades económicas, que para los efectos de este trabajo se obtuvieron de DENUE de INEGI. Otras bases de datos para otras poblaciones pueden utilizarse para conformar *W*.
- Simplificación, La ontología simplifica la realidad al usar conceptos para describir a los municipios. Un municipio *m* puede ser descrito en términos de los conceptos que contienen sus unidades económicas (por ejemplo, marisquería, servicios de consultoría, abogados, universidad, etc.) y por tanto un municipio *m* puede ser definido con una métrica que determine su nivel de desarrollo en dicha dimensión *d* mediante la fórmula: $M_d = \frac{|N_m|}{|L_d|}$, donde: $N_m = \{w|\exists in(w,d), tiene(m,w)\}$ y $L_d = \{w|\exists in(w,d\}$, donde $w \in W, d \in D$, y *m* es un municipio considerado en la base de datos de donde se han tomado los conceptos.
  En particular, decimos que un municipio Mx_d tiene más desarrollo que un municipio My_d en una dimensión *d* si se cumple que: $Mx_d > My_d$. Como nota, $0 \leq M_d \leq 1$.

- Abstracción: la ontología abstrae una parte de la realidad al considerar a las palabras del conjunto *W* como elementos de una dimensión determinada *d*.
- Propósito específico de la ontología: poder describir actividades económicas preponderantes de los municipios de México, agrupadas por 12 dimensiones (conceptos).
- Testabilidad: Si bien la ontología no pretende predecir una determinada variable, si puede describir, por un lado, municipios con un grado de desarrollo similar al momento de tener valores *M_d* similares para una o más dimensiones *d*.
- Verificabilidad: al describir un municipio, se puede verificar los conceptos *w* que contiene a partir de la base de datos que se considere, en este caso, la base de datos DENUE de INEGI.
- Escalabilidad: La ontología puede crecer en conceptos *w* y en dimensiones *d* al contar con una base de datos de conceptos de las unidades económicas que sea más grande, por lo que sí puede escalar a una conceptualización más amplia (con una jerarquía de conceptos más grande). En principio la cantidad de conceptos puede rondar el orden de mil a tres mil conceptos (un diccionario en español maneja 90 mil palabras aproximadamente). Si bien en teoría no hay límite, un número cercano al total de palabras de diccionario podría suponer un problema de alta dimensionalidad haciendo más lento el cálculo de las dimensiones o clústeres. De cualquier manera su cálculo está en un orden polinomial ($O(n^3)$).
- Transparencia: Las palabras son visualizables en la jerarquía de la ontología y cada relación es conocida a partir de los datos. Más adelante se detallan aspectos de la clusterización.
- Comprensibilidad: La ontología se puede visualizar como una estructura jerárquica (ver figura 10) por lo que es comprensible para personas del área de Inteligencia Artificial como no personas cercanas a esta área.
- Generalidad: La ontología generaliza la descripción de un municipio a sus aspectos más relevantes que provienen de sus unidades económicas. Naturalmente, se pierden algunos otros elementos que en la literatura sí son considerados para construir otras ontologías que describen ciudades o municipios, por ejemplo: población, extensión territorial, encuestas de satisfacción de la población, organizaciones civiles en el municipio, etc. Sin embargo, consideramos que la generalización realizada captura información, si bien con énfasis en lo económico, que permite describir un grado de desarrollo en el municipio.
- Especificidad: un municipio es describible por un conjunto $r_m \subseteq R$ que permite determinar su grado de desarrollo por las fórmulas propuestas anteriormente. Este grado de desarrollo sirve para propósitos comparativos. No para determinar una posición absoluta de un municipio en esta métrica.
- Adaptabilidad: La ontología propuesta es adaptable a otros contextos, previo re-entrenamiento, es decir, el aprendizaje de conceptos diferentes, con alguna base de datos diferente.

- Parámetros: los parámetros que dan construcción a la ontología son: n_c que es el número de clústeres que determina las dimensiones, n_w un número que describa la frecuencia mínima que una palabra debe cumplir para ser incluida en la ontología. En nuestra propuesta, este número es 0 dado que se aceptaron todas las palabras contenidas en DENUE, eliminando solamente palabras vacías, como conjunciones, preposiciones, números y artículos.
- Variables: En la conceptualización de la ontología, las variables consideradas fueron: *d* (dimensiones), *w* (palabras que aparecen en las unidades económicas), *m* (municipios) y las relaciones que involucran a estas variables.
- Suposiciones: Varias suposiciones se hicieron para la construcción de la ontología:
    - Palabras cercanas semánticamente deben representar a una misma dimensión.
    - Las palabras de las unidades económicas que sí aparecen en un municipio sugieren un grado de desarrollo en ese municipio. Palabras que no aparecen, sugieren una aparición de esa unidad económica en ese municipio.
    - Si una palabra aparece muchas veces en un municipio, sugiere que la dimensión *d* a la que pertenece esa palabra, tiene más presencia en dicho municipio.
- Limitaciones: El modelo está limitado por las palabras aprendidas de la base de datos que se utilice. Igualmente, el modelo no describe aspectos que se encuentren fuera de las posibilidades de estas palabras, tal como demografía, recursos hídricos, grado de contaminación del municipio y sus compuestos químicos, etc. Estas otras variables quedan fuera de los alcances descriptivos de la ontología propuesta. En este sentido la ontología diagnóstica las actividades más preponderantes, pero no predice alguna variable a partir de la información de entrada.

Bajo estas consideraciones para la formulación de la ontología, se describen a continuación algunos aspectos del método seguido para la construcción de esta ontología.

3.2. Fuente de Datos

Para recuperar información sobre las actividades económicas de las ciudades, se trabajó con datos del INEGI (Base de Datos DENUE (Directorio Estadístico Nacional de Unidades Económicas) del año 2023 (INEGI, 2023). Esta base de datos contiene información textual de empresas de diversos sectores (comercial, industrial en México, incluyendo información georreferenciada que favorece la descripción, agricultura) a escala municipal.

Para construir una representación útil para los propósitos de este trabajo, se unieron diferentes bases de datos obtenidas de DENUE, teniendo un conjunto de datos M_{fxc}, donde f representa la totalidad de los municipios de méxico y c representa a diversos conceptos económicos y demográficos (palabras) que ese municipio posee a partir de un análisis y limpieza de datos de la base de datos del DENUE.

- Proceso de Recolección

- Descarga de Datos: se accedió a los datos del DENUES a través de su plataforma en línea, descargando conjuntos de datos relevantes para los municipios en estudio.
- Selección de Variables: se seleccionaron variables clave que caracterizan las actividades económicas y otros aspectos relevantes para definir una ciudad inteligente, como infraestructura, servicios públicos, y tecnología. Estas variables fueron guardadas como tokens o conceptos en formato de una sola palabra.

- Limpieza y Preparación de los Datos
  - Eliminación de Datos Duplicados: se identificaron y eliminaron registros duplicados para asegurar la unicidad de las entradas.
  - Manejo de Datos Faltantes: se implementaron métodos para tratar datos faltantes, incluyendo la imputación de valores y la eliminación de registros con datos incompletos.
  - Normalización y Estandarización: los datos fueron normalizados y estandarizados para garantizar consistencia y facilitar análisis posteriores.

