UNIVERSIDAD EUROPEA DE MADRID

ESCUELA DE ARQUITECTURA, INGENIERÍA Y DISEÑO

MÁSTER UNIVERSITARIO EN ANÁLISIS DE DATOS MASIVOS

TRABAJO FIN DE MÁSTER

ANÁLISIS DE LA CORRESPONDENCIA ENTRE LAS INICIATIVAS COMUNITARIAS PDET Y LA INVERSIÓN PÚBLICA EN EL MARCO DEL ACUERDO DE PAZ EN COLOMBIA: UN ENFOQUE BASADO EN CLUSTERING DIFUSO

CARLOS ANDRES BALLESTEROS CASTAÑEDA

Dirigido por

Jose Manuel Lopez Lopez

CURSO 2024 - 2025

TÍTULO: ANÁLISIS DE LA CORRESPONDENCIA ENTRE LAS INICIATIVAS COMUNITARIAS PDET Y LA INVERSIÓN PÚBLICA EN EL MARCO DEL ACUERDO DE PAZ EN COLOMBIA: UN ENFOQUE BASADO EN CLUSTERING DIFUSO

AUTOR: CARLOS ANDRES BALLESTEROS CASTAÑEDA

TITULACIÓN: MÁSTER UNIVERSITARIO EN ANÁLISIS DE DATOS MASIVOS

DIRECTOR DEL PROYECTO: Jose Manuel Lopez Lopez

FECHA: [JULIO] de 2025

RESUMEN

Este Trabajo de Fin de Máster aborda el reto de identificar correspondencias temáticas entre las iniciativas comunitarias formuladas en el marco de los Programas de Desarrollo con Enfoque Territorial (PDET) y los proyectos de inversión pública registrados a nivel nacional en Colombia. El objetivo principal fue analizar si las demandas expresadas por las comunidades han sido atendidas por la acción institucional, a través de herramientas de análisis semántico y agrupamiento no supervisado.

El proyecto se desarrolló de manera independiente, sin colaboración directa con empresas. Se utilizaron tecnologías de procesamiento de lenguaje natural y modelos de representación vectorial, como Word2Vec y TF-IDF, para calcular similitudes entre textos de iniciativas y proyectos, posteriormente, se aplicó el algoritmo de clustering difuso Fuzzy C-Means, permitiendo identificar tipologías temáticas emergentes.

Entre los principales resultados se destaca la construcción de un modelo capaz de emparejar iniciativas y proyectos en función de su proximidad semántica, así como la identificación de patrones de atención institucional diferenciada por tema y territorio. Demostrando el potencial de los datos al momento de contruir política pública fortaleciendo la planificación participativa y mejorar la alineación entre las necesidades de la comunidad y la inversión pública.

Palabras clave: procesamiento de lenguaje natural, clustering difuso, Word2Vec, inversión pública, análisis semántico, PDET.

ABSTRACT

This Master's Thesis explores the challenge of identifying thematic correspondences between community-driven initiatives formulated within the framework of the Development Programs with a Territorial Focus (PDET) and public investment projects registered at the national level in Colombia. The main objective was to analyze whether the demands expressed by local communities have been addressed by institutional action, using semantic analysis tools and unsupervised clustering techniques.

The project was developed independently, without direct collaboration with companies. Natural language processing technologies and vector representation models, such as Word2Vec and TF-IDF, were used to calculate semantic similarities between initiative and project texts. Subsequently, the Fuzzy C-Means clustering algorithm was applied to identify emergent thematic typologies.

Among the main outcomes is the construction of a model capable of matching initiatives and projects based on semantic proximity, as well as the identification of differentiated patterns of institutional response according to topic and territory. This work demonstrates the potential of data-driven approaches to strengthen participatory planning and improve the alignment between community needs and public investment.

Keywords: natural language processing, fuzzy clustering, Word2Vec, public investment, semantic analysis, PDET.

Índice general

Índice de Figuras

Índice de Tablas

Lista de Acrónimos

Sigla	Significado
ART	Agencia de Renovación del Territorio
DANE	Departamento Administrativo Nacional de Estadística
DNP	Departamento Nacional de Planeación
FARC	Fuerzas Armadas Revolucionarias de Colombia
FCM	Fuzzy C-Means
MGA	Metodología General Ajustada
PATR	Planes de Acción para la Transformación Regional
PDET	Programas de Desarrollo con Enfoque Territorial
PISDA	Planes integrales comunitarios y municipales de sustitución y desarrollo alternativo
RRI	Reforma Rural Integral
TF-IDF	Term Frequency – Inverse Document Frequency
TFM	Trabajo Final de Máster
AWS	Amazon Web Services

Tabla 1. Lista de acrónimos utilizados en el documento.

Capítulo 1. INTRODUCCIÓN

1.1 Contexto y justificación

Colombia, ha estado inmersa en diversas situaciones de conflicto armado interno, economías ilegales y diferentes formas de violencia que han afectado a su población, sus instituciones e incluso su relacionamiento con otros países de la región. Con el fin de lograr la solución al conflicto interno, diferentes gobiernos han tratado e iniciado diálogos de consenso con los actores armados, presentes en el país. Por lo anterior el gobierno del presidente Juan Manuel Santos Calderón (2010 - 2018), en febrero del año 2012; inició la fase de diálogos exploratorios para iniciar un proceso de paz, con el grupo armado ilegal denominado: Fuerzas Armadas Revolucionarias – (FARC).

La cronología más relevante del proceso de paz es la siguiente [1]:

- **23 de febrero de 2012:** Comienza la fase exploratoria entre el Gobierno de Juan Manuel Santos y la extinta guerrilla de las FARC.
- **26 de agosto de 2012:** Se firma el acuerdo general que sienta las bases para la negociación.
- **18 de octubre de 2012:** Se instala la Mesa de Diálogo de Paz en Oslo, Noruega, y se dan a conocer los nombres de los negociadores por cada una de las partes.
- **26 de septiembre de 2016:** Firma oficial del acuerdo de paz por parte del presidente de Colombia, Juan Manuel Santos, y el líder de las extintas FARC, Rodrigo Londoño, alias Timochenko.

El suscrito acuerdo de paz estableció, la implementación de 5 grandes puntos, entre otros aspectos relevantes, así [2]:

Figura 1.1. Puntos del acuerdo de paz. Fuente: La ciencia de la justicia [3].

Los anteriores puntos se describen a continuación:

1. **Reforma rural integral:** se orienta a erradicar la pobreza, promover la igualdad y asegurar el pleno disfrute de derechos básicos en el campo.

2. **Participación política:** se trata de fortalecer la integración política de las regiones y de asegurar la representación y promoción de los intereses de sus pobladores. Y se trata

también de una medida de reparación política para las poblaciones y comunidades que más han sufrido el conflicto.

3. **Fin del conflicto:** este acuerdo estableció los términos en que se dará el fin de las confrontaciones con las FARC mediante un cese al fuego y de hostilidades bilateral y definitivo.
4. **Solución al problema de drogas:** propone una nueva visión para abordar el problema de las drogas ilícitas, diferenciando entre el consumo, los cultivos de uso ilícito y la criminalidad organizada asociada al narcotráfico. Se asegura un enfoque de derechos humanos y salud pública, con programas de sustitución de cultivos y fortalecimiento de la lucha contra las finanzas ilícitas.
5. **Víctimas del conflicto:** establece el Sistema Integral de Verdad, Justicia, Reparación y No Repetición, que combina mecanismos judiciales y extrajudiciales para esclarecer la verdad, buscar a las personas desaparecidas y reparar a las víctimas del conflicto. Este sistema incluye la Comisión para el Esclarecimiento de la Verdad, la Unidad Especial para la Búsqueda de Personas Desaparecidas y la Jurisdicción Especial para la Paz.

Para la finalización del año 2016, el gobierno colombiano, inició con la implementación del acuerdo de paz suscrito con las extintas FARC, por lo anterior se requirió de un marco institucional y administrativo que llevará a cabo el despliegue de diferentes programas y proyectos que permitieran la implementación y el cumplimiento de lo establecido en el referido acuerdo de paz.

Para ello, se creó una entidad pública del orden nacional, denominada: Agencia de Renovación del Territorio – ART-, cuya misión principal se centraba en la implementación del punto No 1 Reforma rural integral del acuerdo de paz.

Para la implementación de lo pactado, en el punto No 1, se crearon los denominados Programas de Desarrollo con Enfoque Territorial – PDET-, los cuáles son confluyentes en términos de política pública para enfoques territoriales, poblaciones y de reparación a las víctimas del conflicto armado.

Así entonces, los PDET, son un Instrumento especial de planificación y gestión para implementar la Reforma Rural Integral – RRI en 170 municipios priorizados por los siguientes criterios [4]:

- Afectación del conflicto
- Niveles de pobreza
- Debilidad institucional
- Presencia de economías ilícitas

Los PDET son instrumentos participativos que recogen las demandas comunitarias de las poblaciones, asociaciones y diferentes actores que habitan y confluyen en los territorios más apartados y que han sido históricamente afectados por el conflicto armado.

Para la construcción de los PDET, se definió una metodología participativa de amplio espectro que convocaba a la acción de participación a todos los actores territoriales. Dicha metodo-

logía se implementó bajo un enfoque *Bottom-Up*, que iniciaba desde el nivel más bajo de la estructura territorial del país e iba ascendiendo hasta llegar a instancias superiores. Partiendo del nivel veredal (mínima unidad regional en el Estado colombiano), pasando por el nivel municipal (entidad territorial fundamental de la división político-administrativa del Estado colombiano con autonomía política, fiscal y administrativa) y llegando al nivel subregional (unidad de agregación de diferentes municipios en el Estado colombiano). Esta misma forma de agrupación territorial se plasmaba en una agregación de información, la cual debía ser consolidada e ir transitando por los diferentes niveles de formulación participativa del PDET.

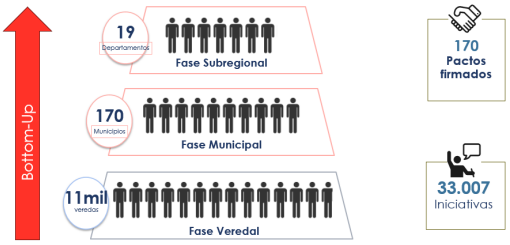


Figura 1.2. Metodología participativa PDET. Fuente: adaptado de Agencia de Renovación del Territorio [5].

Al final del ejercicio participativo, se obtuvieron los siguientes resultados [6]:

- 170 pactos municipales para la transformación regional.
- 16 Planes de Acción para la Transformación Regional- PATR, para los cuáles su ejecución será a 10 años, más 5 años de transición y sostenibilidad.
- 33.007 iniciativas o demandas comunitarias, para realizar su implementación, las cuales obedecen a las demandas territoriales de la tercera parte del territorio nacional donde residen más de 6 millones de colombianos.

Toda esta información fue recolectada y procesada, en términos de narrativa o textos, expresados por las comunidades y en donde, estas, planteaban la solución a sus necesidades territoriales, mediante la definición de iniciativas con el propósito de ser implementadas por el gobierno nacional en sus territorios.

Por otro lado, el gobierno nacional en su función constitucional y de planeación presupuestal; realiza diferentes inversiones en el ámbito territorial tomando como base diferentes mecanismos de planeación, no necesariamente articulados con los denominados PDET. En este sentido, el gobierno nacional busca cerrar las brechas de desigualdad en todo el territorio colombiano, y utiliza instrumentos de planeación como:

- Plan nacional de desarrollo: Política nacional.
- Planes nacionales sectoriales: Políticas nacionales.
- Plan municipal de desarrollo: Necesidades municipales.
- Plan departamental de desarrollo: Necesidades departamentales.
- Focalización de los sectores públicos en Colombia: Necesidad sectoriales. Entre otros.

Si bien el gobierno nacional, realiza un esfuerzo para garantizar que los instrumentos anteriormente descritos, tengan una alineación real con los PDET, estos no necesariamente están siendo formulados y estructurados con el insumo de las iniciativas determinadas por

los actores comunitarios en los procesos participativos. Estos proyectos, los definidos por el gobierno nacional, son formulados y estructurados bajo la Metodología General Ajustada (MGA), desarrollada por el Departamento Nacional de Planeación (DNP), garantizando desde la misma formulación del proyecto la definición de un alcance, objetivo y beneficiarios de manera clara[7].

Lo anterior, ha denotado la estructuración de dos conjuntos de planeación territorial diferentes, uno centrado en el deber ser del estado social de derecho y con enfoque desde el centralismo y otro definido mediante demandas comunitarias y con enfoque en las necesidades territoriales.

De igual manera, en términos de la información requerida para la estructuración de los proyectos a implementar, se denotan los mismo dos universos de información, los cuales son disimiles en términos de la agregación de datos y de su misma estructura; una orientada en clave de proyectos y otra orientada en clave de narrativa para soluciones a demandas comunitarias.

Esto, ha generado, reparos al gobierno nacional, con la argumentación en el no cumplimiento del referido acuerdo de paz, por no considerar las iniciativas comunitarias en la formulación de los proyectos.

Por lo expuesto, la presente propuesta de trabajo final de master permitirá identificar los esfuerzos que desde el gobierno nacional se han realizado y se están realizando a través de proyectos de inversión pública para el cumplimiento de las iniciativas o demandas comunitarias, contenidas en los PDET.

En tal sentido, el propósito del presente trabajo académico se centra en buscar los grados de pertenencia y correlación de los proyectos formulados por el gobierno nacional y las iniciativas o demandas comunitarias contenidas en los PDET, para que basados en la evidencia lograr concluir si efectivamente o no, existen avances en la implementación de los PDET.

Es importante resaltar que lo acá contenido, no pretender cambiar o cuestionar los diferentes procesos de planeación, su disposición de recursos, su formulación en términos de proyecto y su respectiva implementación. Para el anterior propósito, se plantea una identificación de similitud entre las iniciativas y los proyectos, utilizando técnicas de procesamiento de lenguaje natural (NLP) y algoritmos de clustering o agrupamiento.

1.2 Planteamiento del problema

Las diferentes formas e instrumentos de planeación definidos y ejecutados desde el gobierno nacional, para la implementación del punto No1 del acuerdo de paz, han generado silos de información que no permiten con claridad generar métricas en la implementación del referido punto No1 y que al mismo tiempo causan deterioro de la confianza por la posible no incorporación de las iniciativas o demandas comunitarias en los proyectos formulados desde el nivel nacional. Lo anterior se traslada directamente a la entidad encargada de la implementación del punto No1, la cual como ya se ha mencionado, es la Agencia de Renovación del Territorio – ART-, generando situaciones como las siguientes:

- Dificultad para evaluar el cumplimiento del Acuerdo de Paz en especial en el punto 1, correspondiente a la reforma rural integral.
- Desconexión entre las iniciativas PDET y los proyectos de inversión pública, derivándose en inversiones no focalizadas sin atender las demandas comunitarias.
- Poca eficiencia en la toma de decisiones, debido a la manualidad de los procesos de vinculación.

Por lo anterior el problema que se plantea solucionar con el presente TFM, es diseñar y desarrollar un modelo analítico automatizado que permita identificar patrones de pertenencia entre las iniciativas comunitarias y los proyectos gubernamentales. Este análisis permitirá medir el cumplimiento del gobierno colombiano respecto a las demandas comunitarias establecidas en el Acuerdo de Paz de 2016, contribuyendo a la evaluación de la efectividad de los PDET y proporcionando información clave para la mejora de su implementación.

Buscando dar respuesta a la siguiente pregunta:

¿En qué medida la inversión pública realizada en los territorios refleja o responde a las prioridades expresadas por las comunidades?

