UNIVERSIDAD NACIONAL DEL ALTIPLANO

ESCUELA DE POSGRADO

MAESTRÍA EN INGENIERÍA DE SISTEMAS

TESIS

MODELO DE ANÁLISIS DE TÓPICOS EN ORDENANZAS REGIONALES DEL GOBIERNO REGIONAL DE PUNO DE 2010 A 2024 UTILIZANDO TÉCNICAS DE PROCESAMIENTO DE LENGUAJE NATURAL

PRESENTADA POR:

MARCOS DENYS CHOQUE CASTRO

PARA OPTAR EL GRADO ACADÉMICO DE:

MAESTRO EN INGENIERÍA DE SISTEMAS

PUNO, PERÚ

2024

MARCOS DENYS CHOQUE CASTRO

MODELO DE ANÁLISIS DE TÓPICOS EN ORDENANZAS REGIONALES DEL GOBIERNO REGIONAL DE PUNO DE 2010 A...

23.- INGENIERÍA DE SISTEMAS

MAESTRIAS

Universidad Nacional del Altiplano

Detalles del documento

Identificador de la entrega
trn:oid:::8254:409374231

Fecha de entrega
25 nov 2024, 12:26 p.m. GMT-5

Fecha de descarga
25 nov 2024, 12:32 p.m. GMT-5

Nombre de archivo
MODELO DE ANÁLISIS DE TÓPICOS EN ORDENANZAS REGIONALES DEL GOBIERNO REGIONAL DE ....pdf

Tamaño de archivo
11.7 MB

106 Páginas

22,720 Palabras

125,763 Caracteres

Dr. Miguel Romilio Aceituno Rojo
INGENIERO DE SISTEMAS
CIP. 169010

UNIVERSIDAD NACIONAL DEL ALTIPLANO - PUNO-PERU - ESCUELA DE POSGRADO - COORDINACIÓN DE INVESTIGACIÓN

Ing. Estadístico e Informático
CIP. 116625

10% Similitud general

El total combinado de todas las coincidencias, incluidas las fuentes superpuestas, para ca...

Filtrado desde el informe

- Bibliografía
- Texto citado
- Texto mencionado
- Coincidencias menores (menos de 12 palabras)

Fuentes principales

8% Fuentes de Internet
2% Publicaciones
5% Trabajos entregados (trabajos del estudiante)

Marcas de integridad

N.º de alertas de integridad para revisión

No se han detectado manipulaciones de texto sospechosas.

Los algoritmos de nuestro sistema analizan un documento en profundidad para buscar inconsistencias que permitirían distinguirlo de una entrega normal. Si advertimos algo extraño, lo marcamos como una alerta para que pueda revisarlo.

Una marca de alerta no es necesariamente un indicador de problemas. Sin embargo, recomendamos que preste atención y la revise.

Dr. Miguel Romilio Aceituno Rojo
INGENIERO DE SISTEMAS
CIP. 169010

UNIVERSIDAD NACIONAL DEL ALTIPLANO
ESCUELA DE POSGRADO
COORDINACIÓN DE INVESTIGACIÓN
PUNO - PERÚ

Ing. Jared Luque Coyla
ING. ESTADÍSTICO E INFORMÁTICO
CIP. 116625

UNIVERSIDAD NACIONAL DEL ALTIPLANO

ESCUELA DE POSGRADO

MAESTRÍA EN INGENIERÍA DE SISTEMAS

TESIS

MODELO DE ANÁLISIS DE TÓPICOS EN ORDENANZAS REGIONALES DEL GOBIERNO REGIONAL DE PUNO DE 2010 A 2024 UTILIZANDO TÉCNICAS DE PROCESAMIENTO DE LENGUAJE NATURAL

PRESENTADA POR:

MARCOS DENYS CHOQUE CASTRO

PARA OPTAR EL GRADO ACADÉMICO DE:

MAESTRO EN INGENIERÍA DE SISTEMAS

APROBADA POR EL JURADO SIGUIENTE:

PRESIDENTE

.....
Dra. MILDER ZANABRIA ORTEGA

PRIMER MIEMBRO

.....
M.Sc. MARGA ISABEL INGALUQUE ARAPA

SEGUNDO MIEMBRO

.....
M.Sc. MAGALI GIANINA GONZALES PACO

ASESOR DE TESIS

.....
Dr. MIGUEL ROMILIO ACEITUNO ROJO

Puno, 28 de octubre de 2024.

ÁREA: Ingeniería de software e inteligencia artificial.
TEMA: Análisis documental en ordenanzas regionales.
LÍNEA: Sistemas, computación e informática.

DEDICATORIA

A mis queridos padres, Denis Choque y Frida Castro, quienes me han apoyado incondicionalmente en cada paso de mi vida. Su paciencia y sacrificio son la luz que guía mis pasos en este largo camino del aprendizaje.

Marcos Denys Choque Castro.

AGRADECIMIENTOS

A la Universidad Nacional del Altiplano, a la Maestría en Ingeniería de Sistemas por darme la oportunidad de expandir mis conocimientos. Así mismo, a los docentes, su apoyo y orientación han sido fundamentales para mi formación profesional.

A mis jurados Dra. Milder Zanabria, M.sc. Marga Ingaluque, M.sc. Magali Gonzales y especialmente a mi asesor Dr. Miguel Aceituno por su paciencia y dedicación durante la elaboración de esta tesis, así como por los alcances y correcciones.

Marcos Denys Choque Castro.

ÍNDICE GENERAL

ÍNDICE DE TABLAS

Pág.

ÍNDICE DE FIGURAS

Pág.

ÍNDICE DE ANEXOS

Pág.

ACRÓNIMOS

BERT	:	Representaciones de Codificadores Bidireccionales de Transformadores
BS4	:	Biblioteca de Manipulación de documentos HTML y XML en Python
KDD	:	Descubrimiento de Conocimientos en Base de datos
LDA	:	Asignación de Dirichlet Latente
NLTK	:	Kit de herramientas de lenguaje natural
OCR	:	Reconocimiento Óptico de Caracteres
PDF	:	Archivo de documento portable
NLP	:	Procesamiento de Lenguaje Natural
TF-IDF	:	Frecuencia de Término-Frecuencia Inversa de Documento

RESUMEN

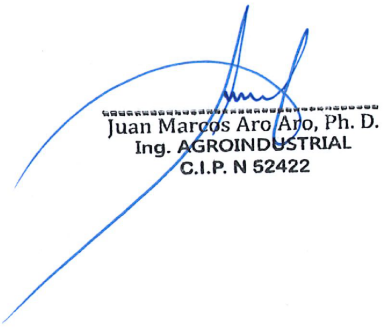
Los gobiernos regionales generan una vasta cantidad de documentos oficiales que contienen decisiones, normativas y políticas públicas fundamentales para el desarrollo de sus regiones. Esta investigación tuvo como objetivo determinar un modelo de análisis que optimice la identificación de tópicos en las ordenanzas regionales emitidas por el Gobierno Regional Puno durante el periodo del mes de febrero de 2010 al mes de julio del 2024. La investigación fue de tipo no experimental con un enfoque cuantitativo, por lo cual se utilizó técnicas de extracción de información, técnicas de procesamiento de lenguaje natural (NLP) y los modelos LDA y BERTopic. Como resultado se obtuvo un corpus de 249 ordenanzas regionales pre procesadas y un modelo de identificación de tópicos BERTopic con una precisión de 0,67 en la métrica de coherencia, lo que indica una buena capacidad para capturar tópicos coherentes y significativos en el corpus de ordenanzas, por otro lado, el modelo LDA ofrece una métrica de coherencia de 0,57 lo que evidencia una capacidad menor para identificar tópicos coherentes en el contexto específico de los documentos analizados. Se concluye que el modelo propuesto permitió identificar temas como transporte, cultura, salud, medio ambiente y presupuesto; que proporcionan una comprensión más profunda de los temas priorizados en las ordenanzas regionales de Puno.

Palabras Clave: BERTopic, LDA, modelado de tópicos, NLP, ordenanzas regionales.