3.3. Entorno de desarrollo

Se utilizó Python versión 3.11 con librerías para manejo de datos y uso de algoritmos de aprendizaje automático de tipo no supervisado. Se escogió este entorno por su amplio uso en diversos campos como aplicaciones de escritorio, desarrollo web y ciencia, Python resultó una opción versátil por el manejo de diferentes contextos (Jalolov, 2023).

3.4. Incrustación de palabras

Incrustación léxica o mejor conocida como *Word Embedding* (Karani, 2018), es una técnica que permite representar una palabra (o un concepto) de un idioma determinado (español, inglés, etc.) en una representación de un vector $\omega^n, n \in N$ tal que, debido a su proximidad a otras palabras en un corpus de un texto determinado, supone que están cerca en dicho espacio *n-dimensional*. Si consideramos un corpus suficientemente grande que incluya todos los diccionarios y textos de una lengua disponibles públicamente, podemos suponer que las palabras sinónimas se encontrarán muy cerca en dicho espacio, y las palabras antónimas se encontrarán igualmente distantes. El modelo Word2Vec (Karani, 2018), es una representación de las palabras del idioma inglés siguiendo esta idea.

3.5. Propuesta de Ontología de Dimensiones

Una ontología se define como una descripción formal y explícita de los conceptos presentes en un determinado campo del conocimiento, así como de las relaciones que existen entre ellos (Rahayu et al., 2022). En definitiva, constituye una representación estructurada de los términos y conexiones que caracterizan un dominio específico.

Estas ontologías permiten la comprensión y el procesamiento jerárquico de información textual o conceptual mediante sistemas informáticos. Al proporcionar un marco semántico y estructurado, facilitan la interacción y la interpretabilidad de los sistemas, así como la búsqueda y recuperación de datos según las reglas definidas por la ontología. Esto permite la reutilización del conocimiento, su uso como base de datos y base de conocimiento, y facilita su explotación por sistemas basados en inteligencia artificial.

Las ontologías suelen expresarse en lenguajes formales como OWL (Web Ontology Language) o RDF (Resource Description Framework) (Malkawi *et al.,* 2022), lo que permite una descripción precisa y estructurada de conceptos, propiedades y relaciones. En este trabajo buscamos crear una ontología que se asemeje a las descripciones ontológicas realizadas en la literatura sobre áreas económicas, llamadas dimensiones y categorías dentro de cada una de estas dimensiones.

Esta ontología pretende mantener coherencia con el trabajo sobre ciudades inteligentes y basarse en datos observados en los municipios o ciudades que conforman un país, teniendo a México como caso de estudio, es por ello que se utilizará una ontología de dominio, éstas están diseñadas para modelar el conocimiento en un dominio específico de interés, proporcionando una representación detallada de conceptos y relaciones relevantes para este campo, así como una combinación con una ontología de representación de datos, la cual se utiliza para describir la estructura y las restricciones de los datos. Son importantes para la integración y la interoperabilidad de datos procedentes de diversas fuentes.

3.6. Clustering

Un *clúster* en el contexto del análisis de datos y el aprendizaje automático es un conjunto de instancias de datos representadas por vectores del mismo tamaño ($V^n, n \in N$), donde n representa la dimensión del vector. Estos vectores (registros en el contexto de bases de datos) se agrupan porque contienen una similitud de distancia (como la euclidiana) entre ellos. Este umbral de distancia puede definirse *a priori* o mediante técnicas cuantitativas y automáticas. La tarea de agrupar bajo el enfoque de aprendizaje no supervisado es agrupar un conjunto de datos en varios grupos de modo que los puntos dentro de cada grupo sean más similares entre sí que a los puntos de otros grupos. Los clústeres pueden representar grupos naturales o patrones intrínsecos en los datos y pueden usarse para explorar y comprender la estructura subyacente de un conjunto de datos, ya sea de baja o alta dimensión.

Uno de los enfoques más comunes para la agrupación es el algoritmo *K-medias*. En *K-medias*, el algoritmo busca minimizar la suma de las distancias al cuadrado de cada punto de datos al centroide de su grupo asignado. El centroide representa un punto sintético que describe su grupo de vectores más similar.

El método *K-medias* consiste en clasificar objetos dados en k grupos diferentes mediante iteraciones, convergen a un mínimo local. El algoritmo consta de dos fases separadas. La primera fase selecciona k centros aleatoriamente, donde el valor k se fija de antemano.

La siguiente fase es llevar cada objeto de datos al centro más cercano. Generalmente, se considera la distancia L_1 o euclidiana para determinar la distancia entre cada objeto de datos y los centros del grupo. Cuando todos los objetos de datos están incluidos en algunos grupos, se completa el primer paso y se realiza la agrupación temprana. Vuelve a calcular el promedio de los primeros grupos formados. Este proceso iterativo continúa repetidamente hasta que la función de criterio se vuelve mínima. Suponiendo que el objetivo es x, x_i denota el promedio del grupo C_i, la función de criterio se define de la siguiente manera: $E = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2$, donde *E* es la suma del error al cuadrado de todos los objetos x en la base de datos, μ_i son las coordenadas del centroide *i*-ésimo y k es el número total de clústeres del algoritmo *K-medias*. La distancia de la función de criterio es la distancia euclidiana, que se utiliza para determinar la distancia más cercana entre cada objeto de datos y el centro del grupo. La distancia euclidiana entre un vector $x = x_1, x_2, \ldots, x_n$ y otro vector $y = y_1, y_2, .., y_n$ se puede obtener de la siguiente manera: $d(x, y) = [\sum_{i=1}^{n} (x_i - y_i)^2]^{\frac{1}{2}}$. El algoritmo de agrupamiento de *K-medias* siempre converge al mínimo local. Antes de que el algoritmo *K-medias* converja, se realizan cálculos de distancia y centro del grupo mientras los bucles se ejecutan varias veces, donde el entero positivo *t* se conoce como el número de iteraciones de *K-medias*. El valor preciso de *t* varía dependiendo de los centros iniciales del grupo inicial. La distribución de puntos de datos tiene una relación con el nuevo centro de agrupamiento, por lo que la complejidad del tiempo computacional del algoritmo *K-medias* es $O(nkt)$. Donde *n* es el número de todos los objetos de datos, *k* es el número de grupos, *t* son las iteraciones del algoritmo. Generalmente requiere $k \ll n \; y \; t \ll n$.

En cuanto a las funciones asociadas al clustering, incluyen:

- Preprocesamiento de datos: Este paso implica la limpieza y transformación de datos para prepararlos para el clustering. Esto puede incluir la eliminación de valores atípicos, la normalización o estandarización de características, la codificación de variables categóricas y la reducción de dimensionalidad si es necesario.
- Selección de algoritmo de clustering: Seleccionar el algoritmo de clustering adecuado para el conjunto de datos y el problema en cuestión es crucial. Algunos de los algoritmos comunes incluyen *K-medias*, *Agglomerative Clustering, DBSCAN, Mean Shift*, y otros. La elección del algoritmo puede depender de factores como la distribución de los datos, el tamaño del conjunto de datos y la estructura esperada de los clústeres.
- Definición de parámetros: Una vez seleccionado el algoritmo de clustering, es importante definir los parámetros necesarios para ejecutar el algoritmo. Por ejemplo, en *K-medias*, se debe especificar el número de clústers, mientras que en *DBSCAN* se debe definir el radio y el número mínimo de puntos.
- Entrenamiento del modelo: Una vez que se han definido los parámetros, se entrena el modelo de clustering utilizando el conjunto de datos preparado.