El TFM se ha definido para investigar y resolver un problema específico en el ámbito científico-técnico. Se propone el uso de algoritmos de clustering para identificar el grado de pertenencia entre los proyectos de inversión pública y las iniciativas PDET.

1.3 Objetivos del proyecto

Objetivo general

Diseñar e implementar un modelo predictivo, utilizando algoritmos de clustering para vincular proyectos de inversión pública en Colombia con las demandas territoriales o iniciativas comunitarias incluidas en los PDET.

Objetivos específicos

1. Analizar y documentar los diferentes esquemas de datos y fuentes de información relevantes para el estudio.
2. Establecer y definir el modelo o modelos aplicables, de acuerdo con la problemática identificada con el fin de automatizar el proceso de vinculación de iniciativas PDET a proyectos.
3. Desarrollar y entrenar el modelo analítico que permita automatizar el proceso de vinculación de iniciativas PDET a proyectos.
4. Evaluar los resultados obtenidos de la implementación del modelo analítico.

1.4 Motivación para el desarrollo del proyecto

La motivación para el desarrollo de este Trabajo Final de Master por un lado responde a un compromiso social con las comunidades ubicadas en los territorios con más desigualdad en Colombia, los cuales han sufrido por más de 20 años las consecuencias del conflicto armado.

Muchas de estas comunidades participaron en la implementación de los Acuerdos de Paz, principalmente en la construcción de los Programas de Desarrollo con Enfoque Territorial – PDET, confiando en que ese esfuerzo colectivo se traduciría en transformaciones territoriales reales y tangibles en sus territorios. Sin embargo, a pesar de el esfuerzo del gobierno nacional en cuanto a inversión a través de proyectos, el sentimiento de estas comunidades es de desesperanza no solo frente al estado sino al mismo proceso de paz que ayudaron a construir.

Por otra parte, durante mi experiencia profesional como servidor público he visto en ocasiones que el diseño y la implementación de políticas públicas están condicionadas por intereses particulares o decisiones intuitivas, más que por un análisis riguroso de datos. Soy un convencido que las decisiones basadas en los datos permitirán realizar intervenciones en los territorios de manera más efectivas y acordes a lo que requieren las comunidades.

Capítulo 2. ANTECEDENTES / ESTADO DEL ARTE

Para una mejor comprensión de la problemática presentada y con el fin de identificar posibles soluciones a esta, se hace necesario analizar con detalle el contenido y las características de las fuentes de datos que serán objeto de estudio del presente TFM, el proceso de vinculación actual de las iniciativas PDET y de proyectos de inversión pública y una revisión de enfoques y algoritmos computacionales que han sido aplicados en contextos similares.

Los resultados presentados en este Capítulo, fueron desarrollados mediante scripts de Python, utilizando Jupiter Notebook como entorno de trabajo. El detalle del código fuente, así como la estructura del notebook se encuentra disponible en el Anexo 6.1 del presente TFM.

2.1 Fuentes de datos

El presente Trabajo de Fin de Máster (TFM) se fundamenta en el acceso y análisis de información catalogada como pública, de conformidad con el ordenamiento jurídico colombiano. Dicha información es recopilada a partir de fuentes oficiales abiertas, siendo cada entidad pública responsable de su producción, custodia y publicación.

En Colombia, el derecho de acceso a la información pública es un derecho fundamental, consagrado en el Artículo 74 de la Constitución Política, el cual establece que “todas las personas tienen derecho a acceder a los documentos públicos salvo los casos que establezca la ley” [8]. Este principio se desarrolla y reglamenta en la Ley 1712 de 2014, conocida como la Ley de Transparencia y del Derecho de Acceso a la Información Pública Nacional [9], que establece que la información pública debe estar disponible para toda persona, sin discriminación ni necesidad de justificación, garantizando su consulta libre y gratuita.

2.1.1. Iniciativas PDET

Como se ha indicado con anterioridad, las iniciativas PDET constituyen las necesidades comunitarias definidas de forma participativa en los 170 municipios priorizados por el acuerdo de paz, las cuales fueron agrupadas temáticamente en ocho pilares.

- 1. Ordenamiento social de la propiedad rural y uso del suelo.
- 2. Infraestructura y adecuación de tierras.
- 3. Salud rural.
- 4. Educación rural y primera infancia rural.
- 5. Vivienda rural, agua potable y saneamiento básico rural.
- 6. Reactivación económica y producción agropecuaria.
- 7. Sistema para la garantía progresiva del derecho a la alimentación.
- 8. Reconciliación, convivencia y construcción de paz.

Estas iniciativas están disponibles públicamente en el portal de información de la Agencia de Renovación del Territorio[10].

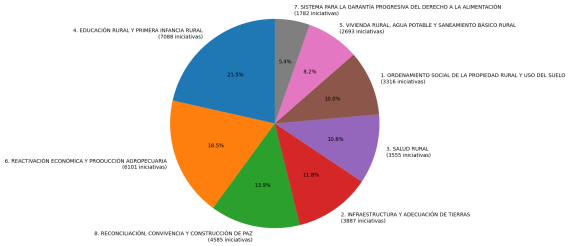


Figura 2.1. Distribución iniciativas PDET por pilar (elaboración propia).

Diccionario de datos

Es importante resaltar que, al momento de la definición de las iniciativas por parte de la comunidad, se definieron una serie de atributos que pueden ser relevantes en la identificación de patrones y compresión de las temáticas de estas. De igual forma, resulta pertinente indicar que se trabajará con un dataset de 33.007 iniciativas.

Campo/Atributo	Descripción
Subregión	Nombre de la Subregión PDET. Ubicación territorial de las iniciativas, siguiendo las divisiones territoriales definidas en el acuerdo de paz.
Municipio/Sujeto Concertación	Nombre del municipio PDET o Sujeto de Concertación en el cual se definió la iniciativa. Para iniciativas del orden Subregional este atributo es NULL.
Código DANE	Codificación de los municipios según el DANE. Para iniciativas Subregionales o sujetos de concertación, el valor es NULL.
Código iniciativa	Código único de la iniciativa PDET.
Pilar	Pilar o temática asociada a la iniciativa.
Título de la iniciativa	Título en lenguaje natural que indica el nombre de la iniciativa PDET.
Descripción de la iniciativa	Texto en lenguaje natural que detalla el propósito de la iniciativa.
Estrategia	Línea de acción dentro del pilar, imputada a juicio de expertos.
Código de la categoría	Código asociado a la categoría de la iniciativa.
Categoría	Subdivisión dentro de una estrategia, definida a juicio de expertos.
Número producto	Número o código del producto asociado a la iniciativa.
Producto	Describe lo que se espera entregar, construir o ejecutar dentro de la iniciativa.
Proyecto/Gestión	Tipo de acción esperada (proyecto o gestión).
Clasificación iniciativa	Indica si la iniciativa es étnica, no étnica o común étnica.
Veredas	Nombre de las veredas cubiertas por la iniciativa.
Código Veredas	Código DANE de las veredas asociadas.
Etiquetas	Clasificaciones otorgadas por las comunidades.
PISDA	Indica si está incluida en los planes de sustitución y desarrollo alternativo.
Fase de formulación	Nivel territorial en que fue formulada (municipal o subregional).

Tabla 2.1. Descripción de los campos de la base de datos de iniciativas PDET (elaboración propia).

Análisis exploratorio

Con el objetivo de comprender la estructura y distribución de las iniciativas PDET, se presenta un análisis exploratorio del conjunto de datos.

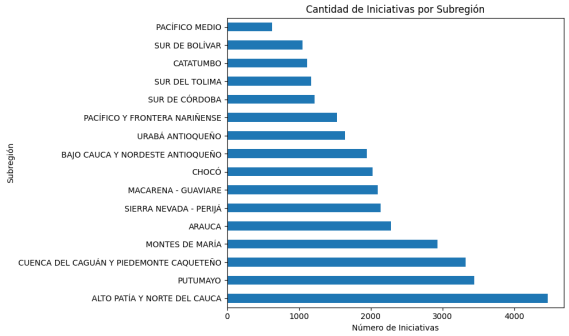


Figura 2.2. Cantidad de iniciativas por Subregión (elaboración propia).

La distribución de cantidad de iniciativas por Subregión muestra un mayor número de iniciativas en la Subregión de Alto Patía y Norte del Cauca con más de 4.000 mil iniciativas, seguidas por Putumayo y la Cuenca del Caguán y Pidemonte Caqueteño. En contraste subregiones como Pacifico Medio, Sur de Bolivar y Catatumbo registran iniciativas en menor proporción, esta variabilidad puede obedecer a la cantidad de municipios en la subregión, lo cual tiene una relación directa con la cantidad de demandas comunitarias (iniciativas) en cada uno de los municipios que conforman la subregión.

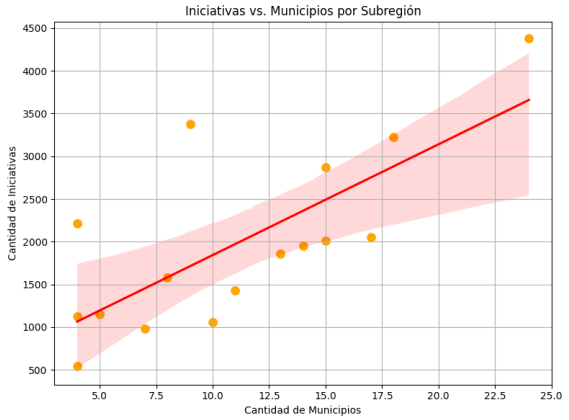


Figura 2.3. Relación número de municipios y la cantidad de iniciativas PDET por Subregión (elaboración propia).

Con el fin de verificar la observación previamente indicada, se realizó un análisis de correlación entre la cantidad de municipios por subregión y la cantidad de iniciativas de cada subregión, confirmando lo anteriormente expuesto. La subregiones con mayor número de iniciativas tienden a tener un mayor número de municipios de acuerdo con la composición geográfica determinada en el acuerdo de paz. Esto tiene una relación directa con las dinámicas y problemáticas territoriales y en donde a mayor agregación de municipios las carencias sociales, económicas, de violencia y de pobreza se incrementan.

Ahora bien, otro tipo de análisis exploratorio a realizar sobre la fuente de información es el orientado a determinar la distribución de las temáticas (pilares) en las iniciativas PDET por subregión.

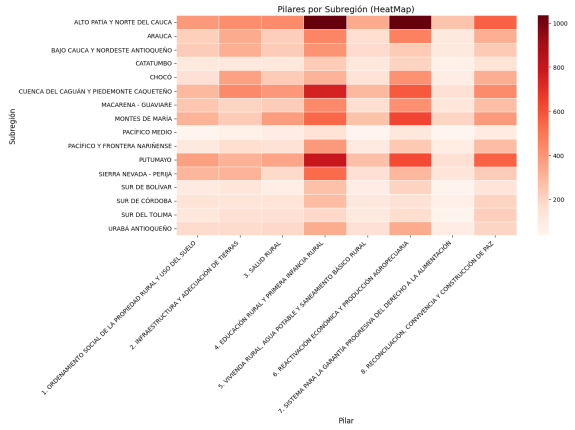
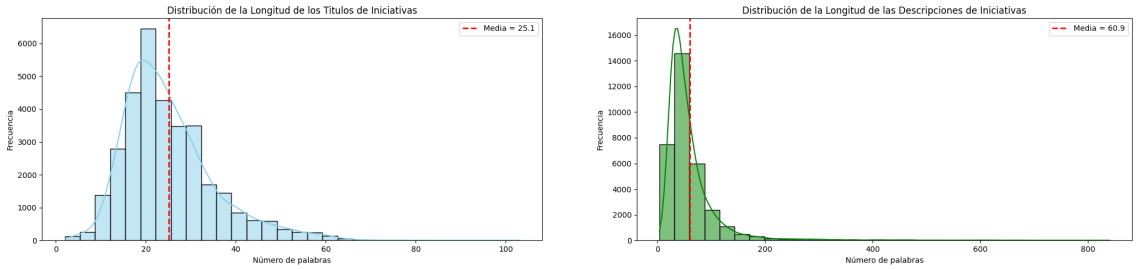


Figura 2.4. Distribución Iniciativas PDET por Pilar y Subregión (elaboración propia).

El Heatmap o mapa de calor evidencia que el pilar con mayor cantidad de iniciativas es el pilar 4. Educación rural y primera infancia rural, siendo predominante en la mayoría de las subregiones, es importante resaltar el pilar 6. Reactivación económica y producción agropecuaria, esto puede interpretarse que la comunidad concibe la reactivación económica como un motor de desarrollo y como una vía para disminuir la dependencia de economías ilícitas predominantes en estos territorio.

De igual manera, otro de los análisis relevantes sobre la fuente de datos, es la volumetría en las dos variables con mayor relevancia para la compresión de las demandas comunitarias en los territorios PDET, estas son: Titulo de la iniciativa y Descripción de estas. Esto obedece a que en estas dos variables se almacenaron en narrativa de la comunidad la expresión de sus necesidades.



Título Descripción

Figura 2.5. Distribución de longitud de los campos de titulo y descripción en las iniciativas PDET (elaboración propia).

Los dos histogramas presentados evidencian distintos patrones en la formulación de las necesidades, primero tenemos los títulos de las iniciativas con una media de 25,1 en su longitud, lo que sugiere una claridad y síntesis de la necesidad de la comunidad. De otra parte, tenemos las descripciones de las iniciativas con una media de 60,9 lo que refleja un mayor nivel de detalle de la necesidad. Partiendo esto, resulta pertinente realizar una nube de palabras asociadas a la descripción de las iniciativas y segmentadas por pilar, eliminando previamente stopwords, para con ello lograr inferir temáticas recurrentes en cada uno de los pilares.

Figura 2.6. Nubes de palabras por pilar de las iniciativas PDET (elaboración propia).

El análisis de nubes de palabras muestra una diversidad semántica en las descripciones de iniciativas agrupadas por pilares, ejemplo en el Pilar 3 (Salud rural) se destacan términos como ’prestación de servicios’ y ’puestos salud’, lo cual coloca de presente las necesidades y carencias enfatizadas por las comunidades en relación con la cobertura en servicios de salud

y una posible insuficiencia en profesionales de la salud.

Para el Pilar 2 (Infraestructura y adecuación de tierras) predominan palabras como 'KM' y 'realizar estudios', 'diseños construcción' con lo cual se puede inferir demandas comunitarias en el sentido de contar con estudios, diseños para la construcción de obras civiles que mejoren la conectividad terrestre de los territorios.

Esta diferenciación sugiere la posibilidad de agrupar iniciativas a través de otra clasificación distinta a la ya aplicada (pilares), y a través de la identificación de similitudes en sus descripciones, lograr encontrar patrones asociados a las demandas comunitarias; las cuales a través de la aplicación de algoritmos de clustering se pudiesen detectar.

2.1.2. Proyectos de Inversión Pública

Diccionario de datos

Los proyectos de inversión pública están disponibles en el portal de información de información Mapa Inversiones del Departamento Nacional de Planeación – DNP [11]. Es importante indicar que se trabajará con un dataset de 68.729 proyectos con corte a 20 de febrero del 2025.

Así mismo, es relevante clarificar que dicha cifra de proyectos varia en aumento debido a la formulación y ejecución de recursos públicos en los territorios.