ABSTRACT

Regional governments generate a vast amount of official documents containing decisions, regulations and public policies crucial for the development of their regions. This research aimed to determine an analysis model that optimizes topic identification in the regional ordinances issued by the Government Puno during the period from February 2010 to July 2024. The research was non-experimental with a quantitative approach, for which techniques used included information extraction, natural language processing (NLP) techniques, and the LDA and BERTopic models. As a result, a corpus of 249 pre-processed regional ordinances and a BERTopic topic identification model with an accuracy of 0.67 in the coherence metric were obtained, indicating a good ability to capture coherent and significant topics in the corpus of ordinances. On the other hand, the LDA model offers a coherence metric of 0.57, which shows a lower capacity to identify coherent topics in the specific context of the analyzed documents. It is concluded that the proposed model allowed to identify topics such as transportation, culture, health, environment and budget; which provide a deeper understanding of the prioritized topics in the regional ordinances of Puno.

Keywords: BERTopic, LDA, NLP, regional ordinances, topic modeling.



INTRODUCCIÓN

En la era de la información, los gobiernos regionales generan una vasta cantidad de documentos oficiales que contienen decisiones, normativas y políticas públicas fundamentales para el desarrollo de sus comunidades. Sin embargo, el acceso y análisis eficiente de estos documentos puede ser una tarea compleja debido al volumen y la diversidad de los datos. Este desafío es particularmente relevante en el caso del Gobierno Regional Puno, que ha emitido numerosas ordenanzas entre los años 2010 y 2024. La identificación de temas subyacentes en estas ordenanzas puede proporcionar una visión integral de las prioridades y enfoques adoptados por el gobierno regional a lo largo de este período.

El presente trabajo de investigación se centra en el desarrollo de un modelo de identificación de tópicos en las ordenanzas regionales del Gobierno Regional Puno utilizando técnicas avanzadas de Procesamiento de Lenguaje Natural (PLN). El objetivo principal es descubrir y analizar los temas recurrentes y emergentes en estas ordenanzas para ofrecer una comprensión más profunda y sistemática del contenido y las tendencias normativas. Mediante la aplicación de algoritmos como Latent Dirichlet Allocation (LDA) y BERTopic, se pretende estructurar la información de manera que facilite su interpretación y uso práctico. La metodología propuesta incluye la recolección y preparación de un corpus de ordenanzas desde el sitio web oficial del Gobierno Regional Puno. Este corpus abarcará documentos emitidos durante el periodo de 2010 a 2024 y será sometido a procesos de pre procesamiento, tokenización y limpieza para garantizar la calidad de los datos. Posteriormente, se implementarán y evaluarán diferentes modelos de tópicos, comparando sus resultados para seleccionar el más adecuado para el análisis de este conjunto de datos. Esta investigación no solo contribuirá al campo del PLN aplicando técnicas de análisis de tópicos a un contexto regional y normativo, sino que también proporcionará herramientas valiosas para los tomadores de decisiones, investigadores y ciudadanos interesados en el desarrollo y la gestión pública de la región Puno. Al identificar y analizar los temas clave en las ordenanzas regionales, se espera promover una mayor transparencia, eficiencia y eficacia en la formulación de políticas públicas.

CAPÍTULO I

REVISIÓN DE LITERATURA

1.1 Marco teórico

1.1.1 Inteligencia artificial (IA)

La Inteligencia Artificial (IA) es un campo multidisciplinario que abarca al desarrollo de sistemas con capacidad de realizar tareas que, tradicionalmente, requieren inteligencia humana. Estas tareas incluyen el reconocimiento de voz, la toma de decisiones, el aprendizaje y la resolución de problemas. Desde la década de 1950, la IA ha evolucionado significativamente, influenciada por avances en matemáticas, estadística, informática y neurociencia. Según Russell y Norvig (2020) la IA es entendida como la simulación de procesos cognitivos humanos mediante máquinas con el propósito de realizar tareas complejas de manera autónoma. Una de las áreas fundamentales dentro de la IA es el aprendizaje automático (machine learning), que permite a las máquinas aprender a partir de datos sin ser explícitamente programadas para realizar una tarea específica.

Un aspecto crucial en la IA es la ética y la responsabilidad, especialmente a medida que la tecnología se integra en aspectos cotidianos de la vida. La implementación de sistemas de IA plantea desafíos éticos relacionados con la privacidad, la equidad y la transparencia. La capacidad de los sistemas de IA para tomar decisiones automatizadas requiere un marco ético sólido para garantizar que estas tecnologías se utilicen de manera justa y responsable, evitando sesgos y protegiendo los derechos individuales (Binns, 2018).

El impacto de la IA en la sociedad y la economía está siendo cada vez más evidente, la inteligencia artificial y la automatización están transformando el mercado laboral, generando nuevas oportunidades y desafíos. La integración de la IA en diversas industrias está impulsando la eficiencia y la innovación, pero también plantea preguntas sobre el futuro del trabajo y la necesidad de políticas que apoyen la transición hacia un entorno económico cada vez más automatizado (Brynjolfsson y McAfee, 2014).

1.1.2 Procesamiento de lenguaje natural

El Procesamiento del Lenguaje Natural es un subcampo de la inteligencia artificial y lingüística dedicado a hacer que las computadoras comprendan declaraciones o palabras escritas en lenguajes humanos. Nació para facilitar el trabajo del usuario y satisfacer el deseo de comunicarse con la computadora en lenguaje natural (Khurana et al., 2023), cuyo propósito principal es de permitir que las computadoras comprendan, interpreten y generen texto o habla de manera similar a como lo haría un ser humano (Rawat et al., 2022).

El procesamiento de lenguaje natural (NLP) es una disciplina fundamental en la inteligencia artificial que permite a las máquinas comprender y generar lenguaje humano. Su importancia radica en la capacidad de automatizar y mejorar la interacción entre las personas y los sistemas computacionales. Desde la traducción automática hasta los asistentes virtuales, el NLP facilita la comunicación más eficiente y accesible, transformando la manera en que las personas interactúan con la tecnología. Además, el NLP es crucial en el análisis y extracción de información valiosa a partir de grandes volúmenes de texto. Esto es particularmente relevante en el ámbito empresarial, donde las organizaciones pueden utilizar NLP para analizar opiniones de clientes, detectar tendencias en redes sociales, y optimizar la toma de decisiones basada en datos textuales. La capacidad de procesar y comprender lenguaje no estructurado a gran escala proporciona a las empresas una ventaja competitiva significativa.

El NLP tiene un impacto profundo en la democratización del conocimiento y la accesibilidad. Herramientas como la traducción en tiempo real, la síntesis de voz para personas con discapacidades visuales, y los sistemas de aprendizaje de idiomas están haciendo que el conocimiento y la tecnología sean más accesibles para todos, independientemente de las barreras idiomáticas o físicas. Esto subraya la importancia de NLP no solo como una herramienta tecnológica, sino también como un motor de inclusión social y cultural en la era digital.

El procesamiento de lenguaje natural (NLP) principalmente se clasifica en dos tareas principales, en la tarea de comprensión del lenguaje natural y la tarea de generación del lenguaje natural (Hamilton y Lahne, 2022).

A. Comprensión del lenguaje natural (NLU)

Se refiere a la capacidad de una máquina para interpretar y comprender el significado del lenguaje humano. Este proceso implica analizar texto o discurso para extraer información, comprender intenciones, y manejar la ambigüedad y las sutilezas del lenguaje. El objetivo principal de la NLU es convertir el lenguaje natural en una representación que las máquinas puedan procesar y utilizar para tomar decisiones o realizar acciones. NLU abarca tareas como el análisis sintáctico (parsing), el análisis semántico, la resolución de referencias y la comprensión de intenciones. Las aplicaciones de NLU Son:

- **Sistemas de respuesta a preguntas:** Estos sistemas, como los implementados en asistentes virtuales (por ejemplo, Siri, Alexa), dependen de la NLU para interpretar las preguntas del usuario y proporcionar respuestas precisas.
- **Chatbots y atención al cliente:** Los chatbots utilizan técnicas de NLU para comprender las consultas de los usuarios y generar respuestas coherentes.
- **Análisis de opiniones y sentimientos:** NLU es fundamental para analizar grandes volúmenes de texto en redes sociales o reseñas de productos para entender las percepciones de los usuarios.

La comprensión del lenguaje natural enfrenta desafíos como la ambigüedad, las figuras retóricas y la ironía, que son difíciles de manejar para los modelos actuales. Uno de los mayores retos es la necesidad de desarrollar modelos que comprendan el significado pragmático, es decir, cómo se utiliza el lenguaje en diferentes contextos sociales y culturales.