- Evaluación del modelo: Después de entrenar el modelo, es importante evaluar su rendimiento utilizando métricas adecuadas para el clustering. Algunas métricas comunes incluyen la inercia, el coeficiente de silueta, la pureza de clúster, entre otros.
- Visualización de resultados: Finalmente, es útil visualizar los resultados del clustering para comprender mejor la estructura de los clústers y la distribución de los datos. Esto puede incluir visualizaciones como gráficos de dispersión, diagramas de dendrogramas, y mapas de calor.

Generalmente son funciones que ayudan a calcular y evaluar los clústers generados por un algoritmo de agrupamiento. Algunas de las funciones que se utilizaron en el estudio incluyen:

- Funciones de distancia: Estas funciones calculan la distancia entre dos puntos de datos en el espacio de características, en nuestro caso de estudio se utilizará la distancia euclidiana en donde se necesitan medidas de similitud o distancia entre puntos en un espacio *n-dimensional*.
- Funciones de evaluación: Estas funciones se utilizan para evaluar la calidad de los clústeres generados por un algoritmo de clustering. Utilizaremos el coeficiente de silueta, la medida de Davies-Bouldin, la homogeneidad, entre otros para poder comparar la mejor cantidad de clústeres.
- Funciones de inicialización: En el caso de algoritmos de clustering como *K-medias*, se utilizan funciones de inicialización para seleccionar los centroides iniciales antes de comenzar el proceso de clustering (Na *et al.,* 2010).

En la Figura 1 se muestra un diagrama bloques de los pasos que sigue la metodología propuesta que permite llegar a descubrir una ontología de conceptos económicos a partir de la base de datos del DENUE. Esta metodología puede generar distintas metodologías si en versiones sucesivas se amplían los datos disponibles como fuente.

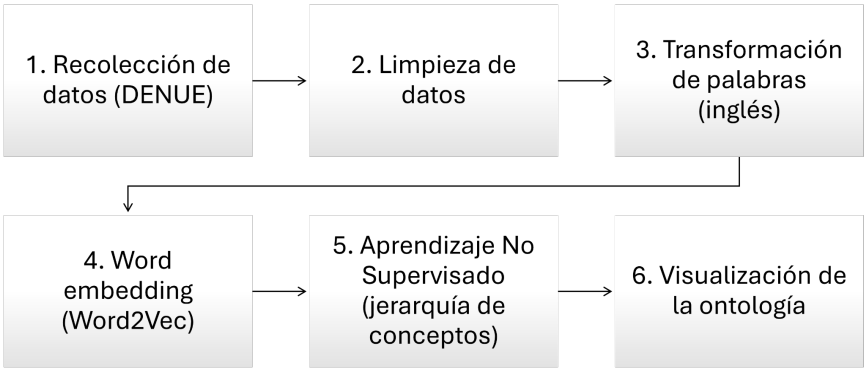

Figura 1. Diagrama de bloques de la metodología seguida para el descubrimiento y visualización de una ontología aplicable al contexto de municipios de México. La metodología propuesta admite otros contextos en la medida en que se recolecten datos de otras ciudades.

4. Resultados

Se llevó a cabo la recolección de información relacionada, inicialmente, con la revisión de la literatura de trabajos que sugieren algunas dimensiones (jerarquías de conceptos) que caracterizan una ciudad inteligente. A partir de esto, se optó por tomar los registros del Directorio Estadístico Nacional de Unidades Económicas (DENUE), una base pública proporcionada por el Instituto Nacional de Estadística y Geografía (INEGI) de México. Esta fuente incluye información textual sobre empresas de diversos sectores (comercial, industrial, agrícola, entre otros) en México, con detalles georreferenciados que facilitan su análisis a escala municipal.

Se realizó la estandarización, cuantificación y conversión de los valores a formatos categóricos o numéricos, entre otros procesos. Asimismo, se aplicaron técnicas para identificar posibles duplicados y determinar los atributos que presentan carencias de información, con el propósito de caracterizar los conjuntos analizados. En la Figura 2 se muestra un mapa de calor que facilita la visualización de estas carencias, indicando rápidamente las áreas del conjunto donde falta información. En este gráfico, las barras negras representan valores existentes, mientras que las blancas señalan los vacíos detectados. En el caso trabajado, los vacíos se relacionan principalmente con datos como las páginas web o correos electrónicos de los negocios descritos. Sin embargo, también se observaron algunos huecos relacionados con códigos postales o tipos de vialidad de las ubicaciones comerciales. Al finalizar esta etapa, se obtuvo un modelo tabular que integra toda la información a escala municipal, complementado con datos de apoyo local, estatal y nacional, como se ilustra en la Figura 3.

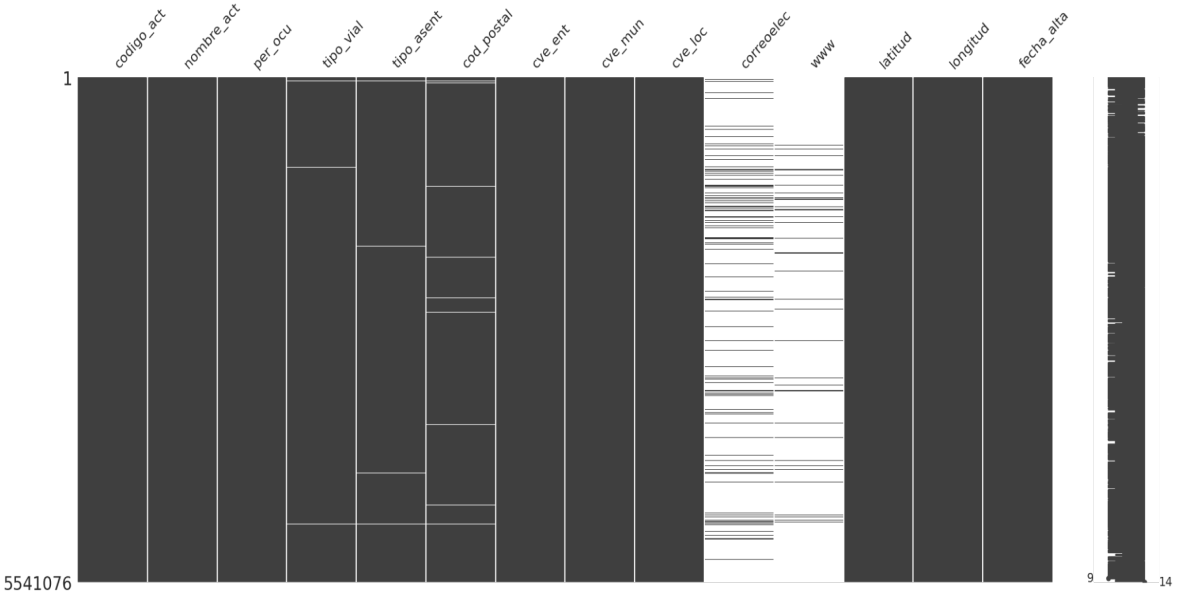

Figura 2. Mapa de calor de datos faltantes del conjunto de datos. Se puede observar en la figura que existen 5541076 registros de la base de datos del DENUE.