Campo/Atributo	Descripción
Tipo	Tipo de fuente de financiación del proyecto (SGR, TERRITORIO).
Id	Identificador único del proyecto.
Bpin	Código BPIN del proyecto según el Banco de Proyectos de Inversión del Gobierno Nacional.
CodigoEntidad	Código de la entidad formuladora o ejecutora del proyecto (por ejemplo, código DANE del municipio, departamento o entidad nacional).
Descripcion	Descripción técnica del proyecto.
Nombre	Nombre o título del proyecto.
NombreEntidad	Nombre de la entidad formuladora o ejecutora.
ObjetivoGeneral	Objetivo general del proyecto.
Periodo	Periodo del plan de desarrollo o estrategia a la que está asociado el proyecto.
Problema	Problema identificado que el proyecto busca resolver.
ProgramaPND	Programa del Plan Nacional de Desarrollo al que se asocia el proyecto.
VigenciaFinal	Año de finalización del proyecto.
VigenciaInicio	Año de inicio del proyecto.
TotalProyecto	Valor total del proyecto en pesos colombianos.
Habilitado	Indica si el proyecto está habilitado en el sistema (1: Sí, 0: No).
Validado	Estado de validez del proyecto en su etapa de formulación.
Fase	Fase del proyecto según el ciclo de formulación.
EstadoViabilizado	Estado del proyecto en términos de viabilidad técnica o financiera.
EstadoUnidadProyecto	Estado actual del proyecto dentro de la unidad ejecutora.
Idestadohomologado	Código del estado del proyecto.
EstadoHomologado	Estado del proyecto.
Última Actualización	Fecha de la última actualización registrada del proyecto.

Tabla 2.2. Diccionario de datos de la base de proyectos de inversión pública (elaboración propia).

Análisis exploratorio

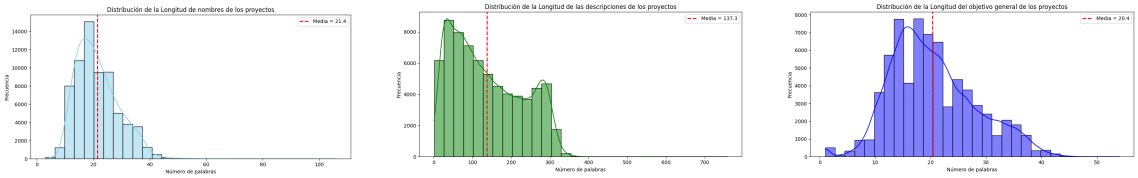
El análisis exploratorio de los proyectos de inversión pública permite comprender la magnitud y diversidad de esta. Los proyectos identificados registran un rango de vigencia de estos (año de inicio, año de finalización), así como montos de variables de recursos invertidos. Los proyectos se encuentran en distintos estados de avance, desde la estructuración, ejecución, y terminación. No obstante, es importante indicar que el alcance de este TFM se limita al análisis de alineación semántica y temática entre las iniciativas PDET y los proyectos, independientemente de su estado actual o nivel de ejecución financiera.

Debido a que no existe en la fuente de datos un segmentación o temática asociada al proyecto de manera específica, se realiza una nube de palabras con el fin de identificar rápidamente conceptos recurrentes en la formulación de proyectos.

Figura 2.7. Nube de palabras proyectos de inversión pública (elaboración propia).

Si bien se identifican temáticas que podrían coincidir con las demandas comunitarias expresadas en las iniciativas PDET -como instituciones educativas, salud pública, entre otras-, sin una segmentación previa o agrupación temática, no es posible establecer una relación clara y precisa entre ambas fuentes de información.

Al igual que con las iniciativas PDET, un análisis relevante de la fuente de datos es el asociado a la volumetría en las tres variables con mayor relevancia para la compresión de la necesidad que desea solucionar con el proyecto de inversión, estas son: Nombre del proyecto, Descripción del proyecto y Objetivo general.



Nombre *Descripción* *Objetivo*

Figura 2.8. Distribución de longitud de los campos de nombre, descripción y objetivo de los Proyectos de Inversión Pública en Colombia (elaboración propia).

Los resultados de los histogramas permiten observar que no todos los campos textuales potenciales para el análisis de similitud de la base de proyectos de inversión pública aportan el mismo nivel de detalle del proyecto. La descripción es el campo con mayor potencial analítico, dado por su extensión —media de 137 palabras—. En contraste, el nombre del proyecto y el objetivo general presentan extensiones similares, alrededor de 20 palabras, lo cual limitaría un ejercicio de identificación de patrones o agrupamientos de información.

2.2 Mecanismo actual de vinculación entre iniciativas PDET y proyectos de inversión pública

Actualmente, el proceso mediante el cual se vinculan las iniciativas comunitarias priorizadas en los Programas de Desarrollo con Enfoque Territorial (PDET) con los proyectos de inversión pública sigue siendo mayoritariamente manual. Esta tarea recae en equipos técnicos o expertos temáticos, quienes, a partir de la revisión de los documentos de formulación de los proyectos, determinan de forma cualitativa si existe o no correspondencia con alguna de las iniciativas incluidas en los Planes de Acción para la Transformación Regional (PATR).

Este procedimiento está formalizado mediante la Resolución Número 000141 de 2024[12], en la cual se establece que la identificación de la correspondencia entre las iniciativas PDET y los proyectos de inversión pública es responsabilidad de la entidad formuladora. Esta debe realizar un análisis detallado del contenido del proyecto y justificar manualmente su relación con una o varias iniciativas priorizadas, sustentando dicha correspondencia a través de una matriz técnica. Actualmente, no existen herramientas automatizadas que faciliten este cruce de forma sistemática, por lo que el proceso depende en gran medida del criterio técnico, la experiencia del equipo formulador y la interpretación del evaluador designado por la ART.

Aunque el procedimiento cuenta con formatos y lineamientos definidos, no se dispone de sistemas que faciliten la vinculación masiva entre proyectos y las más de 33.000 iniciativas priorizadas en los territorios PDET. Esta limitación genera retos importantes en términos de trazabilidad, consistencia y eficiencia. Además, al tratarse de un proceso basado en juicio técnico, se introduce inevitablemente un grado de subjetividad. La forma en que se interpreta el alcance del proyecto, su pertinencia temática o su adecuación territorial puede variar entre profesionales, generando resultados dispares para casos similares. A esto se suma el tiempo y esfuerzo que requiere cada análisis individual, lo cual restringe la escalabilidad del proceso y dificulta su aplicación en contextos donde se necesita acelerar la ejecución de proyectos.

Más allá de automatizar el procedimiento actual, este TFM plantea una alternativa metodológica basada en técnicas de aprendizaje automático. En principio, podría considerarse un enfoque supervisado, aprovechando los ejemplos previos de vinculación realizados por expertos como base para entrenar modelos con datos etiquetados. Sin embargo, este enfoque podría replicar los mismos sesgos y omisiones que se derivan del proceso manual.

Por esa razón, se optó por explorar un enfoque no supervisado, particularmente a través del uso de algoritmos de agrupamiento difuso (fuzzy clustering). El objetivo no es reemplazar el juicio técnico, sino complementarlo con una herramienta que permita identificar patrones en los datos que podrían pasar desapercibidos en el modelo actual. A diferencia de los algoritmos supervisados, los modelos no supervisados no requieren etiquetas previas y operan a partir de la estructura misma de los datos para identificar agrupamientos con características comunes.

El uso de fuzzy clustering ofrece, además, una ventaja adicional, permite que un proyecto esté asociado simultáneamente a más de una iniciativa PDET, reflejando la complejidad temática y sectorial que caracteriza a los proyectos de inversión pública en Colombia. Esto último resulta útil en proyectos que no responden a una única necesidad o línea de acción, sino que abordan de manera trasversal diversas problemáticas, implementando acciones integrales en el territorio.

2.3 Experiencias previas y abordaje metodológico al proyecto

Para el desarrollo del presente TFM, es importante considerar tanto un enfoque técnico como la revisión de experiencias previas en la aplicación de algoritmos de clustering no supervisado a problemáticas similares. Estas experiencias permiten contextualizar y fundamentar conceptualmente el trabajo, evidenciando el potencial que tienen estas técnicas para abordar problemas asociados al análisis de datos.

A continuación, se presentan estudios que sirven como antecedentes metodológicos para la problemática a resolver.

- "An Unsupervised Narrative Detection Framework to Demystify Online Protest Composition": Propone un estudio para detectar narrativas en protestas en línea a partir de datos de Twitter, utilizando cluster no supervisados como parte central de la metodología[13].
- "A Fuzzy Based Approach To Text Mining And Document Clustering": El estudio utiliza el algoritmo Fuzzy C-Means para clasificar documentos en dos categorías, indicando claramente los pasos que se siguieron para la obtención del resultado[14].

El clustering difuso es tipo de algoritmo que permite que un mismo dato pertenezca a múltiples clústeres con diferente grado de pertenencia. Esto es particularmente útil cuando los límites entre las categorías no son claramente definidos, que para nuestro caso en particular aplica dado la naturaleza de las iniciativas PDET y proyectos de inversión.

A continuación, se establecen los siguientes pasos metodológicos:

1. **Procesamiento de los datos:** En este paso se plantea una exploración preliminar de las dos fuentes principales de datos: Las iniciativas PDET y los Proyectos de Inversión Pública. El objetivo principal será identificar posibles relaciones entre ambas fuentes, reconociendo patrones comunes de ubicación geográfica, que permitan establecer un marco de análisis relacional.

   El resultado de este paso será un conjunto de datos más robusto y territorialmente alineado, sobre el cual se construirán las representaciones textuales y los modelos de similitud semántica en los pasos siguientes.

2. **Preparación semántica de los textos:** Este paso se enfoca en representar el contenido textual de las Iniciativas PDET y los Proyectos de Inversión Pública en forma de vectores dentro de un espacio semántico, lo que facilita las operaciones como la comparación, agrupamiento y cálculo de similitudes.

   Inicialmente se aplican técnicas de procesamiento de lenguaje natural (NLP) que buscan reducir la variabilidad léxica y homogenizar los textos (tokenización, lematización, eliminación de stopwords).

3. **Representación vectorial de los textos:** Este paso plantea representar la información textual de las iniciativas y de los proyectos en vectores numéricos utilizando las técnicas de *Term Frequency–Inverse Document Frequency (TF-IDF)* y *Word2Vec*, de secuencial, primero se ejecutará el TF-IDF y posteriormente el Word2Vec, con el fin de evaluar su comportamiento sobre los mismos conjuntos de datos.

   La selección de TF-IDF y Word2Vec obedece a la capacidad para representar textos breves sin etiquetas predefinidas y a su bajo costo computacional. Estas técnicas permiten generar vectores numéricos comparables desde dos enfoques distintos: uno basado en la importancia relativa de los términos (TF-IDF), y otro en la cercanía semántica contextual (Word2Vec), lo cual se ajustan a las características de las iniciativas y proyectos analizados[15].

Una vez obtenidas estas representaciones vectoriales, se requerirá una métrica que permita comparar dichos vectores y evaluar el grado de similitud entre textos. Para ello, se emplea la *similitud del coseno*, una medida utilizada en análisis de textos, que permite cuantificar el grado de alineación semántica entre documentos mediante el ángulo entre sus respectivos vectores.

TF-IDF

La técnica de *Term Frequency–Inverse Document Frequency* (TF-IDF) tiene como propósito capturar la relevancia de cada término dentro de un documento, considerando tanto su frecuencia relativa como su especificidad dentro del corpus completo. Se calcula mediante la siguiente expresión [15]:

$$\mathrm{TFIDF}(t, d, D) = \mathrm{tf}(t, d) \times \mathrm{idf}(t, D)$$

Donde:

- $\mathrm{tf}(t, d)$ es la frecuencia del término t en el documento d, calculada como la proporción entre las veces que aparece el término y el número total de términos del documento.
- $\mathrm{idf}(t, D)$ es la frecuencia inversa de documentos, definida como:

  $$\mathrm{idf}(t, D) = \log\left(\frac{N}{1 + n_t}\right)$$

  siendo N el número total de documentos del corpus D y n_t el número de documentos que contienen el término t.

Esta técnica permite asignar un mayor peso a los términos que se repiten dentro de un documento específico, pero que no son comunes en el resto del corpus. Es decir, resalta aquellas palabras que, por su uso poco frecuente en otros textos, aportan información más representativa del contenido del documento en el que aparecen. Por el contrario, las palabras que están presentes en muchos documentos tienden a recibir un peso menor, ya que contribuyen menos a caracterizar el contenido específico del texto.

Word2Vec

Word2Vec es una técnica de representación semántica que transforma cada palabra en un vector numérico de dimensión fija, entrenado para reflejar similitudes de uso en el lenguaje. Estos vectores, conocidos como *embeddings*, están compuestos por valores reales ($\mathbb{R}^d$) y se ubican en un espacio vectorial donde la proximidad entre palabras indica cercanía semántica [15].

El modelo cuenta con dos arquitecturas principales:

- **CBOW (Continuous Bag-of-Words)**: predice una palabra objetivo a partir de las palabras que la rodean en una ventana de contexto.
- **Skip-gram**: predice las palabras del contexto a partir de una palabra central.

En este TFM se emplea la arquitectura **Skip-gram**, debido a su capacidad para modelar relaciones semánticas más precisas, especialmente útil cuando se dispone de un corpus pequeño o con términos poco frecuentes, como ocurre en las narrativas de iniciativas comunitarias y proyectos institucionales.

El modelo se entrena optimizando la siguiente función de probabilidad condicional:

$$P(w_{t+j} \mid w_t) = \frac{\exp\left(\vec{v'}_{w_{t+j}} \cdot \vec{v}_{w_t}\right)}{\sum_{w=1}^{V} \exp\left(\vec{v'}_{w} \cdot \vec{v}_{w_t}\right)}$$

donde $\vec{v}_{w_t}$ y $\vec{v'}_{w_{t+j}}$ son los vectores de entrada y salida asociados a las palabras w_t y w_{t+j}, respectivamente, y V es el tamaño del vocabulario.

Para representar textos completos (iniciativas o proyectos), se promedia el conjunto de vectores correspondientes a las palabras del documento:

$$\vec{d} = \frac{1}{n} \sum_{i=1}^{n} \vec{v}_{w_i}$$

Similitud del Coseno

La *Similitud del Coseno*, esta métrica resulta adecuada para comparar tanto vectores generados por TF-IDF como vectores generados por Word2Vec.

La similitud del Coseno permite medir la similitud entre los vectores numéricos utilizando métricas espaciales, la cual calcula el coseno del ángulo entre dos vectores en un espacio multidimensional.

La similitud del coseno entre dos documentos d_1 y d_2, representados por sus vectores $\vec{v}_1$ y $\vec{v}_2$, se define como [15]:

$$\text{sim}_{\cos}(\vec{v}_1, \vec{v}_2) = \frac{\vec{v}_1 \cdot \vec{v}_2}{\|\vec{v}_1\| \cdot \|\vec{v}_2\|}$$

Donde:

- $\vec{v}_1 \cdot \vec{v}_2$ es el producto punto entre ambos vectores.
- $\|\vec{v}_i\|$ representa la norma (magnitud) del vector i.

El valor resultante de esta operación varía en el rango $[0, 1]$, donde:

- Un valor cercano a 1 indica que los documentos comparten una alta proporción de términos relevantes y, por tanto, son semánticamente similares.

- Un valor cercano a 0 indica que los documentos no tienen términos en común o están en direcciones casi ortogonales en el espacio vectorial.

4. **Reducción de dimensionalidad para representación y agrupamiento semántico:**

   Antes de aplicar técnicas de agrupamiento semántico sobre las iniciativas PDET, se debe implementar un proceso de reducción de dimensionalidad, con el fin de facilitar la identificación de patrones temáticos y mejorar la capacidad de segmentación del algoritmo Fuzzy C-Means.