B. Generación del lenguaje natural (NLG)

Es la tarea de crear texto o discurso en lenguaje humano a partir de datos estructurados o representaciones semánticas. A diferencia de la NLU, que se centra en comprender el lenguaje, la NLG se enfoca en producir texto coherente y significativo. NLG tiene como propósito

transformar datos o representaciones de conocimiento en un texto legible y comprensible por humanos. La NLG involucra varias subtareas, como la planificación de contenido, la estructuración de frases, y la realización lingüística. Las aplicaciones de NLG son:

- **Generación automática de contenidos:** NLG se utiliza en la creación de resúmenes automáticos, generación de noticias y reportes financieros, donde grandes cantidades de datos estructurados se convierten en texto narrativo.
- **Sistemas de conversación:** En aplicaciones como los chatbots, NLG es fundamental para generar respuestas naturales y coherentes en las interacciones con los usuarios.
- **Narrativas personalizadas:** En educación y marketing, NLG permite la creación de narrativas personalizadas basadas en el perfil y preferencias del usuario.

A pesar de los avances, la generación de lenguaje natural presenta desafíos significativos. La coherencia global del texto, la capacidad de mantener un estilo consistente, y la generación de textos creativos que no se limiten a los datos de entrenamiento siguen siendo problemas difíciles. Además, la generación de contenido ético y libre de sesgos es un reto crucial en aplicaciones reales de NLG.

1.1.3 Modelos de lenguaje

Los modelos de lenguaje han evolucionado desde enfoques estadísticos simples hasta complejos sistemas de aprendizaje profundo que son capaces de captar las complejidades del lenguaje humano. Estos modelos no solo han transformado la forma en que interactuamos con la tecnología, sino que también han abierto nuevas posibilidades en áreas como la asistencia virtual, la traducción automática, la generación creativa de contenido y la investigación en lingüística computacional. A medida que los modelos de lenguaje continúan evolucionando, se espera que sigan desempeñando un papel crucial en la mejora de la interacción entre humanos y máquinas, así como en la expansión de las capacidades de la inteligencia artificial.

Los modelos de lenguaje son un pilar fundamental en el procesamiento de lenguaje natural (NLP), y se han convertido en una herramienta esencial para diversas aplicaciones de inteligencia artificial. Estos modelos son sistemas que, a partir de un gran corpus de texto, aprenden a predecir la probabilidad de una secuencia de palabras, permitiendo generar, comprender y manipular texto en lenguaje natural. El objetivo principal de un modelo de lenguaje es captar las complejidades del idioma, desde la gramática hasta las sutilezas semánticas y contextuales, para realizar tareas como la predicción de la próxima palabra en una frase, la generación de texto coherente y la traducción automática (Khurana et al., 2023).

En sus inicios, los modelos de lenguaje estaban basados en enfoques estadísticos, como los modelos n-grama, que predecían la próxima palabra en una secuencia basándose en una ventana limitada de palabras anteriores. Aunque efectivos para ciertos propósitos, estos modelos tenían importantes limitaciones, como la incapacidad de captar dependencias a largo plazo o la sensibilidad a problemas de escasez de datos. Con el tiempo, estos enfoques dieron paso a métodos más sofisticados, como las redes neuronales recurrentes (RNNs), que mejoraron la capacidad de los modelos para capturar relaciones más complejas entre las palabras, aunque todavía enfrentaban desafíos en el manejo de secuencias largas.

La verdadera revolución en los modelos de lenguaje llegó con el desarrollo de la arquitectura Transformer, que se basa en mecanismos de atención para procesar secuencias de palabras en paralelo, en lugar de hacerlo de manera secuencial como en las RNNs. Modelos como GPT (Generative Pre-trained Transformer) y BERT (Bidirectional Encoder Representations from Transformers), ambos basados en esta arquitectura, han demostrado un rendimiento superior en una amplia gama de tareas de NLP. GPT, por ejemplo, es capaz de generar texto de manera coherente y contextualmente adecuada, mientras que BERT ha establecido nuevos estándares en tareas como la comprensión de texto y la clasificación de sentencias, gracias a su capacidad para analizar el contexto bidireccionalmente.

La característica clave de estos modelos es su capacidad para ser preentrenados en grandes volúmenes de texto y luego ajustados a tareas específicas con una cantidad mucho menor de datos. Este enfoque de preentrenamiento y ajuste fino ha permitido a los modelos de lenguaje no solo generalizar bien en una variedad de dominios, sino también adaptarse rápidamente a nuevas tareas con alta precisión. Además, la escalabilidad de estos modelos, que pueden ser entrenados con cantidades masivas de datos y en arquitecturas con miles de millones de parámetros, ha llevado a un progreso sin precedentes en la capacidad de las máquinas para entender y generar lenguaje natural (Hamilton y Lahne, 2022).

1.1.4 Representación de texto

La representación del texto es un aspecto clave en el procesamiento de lenguaje natural (NLP), ya que permite a las máquinas interpretar y manipular información textual de manera efectiva. Las palabras en un texto deben ser transformadas en una forma que los algoritmos puedan procesar, lo cual se logra a través de diversas técnicas de representación. Una de las formas más básicas es el modelo de Bolsa de Palabras (Bag of Words, BoW), donde el texto se representa como un conjunto de palabras, ignorando el orden, pero manteniendo la frecuencia de cada una. Aunque es sencillo y útil en muchos casos, este enfoque tiene limitaciones, como la incapacidad de capturar el contexto y el significado semántico de las palabras.

Para superar las limitaciones de BoW, se han desarrollado técnicas más avanzadas como las representaciones distribuidas o embeddings. Los embeddings, como Word2Vec, GloVe o FastText, asignan a cada palabra un vector en un espacio continuo de alta dimensión, donde la proximidad entre vectores refleja similitudes semánticas. Esto significa que palabras con significados similares tendrán representaciones numéricas cercanas en el espacio vectorial, lo que mejora significativamente el rendimiento de los modelos en tareas como la clasificación de texto, el análisis de sentimientos y la traducción automática. Los embeddings también permiten capturar información contextual, una ventaja crucial en el procesamiento del lenguaje natural.

En los últimos años, la aparición de modelos de lenguaje preentrenados como BERT y GPT ha llevado la representación del texto a un nuevo nivel. Estos modelos no solo generan representaciones distribuidas de palabras, sino que también consideran el contexto completo en el que una palabra aparece, lo que les permite captar matices más profundos y relaciones complejas entre las palabras. Como resultado, las tareas de NLP que requieren una comprensión más sofisticada del lenguaje, como la respuesta a preguntas y la generación de texto, han experimentado avances significativos. La representación efectiva del texto es, por lo tanto, fundamental para el éxito de cualquier aplicación de NLP.

La representación del texto es una tarea fundamental en el procesamiento de lenguaje natural (NLP) y en la modelización de texto. Existen diversas técnicas para convertir texto en representaciones numéricas que los algoritmos de machine learning puedan utilizar. Dos enfoques prominentes son el Bag of Words (BoW) y los embeddings (Abdelrazek et al., 2023).

A. Bag of words (BoW)

La representación de texto mediante el modelo Bolsa de Palabras (Bag of Words, BoW) es una técnica fundamental en el procesamiento de lenguaje natural (NLP). En este enfoque, un documento de texto se convierte en un vector en el que cada dimensión corresponde a una palabra única del vocabulario general, y el valor de cada dimensión indica la frecuencia con la que esa palabra aparece en el documento. Este método es sencillo y fácil de implementar, permitiendo a los modelos de aprendizaje automático procesar grandes cantidades de texto y realizar tareas como la clasificación de documentos, análisis de sentimientos y detección de spam (Hamilton y Lahne, 2022).

Sin embargo, la simplicidad de BoW también trae consigo ciertas limitaciones. Al ignorar el orden y la relación contextual de las palabras, BoW pierde información valiosa sobre la estructura del lenguaje. Por ejemplo, las frases “no me gusta” y “me gusta” tendrían representaciones muy similares, a pesar de tener significados opuestos. Además, al tratar cada palabra de forma independiente, BoW no captura las relaciones

semánticas entre palabras, lo que puede llevar a que palabras con significados similares se traten como completamente distintas en el modelo.

A pesar de estas limitaciones, BoW sigue siendo una herramienta poderosa, especialmente cuando se combina con técnicas adicionales como la ponderación TF-IDF (Term Frequency-Inverse Document Frequency), que ayuda a reducir la importancia de palabras comunes y resaltar términos más relevantes para la tarea en cuestión. Aunque en la actualidad se prefieren métodos más avanzados, como las representaciones distribuidas, BoW sigue siendo una base esencial en el campo de NLP y continúa siendo utilizada en muchos sistemas debido a su simplicidad y eficiencia (Kherwa y Bansal, 2020).