	codigo_act	nombre_act	per_ocu	tipo_vial	tipo_asent	cod_postal	cve_ent	cve_mun	cve_loc	latitud	longitud	fecha_alta
0	115119	Otros servicios relacionados con la agricultura	6 a 10 personas	AVENIDA	COLONIA	20130.0	1	1	1	21.906992	-102.309807	2019-04
1	115119	Otros servicios relacionados con la agricultura	0 a 5 personas	CALLE	FRACCIONAMIENTO	20040.0	1	1	1	21.889926	-102.314009	2019-11
2	112512	Piscicultura y otra acuicultura, excepto camar...	0 a 5 personas	CALLE	LOCALIDAD	20437.0	1	7	239	22.167778	-102.345556	2010-07
3	112512	Piscicultura y otra acuicultura, excepto camar...	6 a 10 personas	CARRETERA	LOCALIDAD	20336.0	1	10	272	21.956298	-101.997312	2014-12
4	112512	Piscicultura y otra acuicultura, excepto camar...	0 a 5 personas	PROLONGACION	NaN	20000.0	1	3	56	21.837255	-102.710931	2014-12
...	...	...	...	...	...	...	...	...	...	...	...	...
5541071	337120	FabricaciÃ³n de muebles, excepto cocinas integ...	0 a 5 personas	CALLE	BARRIO	99700.0	32	48	1	21.783809	-103.299234	2020-11
5541072	337120	FabricaciÃ³n de muebles, excepto cocinas integ...	0 a 5 personas	CALLE	COLONIA	99100.0	32	42	1	23.647130	-103.648225	2010-07
5541073	337120	FabricaciÃ³n de muebles, excepto cocinas integ...	0 a 5 personas	CALLE	COLONIA	99545.0	32	55	1	22.353532	-102.870949	2023-11
5541074	332320	FabricaciÃ³n de productos de herrerÃa	0 a 5 personas	CALLE	COLONIA	99960.0	32	23	1	21.415183	-103.118349	2010-07
5541075	339950	FabricaciÃ³n de anuncios y seÃ±alamientos	0 a 5 personas	CALZADA	FRACCIONAMIENTO	98610.0	32	17	1	22.758608	-102.531529	2019-11

5541076 rows × 12 columns

Figura 3. Modelo tabular de los datos recolectados.

Para visualizar de mejor manera las variables o indicadores que se usarán, es necesario construir un esquema de dimensiones. Existen múltiples elementos y factores que conforman y se entrelazan en estos modelos, siendo los siete más relevantes, éstos son: vivienda, economía, gestión pública, movilidad y transporte, inclusión social, medioambiente, energías renovables y ciudadanía, como se muestra en la Figura 4, con la premisa fundamental de poner al ciudadano en el centro, involucrándose en la toma de decisiones, en la construcción y operación de la ciudad en su totalidad. Esta ontología es un diagrama propuesto mediante la observación de los datos de forma manual, esta ontología nos servirá para tener una noción de las dimensiones que conforman una ciudad inteligente.

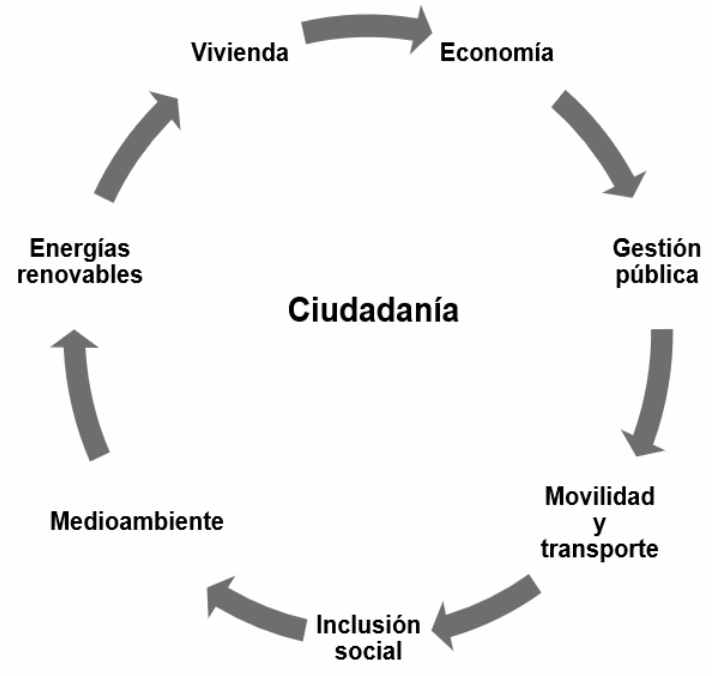

Figura 4. Ontología propuesta a partir de la revisión de la literatura.

Para poder contemplar el desarrollo de una ontología que se base en los datos, fue necesario traducir todos los datos a inglés dado que en un inicio los datos recolectados están en español, es por ello que para generar la ontología, nos enfocamos en los datos que muestran las actividades realizadas en cada municipio, para ello, mediante el procesamiento de lenguaje natural (PLN), se filtraron todas las palabras que conforman la columna de las actividades, eliminando las *stopwords*, estas "palabras vacías", son palabras comunes que suelen ser omitidas, de modo que aparecen con mucha frecuencia en un idioma pero no aportan mucho significado al contenido de un texto. Ejemplos típicos de *stopwords* en español incluyen "el", "la", "los", "de", "y", "a", entre otros, una vez filtradas estas palabras se generó una nube de palabras, algunas de éstas son: comida, diseño, agrícola, preparación, eléctrico, público, construcción, trabajo, joyería, computación, extranjeros, especializado, financiero, entre otras, como se muestra en la Figura 5, con esto se puede ver que la ontología propuesta en un inicio concuerda con los datos.

Figura 5. Nube de todas las palabras filtradas y traducidas. Las palabras más grandes representan una mayor frecuencia de aparición entre los municipios.

Para poder trabajar los datos de una forma más sencilla en donde algún modelo de agrupamiento se ejecute sin problema, se incrustó léxicamente todas las palabras filtradas, para encontrar los vectores cercanos, se utilizó en conjunto un modelo pre-entrenado de word embedding Word2Vec.

Habiendo obtenido los vectores resultantes, se procesaron en un modelo de *K-medias,* con el que se generaron 20 clústeres o agrupaciones, cada clúster tiene información latente de cada palabra en un espacio n-dimensional obtenido por Word2Vec, es decir, cada clúster almacena aquellas palabras que son semánticamente parecidas, de acuerdo a Word2vec, en

la Figura 6 se muestra la distribución de cada clúster, con esto podemos observar que los datos dentro de cada grupo son similares entre sí (que se denomina similaridad intraclase), y los datos de diferentes grupos son más distintos entre sí (disimilaridad interclase), aunque en general comparten ciertas características (aspectos económicos, por lo general). Esto se puede visualizar en la Figura 7, en donde esta visualización ayudó a descubrir agrupamientos naturales en los datos sin ningún conocimiento previo de las categorías.