   Es por esto que en este paso, se aplican dos técnicas complementarias: el Análisis de Componentes Principales (PCA), una técnica lineal orientada a la preservación de la varianza global; y t-Distributed Stochastic Neighbor Embedding (t-SNE), una técnica no lineal diseñada para preservar las relaciones locales de vecindad entre observaciones.

PCA

El Análisis de Componentes Principales (PCA, por sus siglas en inglés) es una técnica estadística de reducción de dimensionalidad que transforma un conjunto de variables posiblemente correlacionadas en un nuevo conjunto de variables linealmente independientes denominadas componentes principales. Su objetivo es capturar la mayor cantidad de varianza posible del conjunto original utilizando un número reducido de dimensiones[16].

t-SNE

t-distributed Stochastic Neighbor Embedding (t-SNE) es una técnica de reducción de dimensionalidad no lineal diseñada específicamente para visualizar y analizar datos de alta dimensión al proyectarlos en espacios de dos o tres dimensiones, conservando la estructura de vecindad entre los datos.

A diferencia de PCA, que se basa en transformaciones lineales, t-SNE transforma las distancias euclidianas entre puntos en probabilidades condicionales[17].

5. **Tipificación semántica de las iniciativas mediante clustering difuso:**

   Las técnicas de clustering no supervisado permiten identificar patrones en los datos sin requerir etiquetas predefinidas. En el caso de las iniciativas PDET, cuyo contenido frecuentemente abarca dimensiones temáticas diversas y superpuestas, se requiere un enfoque flexible que permita representar esta particularidad semántica. Por ello, se opta por el algoritmo *Fuzzy C-Means* (FCM), una técnica de clasificación difusa que permite a cada observación pertenecer simultáneamente a múltiples grupos, con diferentes grados de pertenencia.

   El FCM se basa en la minimización iterativa de una función objetivo que pondera la distancia entre los datos y los centroides de los clusters en función del grado de pertenencia de cada dato a cada grupo. Esta función se expresa como [18]:

$$J_m = \sum_{i=1}^{c} \sum_{j=1}^{n} u_{ij}^m \cdot \|x_j - c_i\|^2$$

donde:

- $u_{ij} \in [0, 1]$ representa el grado de pertenencia del dato x_j al cluster i,
- c_i es el centroide del cluster i,
- x_j es el vector de características de la observación j,
- $m > 1$ es el parámetro de difusividad, que regula el grado de borrosidad en la asignación.

El procedimiento de minimización alterna entre dos pasos principales: la actualización de los centroides (fijando los valores de pertenencia) y la actualización de la matriz de pertenencias (fijando los centroides), hasta alcanzar la convergencia.

Capítulo 3. DESARROLLO DEL PROYECTO

Con el propósito de avanzar en la solución planteada en el TFM y de poner en práctica el enfoque metodológicos descrito anteriormente, este capítulo presenta el desarrollo técnico del TFM. Para ello, se describen los pasos adelantados.

Los resultados presentados en este Capítulo, fueron desarrollados mediante scripts de Python, utilizando Jupiter Notebook como entorno de trabajo. El detalle del código fuente se encuentran en los anexos 6.2, 6.3, 6.4, 6.5 del presente TFM.

3.1 Procesamiento y enriquecimiento de los datos de iniciativas y proyectos de inversión

Como se identificó en el análisis exploratorio, no existe una relación directa entre los Proyectos de Inversión Pública y las Iniciativas PDET, que permita determinar el grado de alineación o pertenencia entre ambos conjuntos de datos, es por esto que, se requiere en primer lugar definir algún tipo de vínculo estructural entre los dos universos, que permita luego avanzar hacia ejercicios analíticos más complejos, como la clusterización basada en similitud semántica.

Uno de los vínculos estructurales para establecer esa relación, es la ubicación geográfica. Es importante resaltar que las iniciativas PDET fueron definidas exclusivamente en 170 municipios priorizados del territorio colombiano, mientras que los proyectos de inversión pública abarcan la totalidad del país. Partiendo de esta diferencia, se propone a continuación un ejercicio de cruce geográfico con el objetivo de segmentar los datos y acotar el universo de análisis a los territorios PDET, lo que permitirá una comparación más precisa entre las iniciativas comunitarias y los proyectos institucionales ejecutados en los mismos territorios.

Por lo anterior, se desarrolla una estrategia de depuración e imputación geográfica. A continuación, se detallan las acciones realizadas, acompañadas de los resultados cuantitativos y cualitativos obtenidos:

3.1.1. Carga y normalización de datos

Cargue de las fuentes de información iniciativas PDET y proyectos de inversión pública y apoyado en una nueva fuente de información que, si bien no es relevante para el ejercicio propuesto en el TFM, es una información que permite estandarizar los datos geográficos de los proyectos e iniciativas, esta base fuente de datos es la codificación territorial oficial de municipios y departamentos de Colombia según DIVIPOLA [19].

3.1.2. Imputación territorial de iniciativas

Teniendo en cuenta la metodología participativa PDET y sus fases de formulación de las iniciativas, existen 1.337 iniciativas denominas Subregionales, las cuales tienen cobertura territorial en todos los municipios que hacen parte de la Subregión. Es por esto que es necesario

imputarle el o los municipios que hacen parte de ella, para lograr identificar la ubicación geográfica de la iniciativa dentro de la Subregión. Esto permite garantizar que todas las iniciativas contaran con al menos una ubicación geográfica a nivel municipal.

3.1.3. Imputación de ubicación geográfica del proyecto a través del código de la entidad

Se eliminaron los proyectos que no se desarrollan en municipios PDET, esto permite que el análisis se enfoque únicamente en el 170 municipios priorizado del PDET.

3.1.4. Imputación de ubicación geográfica por análisis de texto del campo Nombre

Para los proyectos en los cuales no fue posible realizar la imputación geográfica de la acción anterior, se aplicaron técnicas de análisis textual sobre los campos de texto, buscando nombres de municipios o departamentos sobre ellos.

3.1.5. Eliminación de proyectos sin referencia territorial

Tras la imputación de los municipios a través de los métodos descritos anteriormente, se procedió a eliminar aquellos proyectos que no contenían ningún tipo de referencia a municipios PDET

3.1.6. Consolidación final

Como resultado del proceso, se obtuvo una base final con **63.563 proyectos**

Descripción	Número de proyectos
Universo original de proyectos – Carga y normalización de datos	68.729
Imputación de ubicación geográfica del proyecto a través del código de la entidad – Ubicados en municipios PDET	45.960
Imputación de ubicación geográfica del proyecto a través del código de la entidad – Ubicados en municipios NO PDET (eliminados)	1.926
Imputación por análisis del campo Nombre	17.603
Proyectos eliminados sin referencia territorial	3.240
Total de proyectos localizados en municipios PDET	63.563

Tabla 3.1. Resumen de filtrado y localización geográfica de proyectos de inversión (elaboración propia).

Si bien el objetivo del TFM no es establecer una relación causal entre la formulación de iniciativas PDET y la cantidad de proyectos en los municipios; se plantea una primera aproxima-

ción exploratoria orientada a identificar posibles patrones de similitud territorial entre ambas dimensiones.

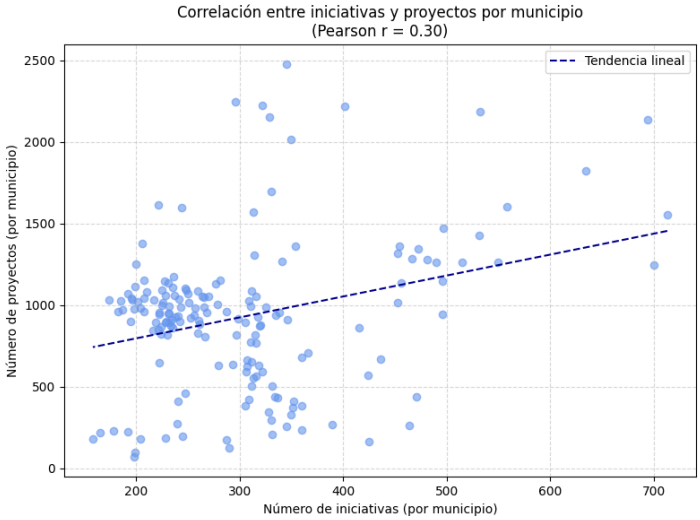


Figura 3.1. Correlación geográfica entre Iniciativas PDET y Proyectos de Inversión (elaboración propia).

A nivel municipal, se identifica una correlación débil positiva (r = 0.30), indicando que en muchos municipios con mayor número de iniciativas también hay una mayor presencia de proyectos. Sin embargo, esta relación no es uniforme, existen municipios con un número elevado de iniciativas y pocos proyectos y viceversa. Esto podría tener una relación directa con las limitadas capacidades institucionales en los municipios PDET, brechas estructurales (educación, seguridad, salud, etc.) no solucionadas históricamente, así como la ausencia del gobierno colombiano en estos territorios.

3.2 Preparación semántica de los textos

Con la base de datos depurada de iniciativas y proyectos georreferenciados en municipios PDET, el siguiente paso consiste en la preparación de los campos textuales que serán utilizados en los análisis de similitud semántica.

Tal como se identificó en el análisis exploratorio de las fuentes de información realizado en el Capitulo2, se seleccionaron los siguientes campos textuales relevantes: *'título iniciativa'* y *'descripción iniciativa'* en el caso de las iniciativas PDET, y *'nombre'*, *'descripcion' 'objetivo del proyecto'* en el caso de los proyectos de inversión pública.

La preparación semántica se desarrolló mediante técnicas de procesamiento de lenguaje natural (PLN), con el objetivo de limpiar, normalizar y enriquecer el texto, siguiendo el enfoque propuesto por Jurafsky y Martin[15],así:

- Conversión a minúsculas.
- Eliminación de signos de puntuación y stopwords, normalización ortográfica mediante la eliminación de tildes.
- Lematización utilizando el modelo en español de SpaCy (`es_core_news_sm`)[20].
- Generación de bigramas frecuentes con Gensim[21].

Texto original (título + descripción)	Texto procesado
Adecuar y optimizar la infraestructura existente del acueducto veredal en la zona rural, de los caseríos y centros poblados, para el abastecimiento permanente del agua, en cantidad, óptima y de calidad para el consumo humano y colectivo y demás necesidades básicas, en el municipio de San Calixto, Norte de Santander. Se requiere la adecuación y la optimización de la infraestructura existente del acueducto, en un plazo no mayor a 2 años, teniendo en cuenta que la población requiere el abastecimiento permanente del agua, en cantidad, con calidad y óptima para el consumo humano y colectivo y demás necesidades básicas, para el núcleo Arboledas: veredas el Silencio, la Torcoroma, Piletas y Balsamina; núcleo Banderas: veredas Banderas y la Laguna; núcleo Cabecera: veredas la Marina, Casas Viejas, Cucurina y la Quina; núcleo San José de la Sabana: veredas Yerbabuena, Vista Hermosa, Fátima, San José de la Sabana, Mesallana, Lagunetas, Chimenecas, Algarrobos, municipio de San Calixto, Norte de Santander; teniendo en cuenta el mejoramiento y la optimización de las obras de la bocatoma de captación del agua, desarenador, obras de aducción o conducción, planta de tratamiento, tanques de almacenamiento y sistemas de distribución.	adecuar optimizar infraestructura existente acueducto_veredal zona_rural caserio centro_poblado abastecimiento permanente agua cantidad optimo_calidad consumo_humano colectivo necesidad_basico municipio san_calixto norte_santander requerir adecuacion optimizacion infraestructura existente acueducto plazo_ano tener poblacion requerir abastecimiento permanente agua cantidad_calidad optimo consumo_humano colectivo necesidad_basico nucleo arboleda vereda_silencio torcoroma pileta balsamina nucleo bandera vereda bandera laguna nucleo_cabecerar vereda marina casa_viejo cucurin quina nucleo san_jose sabana vereda yerbabuen vista_hermoso fatimo san_jose sabana mesallan lagunetas chimeneca algarrobo municipio san_calixto norte_santander tener mejoramiento optimizacion obra bocatoma_captacion agua desarenador obra aduccion_conduccion planta_tratamiento tanque_almacenamiento sistema distribucion

Tabla 3.2. Muestra de texto original y texto procesado en iniciativas (elaboración propia).

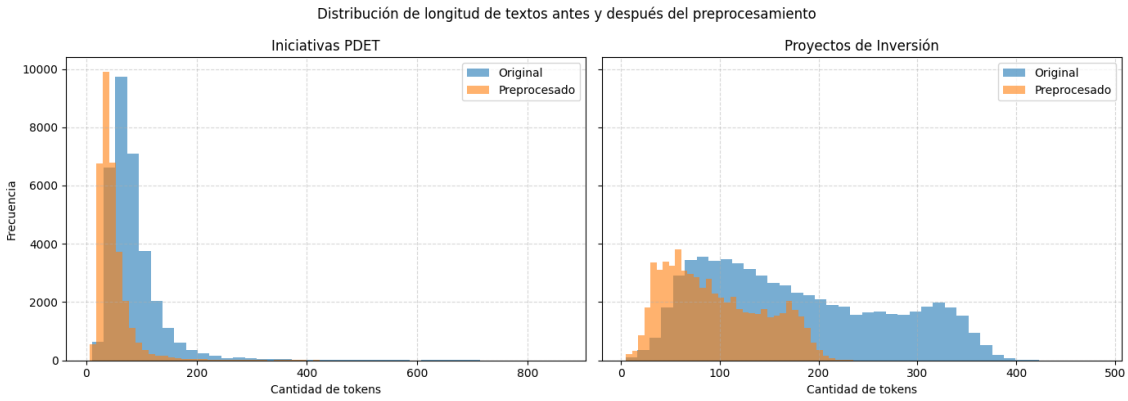


Figura 3.2. Distribución de la cantidad de palabras (tokens) antes y después del preprocesamiento en los textos de iniciativas PDET y proyectos de inversión pública (elaboración propia).

La Figura 3.2 muestra cómo el procesamiento de los datos redujo la longitud de los textos tanto en iniciativas PDET como en proyectos de inversión. En ambos casos, se observa una concentración más estrecha y desplazada hacia la izquierda, lo que indica la eliminación de ruido textual y una mayor compactación semántica.

3.3 Representación vectorial de los textos

Una vez preprocesados los textos de iniciativas PDET y de los Proyectos de Inversión, el siguiente paso corresponde a su transformación en un formato numérico que permita su análisis computacional.

Como se indicó en el capitulo 2 en su abordaje metodológico, se realizará la representación vectorial de los textos a través de dos técnicas *TF-IDF* y *Word2Vec*. No obstante, antes de aplicar cualquiera de estas técnicas, es necesario realizar un paso intermedio generando una versión extendida de las bases de datos (desnormalizar), en la cual cada fila representa una relación única con un municipio. En tal sentido una iniciativa se repetirá tantas veces como proyectos se desarrollen en el municipio de ella. Lo que amplia nuestro universo de análisis a 51.750 iniciativas y 158.230 proyectos

Código DANE	Código Iniciativa	Texto iniciativa (procesado)	BPIN Proyecto	Texto proyecto (procesado)
54670	454670190137	adecuar optimizar infraestructura existente acueducto_veredal zona_rural caserio centro_poblado abastecimiento...	2021004540147	apoyo fortalecer_gobernabilidad gestion_riesgo desastre departamento norte_santander gestion_riesgo ...
54670	454670190137	adecuar optimizar infraestructura existente acueducto_veredal zona_rural caserio centro_poblado abastecimiento...	2019004540143	desarrollo fase_departamental juegos_deportivo nacional magisterio encuentro_folclorico cultural nor...
54670	454670190137	adecuar optimizar infraestructura existente acueducto_veredal zona_rural caserio centro_poblado abastecimiento...	2021004540036	apoyo mejoramiento conocimiento_riesgo_desastre norte_santander alternativa aumentar conocimiento ri...
54670	454670190137	adecuar optimizar infraestructura existente acueducto_veredal zona_rural caserio centro_poblado abastecimiento...	2017004540024	fortalecimiento centro rehabilitacion_cardioneuromuscular cucutir actividad comprender fortalecimien...
54670	454670190137	adecuar optimizar infraestructura existente acueducto_veredal zona_rural caserio centro_poblado abastecimiento...	2021004540153	desarrollo accion mitigacion_adaptacion cambio_climatico departamento norte_santander norte_santande...