La construcción de la técnica se realiza:

- **Construcción del vocabulario:** Se construye un vocabulario de todas las palabras únicas que aparecen en el corpus. Cada palabra en el vocabulario se asigna a un índice único.
- **Vectorización:** Cada documento se convierte en un vector en el espacio de características, donde cada dimensión del vector corresponde a una palabra del vocabulario. El valor de cada dimensión es la frecuencia de la palabra correspondiente en el documento.

B. Embeddings

Los embeddings de palabras son una técnica avanzada de representación de texto que ha revolucionado el campo del procesamiento de lenguaje natural (NLP). A diferencia del modelo Bag of Words, que representa las palabras como vectores esparsos y desconectados, los embeddings representan cada palabra como un vector denso en un espacio continuo de alta dimensión. En este espacio, las palabras que comparten similitudes semánticas están ubicadas cerca unas de otras. Por ejemplo, en un embedding, las palabras “rey” y “reina” estarán más

próximas entre sí que "rey" y "mesa", capturando así relaciones de significado de manera mucho más efectiva (Silveira et al., 2021).

Los embeddings se entrenan utilizando grandes corpus de texto, lo que permite que los vectores resultantes reflejen el contexto en el que las palabras aparecen. Modelos como Word2Vec, GloVe y FastText son algunos de los más conocidos en este campo. Estos modelos son capaces de capturar complejas relaciones semánticas y sintácticas entre las palabras. Esta capacidad de generalización ha permitido que los embeddings mejoren significativamente el rendimiento de los modelos de NLP en diversas tareas, como la clasificación de texto, el análisis de sentimientos y la traducción automática.

En los últimos años, los avances en embeddings han ido aún más allá con el desarrollo de modelos de lenguaje preentrenados como BERT, GPT y ELMo. Estos modelos no solo generan embeddings para palabras individuales, sino que también consideran el contexto completo en el que una palabra aparece, lo que les permite capturar matices más profundos del significado. Estos embeddings contextuales han llevado a mejoras significativas en tareas que requieren una comprensión más precisa del lenguaje, como la respuesta a preguntas, el resumen automático y la generación de texto.

B.1 Técnicas comunes de embeddings

- **Word2Vec:** Utiliza redes neuronales para aprender representaciones densas de palabras basadas en su contexto en grandes corpus de texto. Existen dos modelos principales: Continuous Bag of Words (CBOW) y Skip-gram (Mikolov et al., 2013).
- **GloVe:** Global Vectors for Word Representation (GloVe) es una técnica que utiliza estadísticas de co-ocurrencia globales para aprender representaciones de palabras (Pennington et al., 2014).
- **FastText:** Es una técnica de representación de texto desarrollada por Facebook AI Research que mejora las limitaciones de otros

modelos de embeddings como Word2Vec. A diferencia de los modelos tradicionales que representan cada palabra como una única unidad, FastText descompone las palabras en n-gramas de caracteres, lo que permite capturar información morfológica y manejar mejor las palabras raras o desconocidas (out-of-vocabulary). Esta capacidad de representar palabras a nivel de sub-palabra hace que FastText sea especialmente eficaz en lenguajes con rica morfología y en escenarios multilingües, proporcionando embeddings más robustos y precisos para una amplia gama de aplicaciones en procesamiento de lenguaje natural.

- **BERT:** Bidirectional Encoder Representations from Transformers (BERT) utiliza transformadores para capturar el contexto bidireccional de las palabras y ha demostrado un rendimiento superior en una variedad de tareas de NLP (Devlin et al., 2018).

Los embbeding captura relaciones semánticas y contextuales entre palabras, lo que mejora la calidad de las representaciones. Los embeddings pueden ser densos y de menor dimensionalidad comparado con BoW. Sin embargo, requiere grandes corpus de texto para entrenar. Los modelos pre-entrenados pueden ser costosos en términos de recursos computacionales.

El enfoque Bag of Words es útil para tareas básicas y proporciona una representación intuitiva y directa del texto. Sin embargo, embeddings ofrecen una representación más rica y contextualizada, capturando las relaciones semánticas entre palabras y frases. La elección entre estas técnicas depende del problema específico y de los requisitos de la tarea de NLP.

1.1.5 Modelado de tópicos

Los modelos de tópicos ofrecen una forma sencilla de analizar grandes volúmenes de texto sin etiquetar. Un tópico consiste en un grupo de palabras que frecuentemente aparecen juntas (Mifrah y Benlahmar, 2022). El modelado de tópicos es la nueva revolución en la minería de textos, es una técnica para revelar la estructura semántica subyacente en una gran colección de documentos

(Kherwa y Bansal, 2020). El modelado de tópicos es una forma de aprendizaje no supervisado, se utiliza para categorizar los documentos de un corpus en grupos basados en temas semánticamente interpretables. Este método tiene una serie de aplicaciones potenciales en el contexto de la investigación en ciencias sociales (Péter et al., 2022).

Los modelados de temas intentan modelar tres entidades: constructos, colecciones y temas.

- Los constructos son los elementos que se unen para formar una colección. En los datos textuales, los constructos son palabras que se agrupan para constituir un documento o una colección de palabras.
- Un tema es un conjunto de constructos que juntos describen un significado semántico. Matemáticamente, un tema se describe como una distribución de probabilidad sobre los constructos.

A. Asignación latente de dirichlet (LDA)

El modelado de tópicos es una técnica de análisis de texto que permite identificar temas ocultos o tópicos dentro de un gran corpus de documentos. Esta técnica es esencial en el procesamiento de lenguaje natural (NLP) y se utiliza para descubrir la estructura subyacente en conjuntos de datos textuales no estructurados.

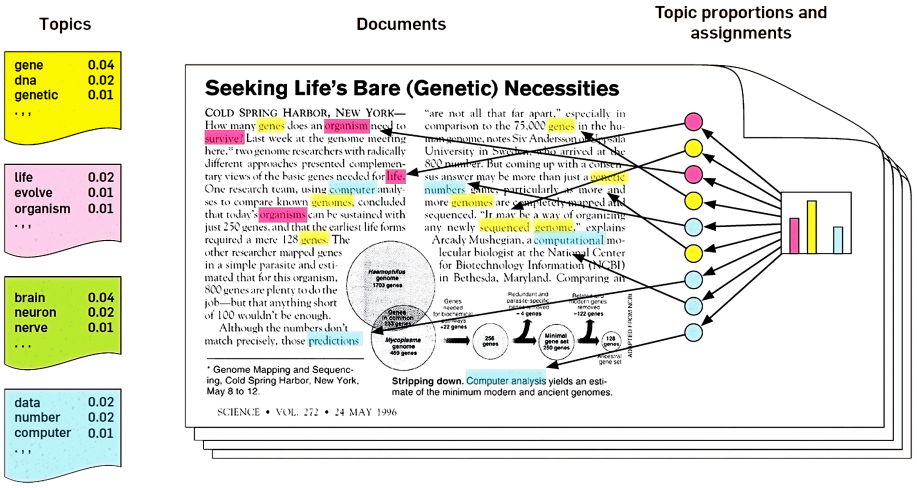
A.1 Intuición

El Latent Dirichlet Allocation (LDA) es uno de los modelos de tópicos más reconocidos y utilizados en este campo. La intuición detrás de LDA es que los documentos exhiben múltiples temas. Por ejemplo, considere el artículo en la Figura 1. Se encuentra resaltado diferentes palabras que se utilizan en el artículo. Las palabras sobre análisis de datos, como “computer” y “prediction”, están resaltadas en azul; las palabras sobre biología evolutiva, como “life” y “organism”, están resaltadas en rosa; las palabras sobre genética, como “sequenced” y “genes”, están resaltadas en amarillo. Si se toma el tiempo de resaltar cada palabra del artículo, se ve que el artículo combina genética, análisis

de datos y biología evolutiva en diferentes proporciones que ayuda a ubicarlo en una colección de artículos científicos (Blei et al., 2010).

Figura 1

Intuición detrás de LDA



Nota. Extraído de Blei et al. (2010).

Un tópico, formalmente se define como una distribución sobre un vocabulario fijo. Por ejemplo, el tema de genética tiene palabras sobre genética con alta probabilidad y el tema de biología evolutiva tiene palabras sobre biología evolutiva con alta probabilidad. Suponemos que estos temas se especifican antes de que se generen los datos. Para cada documento de la colección, generamos las palabras en un proceso de dos pasos.