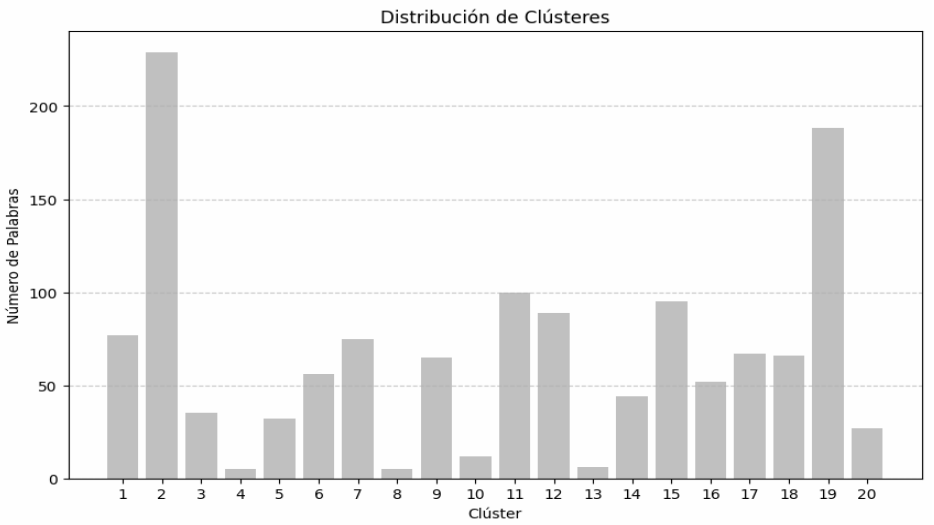

Figura 6. Distribución de clústeres. Los clústeres presentan una distribución desigual de palabras. Los clústeres pequeños sugieren que algunas palabras son atípicas con respecto del resto de palabras encontradas.

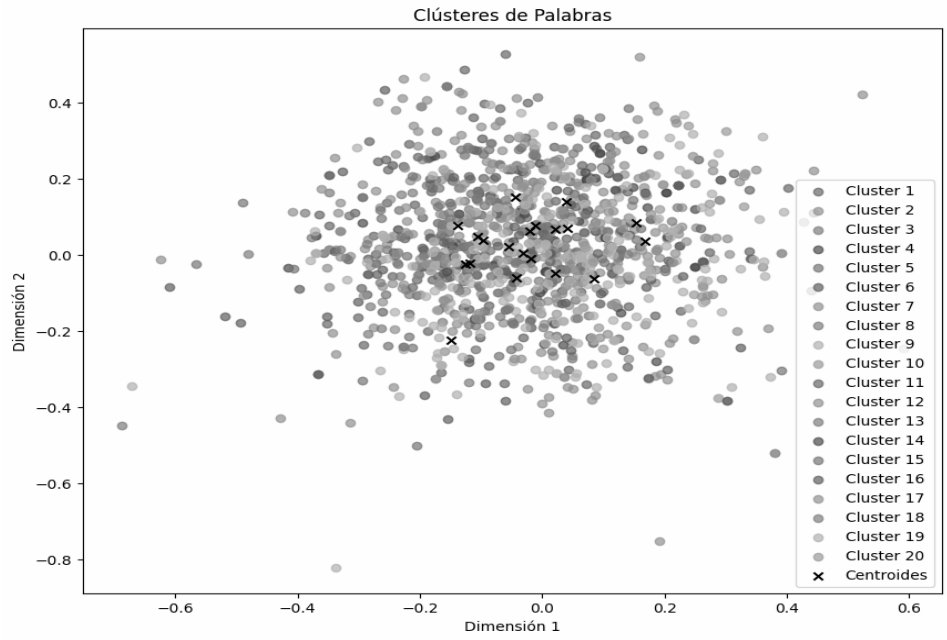

Figura 7. Dispersión de clústeres. Visualización de datos en un diagrama de dispersión en 2D mediante sus dos componentes principales. Se observa en lo general que es aún difícil determinar de esta manera una separabilidad de los centroides de los clústeres.

También fue necesario obtener el contenido de cada clúster para así tener una mejor categorización, es por ello que se generaron 20 dendrogramas como el que se muestra en la Figura 8, el cual es la representación de uno de los clústeres, aquí se pueden observar cómo las palabras al tener características similares se unen entre sí. Para poder visualizar de mejor forma, se ha generado una ontología reducida, como se muestra en la Figura 9, ésta contiene algunas palabras obtenidas del dendrograma de ejemplo, así en la categoría de industria, se obtuvieron palabras como minería, maquinaria, refinería, generadores, textiles, plásticos, máquinas, petroquímica.