Tabla 3.3. Muestra extensión de base de datos por municipio.

3.3.1. TF-IDF y Similitud del Coseno

Esta técnica permite transformar cada documento textual (iniciativas y proyectos) en un vector numérico que cuantificando la importancia relativa de cada término en el contexto del corpus, penalizando los términos comunes y resaltando los distintivos.

Procedimiento técnico

Para cada municipio, se construyó un corpus compuesto por todas las iniciativas y todos los proyectos asociados territorialmente. A cada texto se le aplicó el vectorizador TF-IDF y, posteriormente, se calculó la similitud del coseno entre todos los pares posibles de iniciativa y proyecto dentro de ese municipio.

Resultados y visualización

A continuación se presentan los resultados algunas visualizaciones que permiten observar los valores de las similitudes obtenidos y su distribución.

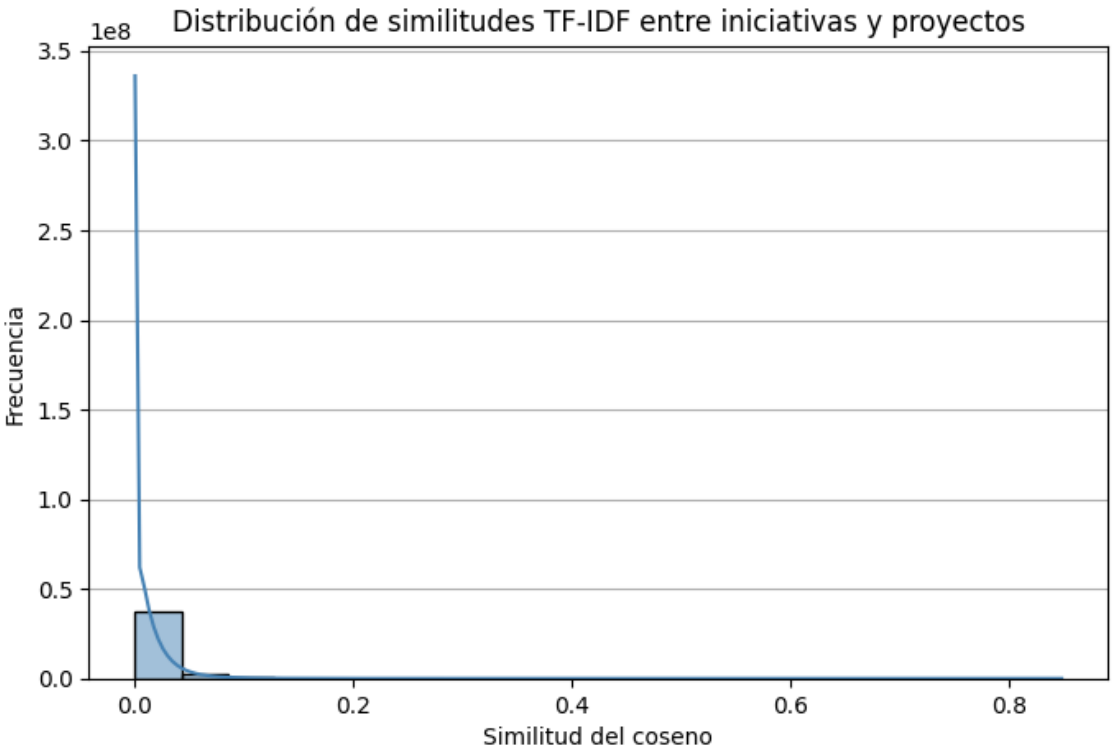


Figura 3.3. Distribución de similitudes entre proyectos e iniciativas (elaboración propia).

La Figura 3.3 muestra una distribución extremadamente sesgada a la derecha. A medida que aumenta la similitud, la frecuencia decrece. La mayoría de las coincidencias se concentran en valores extremadamente bajos, dificultando la identificación de emparejamientos temáticos válidos.

Debido a esta limitación, se exploró una alternativa metodológica basada en modelos de representación semántica más avanzados. Esta nueva aproximación se presenta más adelante en la Figura 3.5, donde se utiliza el modelo Word2Vec para obtener una distribución mucho más informativa y adecuada para el análisis de correspondencias.

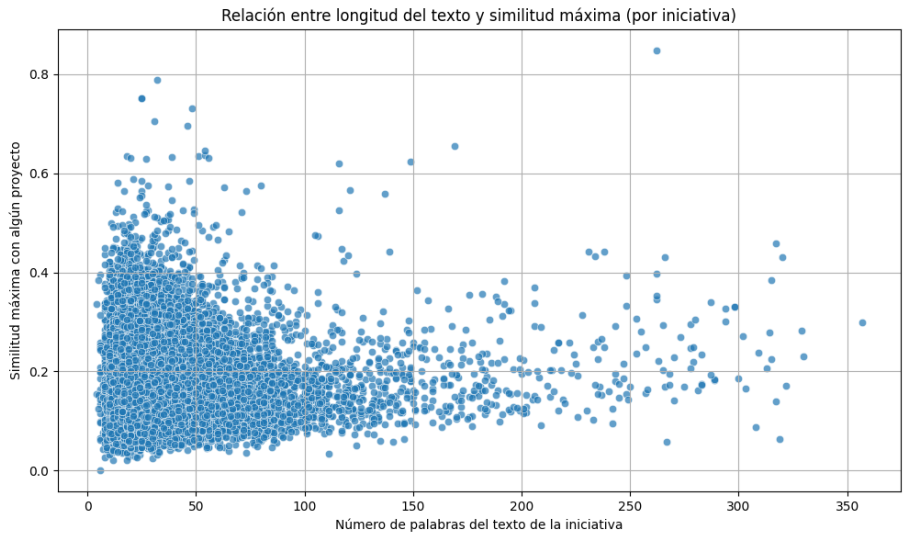


Figura 3.4. Relación entre longitud texto y similitud (elaboración propia).

Adicionalmente, se examinó la relación entre la longitud del texto de las iniciativas y la similitud máxima obtenida con algún proyecto (Figura 3.4). El gráfico evidencia que los textos cortos presentan una mayor dispersión en los valores de similitud, incluyendo coincidencias atípicamente altas, mientras que los textos largos tienden a estabilizarse en niveles bajos de similitud. Esto refuerza la hipótesis de que TF-IDF favorece coincidencias superficiales en documentos breves.

Como complemento al análisis presentado, se realizó una revisión cualitativa del TOP 3 de correspondencias con mayor similitud.

Inicativa	**Proyecto**	**Similitud Coseno**	**Análisis**
construir_alcantarilla vereda_nucleo municipio valle_guamuez departamento_putumayo construir_alcantarilla requerida vereda_nucleo municipio valle_guamuez departamento_putumayo donir vinculir siguiente obra descrito nucleo_cairo unidad construccion_alcantarilla hormiga cairo unidad cairo_delicia unidad...	fortalecimiento pequeno emprendimiento marco implementacion posconflicto municipio valle_guamuez suministro emprendedor comprender vegetal litro kg_kg unidad kg unidad unidad cubierto unidad cm unidad unidad unidad unidad unidad unidad puesto unidad liquido litro barra_barra...	0.848403	A pesar del alto valor de similitud, este caso representa una correspondencia espuria.
implementar acceso beca_estimulo subsidio_credito educacion tecnico_tecnologico profesional manutencion comunidad rural poblacion_victima municipio_briceno departamento_antioquia implementar acceso beca_estimulo subsidio_credito educacion tecnico_tecnologico profesional manutencion comunidad rural ...	implementacion acceso beca_estimulo subsidio_credito educacion tecnico_tecnologico profesional manutencion comunidad rural poblacion_victima municipio_briceno...	0.788412	La coincidencia literal en términos y estructura evidencia que el proyecto responde directamente a la iniciativa
estudio_diseno construccion_dotacion centro atencion_integral adulto persona capacidad especial municipio_tambo cauco estudio_diseno construccion_dotacion centro atencion_integral adulto persona capacidad especial...	construccion_dotacion centro atencion_integral adulto persona capacidad especial municipio_tambo construccion_dotacion centro atencion_integra...	0.750548	Ambos textos abordan la creación de un centro de atención integral para infancia y adolescencia.

Tabla 3.4. TOP 3 correspondencia semántica TF-IDF.

Conclusión TF-IDF y Similitud del Coseno

Los resultados confirman que, si bien *TF-IDF* ofrece una herramienta útil para filtrar y priorizar candidatos de vinculación, su alcance es limitado para capturar la complejidad semántica de las narrativas, lo que justifica la necesidad de técnicas más sofisticadas, como los modelos de representación distribuida tipo *Word2Vec*.

3.3.2. Word2Vec y Similitud del Coseno

Tal como fue planteado en el enfoque metodológico, esta sección implementa el modelo Word2Vec para obtener representaciones vectoriales densas de las iniciativas y los proyectos.

A partir de estas representaciones, se calcula la similitud del coseno entre textos vinculados geográficamente, con el fin de explorar relaciones semánticas contextuales.

Procedimiento técnico

Se entrenó un modelo *Skip-Gram*, el cual predice el contexto de una palabra dada, permitiendo capturar relaciones semánticas más robustas, especialmente útiles en corpus pequeños o especializados. El corpus se construyó uniendo los tokens de todas las iniciativas y proyectos. Al igual que *TF-IDF* para asegurar la relevancia contextual de las comparaciones, se restringió el análisis a iniciativas y proyectos ubicados en los mismos municipios,

Resultados y visualización

A continuación se presentan los resultados de la aplicación de *Word2Vec* sobre las Iniciativas PDET y Proyectos de Inversión Pública.

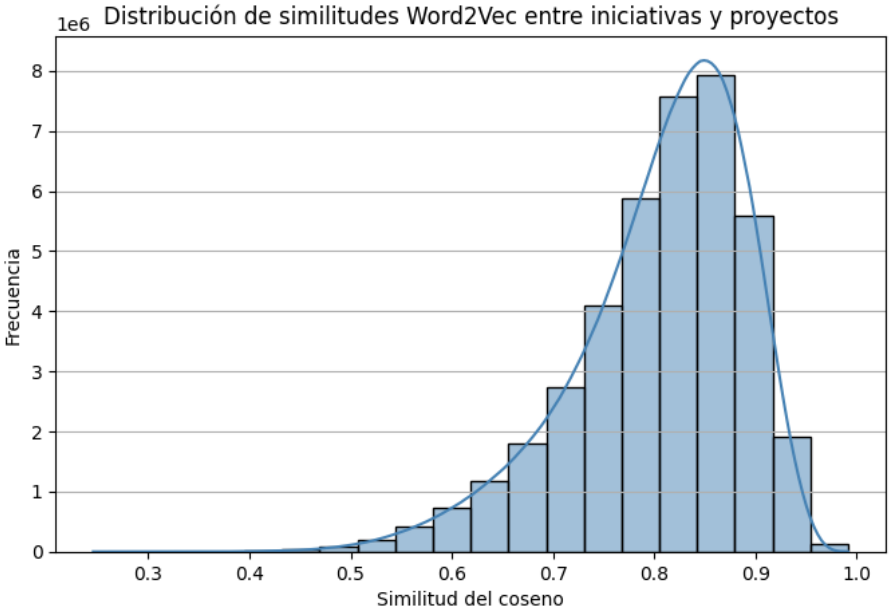


Figura 3.5. Distribución de similitudes entre iniciativas y proyectos (elaboración propia).

A diferencia de lo observado con TF-IDF Figura 3.3, cuyos valores de similitud se concentraban en niveles extremadamente bajos, el modelo Word2Vec permite capturar relaciones semánticas mucho más profundas y contextuales entre los textos de iniciativas y proyectos, mostrando una distribución asimétrica de las similitudes del coseno, con alta concentración entre 0.75 y 0.90, lo que refleja una fuerte afinidad semántica entre iniciativas y proyectos, atribuirse no solo a la capacidad del modelo *Word2Vec* para capturar relaciones contextuales entre palabras, sino también al hecho de que ambos conjuntos de textos comparten un marco narrativo común, derivado del contexto participativo y territorial del que emergen. En efecto, tanto las iniciativas como los proyectos institucionales tienden a incorporar un léxico vinculado a *zonas rurales*, *comunidades*, *fortalecimiento*, *necesidades*, *carencias* y *soluciones*, lo cual facilita la generación de similitudes semánticas, incluso en ausencia de coincidencias literales.

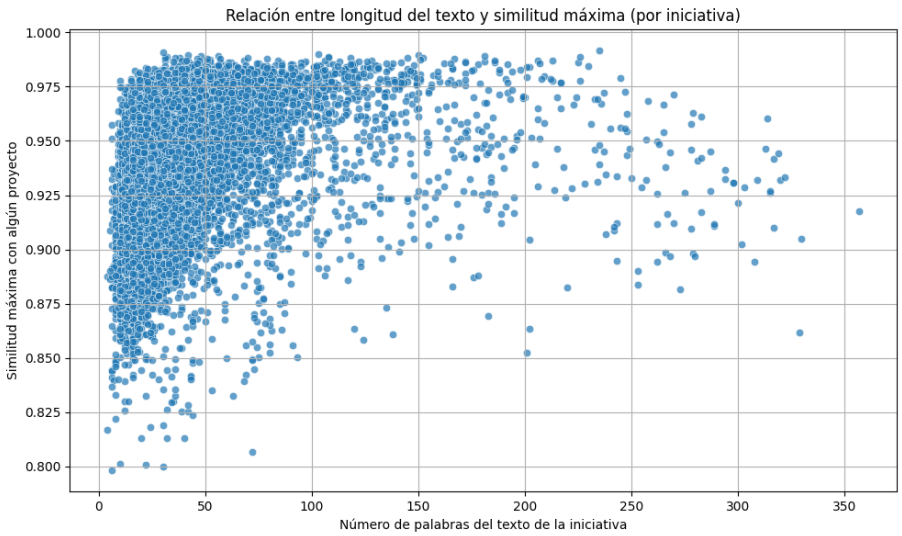


Figura 3.6. Relación entre longitud texto y similitud (elaboración propia).

La Figura 3.6 evidencia que existe una alta densidad de iniciativas con textos cortos que aún así logran valores altos de similitud (mayores a 0.90). Asimismo, se observa una amplia dispersión entre las iniciativas de mayor longitud. De igual manera, algunas iniciativas con más de 150 palabras presentan incluso similitudes más bajas que textos mucho más breves, confirmando que el modelo logra identificar patrones semánticos significativos sin requerir textos extensos.

Como complemento al análisis presentado, se realizó una revisión cualitativa del TOP 3 de corresponencias con mayor similitud.