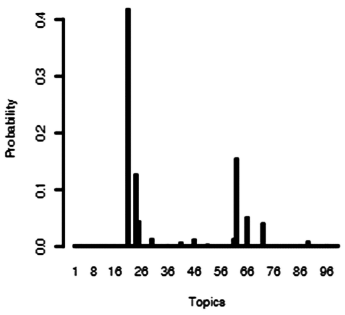
1. Elegimos aleatoriamente una distribución sobre temas.
2. Para cada palabra del documento:
   a) Elegimos aleatoriamente un tema de la distribución sobre temas del paso 1.
   b) Elegimos aleatoriamente una palabra de la distribución correspondiente sobre el vocabulario.

Este modelo estadístico refleja la intuición de que los documentos presentan múltiples temas. Cada documento presenta los temas con diferentes proporciones (paso 1); cada palabra en cada documento se extrae de uno de los temas (paso 2b), donde el tema seleccionado se elige

de la distribución por documento entre los temas (paso 2a). Esta inferencia se ilustra en la Figura 2.

Figura 2

Inferencia real con LDA



"Genetics"	"Evolution"	"Disease"	"Computers"
human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living	infectious	parallel
information	diversity	malaria	methods
genetics	group	parasite	networks
mapping	new	parasites	software
project	two	united	new
sequences	common	tuberculosis	simulations

Nota. Extraído de Blei et al. (2010).

A.2 Fundamentos

LDA asume que los documentos dentro de un corpus son generados de acuerdo a un proceso probabilístico, donde los tópicos son distribuciones sobre palabras. LDA supone que existen K tópicos en el corpus, y cada tópico es una distribución sobre el vocabulario. Cada documento es visto como una mezcla de K tópicos, donde cada palabra en el documento es atribuida a uno de esos tópicos de acuerdo a una distribución probabilística.

El modelo LDA se basa en dos distribuciones Dirichlet:

1. **Distribución Dirichlet sobre tópicos:** Para cada documento, LDA asume una distribución Dirichlet sobre los tópicos.
2. **Distribución Dirichlet sobre palabras:** Para cada tópico, se asume una distribución Dirichlet sobre el vocabulario.

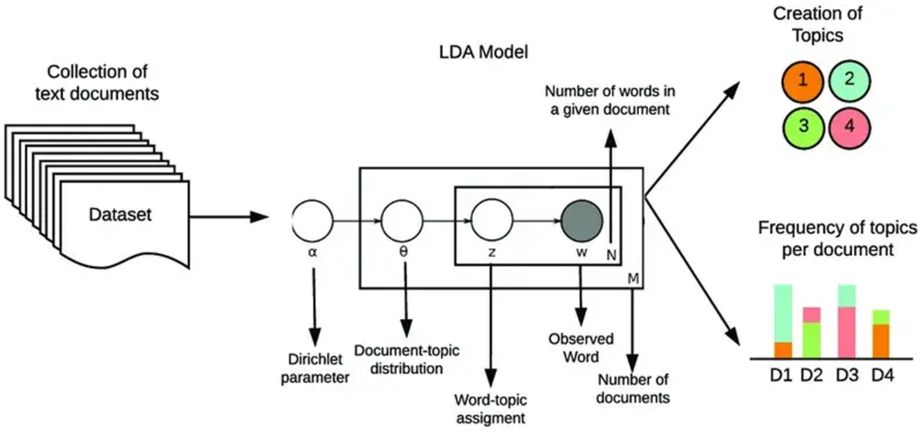
La asignación latente de Dirichlet (LDA), es un modelo probabilístico generativo de un corpus, que consiste en que los documentos se representen como mezclas aleatorias de tópicos latentes, donde un tópico se caracteriza por una distribución de palabras. Para Blei et al. (2010) es uno de los métodos más populares en modelado de

tópicos. LDA representa tópicos mediante probabilidades de palabras. Las palabras con mayores probabilidades en cada tópico suelen dar una buena idea de cuál es la palabra del tópico.

Es una técnica de modelado de tópicos que asume que los documentos son mezclas de una serie de tópicos y que cada tópico está caracterizado por una distribución de palabras. El objetivo de LDA es descubrir estos tópicos ocultos en un conjunto de documentos como se muestra en la Figura 3.

Figura 3

Modelado de tópicos basado en LDA



Nota. Extraído de Buenaño-Fernandez et al. (2020).

Los tópicos son $\beta_{1:k}$, donde cada β_k es una distribución sobre el vocabulario. Las proporciones de tópicos del *dth* el documento es θ_d, donde $\theta_{d,k}$ es la proporción de tópicos para el tema k en el documento d. Las asignaciones de tópicos para el documento d es z_d, donde $z_{d,n}$ es la asignación de tópicos para la palabra n-ésima en el documento d. Finalmente, las palabras observadas para el documento d son w_d, donde $w_{d,n}$ es la palabra n en el documento d, que es un elemento del vocabulario fijo.

LDA es ampliamente utilizado en diversas áreas, incluyendo la minería de textos, análisis de sentimientos, clasificación de documentos, y recomendación de contenido. Según Blei et al. (2010) LDA permite identificar tópicos ocultos en grandes colecciones de textos, lo que lo convierte en una herramienta poderosa para la exploración de datos

textuales.

B. BERTopic

BERTopic es una herramienta avanzada para la identificación de tópicos en grandes corpus de texto, que combina técnicas modernas de procesamiento de lenguaje natural con métodos de reducción dimensional y clustering. Los modelos de tópicos pueden ser herramientas útiles para descubrir temas latentes en colecciones de documentos. BERTopic, un modelo de temas que extiende este proceso mediante la extracción de una representación coherente de temas a través del desarrollo de una variación basada en clases de TF-IDF. Más específicamente, BERTopic genera incrustaciones de documentos con modelos de lenguaje basados en transformadores entrenados previamente, agrupa estas incrustaciones y, finalmente, genera representaciones de temas con el procedimiento TF-IDF basado en clases. BERTopic genera temas coherentes y sigue siendo competitivo en una variedad de puntos de referencia que involucran modelos clásicos y aquellos que siguen el enfoque de agrupamiento más reciente del modelado de temas.

Una de las innovaciones clave que BERTopic incorpora es el uso de *embeddings* contextuales para representar textos. Modelos de lenguaje como BERT (Bidirectional Encoder Representations from Transformers), introducidos por Devlin et al. (2018) han revolucionado el procesamiento de lenguaje natural al proporcionar representaciones más ricas y contextuales de las palabras. BERT utiliza una arquitectura de transformadores para capturar la relación entre las palabras en un contexto bidireccional, mejorando significativamente la calidad de los *embeddings*.

B.1 Reducción dimensional y clustering

BERTopic utiliza técnicas de reducción dimensional para simplificar las representaciones vectoriales de alto nivel generadas por BERT. UMAP (Uniform Manifold Approximation and Projection),

desarrollado por McInnes et al. (2018) es una técnica de reducción dimensional no lineal que preserva la estructura global y local de los datos. UMAP es eficaz para la visualización de datos y la preparación para técnicas de clustering. El clustering en BERTopic se realiza utilizando HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) extiende el enfoque DBSCAN (Density-Based Spatial Clustering of Applications with Noise) al considerar una jerarquía de clusters y permite identificar clusters de forma jerárquica basada en la densidad de los datos.

B.2 Extracción e interpretación de tópicos

La extracción de tópicos en BERTopic se basa en la representación de los clústeres generados a partir de los embeddings y la reducción dimensional. Cada clúster representa un tópico, y los términos más representativos de cada clúster se extraen para describir el tópico. Esto se hace mediante el análisis de palabras clave que mejor representan el contenido del clúster (Rudolph y Feldman, 2018).