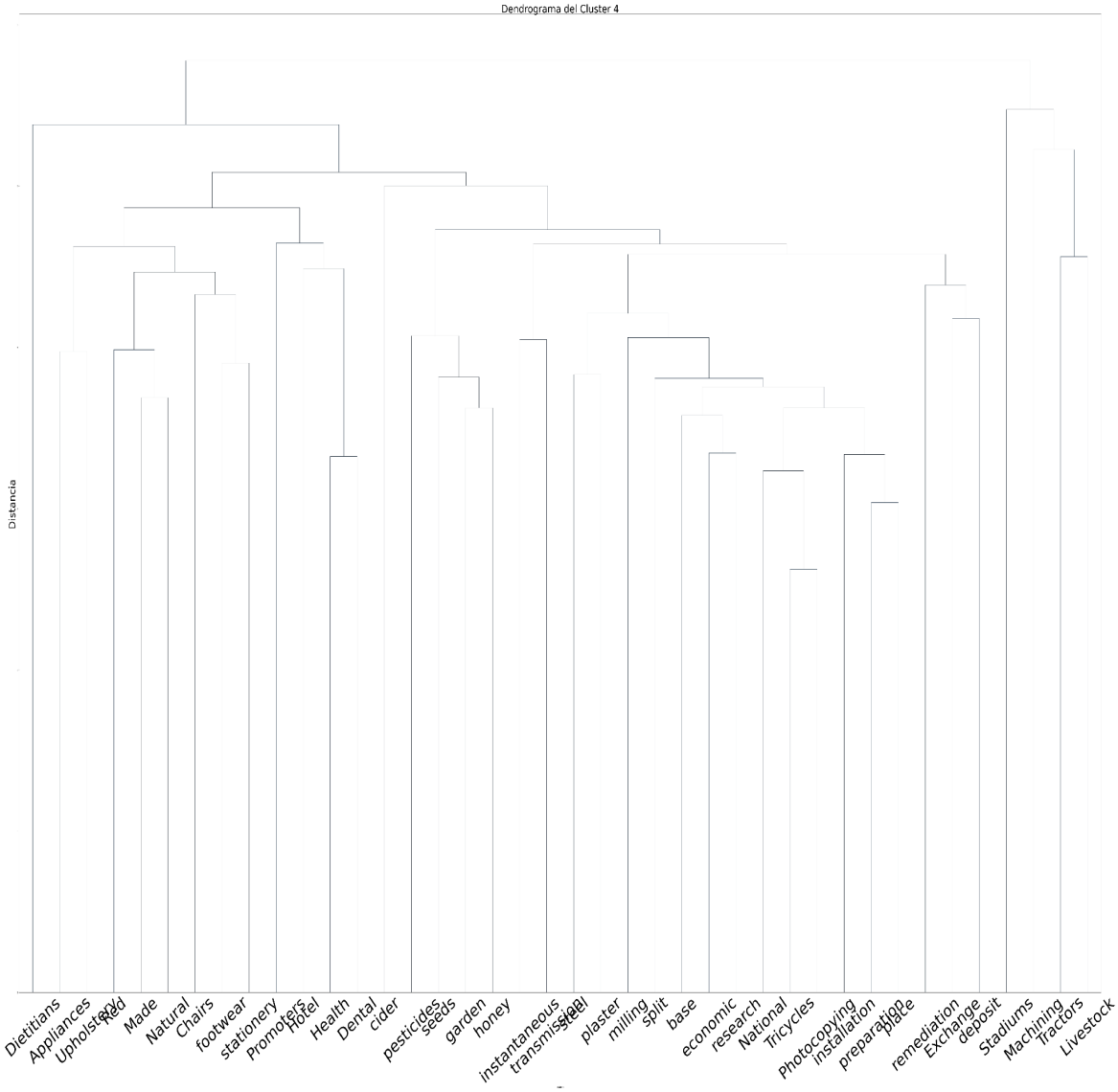

Figura 8. Representación del dendrograma del clúster no. 4. En la imagen se muestran todas las palabras que complementan el clúster. La letra es pequeña debido a que las estructuras jerárquicas de los dendrogramas son típicamente muy grandes.

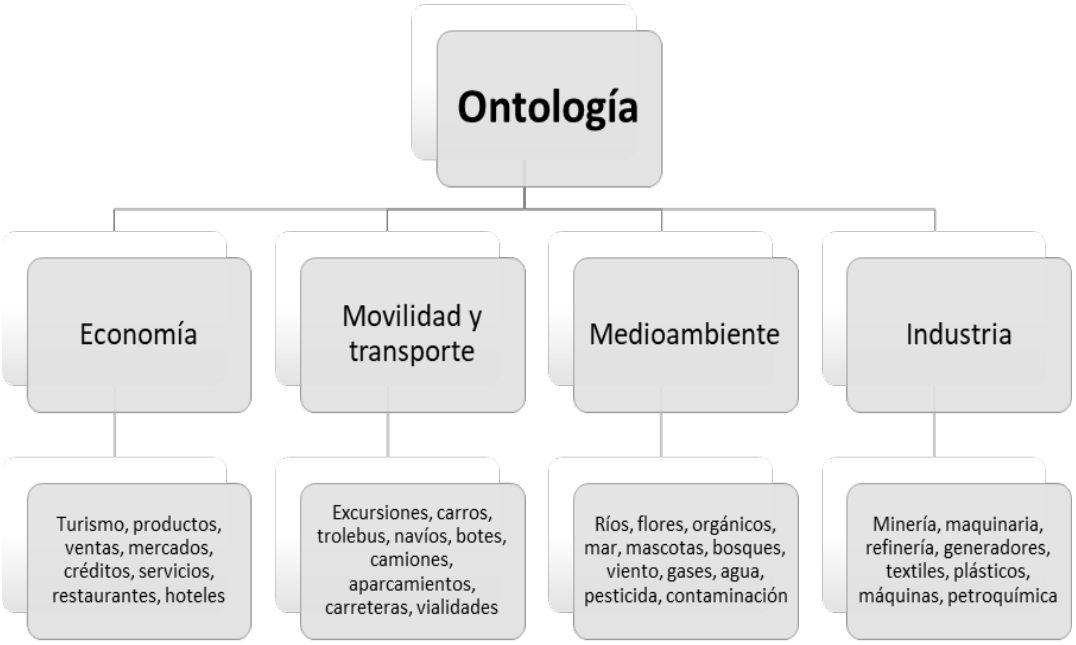

Figura 9. Ontología reducida del clúster no. 4, obtenida del dendrograma. En la imagen se muestran algunas palabras que complementan el clúster, éstas son tomadas del dendrograma que se obtiene del clúster.

Asimismo, se categorizaron los datos, obteniendo 12 dimensiones que entre sí comparten características, las cuales se categorizan como: industria, economía, salud, alimentación, comunicación, cultura, medioambiente, movilidad y transporte, ciudadanía, vivienda, gobernanza y educación, es por eso que en la Figura 10, se representa la ontología que se obtiene al procesar los datos, para el caso mexicano. Con ésta, podemos obtener algunos descubrimientos en el ámbito del análisis de datos, así como la posibilidad de aplicar técnicas de inteligencia artificial y gestión del conocimiento posteriormente, pues la ontología por sí misma proporciona una estructura definida de los datos y que mediante un sistema de razonamiento, proporcionando una base para entender y procesar el conocimiento, haciendo posible el desarrollo de algoritmos que estiman el grado de desarrollo económico a partir de esta ontología.

Finalmente, en este estudio, se realizó un análisis exploratorio de la frecuencia de términos asociados a diferentes municipios de México para identificar aquellos con una mayor diversidad de características documentadas. Este estudio, se ha centrado en observar patrones en la representación de datos de los municipios y relacionar estas características con la presencia de información relevante en diversas fuentes.

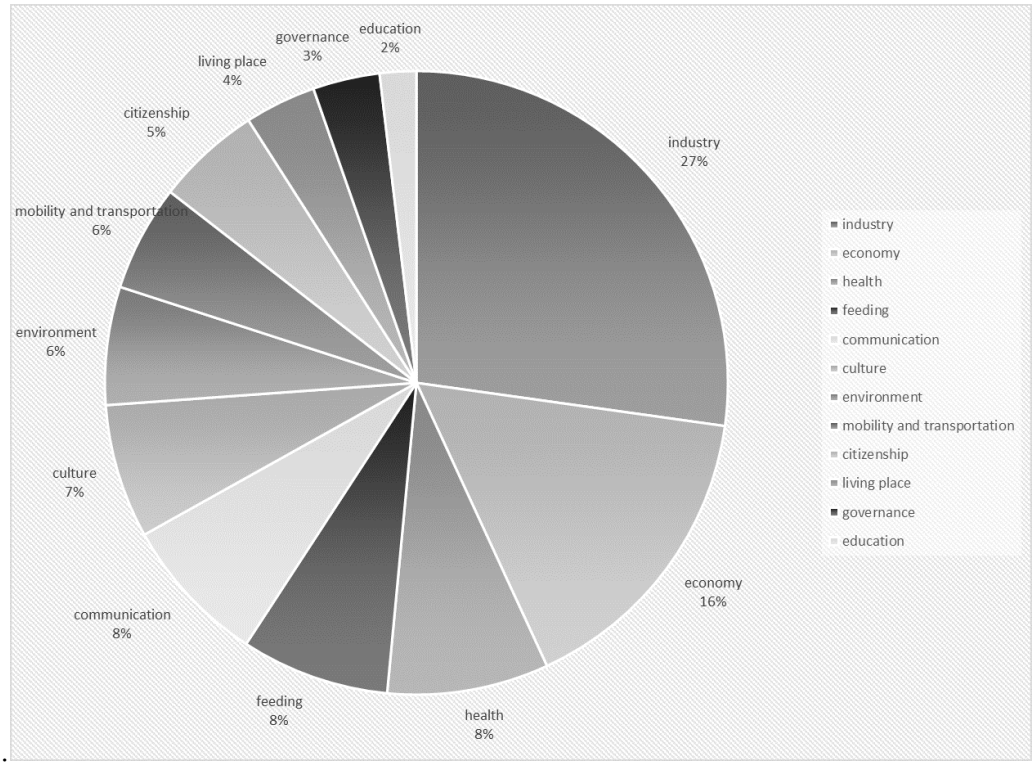

Figura 10. Ontología obtenida del procesamiento de los datos. Los conceptos obtenidos, hacen referencia a las dimensiones de la ontología obtenida. cada dimensión es un concepto que contiene un conjunto de otros conceptos que se encuentran vinculados de manera estrecha de acuerdo el encaje léxico realizado (word embedding vía Word2Vec).