Inicativa	Proyecto	Similitud Coseno	Análisis
fortalecer linea_productivo cacao establecimiento mantenimiento cultivo comercializacion cacao_tostado molido municipio_mercader cauca fortalecer linea_productivo cacao establecimiento mantenimiento cultivo comercializacion cacao_tostado molido ...	fortalecimiento produccion mora_castilla alto calidad departamento_caucir proyecto objetivo mejorar competitividad productor mora_castilla departamento_cauca suministro_material insumos proceso productivoc...	0.9918	Alta coherencia productiva, aunque varía el cultivo. Enfoque similar en fortalecimiento y comercialización
fortalecer_cadena productivo especie_menor caprino_avicola municipio fundacion_magdalena estudio_viabilidad factibilidad implementacion proyecto productivo especie_menor tal gallina carnero pequeno_productor...	...familia rural hectarea_cacao cultivo_platano cria_especie menor_ganadeer ovina porcino gallina_ponedora cria_especie menor base producto infraestructura productivo_pequeno ganadero_bovino ovino establecimiento banco_semilla forraj mantenimiento ampliacion...	0.9910	Coincidencia directa en especie menor, enfoque técnico y de capital semilla
implementar proyecto piscicultura_integracion productor_comerciante apropiacion_tecnologia municipio valle_guamuez departamento_putumayo implementar proyecto piscicultura_integracion productor_comerciante apropiacion_tecnologia extension_agropecuario reconocimiento_saber proceso fortalecimiento unidad_productivo existente ...	contribucion promocion forma_asociativo agropecuaria fortalecimiento productivo comercializacion entrega_insumo capacitacion municipio valle_guamuez putumayo solitud allegado analisis problematica_existente alternativa contribuir_mejoramiento productivo agropecuario comercializacion_producto apoyo forma_organizativo insumos material capacitacion...	0.9910	Coherencia alta en temática en comercialización de productos e inclusión comunitaria.

Tabla 3.5. TOP 3 correspondencia semántica Word2Vec.

Conclusión Word2Vec y Similitud del Coseno

La aplicación del modelo *Word2Vec* arrojó mejores resultados que *TF-IDF* en la identificación de similitudes entre iniciativas comunitarias y proyectos de inversión. Mientras TF-IDF permitió una primera aproximación basada en coincidencias léxicas, sus limitaciones para capturar relaciones semánticas más profundas resultaron evidentes, a diferencia de *Word2Vec* el cual mostró una mayor capacidad para identificar patrones de coincidencia conceptual, al construir representaciones del significado de las palabras en función de su contexto de uso.

Esta diferencia quedó clara al analizar la Tabla 3.5, donde se identificaron correspondencias temáticamente coherentes, incluso cuando el objeto productivo variaba —por ejemplo, cacao versus mora castilla— siempre y cuando el enfoque de intervención fuera compartido, como en casos de fortalecimiento productivo, transformación o comercialización.

Ahora bien, con el objetivo de definir un umbral que permita distinguir pares con correspondencia significativa, se recurrió a dos visualizaciones complementarias.

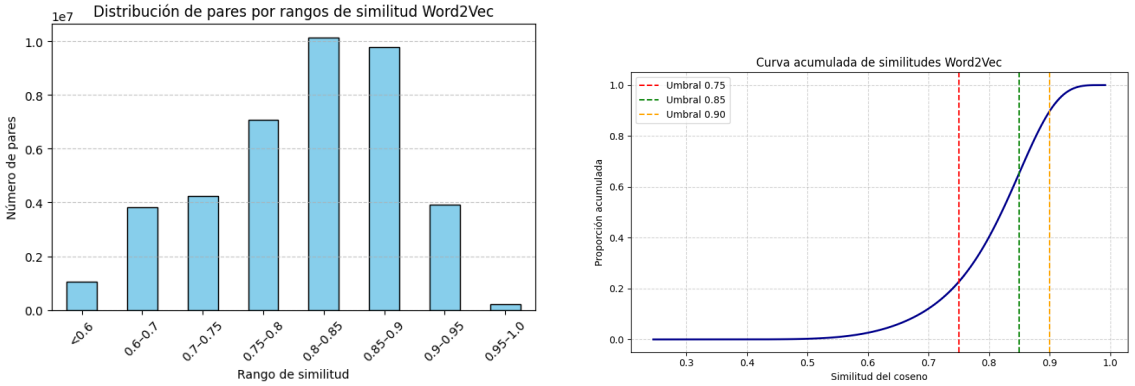


Figura 3.7. Distribución de similitudes entre iniciativas y proyectos (elaboración propia).

Por un lado, el histograma de frecuencias por rangos muestra que más de 10 millones de pares se concentran en el rango de similitud entre 0.80 y 0.85, y cerca de 9.7 millones en el rango de 0.85 a 0.90, lo que evidencia una alta densidad de emparejamientos con afinidad temática. Por otro lado, la curva acumulada de frecuencias muestra que aproximadamente el 75 % de los pares se encuentra por debajo de una similitud de 0.8678, y que los valores superiores a 0.85 corresponden al 25 % más alto, lo que permite considerarlos como correspondencias destacadas.

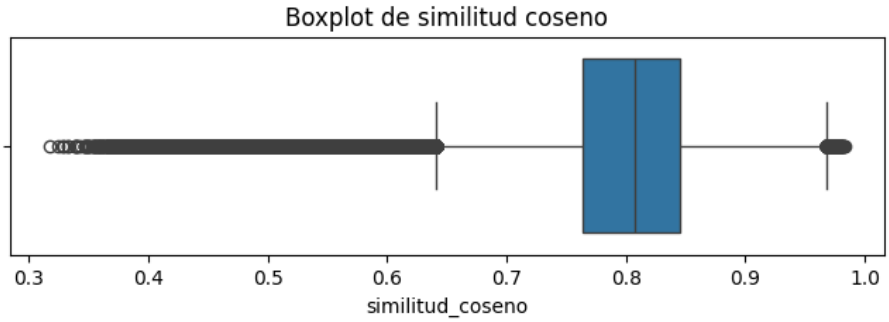


Figura 3.8. Distribución de similitudes entre iniciativas y proyectos (elaboración propia).

Por otra parte, El análisis del boxplot refuerza esta observación, mostrando que la mayoría de los valores se concentran en el rango intercuartil (IQR), aproximadamente entre 0.74 (Q1) y 0.87 (Q3), lo que indica que el 50 % de los datos presenta una similitud semántica significativa.

Basado en esto, se establece como ***umbral de referencia el valor de 0.85***, a partir del cual los pares serán considerados correspondencias semánticas fuertes. Asimismo, para los pares con similitudes entre 0.75 y 0.85 como un grupo intermedio, sujeto a análisis cualitativo, y descartar aquellos por debajo de 0.75 debido a su baja densidad y alta probabilidad de representar coincidencias espurias o falsos positivos.

3.4 Agrupamientos semánticos con Fuzzy C-Means

En la sección anterior se implementaron técnicas de representación vectorial para capturar el contenido semántico de las iniciativas PDET, utilizando modelos como TF-IDF y Word2Vec. Aunque el modelo Word2Vec permitió establecer correspondencias semánticas entre iniciativas y proyectos de inversión pública a nivel individual, dicha aproximación no permite identificar patrones temáticos presentes en las iniciativas.

En tal sentido, el clustering difuso mediante Fuzzy C-Means permite identificar tipologías en las demandas comunitaria, agrupando iniciativas con similitudes semánticas en un espacio vectorial común. Permitiendo evaluar qué tan alineadas han estado las inversiones ejecutadas con las prioridades comunitarias expresadas en los PDET, no solo a nivel de correspondencia puntual o individual, sino también desde una perspectiva temática agregada.

Procedimiento técnico

En esta sección se utilizó la implementación provista por la librería `scikit-fuzzy` de Python, siguiendo la guía técnica disponible públicamente [22].

Preparación de los vectores semánticos: El conjunto de datos utilizado en esta fase incluye únicamente las iniciativas PDET previamente preprocesadas, las cuales ya cuentan con limpieza textual, lematización y detección de bigramas, tal como se detalló en la sección anterior. A partir de estos textos, se entrenó un modelo *Word2Vec*.

Determinación del número de clusters: Uno de los aspectos más relevantes al aplicar algoritmos de clustering es la determinación del número óptimo de clusters, con el fin de determinar o capturar patrones comunes. En el caso del clustering difuso con *Fuzzy C-Means*, esta decisión adquiere una dimensión adicional, dado que no se trata únicamente de particionar los datos, sino de asignar grados de pertenencia a múltiples grupos.

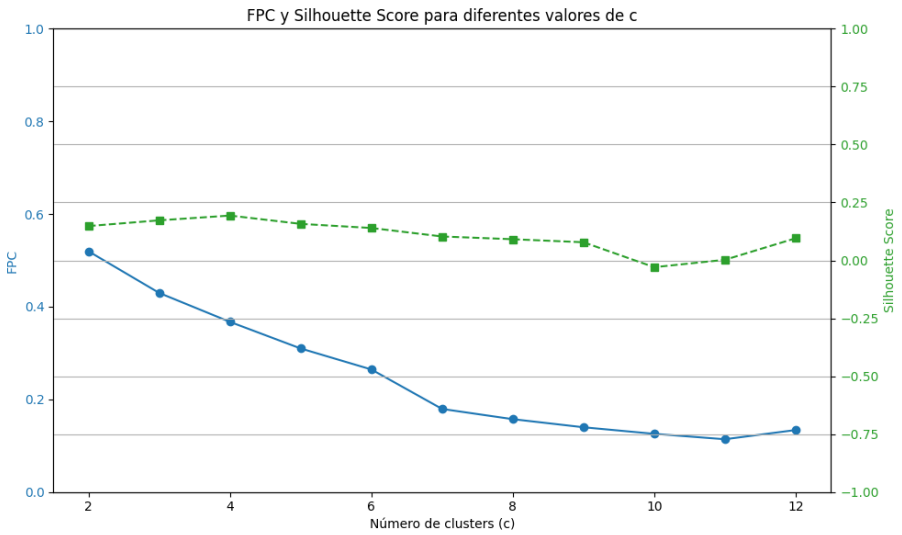


Figura 3.9. FPC vs Número de Clusters (elaboración propia).

Se evaluaron conjuntamente *Fuzzy Partition Coefficient - (FPC)* y *Silhouette Score* para valores de c entre 2 y 12, como se muestra en la figura 3.9. El FPC disminuye progresivamente a medida que aumenta el número de clusters, lo cual es esperable dada la naturaleza difusa del modelo. En paralelo, el Silhouette Score presenta su valor más alto relativo en c=4, sugiriendo que este número de clusters podría ofrecer una separación razonable entre grupos. A partir de estos resultados, se consideraron como candidatos principales los valores c=4 y c=5 para su análisis.

Aplicación del algoritmo sobre el espacio Word2Vec: Los resultados obtenidos tras la aplicación del algoritmo Fuzzy C-Means (FCM) sobre el espacio semántico generado mediante Word2Vec se presentan a continuación. El procedimiento consistió en:

- Generar vectores semánticos promedio para cada iniciativa, a partir de sus tokens lematizados y entrenados en Word2Vec.
- Reducir la dimensionalidad de estos vectores mediante PCA.
- Aplicar el algoritmo FCM con dos configuraciones distintas, c=4 y c=5 clusters.
- Visualizar los resultados mediante t-SNE.

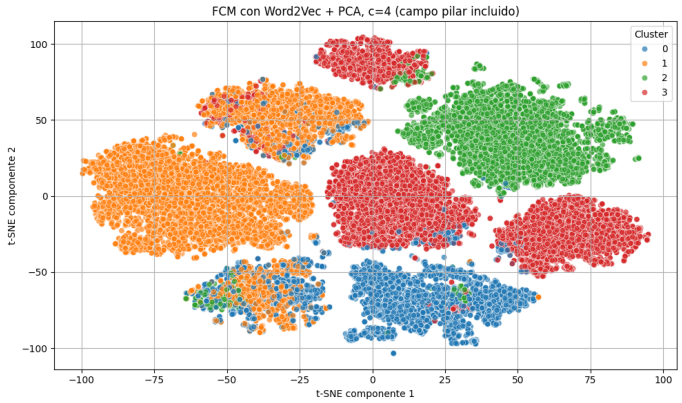


Figura 3.10. Visualización t-SNE de iniciativas agrupadas mediante FCM con $c = 4$ (elaboración propia).

Cluster	Nº iniciativas	Pertenencia promedio	Desv. estándar
0	5.564	0.322664	0.031322
1	11.268	0.515338	0.174085
2	6.558	0.662726	0.136642
3	9.617	0.358421	0.033897

Tabla 3.6. Pertenencia por cluster para FCM con $c = 4$ (elaboración propia).

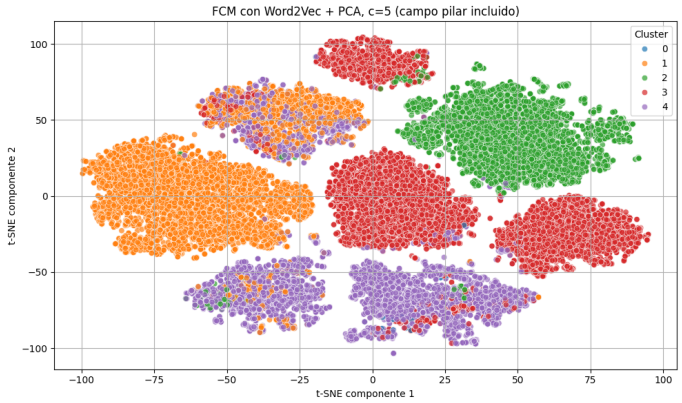


Figura 3.11. Visualización t-SNE de iniciativas agrupadas mediante FCM con $c = 5$ (elaboración propia).

Cluster	Nº iniciativas	Pertenencia promedio	Desv. estándar
0	42	0.261029	0.011724
1	9.287	0.498581	0.173975
2	6.283	0.641686	0.142815
3	9.969	0.269002	0.017577
4	7.426	0.243517	0.018083

Tabla 3.7. Pertenencia por cluster para FCM con $c = 5$ (elaboración propia).

Ambas configuraciones (c=4 y c=5) presentan niveles aceptables de pertenencia promedio, pero el escenario con c=5 generó un cluster con muy baja representatividad (17 iniciativas), lo cual sugiere una posible sobrefragmentación. Dado que el objetivo del análisis es identificar tipologías temáticas generalizables, se opta por **c=4**, que presenta una estructura más equilibrada en tamaño de clusters entre iniciativas.

Evaluación temática de los agrupamientos

Una vez definido el número de clusters mediante Fuzzy C-Means, el siguiente paso es evaluar la coherencia temática de los clusters resultantes. Es por esto que se analizan estos clusters a la luz de los 8 pilares PDET.

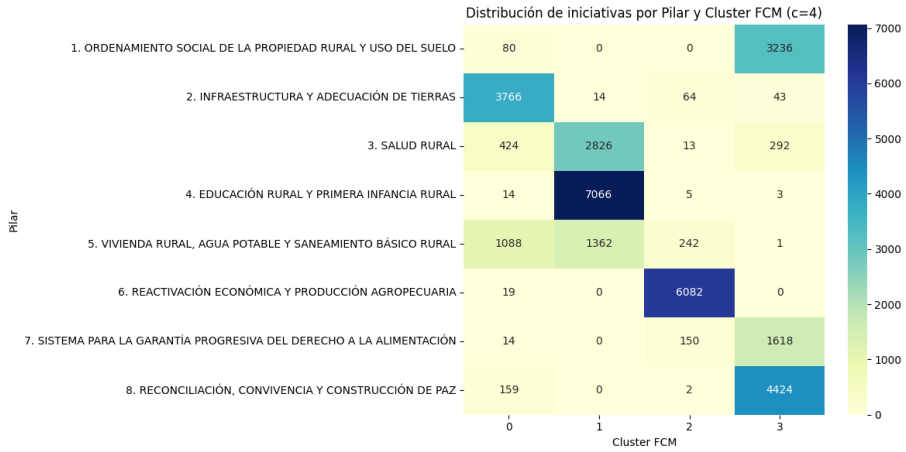


Figura 3.12. Distribución iniciativas por pilar y Clusters (elaboración propia).