B.3 Pasos para la implementación de BERTopic

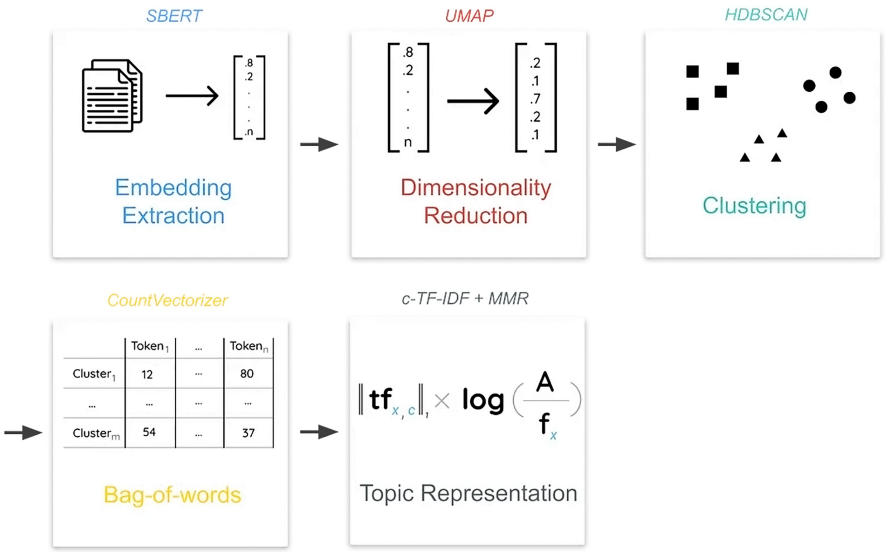
BERTopic produce representaciones temáticas a través de cinco pasos como se muestra en la Figura 4.

1. Cada documento se convierte a su representación incrustada utilizando un modelo de lenguaje previamente entrenado.
2. La dimensionalidad de las incrustaciones resultantes se reduce para optimizar el proceso de agrupación a través de UMAP.
3. El clustering se puede implementar mediante HDBSCAN.
4. La tokenización es a través de librerías como *Count vectorizer*.
5. De los grupos de documentos, las representaciones temáticas se extraen utilizando una variación basada en clases de TF-IDF (c-TF-IDF). El procedimiento c-TF-IDF modela la importancia de las palabras en clústeres en lugar de documentos individuales, lo que permite generar distribuciones de tópicos y palabras para cada

grupo de documentos.

Figura 4

Pasos del modelo de tópicos BERTopic



Nota. Extraído de Grootendorst (2022).

BERTopic representa un avance significativo en el modelado de tópicos al integrar embeddings contextuales, reducción dimensional no lineal y clustering jerárquico. Al aprovechar la representación rica de BERT y la capacidad de UMAP y HDBSCAN para manejar datos complejos, BERTopic ofrece una metodología robusta para la identificación y análisis de tópicos en grandes conjuntos de datos textuales. La combinación de estas técnicas proporciona una visión más precisa y contextualizada de los temas subyacentes en los documentos, superando las limitaciones de los enfoques tradicionales como LDA.

C. Métricas de evaluación de modelos de tópicos

Los modelos de tópicos se pueden aplicar en varios dominios de aplicación. Por lo tanto, pueden evaluarse extrínsecamente según su desempeño en el dominio donde se aplican. También pueden evaluarse intrínsecamente considerando los propios temas generados. La evaluación intrínseca es independiente de cualquier dominio específico y, por tanto, es más general. Los diferentes modelos difieren en simplicidad, eficiencia computacional y supuestos de modelado. En consecuencia, difieren en

cómo se desempeñan en diferentes corpus y diferentes aplicaciones.

C.1 Coherencia

La métrica de Coherencia permite evaluar la calidad semántica de los tópicos individuales. La Coherencia se considera como la manera de medir qué tan relacionadas entre sí están las palabras que integran el tópico y si claramente tiene significado semántico. En términos matemáticos la Coherencia se representa en la ecuación:

$$Coherencia(W) = \sum_{w1,w2 \in W} log\left(\frac{D(w1,w2)+\varepsilon}{D(w2)}\right) \quad (1)$$

Donde:

D(w) y ***D(w1 y w2)*:** son el número de documentos con al menos una instancia de *w*, y de *w1* y *w2*, respectivamente.

ε: es una constante para evitar encontrar un logaritmo de cero.

La métrica de coherencia busca una alta co-ocurrencia de las palabras que integran el tópico, significa que están relacionadas semánticamente.

C.2 Perplejidad

La métrica de perplejidad se utiliza para evaluar qué tan bien el modelo de tópicos explica un conjunto de documentos. Un valor bajo de perplejidad indica que el modelo de temas es bueno para explicar los documentos, lo que sugiere que ha identificado temas coherentes y representativos.

En términos matemáticos, la perplejidad mide la incertidumbre que un modelo tiene sobre los datos. Específicamente, se define como la exponencial del negativo de la probabilidad logarítmica promedio:

$$Perplejidad(D) = exp(-\frac{1}{N}\sum_{i=1}^{N} logP(w_i)) \quad (2)$$

Donde:

D: es el conjunto de datos.

N: es el número total de palabras en el conjunto de datos.

w_i**:** representa la i-ésima palabra en el conjunto de datos.

$P(w_i)$**:** es la probabilidad asignada a la palabra w_i por el modelo.

C.3 Diferenciación de tópicos

Esta métrica mide la diferenciación de tópicos a través de la distancia de coseno entre los vectores de palabras que representan diferentes temas. La distancia de coseno entre dos vectores se define como:

$$Distancia\ de\ coseno = 1 - \frac{v1.v2}{\|v1\|\,\|v2\|} \tag{3}$$

Si la distancia de coseno entre dos temas es cercana a 1, los temas son muy diferentes entre sí, lo que indica una buena diferenciación.

D. Modelado de tópicos en documentos normativos

El derecho computacional aplica métodos computacionales cuantitativos al estudio de las leyes, combinando varios campos de investigación como el procesamiento del lenguaje natural, la ciencia de datos, el aprendizaje automático y los métodos estadísticos (Ashihara et al., 2020).

En el contexto de documentos normativos, que a menudo abordan aspectos legales y regulatorios, el modelado de tópicos se convierte en una herramienta esencial para identificar y comprender los temas críticos que abordan estas normativas (Grajzl y Murrell, 2022).

1.1.6 Herramientas de modelado de tópicos

En este apartado se presentan las bibliotecas ampliamente utilizadas en el procesamiento de lenguaje natural y la selección entre ellas puede depender de la

tarea específica que estés abordando.

A. Beautiful soup (BS4)

Para Richardson (2023) es una biblioteca de Python ampliamente utilizada para el análisis y la manipulación de documentos HTML y XML. Su principal función es facilitar la extracción de datos de páginas web cuyas características principales se presentan a continuación:

- **Intuitiva:** BS4 proporciona una interfaz sencilla y natural para interactuar con el contenido HTML o XML, lo que hace que el proceso de scraping sea más accesible incluso para quienes no tienen una experiencia profunda en programación.
- **Compatibilidad con distintos parsers:** BS4 es compatible con diferentes parsers de HTML, como html.parser, lxml y html5lib.
- **Navegación en el árbol del documento:** Permite buscar elementos específicos del documento usando métodos como find() y find_all(), que permiten localizar etiquetas HTML, clases, identificadores (ID), y otros atributos de manera eficiente. BS4 facilita la modificación del árbol del documento, permitiendo a los usuarios agregar, eliminar o alterar nodos y sus atributos. Esto es útil para limpiar y transformar datos antes de su almacenamiento o análisis.
- **Manejo de HTML malformado:** BS4 puede analizar y manejar documentos HTML que no están bien formados, corrigiendo automáticamente errores comunes para permitir la extracción de datos.
- **Scraping web automatizado:** BS4 se utiliza frecuentemente junto con la biblioteca requests para automatizar el proceso de scraping. Requests se encarga de obtener el contenido de las páginas web, mientras que BS4 se ocupa de analizar y extraer la información deseada.
- **Extracción de datos:** BS4 permite seleccionar elementos usando selectores CSS (select()), proporcionando una forma adicional y más

flexible de acceder a los elementos del documento.

B. OCR (Reconocimiento óptico de caracteres)

Es una herramienta tecnológica que permite convertir texto impreso o manuscrito en un formato digital editable. Esta tecnología funciona mediante el escaneo de documentos físicos y la aplicación de algoritmos que identifican y extraen los caracteres presentes en las imágenes. OCR es utilizado en la digitalización de documentos, facilitando la transformación de documentos en papel en textos que pueden ser editados, buscados y almacenados de manera más eficiente. Esta capacidad es necesaria para la gestión de documentos, ya que permite a las organizaciones preservar registros históricos y mejorar el acceso a información almacenada en formatos físicos. La herramienta OCR ha avanzado significativamente con la integración de técnicas de aprendizaje automático y redes neuronales, lo que ha mejorado su precisión y capacidad para manejar diferentes tipos de documentos y caligrafías.