4.1. Ranking de Municipios con Mayor Total de Palabras

En la tabla 1 de los municipios con mayor frecuencia de palabras, destacan **Monterrey (Nuevo León)**, **Tijuana (Baja California)**, y **Ciudad de México (CDMX)** como los municipios con los totales más altos, superando las 6700 palabras cada uno. Específicamente, **Monterrey** presentó el valor más alto con 8062 palabras, seguido por **Tijuana** con 6996.79 y **Ciudad de México** con 6790.85.

Estos resultados indican que estos municipios concentran una amplia variedad de características y términos asociados, lo cual podría estar relacionado con su importancia económica, demográfica y social en el contexto nacional. La alta representación de palabras sugiere una gran cantidad de datos e información disponibles para estas localidades, lo cual es consistente con el hecho de que son centros urbanos y económicos relevantes en el país.

Tabla 1. Ranking de municipios con mayor total de palabras

MUNICIPIOS	ESTADOS	Total de palabras
Monterrey	Nuevo León	8062.832
Tijuana	Baja California	6996.792
Cuauhtémoc	Ciudad De México	6790.852
Guadalajara	Jalisco	6754.815
Puebla	Puebla	5972.797
León	Guanajuato	5694.746
Chihuahua	Chihuahua	5616.842
Culiacán	Sinaloa	5598.813
Mérida	Yucatán	5452.822
Hermosillo	Sonora	5084.817
Mexicali	Baja California	4932.809
Juárez	Chihuahua	4814.775
Querétaro	Querétaro	4586.818
Centro	Tabasco	4406.800
San Luis Potosí	San Luis Potosí	4332.815
Toluca	México	4106.769
Torreón	Coahuila De Zaragoza	3906.804
Iztapalapa	Ciudad De México	3906.783
Aguascalientes	Aguascalientes	3846.821
Zapopan	Jalisco	3750.823

4.2. Ranking de Municipios con Menor Total de Palabras

En contraste, en la tabla 2, se muestran los municipios con menor frecuencia de palabras mostraron valores extremadamente bajos, llegando a registrar totales de 0 palabras. Todos los municipios con menor número de palabras pertenecen al estado de **Oaxaca**, tales como **San Marcial Ozolotepec**, **San Gabriel Mixtepec**, y **San Ildefonso Amatlán**.

Tabla 2. Ranking de municipios con menor total de palabras

MUNICIPIOS	ESTADOS	Total de palabras
San Miguel Peras	Oaxaca	0.0
San Juan Comaltepec	Oaxaca	0.0
San Miguel Santa Flor	Oaxaca	0.0
San Miguel Suchixtepec	Oaxaca	0.0
San Sebastián Abasolo	Oaxaca	0.0
San Raymundo Jalpan	Oaxaca	0.0
San Martín Lachilá	Oaxaca	0.0

San Martín Peras	Oaxaca	0.0
San Pedro Teutila	Oaxaca	0.0
Santos Reyes Yucuná	Oaxaca	0.0
San Marcial Ozolotepec	Oaxaca	0.0
San Gabriel Mixtepec	Oaxaca	0.0
San Ildefonso Amatlán	Oaxaca	0.0
San Andrés Yaá	Oaxaca	0.0
San Juan Yaeé	Oaxaca	0.0
San Juan Cieneguilla	Oaxaca	0.0
San Juan Yatzona	Oaxaca	0.0
San Lorenzo	Oaxaca	0.0
San Miguel Tecomatlán	Oaxaca	0.0
Santiago Lalopa	Oaxaca	0.0

Este bajo nivel de representación puede ser indicativo de varias situaciones:

1. **Poca disponibilidad de datos**: Los municipios de Oaxaca con menor frecuencia de palabras podrían reflejar una baja cobertura o disponibilidad de información en las bases de datos consultadas.
2. **Pequeña extensión o relevancia económica**: La baja frecuencia puede estar asociada a la menor actividad económica o a ser localidades con poblaciones reducidas, lo que a menudo implica menos eventos o características documentadas.

El análisis muestra una clara disparidad entre los municipios más representados en términos de datos documentados y aquellos con escasa o nula representación. Municipios grandes y relevantes como Monterrey y Ciudad de México presentan una gran diversidad de términos, lo que refuerza su papel como centros de actividad significativa en diversas dimensiones sociales y económicas. Por otro lado, municipios más pequeños o menos conocidos muestran poca diversidad de características, lo que podría sugerir la necesidad de mejorar la recolección de datos en estas áreas menos representadas.

Este análisis ofrece una visión sobre cómo la disponibilidad de información varía significativamente entre diferentes regiones del país, lo cual podría ser útil para estudios futuros centrados en el desarrollo regional, la planificación de políticas públicas y la mejora de los sistemas de información.

5. Discusión

En este estudio, hemos utilizado técnicas de estandarización, cuantificación, conversión de datos y técnicas avanzadas de procesamiento de lenguaje natural (PLN) para construir una ontología detallada de una ciudad inteligente a partir de datos del Directorio Estadístico Nacional de Unidades Económicas (DENUE) del INEGI, los cuales son datos poco estudiados. A continuación, se comparan estos hallazgos con otros trabajos similares en el campo de las ciudades inteligentes.

Estandarización y Tratamiento de Datos

- Hallazgos Propios:
  - Se aplicó un proceso de estandarización, detección y eliminación de duplicados, así como la gestión de datos faltantes, como se muestra en la Figura 1 y Figura 2.
  - La utilización de mapas de calor de datos faltantes permitió identificar áreas con carencias de información como el tipo de vialidad o lugar de asentamiento de los comercios descritos en la base de datos.
- Trabajos Anteriores:
  - Varios estudios en el ámbito de las ciudades inteligentes también han enfatizado la importancia de la estandarización de datos para asegurar la calidad y la interoperabilidad de los datos. (Ageed et al., 2021).
  - En comparación, nuestro enfoque específico en la visualización mediante mapas de calor proporciona un beneficio en la identificación rápida de problemas de calidad de datos, lo cual no siempre está presente en otros trabajos.

Construcción de la Ontología

- Hallazgos Propios:
  - Una ontología inicial propuesta se construyó a partir de la observación de los datos, y se validó mediante una nube de palabras resultante del procesamiento de las actividades municipales, en donde se obtuvieron las principales actividades como agricultura, comercio, joyería, transporte, servicios de alimentos, servicios de salud y educación, entre otros.
  - La figura 7 muestra la ontología final utilizando los datos procesados de la metodología propuesta, que incluye 12 dimensiones relevantes para el caso de una ciudad inteligente en México, las cuales se categorizan como: industria, economía, salud, alimentación, comunicación, cultura, medioambiente, movilidad y transporte, ciudadanía, vivienda, gobernanza y educación.
- Trabajos Anteriores:
  - Investigaciones previas han demostrado la utilidad de las ontologías para modelar conocimientos en dominios específicos como ciudades inteligentes, destacando su aplicación en la planificación urbana y la gestión de servicios públicos. (De Nicola & Villani, 2021).
  - La ontología propuesta es específica para el contexto mexicano y refleja particularidades culturales y económicas que pueden no estar presentes en ontologías desarrolladas en otros estudios.