Como se ha indicado previamente, las iniciativas PDET están clasificadas en ocho pilares como categorías orientadoras, el número de clusters resultantes del análisis no supervisado no tiene por qué coincidir con esta estructura institucional. Mientras que los pilares obedecen a una definición normativa y sectorial de intervención del gobierno Colombiano, el clustering semántico responde a patrones de proximidad en el lenguaje utilizado por las comunidades. Por tanto, la elección de cuatro clusters no busca replicar los ocho pilares, sino identificar agrupaciones del contenido textual de las iniciativas. Esta diferencia era esperable y, de hecho, enriquecedora, ya que permite detectar afinidades temáticas transversales o intersectoriales que pueden no estar explícitas en la clasificación oficial, pero que sí reflejan la lógica narrativa de las comunidades.

Complementando este este análisis y con el fin de profundizar la comprensión del contenido de cada agrupamiento, se utilizaron nubes de palabras, permitiendo identificar focos temáticos.

Figura 3.13. Nubes de palabras por Cluster (elaboración propia).

- **Cluster 0**: Se destacan términos *vía, puente, quebrado, río, construcción, estudio, diseño y kilómetro*.
  Temática sugerida: ***Mejoramiento de infraestructura vial y conectividad territorial***.

- **Cluster 1**: Se destacan términos como *comunitario, mujer, población, capacitación, gestionar, víctima conflicto, participación, paz y territorio*.

  Temática sugerida: ***Fortalecimiento del tejido social y comunitario con enfoque de género, paz y sostenibilidad***.

- **Cluster 2**: Se destacan términos como *proyecto productivo, implementar, producción, productivo integral, comercialización, campesino, mujer, familia y pequeño productor*.

  Temática sugerida: ***Fortalecimiento de sistemas productivos rurales y cadenas de valor***.

- **Cluster 3**: Se destacan términos como *estudiante, población, institución educativa, prestación servicio salud, atención salud, aula y escuela*.

  Temática sugerida: ***Garantía de acceso a servicios básicos de educación y salud***.

Conclusión Agrupamientos semánticos con Fuzzy C-Means

La elección del algoritmo Fuzzy C-Means para la segmentación de las iniciativas PDET se sustenta en la necesidad de abordar un conjunto de datos caracterizado por su complejidad

semántica y alta dimensionalidad. A diferencia de enfoques de clustering tradicionales, que requieren asignaciones exclusivas (una iniciativa en un único grupo), el modelo FCM permite capturar estructuras difusas donde las iniciativas pueden compartir simultáneamente elementos temáticos con múltiples clusters.

3.4.1. Correspondencia entre Proyectos Inversión Pública e Iniciativas PDET

Una vez definidos los agrupamientos semánticos de las iniciativas comunitarias mediante el modelo Fuzzy C-Means, se avanza hacia la identificación de posibles correspondencias entre las iniciativas PDET y los Proyectos de Inversión Pública ejecutados en los municipios PDET priorizados.

Este análisis busca responder a la pregunta de fondo del trabajo: ¿En qué medida la inversión pública realizada en los territorios refleja o responde a las prioridades expresadas por las comunidades?

Metodología de vinculación semántica

Para establecer una correspondencia temática entre las iniciativas priorizadas por las comunidades y los proyectos de inversión pública, se implementó un proceso de vinculación basado en similitud semántica entre textos. El procedimiento fue realizado previamente en la sección 3.3 **Representación vectorial de los textos**, el cual consistió en los siguientes pasos:

- **Representación semántica:** Se utilizó el modelo Word2Vec previamente entrenado sobre el corpus combinado de iniciativas y proyectos. A partir de dicho modelo, se generaron vectores semánticos promedio para cada documento.
- **Filtro geográfico:** Para garantizar la pertinencia territorial de las comparaciones, sólo se evaluaron pares de iniciativas y proyectos que estuvieran localizados en un mismo municipio, utilizando como referencia los códigos DANE presentes en ambos conjuntos de datos.
- **Cálculo de similitud:** La similitud semántica entre cada par de textos fue calculada mediante la métrica de similitud del coseno entre sus respectivos vectores Word2Vec.
- **Relación muchos a muchos:** Una iniciativa puede estar asociada a varios proyectos si se encuentran en el mismo municipio y superan un umbral mínimo de similitud, y viceversa.
- **Umbral de vinculación:** Se definió un umbral de similitud del coseno $\theta = 0{,}85$, a partir del cual se considera que existe una coincidencia semántica suficiente para establecer una posible correspondencia entre iniciativa y proyecto.

El resultado de este procedimiento es un conjunto de pares iniciativa–proyecto acompañados por su puntaje de similitud y su municipio compartido, lo que permite evaluar no sólo la cobertura temática de la inversión pública, sino también la calidad de su alineación narrativa con las prioridades comunitarias.

Adicionalmente, dado que cada iniciativa ya ha sido previamente clasificada en un cluster

temático mediante el modelo Fuzzy C-Means, cada proyecto vinculado hereda de forma indirecta esta clasificación al asociarse con una iniciativa específica.

Presentación y análisis de resultados

En esta sección se presentan los resultados más significativos del emparejamiento de las Iniciativas PDET con Proyectos de Inversión

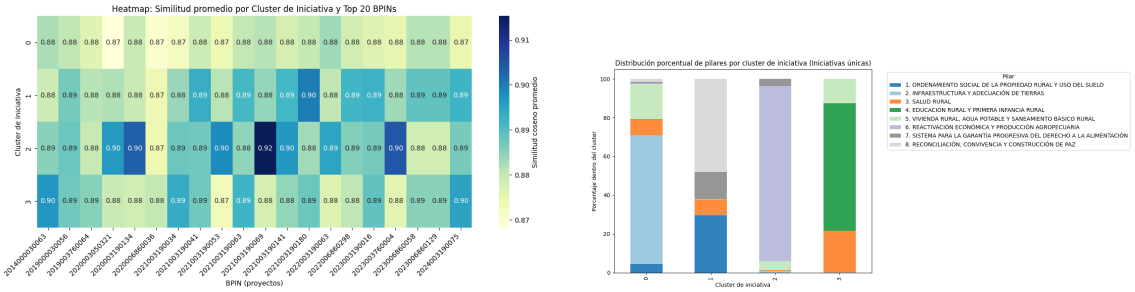


Figura 3.14. Distribución de clusters (elaboración propia).

La primera visualización corresponde a un heatmap de similitud promedio del coseno, que muestra el grado de afinidad semántica entre los cinco clusters de iniciativas y los 20 proyectos de inversión pública (BPIN) más representativos del conjunto vinculado. Esta representación evidencia la capacidad del modelo Fuzzy C-Means para permitir grados de pertenencia difusa: un mismo proyecto puede exhibir niveles de similitud variables con iniciativas de distintos clusters.

Por otra parte, la segunda visualización muestra un gráfico de barras apiladas con la distribución porcentual de los pilares PDET dentro de cada cluster semántico. Es importante mencionar, que aunque las iniciativas fueron previamente clasificadas institucionalmente bajo un único pilar, los resultados revelan que muchos agrupamientos contienen una mezcla significativa de pilares. Este hallazgo refleja la naturaleza multidimensional de las demandas comunitarias, en las que una misma iniciativa puede abordar simultáneamente temas como salud, educación, participación o infraestructura básica. En este sentido, la segmentación "blanda"que ofrece el modelo FCM resulta más adecuada que una clasificación rígida", ya que captura la complejidad y transversalidad de los intereses expresados por las comunidades PDET.

3.5 Aplicación web demostración

Con el propósito de mostrar el funcionamiento de los algoritmos desarrollados y la funcionalidad del modelo propuesto para la vinculación semántica entre Proyectos de Inversión Pública e Iniciativas comunitarias PDET se desarrolló una aplicación web. El codigo fuente de la aplicación desarrollada se encuentra disponible en 6.6.Para facilitar su acceso y prueba, esta fue desplegada en Amazon Web Services (AWS) [23].

Desarrollo de la demo

La aplicación web fue construida como una interfaz de prueba para demostrar proceso de vinculación semántica entre proyectos de inversión pública e iniciativas comunitarias PDET, empleando los modelos y datos procesados en el TFM.

Para su implementación se utilizó el framework Streamlit [24]. La app integra el modelo semántico Word2Vec previamente entrenado, como los vectores generados para cada iniciativa PDET.

Los insumos/entradas utilizados en la aplicación son:

- El modelo Word2Vec entrenado sobre el corpus de iniciativas y proyectos.
- Vectores promedio por iniciativa.
- Datos geográficos de los municipios en Colombia.

El funcionamiento de esta demo se basa en los siguientes pasos:

1. **Autenticación y autorización:** El usuario se autentica mediante un formulario de inicio de sesión en la página web, con el fin de acceder a la funcionalidad.
2. **Ingreso de los datos del proyecto:** El usuario ingresa los datos básicos (nombre, descripción y ubicación) del proyecto que desea formular o implementar en el territorio Colombiano.
3. **Correspondencia geográfica:** Se filtran las iniciativas PDET que que tienen presencia en el o en los municipios seleccionados por el usuario, con el fin de realizar el análisis de correspondencia solo con las iniciativas y proyectos que estén localizados en el mismo territorio.
4. **Procesamiento del texto:** Los datos ingresados son preprocesados mediante el mismo pipeline realizado en el TFM (lematizar, eliminación de stopwords, generación de bigramas).
5. **Cálculo de similitud y asociación de cluster:** Se calcula la similitud del coseno entre el vector del proyecto y los vectores de cada iniciativa localizada, aplicando un umbral de 0.85 (definido previamente).
6. **Presentación de resultados:** Se presentan las iniciativas con similitud con su metadata (subregión, municipio, pilar, clúster temático y nivel de similitud).

Herramientas utilizadas

El desarrollo de la aplicación web se realizó a partir de las herramientas y bibliotecas previamente utilizadas durante el desarrollo del TFM, reutilizando tanto modelos entrenados como estructuras de procesamiento semántico ya validadas. Adicionalmente, se incorporaron nuevos componentes para al despliegue, la interacción con el usuario y componentes de autenticación.

- Framework de interfaz Streamlit : Incorporado como nueva herramienta para la creación de la interfaz web.
- Componente de seguridad Streamlit-authenticator : Libreria incorporada para proteger el acceso a la demo mediante autenticación de usuarios [25].
- Entorno de despliegue: Amazon Web Services (AWS) : Se utilizó una instancia EC2 (Amazon Linux 2023) para el despliegue de la aplicación. Esta instancia se configuró dentro del nivel gratuito, asegurando una ejecución sin costos adicionales.

Arquitectura Cloud

La aplicación web desarrollada fue desplegada en la nube utilizando Amazon Web Services (AWS). La arquitectura empleada es simple, centrada en una única instancia EC2 y no incorpora servicios avanzados como balanceadores, almacenamiento S3 ni bases de datos gestionadas.

Componentes principales

- Instancia EC2 (Amazon Elastic Compute Cloud) : Se utilizó una instancia t2.micro, correspondiente al nivel gratuito, ejecutando Amazon Linux 2023. Esta instancia fue configurada con Python, Streamlit y todas las dependencias necesarias.
- Acceso y seguridad: El acceso a la instancia se realiza mediante el protocolo SSH, protegido con una clave privada en formato .pem. La instancia se asocia a un grupo de seguridad que define las reglas de entrada y salida. Puerto 22 habilitado para conexión SSH desde direcciones IP autorizadas, puerto 8501 habilitado para acceder a la aplicación desde el navegador web, no se utilizó Elastic IP, sino la IP pública dinámica brindada por EC2.
- Archivos y ejecución: Todos los archivos requerido para la ejecucuón fueron cargados directamente en el sistema de archivos de la instancia, no se utilizó almacenamiento S3.
- Red y conectividad: La instancia se desplegó dentro de una subred pública de una VPC predefinida en la región us-east-1. El acceso a internet se habilitó mediante una Internet Gateway.

Esta arquitectura es simple y básica, pero fue suficiente para alojar una aplicación funcional, accesible vía web.

Acceso a la aplicación

Como ya se indicó la aplicación fue desplegada públicamente mediante una instancia EC2 de AWS con dirección IP pública.

Con el fin de restringir el acceso a usuarios no autorizados, se implementó un mecanismo de autenticación utilizando la biblioteca streamlit-authenticator. Se habilitó un único usuario con las siguientes credenciales:

- Usuario: admin

- Contraseña: 8080

La aplicación esta disponible en la URL http://3.148.175.86:8501/

3.6 Recursos requeridos

1. **Recursos técnicos :**
    - Software: Python 3.12.1, Visual Studio Code, Jupyter Notebooks, bibliotecas como pandas, gensim, scikit-learn, matplotlib, scikit-fuzzy, wordcloud, streamlit.
    - Hardware local: Computador con 16 GB de RAM.
    - Servicios en la nube: Amazon Web Services (AWS), específicamente una instancia EC2 con Amazon Linux 2023 para el despliegue de la demo.
2. **Recursos de datos :**
    - Fuentes primarias: Bases de datos de iniciativas PDET, Proyectos de inversión pública y base oficial e municipios DANE.
3. **Recurso humano :**
    - Investigador principal: Analista e implementador de la metodología.
    - Asesor académico: Acompañamiento en la definición metodológica y validación de resultados.

3.7 Presupuesto

Para la definición del presupuesto del Trabajo de Fin de Master – TFM, se establece que el 100 % de las actividades requeridas para el desarrollo del presente TFM serán realizadas por el autor del mismo, basando en los conocimientos adquiridos en el Master Universitario. Acontinuación, se presentan dos escenarios presupuestales: uno básico, correspondiente a una prueba de concepto funcional, y otro extendido, orientado a un entorno mínimo de producción con mayores garantías de disponibilidad, seguridad y accesibilidad.

Escenario 1 – Implementación básica (capa gratuita de AWS)

En este escenario se considera el uso de una instancia EC2 de tipo `t2.micro` dentro de la capa gratuita de Amazon Web Services (AWS). El acceso se realiza mediante IP pública dinámica. Este escenario es adecuado para pruebas de concepto y evaluación académica, aunque presenta limitaciones en estabilidad y escalabilidad.

Tipo de coste	Descripción	Costo estimado
Ordenador personal	Uso del portátil/PC propio	900 €
Software requerido	Python, Jupyter, bibliotecas analíticas (Scikit-learn, Pandas, etc.)	0 €
Acceso a datos	Bases de datos públicas (PDET, Proyectos de Inversión Pública, DANE)	0 €
Recursos humanos	Trabajo del autor como analista e implementador (200 h a 30 €)	6.000€
Infraestructura en la nube	Instancia EC2 `t2.micro` (capa gratuita)	0€
Total estimado		**6.900 €**

Tabla 3.8. Presupuesto estimado del TFM – Escenario básico

Escenario 2 – Implementación extendida (entorno productivo)

Este segundo escenario contempla una configuración más robusta para despliegue público, incluyendo una Elastic IP, certificado SSL y un nombre de dominio personalizado.