C. spaCy

Es una biblioteca de código abierto para procesamiento de lenguaje natural (NLP) en Python que está diseñado para ser un toolkit moderno y eficiente para el procesamiento de lenguaje natural (NLP). Está optimizado para la producción y se enfoca en la velocidad y el rendimiento, permitiendo manejar grandes volúmenes de texto de manera eficiente, mientras que NLTK es una biblioteca orientada al ámbito académico, diseñada inicialmente como una herramienta educativa para aprender y enseñar procesamiento de lenguaje natural. En la Tabla 1, podemos ver una comparativa entre éstas dos herramientas (Honnibal, 2015).

D. Regex (expresiones regulares)

Es una técnica potente utilizada para la manipulación y análisis de texto. Las expresiones regulares son patrones que permiten buscar, extraer

Tabla 1

Comparación de las bibliotecas NLTK y spaCy para el procesamiento de lenguaje natural

Característica	NLTK	spaCy
Facilidad de uso	Compleja, más académica, requiere más configuración.	Intuitiva, diseñada para uso práctico e industrial.
Velocidad	Intuitiva, diseñada para uso práctico e industrial.	Muy rápida, optimizada para producción.
Modelos incorporados	Diversidad de herramientas, pero modelos más básicos.	Modelos pre entrenados de alta calidad para varios idiomas.
Lematización	Necesita configuración adicional (p. ej., WordNet).	Muy sencilla, con modelos pre entrenados.
Tokenización	Flexible, pero puede requerir personalización.	Precisa y optimizada para varios idiomas.
Stopwords	Soporte básico, requiere listas propias para algunos idiomas.	Incorporadas y adaptadas a cada idioma.
NER (Reconocimiento de entidades)	Funciona, pero menos precisa en comparación con spaCy.	Muy precisa y con soporte para múltiples idiomas.
Soporte de Idiomas	Amplio, pero con diferentes niveles de calidad.	Modelos de alta calidad para idiomas clave.
Integración con otras bibliotecas	Bien integrable, pero necesita configuración.	Integración fácil con librerías modernas como TensorFlow, PyTorch.
Comunidad y Documentación	Gran comunidad, documentación extensa pero más académica.	Comunidad en crecimiento, documentación clara y orientada a la industria.

y modificar texto de manera flexible y eficiente. En el contexto de NLP, Regex se emplea para tareas como la extracción de entidades, la limpieza de datos y la normalización del texto. Por ejemplo, se pueden utilizar

expresiones regulares para identificar y extraer direcciones de correo electrónico, números de teléfono o fechas en un corpus de texto. Su capacidad para definir patrones precisos y complejos hace que sea una herramienta valiosa para el pre procesamiento de texto, facilitando la preparación de datos para modelos de lenguaje y otras técnicas de análisis de texto (Murakami y Chakraborty, 2022). Sin embargo, a pesar de su potencia, las expresiones regulares tienen limitaciones en el análisis de lenguaje natural debido a su naturaleza basada en patrones estrictos. No son adecuadas para capturar variaciones lingüísticas complejas o contextos más amplios en los que el significado de las palabras puede cambiar. Por ejemplo, Regex puede tener dificultades para manejar ambigüedades o contextos semánticos profundos que los modelos de lenguaje más avanzados pueden entender mejor. A pesar de estas limitaciones, NLP Regex sigue siendo una herramienta útil cuando se necesita realizar tareas específicas de búsqueda o extracción en texto, complementando otros métodos de procesamiento de lenguaje natural.

E. Gensim

Es una biblioteca de procesamiento de lenguaje natural (NLP) en Python, diseñada para modelar y analizar grandes corpus de texto de manera eficiente. Fundada por Rita M. Schiavi y desarrollada inicialmente por Radim Řehůřek, Gensim se destaca por su capacidad para realizar modelado de tópicos, análisis de similitud y extracción de información utilizando técnicas avanzadas de aprendizaje automático. Entre sus características principales se incluyen la implementación de modelos de tópicos como Latent Dirichlet Allocation (LDA) y Latent Semantic Analysis (LSA), así como herramientas para la representación vectorial de palabras mediante Word2Vec, FastText y Doc2Vec. Gensim es ampliamente apreciada por su eficiencia en el manejo de grandes volúmenes de datos textuales y su enfoque en la escalabilidad, lo que la convierte en una herramienta valiosa para investigadores y profesionales en el campo de la minería de textos y el procesamiento del lenguaje

natural. Su diseño modular y la capacidad de manejar datos de manera eficiente hacen que Gensim sea una opción popular para proyectos de NLP que requieren un procesamiento extensivo y análisis de datos textuales.

1.1.7 Modelado de tópicos en normativas

El modelado de tópicos en el análisis de normativas es una herramienta clave para la organización y comprensión de grandes volúmenes de documentos normativos, como leyes, reglamentos y políticas. Dado que las normativas suelen estar redactadas en un lenguaje técnico y especializado, el modelado de tópicos ayuda a identificar y extraer temas recurrentes y relevantes dentro de estos documentos. Esta técnica permite clasificar y agrupar normativas relacionadas, facilitando la identificación de áreas temáticas dominantes y el análisis de cómo se abordan diferentes aspectos legales y regulatorios en un contexto específico.

Utilizando métodos como Latent Dirichlet Allocation (LDA) o técnicas más avanzadas como BERTopic, los analistas pueden descomponer textos normativos extensos en tópicos coherentes que reflejan las principales áreas de interés y preocupación. Por ejemplo, LDA puede identificar tópicos relacionados con aspectos de seguridad, cumplimiento ambiental o regulación financiera dentro de un conjunto de documentos. Sin embargo, dado que las normativas pueden contener un lenguaje altamente especializado, las técnicas basadas en embeddings de palabras, como las utilizadas en BERTopic, pueden ofrecer una representación más matizada y contextualmente precisa de los temas presentes (Grisales-Aguirre y Figueroa-Vallejo, 2022).

El modelado de tópicos en normativas también facilita la exploración de cambios y tendencias a lo largo del tiempo. Al aplicar estas técnicas a versiones históricas de documentos normativos, los investigadores pueden rastrear cómo los temas y enfoques han evolucionado, proporcionando una visión de la dinámica de las políticas y la adaptación a nuevas circunstancias o desafíos. Esto es particularmente útil en el análisis de reformas legislativas o en la evaluación de la implementación de nuevas regulaciones a lo largo del tiempo. El análisis de tópicos puede mejorar la accesibilidad y comprensión de la normativa para

profesionales y ciudadanos. Al agrupar y simplificar la información en temas clave, el modelado de tópicos facilita la búsqueda de información específica y la navegación por documentos complejos. Esto puede ser especialmente valioso en contextos como la asesoría legal, la educación en políticas públicas o la gestión de cumplimiento normativo, donde una comprensión clara y estructurada de la normativa es esencial para la toma de decisiones informadas y el cumplimiento efectivo (Grisales-Aguirre y Figueroa-Vallejo, 2022).

1.1.8 Políticas públicas en Perú

Las poíticas públicas en Perú desempeñan un papel crucial en la configuración del desarrollo económico y social del país; desde la década de 1990, Perú ha experimentado una serie de reformas políticas y económicas diseñadas para enfrentar los desafíos de crecimiento y estabilidad.

En el ámbito social, las políticas públicas en Perú han abordado diversas áreas como educación, salud y reducción de la pobreza. Gonzales de Olarte (2003) destaca que el gobierno peruano ha implementado programas significativos para mejorar el acceso a la educación y la calidad del sistema educativo, incluyendo la expansión de la cobertura educativa y la mejora de la infraestructura escolar. A pesar de estos avances, persisten desafíos relacionados con la desigualdad en el acceso y la calidad de la educación, especialmente en las regiones más rurales y desfavorecidas del país.

En cuanto a políticas de salud, el Perú ha implementado reformas para fortalecer su sistema de salud pública, incluyendo la expansión de la cobertura de seguros de salud y la mejora de la infraestructura hospitalaria. A pesar de estos esfuerzos, el sistema de salud enfrenta retos significativos como la desigualdad en el acceso a servicios de calidad y la necesidad de una mayor inversión en atención primaria. Estos desafíos han llevado a una constante evaluación y ajuste de las políticas para asegurar una mejor cobertura y equidad en la atención sanitaria (Mendoza y García, 2010).