Procesamiento de Lenguaje Natural y *Word Embedding*

- Hallazgos Propios:
  - Se utilizaron técnicas de PLN para filtrar y procesar las actividades municipales, empleando Word2Vec para crear representaciones vectoriales de las palabras.

- La generación de clústeres mediante *K-medias*, como se muestra en la Figura 5, permitió identificar agrupamientos naturales en los datos sin necesidad de categorización previa.
- El enfoque propuesto destaca por su aplicación específica en el contexto de las ciudades inteligentes y por la generación de clústers que posteriormente se utilizaron para desarrollar una ontología robusta y adaptada al contexto local.

- Trabajos Anteriores:
  - Estudios anteriores han aplicado técnicas de PLN y *word embeddings* para la clasificación y análisis de datos textuales tal como la detección de eventos y la categorización de documentos. (Rahayu et al., 2022).
  - Otros trabajos proponen ontologías que resultan fijas, estáticas, a partir de experiencia en el dominio que podrían depender de los autores para su correspondiente actualización (incorporación de nuevas variables como nuevos negocios y nuevas tecnologías que inciden en la descripción de una ciudad.

Beneficios y Aplicaciones de la Ontología

- Hallazgos Propios:
  - La ontología final ofrece una estructura clara y definida de los datos, facilitando el análisis y la toma de decisiones en el contexto de las ciudades inteligentes.
  - El uso de sistemas de razonamiento basados en la ontología permite desarrollar agentes inteligentes y optimizar la gestión de los servicios urbanos.
- Trabajos Anteriores:
  - La aplicación de ontologías en ciudades inteligentes ha demostrado mejorar la interoperabilidad entre diferentes sistemas y facilitar la integración de datos provenientes de diversas fuentes.
  - Nuestra investigación se alinea con estos beneficios, pero añade valor al enfocarse en un contexto local y específico, aportando *insights* potencialmente aplicables a la planificación urbana y la gestión de servicios en municipios de México.

5.1. Controversias y Desacuerdos

Contextualización y Escalabilidad: Mientras que algunos estudios argumentan que las ontologías deben ser lo más generales posibles para garantizar su aplicabilidad en múltiples contextos, nuestro enfoque muestra que una ontología específica y adaptada al contexto local puede ser más eficaz para ciertos propósitos. Esta especificidad puede no ser compartida por todos los investigadores, quienes podrían preferir una aproximación más universal.

Técnicas de PLN: Existen debates sobre las mejores técnicas de PLN para procesar grandes volúmenes de datos textuales. Aunque Word2Vec ha demostrado ser efectivo en nuestro

estudio, otros métodos como BERT o GPT-3 podrían ofrecer resultados diferentes y potencialmente más precisos, aunque a un costo computacional más alto.

6. Conclusiones

Nuestro trabajo demuestra la eficacia de combinar técnicas de estandarización de datos, procesamiento de lenguaje natural y clustering para desarrollar una ontología detallada y útil para el contexto de las ciudades inteligentes en México. Comparado con otros estudios, nuestra aproximación específica y adaptada al contexto local ofrece *insights* valiosos y aplicaciones prácticas que pueden mejorar significativamente la gestión urbana y la toma de decisiones en entornos urbanos complejos.

7. Agradecimientos

Los autores de este artículo agradecen al Tecnológico Nacional de México (TecNM) y al Consejo Nacional de Humanidades Ciencias y Tecnologías por su apoyo y recursos proporcionados para llevar a cabo esta investigación. Además, expresamos nuestro agradecimiento a todas las personas involucradas en la recolección y procesamiento de datos que hicieron posible este estudio. También queremos reconocer el invaluable apoyo brindado por nuestros colegas y asesores, cuyas contribuciones y orientaciones fueron fundamentales para el desarrollo de este trabajo. Por último, extendemos nuestro agradecimiento a la comunidad académica y científica por su continua inspiración y motivación en la búsqueda de conocimiento y soluciones innovadoras.

8. Referencias bibliográficas

Alti A, Lakehal, A, Laborie S, Roose P. (2016). Autonomic Semantic-Based Context-Aware Platform for Mobile Applications in Pervasive Environments. *Future Internet*, *8*(4), 48. 10.3390/fi8040048

Ageed ZS, Zeebaree SRM, Sadeeq MM, Kak SF, Rashid ZN, Salih A A, Abdullah WM. (2021). A Survey of Data Mining Implementation in Smart City Applications. *Qubahan Academic Journal*, *1*(2), 91-99. https://doi.org/10.48161/qaj.v1n2a52

Chourabi H, Nam T, Walker S, Gil-Garcia JR, Mellouli S, Nahon K, Pardo TA, Scholl HJ. (2012). Understanding Smart Cities: An Integrative Framework. *2012 45th Hawaii International Conference on System Sciences*, 2289-2297. 10.1109/HICSS.2012.615

De Nicola A, Villani M L. (2021). Smart City Ontologies and Their Applications: A Systematic Literature Review. *Sustainability*, *13*(10). 10.3390/su13105578

Djenouri Y, Belhadi H, Akli-Astouati K, Cano A, Chun-Wei Lin J. (2022). An ontology matching approach for semantic modeling: A case study in smart cities. *Computational Intelligence*, *38*(3), 876-902. https://doi.org/10.1111/coin.12474

Instituto Nacional de Estadística y Geografía (INEGI) (2023). Directorio Estadístico Nacional de Unidades Económicas: DENUE Interactivo 11/2023: documento metodológico / Instituto Nacional de Estadística y Geografía.—México, INEGI, c2023.

Jalolov TS. (2023, November 27). Teaching The Basics Of Python Programming. *International Multidisciplinary Journal for Research & Development*, *10*(11). https://www.ijmrd.in/index.php/imjrd/article/view/443

Karani D. (2018, September 01). Introduction to word embedding and word2vec. *Towards Data Science*, *1*. https://towardsdatascience.com/introduction-to-word-embedding-and-word2vec652d0c2060fa

Malkawi O, Obaid N, Almobaideen W. (2022). Toward an Ontological Cyberattack Framework to Secure Smart Cities with Machine Learning Support. *International Journal of Advanced Computer Science and Applications*, *13*(11). 10.14569/IJACSA.2022.0131145

Na S, Xumin L, Yong G. (2010). Research on k-means Clustering Algorithm: An Improved k-means Clustering Algorithm. *2010 Third International Symposium on Intelligent Information Technology and Security Informatics*, 63-67. 10.1109/IITSI.2010.74

Rahayu NW, Ferdiana R, Kusumawardani SS. (2022). A systematic review of ontology use in E-Learning recommender system. *Computers and Education: Artificial Intelligence*, *3*, 100047. https://doi.org/10.1016/j.caeai.2022.100047