Tipo de coste	Descripción	Costo estimado
Ordenador personal	Uso del portátil/PC propio	900 €
Software requerido	Python, Jupyter, bibliotecas analíticas	0 €
Acceso a datos	Bases de datos públicas (PDET, Proyectos de Inversión Pública, DANE)	€
Recursos humanos	Trabajo del autor como analista e implementador (200 h a 30 C/h)	6.000 €
Infraestructura en la nube (1 año)	Instancia EC2, elastic IP, almacenamiento	260 €
Certificado digital	Emisión de certificado SSL (Amazon Certificate Manager o alternativo)	30 €
Dominio web	Registro de dominio por 1 año	10 €
Total estimado		**7.200 €**

Tabla 3.9. Presupuesto estimado del TFM – Escenario extendido

3.8 Viabilidad

La viabilidad del proyecto puede analizarse desde dos dimensiones, un ámbito técnico y un ámbito de aplicación.

Viabilidad técnica

Desde el punto de vista técnico, el desarrollo del presente TFM ha sido factible gracias al uso de herramientas de software libre y lenguaje como Python.

En cuanto a su posible escalabilidad, la solución propuesta puede ser adaptada para procesar volúmenes mayores de datos mediante arquitecturas de cómputo basadas en servicios en la nube. Estas estrategias no solo mejorarían los tiempos de respuesta, sino que permitirían desplegar el modelo en territorios con múltiples subregiones PDET o incluso fuera del contexto original.

Viabilidad de aplicación territorial

En el ámbito de aplicación, la metodología demostrada en este TFM es pertinente para abordar desafíos similares en Colombia relacionados con la planificación participativa. Aunque el presente TFM se centra en las iniciativas comunitarias en el marco de los PDET, existen otros instrumentos de planeación territorial —como los Planes de Desarrollo Municipal, las Agendas de Concertación o las Mesas de Víctimas— que enfrentan el mismo reto estructural: garantizar que las demandas expresadas por las comunidades se traduzcan en proyectos concretos de inversión.

En este sentido, el enfoque puede contribuir a cerrar la brecha entre el diseño participativo y la ejecución programática, mejorando la focalización y la pertinencia de las intervenciones institucionales.

Capítulo 4. CONCLUSIONES

4.1 Conclusiones del trabajo

Este TFM abordó el análisis de la correspondencia temática entre las iniciativas comunitarias PDET y los proyectos de inversión pública, a partir de un enfoque basado en procesamiento del lenguaje natural y aprendizaje no supervisado. El objetivo fue explorar si existen correspondencia temática entre las demandas comunitarias expresadas por las comunidades en los PDET y las acciones gubernamentales implementadas a través de Proyectos de Inversión Pública.

La ruta metodológica no fue lineal. En una primera etapa se aplicaron procesos de limpieza, estructuración de la información y tratamiento de lenguaje natural, a través de eliminación de stopwords, lematización e identificación de biogramas, posterior a esto, se aplicó una representación textual mediante TF-IDF, combinada con la similitud del coseno, para identificar emparejamientos entre iniciativas y proyectos. Sin embargo, esta estrategia resultó limitada en su capacidad para capturar relaciones semánticas. Esto motivó un cambio hacia el modelo Word2Vec, que mejoró la calidad de las correspondencias detectadas, al permitir identificar coincidencias conceptuales más allá de los términos exactos.

Pese a estos avances, el análisis semántico por sí solo no era suficiente para caracterizar estructuralmente la diversidad temática de las iniciativas. Para superar esta limitación, se incorporó el algoritmo de clustering difuso Fuzzy C-Means, que permitió identificar agrupamientos temáticos con pertenencias múltiples.

El uso de técnicas no supervisadas implicó retos importantes. A diferencia de los modelos supervisados, donde existe una base de verdad para orientar el entrenamiento y la validación, aquí la interpretación de los resultados depende de la estructura interna de los datos y del juicio experto. Esto abre la posibilidad de cuestionamiento por expertos sectoriales o institucionales, quienes manejan marcos de clasificación más rígidos.

4.2 Conclusiones personales

Este TFM representó la oportunidad para poner en práctica los aprendizajes adquiridos a lo largo del Máster Universitario, pude aplicar competencias en programación con Python, comprensión y uso de algoritmos de machine learning, análisis de datos, construcción de gráficos interpretativos, formulación de metodologías de investigación y búsqueda sistemática de referencias académicas relevantes.

Más allá de los aspectos técnicos, el proyecto me permitió explorar una problemática que considero fundamental: la necesidad de acercar el diseño de política pública al análisis de datos, aportando evidencia objetiva y estructurada, reduciendo el margen de decisiones influenciadas por ideologías políticas o intereses particulares. Este tipo de enfoque favorece una asignación más eficiente de los recursos públicos, en función de las necesidades expresadas desde el territorio.

Capítulo 5. FUTURAS LÍNEAS DE TRABAJO

Los resultados y aprendizajes obtenidos en el desarrollo de este TFM abren posibilidades de profundización y aplicación futura, tanto desde el plano técnico como institucional y social. A continuación, se destacan algunas líneas de trabajo:

- **Validación institucional con expertos sectorialistas:** Una etapa clave a futuro será la validación de los resultados obtenidos con actores institucionales especializados en educación, salud, infraestructura, reconciliación y desarrollo económico, entre otros. Esto permitiría ajustar el modelo semántico y así fortalecer la herramienta pensada como apoyo en procesos de planificación.
- **Estimación de recursos dirigidos a iniciativas comunitarias:** Una línea futura consiste en estimar cuántos recursos públicos han sido destinados, de manera directa o proporcional, a las iniciativas comunitarias agrupadas temáticamente. En muchos casos, un solo proyecto institucional responde simultáneamente a varias iniciativas y aborda múltiples áreas (por ejemplo, salud, educación o infraestructura). Por tanto, se plantea la posibilidad de distribuir el presupuesto total del proyecto entre las iniciativas a las que se vincula, según su grado de similitud semántica o afinidad temática. Esto permitiría calcular, por ejemplo, que un proyecto asigna un 30 % de su inversión a iniciativas relacionadas con educación y un 70 % a salud. Esta aproximación facilitaría estimar el costo real de la respuesta institucional a cada tipo de demanda y generar análisis más precisos sobre la focalización del gasto público en los territorios.
- **Simulación territorial y visualización del impacto potencial de las iniciativas:** Una línea de desarrollo futuro consiste en la creación de simulaciones dinámicas que permitan visualizar, mediante elementos cartográficos, cómo la implementación de iniciativas comunitarias podría transformar los territorios. Esto implicaría modelar posibles escenarios de impacto sobre indicadores como pobreza, empleo, seguridad o acceso a servicios, con base en datos históricos y proyectivos. A través de estas simulaciones, sería posible no solo observar retrospectivamente qué cambios se han dado tras la ejecución de ciertas iniciativas, sino también anticipar cómo podrían evolucionar los territorios según distintos escenarios de inversión, facilitando la planificación territorial basada en evidencia y enfoque diferencial.
- **Portal de datos abiertos para la ciudadanía:** Como componente clave de transparencia y democratización de la información, se propone la creación de un portal de datos abiertos que permita a la ciudadanía visualizar, en lenguaje sencillo, qué iniciativas comunitarias han sido atendidas, con qué proyectos, cuánto presupuesto se ha ejecutado y qué temas o territorios permanecen sin atención. Este portal no solo funcionaría como una herramienta de consulta histórica de intervenciones realizadas, sino también como un instrumento para que las comunidades puedan identificar brechas persistentes en la inversión pública y orientar sus futuras propuestas de manera más informada dentro de los procesos de planificación participativa.

Bibliografía

[1] *Proceso de paz en Colombia: cronología desde 1982 hasta hoy*, BBVA, 2016. dirección: https://www.bbva.com/es/proceso-paz-colombia-cronologia/ (visitado 01-03-2025).

[2] *Conozca el Acuerdo Final de Paz en 5 puntos*, Canal Institucional, 2022. dirección: https://www.canalinstitucional.tv/noticias/conozca-el-acuerdo-final-de-paz-en-5-puntos (visitado 01-03-2025).

[3] *El Acuerdo de Paz en Colombia*, La ciencia de la justicia, 2018. dirección: http://lacienciadelajusticia.blogspot.com/2018/10/el-acuerdo-de-paz-en-colombia.html (visitado 01-03-2025).

[4] *PDET: una apuesta por el país presente y futuro*, Agencia de Renovación del Territorio, 2022. dirección: https://centralpdet.renovacionterritorio.gov.co/documentos/pdet-una-apuesta-por-el-pais-presente-y-futuro/ (visitado 01-03-2025).

[5] *Informe de gestión: implementación de los Programas de Desarrollo con Enfoque Territorial (PDET)*, Agencia de Renovación del Territorio, 2020. dirección: https://portalparalapaz.gov.co/wp-content/uploads/2023/09/Informe_PDET_-Jun_2020.pdf (visitado 01-03-2025).

[6] *Conoce los PDET*, Agencia de Renovación del Territorio, 2023. dirección: https://centralpdet.renovacionterritorio.gov.co/conoce-los-pdet/ (visitado 01-03-2025).

[7] *Metodología General Ajustada (MGA)*, Departamento Nacional de Planeación, 2023. dirección: https://www.dnp.gov.co/LaEntidad_/subdireccion-general-inversiones-seguimiento-evaluacion/direccion-proyectos-informacion-para-inversion-publica/Paginas/metodologia-general-ajustada-mga.aspx (visitado 01-03-2025).

[8] *Constitución Política de Colombia*, República de Colombia, 1991. dirección: https://www.ramajudicial.gov.co/documents/10228/1547471/CONSTITUCION-Interiores.pdf (visitado 05-05-2025).

[9] *Ley 1712 de 2014: Por medio de la cual se crea la Ley de Transparencia y del Derecho de Acceso a la Información Pública Nacional y se dictan otras disposiciones*, Congreso de Colombia, 2014. dirección: https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=56882 (visitado 05-05-2025).

[10] *Base de datos de Iniciativas PDET*, Agencia de Renovación del Territorio, 2022. dirección: https://centralpdet.renovacionterritorio.gov.co/wp-content/uploads/2022/02/IniciativasPDET.xlsx (visitado 01-03-2025).

[11] *MapaInversiones Colombia: Transparencia en la inversión pública*, Departamento Nacional de Planeación, 2023. dirección: https://mapainversiones.dnp.gov.co/ (visitado 01-03-2025).

[12] *Resolución Número 000141 de 2024*, Agencia de Renovación del Territorio, 2021. dirección: https://portal.renovacionterritorio.gov.co/descargar.php?idFile=32940 (visitado 05-05-2025).

[13] K. Neha, V. Agrawal, S. Chhatani, A. B. Buduru y P. Kumaraguru, *An Unsupervised Narrative Detection Framework to Demystify Online Protest Composition*, 2023. dirección: https://doi.org/10.21203/rs.3.rs-2753534/v1 (visitado 01-03-2025).

[14] S. Goswami y M. S. Shishodia, «A Fuzzy Based Approach to Text Mining and Document Clustering,» 2013. dirección: https://doi.org/10.48550/arXiv.1306.4633.

[15] D. Jurafsky y J. H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, 3rd (draft). Draft online, 2025, Versión preliminar en línea, consultada el 1 de marzo de 2025. dirección: https://web.stanford.edu/~jurafsky/slp3/.

[16] I. T. Jolliffe, *Principal Component Analysis*. 2002. dirección: http://cda.psych.uiuc.edu/statistical_learning_course/Jolliffe%20I.%20Principal%20Component%20Analysis%20(2ed.,%20Springer,%202002)(518s)_MVsa_.pdf (visitado 05-05-2025).

[17] L. van der Maaten y G. Hinton, «Visualizing Data using t-SNE,» *Journal of Machine Learning Research*, vol. 9, págs. 2579-2605, 2008. dirección: http://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf (visitado 05-05-2025).

[18] J. Nayak, B. Naik y H. S. Behera, «Fuzzy C-Means (FCM) Clustering Algorithm: A Decade Review from 2000 to 2014,» en 2015. dirección: https://www.researchgate.net/publication/278660061 (visitado 05-05-2025).

[19] *DIVIPOLA: División político-administrativa de Colombia*, Departamento Administrativo Nacional de Estadística (DANE), 2025. dirección: https://geoportal.dane.gov.co/geovisores/territorio/consulta-divipola-division-politico-administrativa-de-colombia/ (visitado 05-05-2025).

[20] Explosion, *Spanish pipelines · spaCy models documentation*, n.d. dirección: https://spacy.io/models/es (visitado 05-05-2025).

[21] R. Řehůřek, *Gensim: Phrase Detection — models.phrases*, n.d. dirección: https://radimrehurek.com/gensim/models/phrases.html (visitado 05-05-2025).

[22] S.-F. Developers, *Fuzzy c-means clustering — Example from scikit-fuzzy documentation*, 2019. dirección: https://pythonhosted.org/scikit-fuzzy/auto_examples/plot_cmeans.html (visitado 05-05-2025).

[23] A. W. Services, *Amazon EC2 - Virtual Server Hosting*, 2023. dirección: https://aws.amazon.com/ec2/.

[24] S. Inc., *Streamlit: Turn Data Scripts into Shareable Web Apps*, 2024. dirección: https://streamlit.io/.

[25] M. Ebrahim, *Streamlit Authenticator: Simple user login flow for Streamlit apps*, 2024. dirección: https://github.com/mkhorasani/streamlit-authenticator.

Capítulo 6. ANEXOS

6.1 Anexo 1 - Jupyter Notebook (Estado del Arte)

El notebook correspondiente al Capítulo 2 (*Estado del Arte*), en donde se aplican los algoritmos de esta sección, encuentra disponible públicamente en:

- **Repositorio GitHub:** `https://github.com/223b6003/TFM`
- **Archivo:** `Anexo 1 - Jupyter Notebook (Estado del Arte).ipynb`

6.2 Anexo 2 - Jupyter Notebook (Desarrollo del proyecto)

El notebook correspondiente al Capítulo 2 (*Desarrollo del proyecto*), en donde se aplican los algoritmos de procesamiento de los datos y preparación semántica, se encuentra disponible públicamente en:

- **Repositorio GitHub:** `https://github.com/223b6003/TFM`
- **Archivo:** `Anexo 3 - Jupyter Notebook (Desarrollo del proyecto).ipynb`

6.3 Anexo 3 - Jupyter Notebook (Desarrollo del proyecto)

El notebook correspondiente al Capítulo 2 (*Desarrollo del proyecto*), en donde se aplican los algoritmos de TF-IDF y Similitud del Coseno, se encuentra disponible públicamente en:

- **Repositorio GitHub:** `https://github.com/223b6003/TFM`
- **Archivo:** `Anexo 3 - Jupyter Notebook (Desarrollo del proyecto).ipynb`

6.4 Anexo 4 - Jupyter Notebook (Desarrollo del proyecto)

El notebook correspondiente al Capítulo 2 (*Desarrollo del proyecto*), en donde se aplican los algoritmos de Word2Vec y Similitud del Coseno, se encuentra disponible públicamente en:

- **Repositorio GitHub:** `https://github.com/223b6003/TFM`
- **Archivo:** `Anexo 4 - Jupyter Notebook (Desarrollo del proyecto).ipynb`

6.5 Anexo 5 - Jupyter Notebook (Desarrollo del proyecto)

El notebook correspondiente al Capítulo 2 (*Desarrollo del proyecto*), donde se define número de cluster y analizan los clusters y conclusiones finales se encuentra disponible públicamente en:

- **Repositorio GitHub:** `https://github.com/223b6003/TFM`
- **Archivo:** `Anexo 5 - Jupyter Notebook (Desarrollo del proyecto).ipynb`

6.6 Anexo 6 - Demo

Archivo Python correspondiente, donde se desarrolló la demo desplegada en los servicios AWS se encuentra disponible públicamente en:

- **Repositorio GitHub:** `https://github.com/223b6003/TFM`
- **Archivo:** `app.py`