En cuanto a las políticas de desarrollo regional han sido una prioridad en el contexto de la descentralización administrativa promovida por el gobierno

peruano. Según Aráoz (2015) la implementación de políticas para el desarrollo regional busca mejorar la infraestructura, fomentar el crecimiento económico y reducir las disparidades entre las regiones. Estas políticas incluyen programas de inversión en infraestructura, promoción de la inversión privada y apoyo a las pequeñas y medianas empresas. Sin embargo, Aráoz también señala que la efectividad de estas políticas varía considerablemente entre regiones, y es necesario un enfoque más coordinado y efectivo para lograr un desarrollo equitativo en todo el país.

A. Ordenanzas Regionales

Las ordenanzas regionales son normas jurídicas emitidas por las autoridades regionales para regular diversos aspectos de la vida en una región específica. Estas ordenanzas son fundamentales para la administración y el desarrollo local, ya que permiten a los gobiernos regionales adaptar las leyes y regulaciones a las necesidades y particularidades de su jurisdicción. A diferencia de las leyes nacionales, que tienen un alcance general, las ordenanzas regionales se centran en cuestiones locales, abordando temas como la planificación urbana, el medio ambiente, la salud pública y la educación, entre otros. Su propósito es asegurar que las políticas y servicios se ajusten a las características y demandas de la población regional (Lange et al., 2021).

El proceso de formulación de ordenanzas regionales suele involucrar la participación de diversos actores, incluidos funcionarios gubernamentales, expertos en la materia y ciudadanos. Este proceso participativo es crucial para garantizar que las ordenanzas respondan a las necesidades reales de la comunidad y cuenten con el respaldo y la aceptación de quienes se verán afectados por ellas. Las ordenanzas pueden ser propuestas por el ejecutivo regional, revisadas y aprobadas por el consejo regional o asamblea, y finalmente promulgadas por el gobierno regional. Este enfoque democrático y consultivo asegura que las decisiones sean informadas y representativas de los intereses locales.

Una característica importante de las ordenanzas regionales es su

capacidad para ser modificadas o actualizadas en respuesta a cambios en las condiciones locales o en las necesidades de la comunidad. A medida que las regiones evolucionan y enfrentan nuevos desafíos, es esencial que las normas y regulaciones también se ajusten para seguir siendo relevantes y efectivas. Los mecanismos de revisión y actualización permiten que las ordenanzas regionales se mantengan al día con los desarrollos sociales, económicos y ambientales, garantizando que sigan cumpliendo su función de manera adecuada y oportuna. Además, las ordenanzas regionales juegan un papel clave en la implementación de políticas públicas y la promoción del desarrollo regional. Al establecer normas específicas y directrices para diversos aspectos de la vida regional, estas ordenanzas contribuyen a crear un entorno ordenado y predecible en el que se pueden llevar a cabo proyectos y actividades. Esto no solo mejora la calidad de vida de los habitantes, sino que también fomenta la inversión y el crecimiento económico al proporcionar un marco normativo claro y estable. En resumen, las ordenanzas regionales son instrumentos esenciales para la gobernanza local y el desarrollo sostenible de las regiones (Gobierno Regional Puno, 2024).

A.1 Características de las Ordenanzas Regionales

Las características específicas de las ordenanzas regionales pueden variar según el país y su estructura política y administrativa. Algunos ejemplos de áreas que pueden ser reguladas por ordenanzas regionales incluyen:

- **Ordenación del Territorio:** Regulación de la planificación urbana, uso del suelo, construcción de infraestructuras y desarrollo regional.
- **Ambiente:** Normativas relacionadas con la protección del medio ambiente, conservación de recursos naturales y gestión de residuos.
- **Cultura y Educación:** Regulación de actividades culturales, promoción de eventos regionales, y normativas relacionadas con la educación en el ámbito regional.
- **Transporte y Tráfico:** Regulación del transporte público y privado,

control del tráfico, y normas relacionadas con la movilidad en la región.

- **Salud Pública:** Normativas relacionadas con la salud, seguridad y bienestar de la población en la región.
- **Comercio y Economía:** Regulación de actividades comerciales y económicas específicas de la región.

Las ordenanzas regionales no pueden entrar en conflicto con leyes nacionales o constitucionales, ya que están subordinadas a la legislación de nivel nacional. Sin embargo, tienen un papel significativo en adaptar y aplicar normativas a las características y necesidades específicas de una región determinada (Reyes, 2017).

1.2 Antecedentes

Se presenta el estado del arte en relación al modelado de tópicos en el contexto de normativas sociales, entre ellos destacan diversos artículos en el ámbito internacional, nacional y local:

1.2.1 Internacionales

Péter et al. (2022) en su investigación tuvo como objetivo proporcionar una visión general de las oportunidades y limitaciones del modelado de tópicos dentro de un contexto de ciencias sociales llegando a la conclusión que las categorías generadas por los modelos de tópicos son semánticamente interpretables y son relevantes para posibles estudios adicionales en corpus dentro del área de ciencias sociales, pero se limitan a procesamientos preliminares de estructuración de datos.

Dyevre (2021) en su trabajo de investigación tuvo como objetivo explorar un corpus que compila más de 200.000 actos legislativos, 55.000 sentencias y opiniones judiciales y 4.000 artículos de una importante revista jurídica de la Unión Europea para encontrar prioridades legales y políticas, al aplicar una técnica de aprendizaje automático no supervisado conocida como modelado de tópicos, arribando a la conclusión que la integración económica sigue siendo el foco de la legislación de la Unión Europea, pero que los académicos tienden a enfatizar más las cuestiones de derechos e ignorar ciertos temas, como las

regulaciones agrícolas.

Grajzl y Murrell (2022) en su artículo científico tuvieron como objetivo transmitir a los historiadores del derecho tradicional el papel que estas nuevas técnicas computacionales pueden desempeñar en la investigación histórico-jurídica en su investigación utilizaron la técnica de modelado de tópicos en historia jurídica, con una aplicación al caso de Ley de Finanzas; los autores llegaron a la conclusión de que el modelado de tópicos, una técnica de aprendizaje automático no supervisado para el análisis de grandes corpus, puede ser una herramienta poderosa para la investigación histórico-jurídica.

Wendel et al. (2022) realizaron su investigación en el análisis de grandes corpus de textos legales con métodos de minería de textos, en la investigación emplearon el modelado de tópicos, cuyo objetivo fue recuperar los tópicos de un corpus, para identificar palabras relacionadas con ciertas áreas del derecho presentes en la jurisprudencia del Tribunal Constitucional Federal de Alemania (FCC), los autores llegaron a la conclusión que las áreas técnicas y algo inestables del derecho, como el derecho tributario, el derecho social y el derecho de la función pública, están significativamente sobrerrepresentadas en las remisiones para revisión judicial, mientras que las áreas del derecho caracterizadas por una jurisprudencia y una doctrina judicial bien desarrolladas aparecen con mucha más frecuencia en los tribunales constitucionales.

Grisales-Aguirre y Figueroa-Vallejo (2022) en su investigación tuvieron como objetivo analizar el papel del aprendizaje automático de datos en las revisiones sistemáticas de la literatura. Se aplicó la técnica de Procesamiento de Lenguaje Natural denominada modelado de tópicos, a un conjunto de títulos y resúmenes recopilados de la base de datos Scopus. En una de sus conclusiones mencionan que el uso de la técnica LDA permitió identificar los autores y revistas más relevantes en lo referente a la revisión de literatura científica.

Murshed et al. (2023) en su investigación tiene como objetivo examinar el estado actual del arte en algoritmos STTM (Short Text Topic Modeling). Presenta un estudio completo y una taxonomía de los algoritmos STTM para el modelado de tópicos de textos breves, llegando a la conclusión que la revisión de las

técnicas de modelado de temas de texto breve (STTM) abarcó tres categorías amplias de métodos: basados en DMM, basados en coocurrencias globales de palabras y basados en auto agregación.

Lee et al. (2023) se plantearon como objetivo explorar los datos desde una perspectiva de minería de textos utilizando representaciones de codificador bidireccional del tema de transformadores (BERTopic), una técnica de modelado de tópicos de última generación, arribaron a la conclusión que es apropiado aplicar datos de noticias y artículos para comprender la conciencia pública y el interés académico. Además, ESG no recibió atención mediática inmediatamente después de ser presentado por instituciones globales.

Dillan y Fudholi (2023) en su estudio tienen como objetivo desarrollar un software independiente del idioma que agilice el proceso de identificación de investigaciones de vanguardia en diversos tópicos académicos, utilizando asignación latente de Dirichlet (LDA) y representaciones de codificador bidireccional de transformadores (BERT) para descubrir y etiquetar temas automáticamente, llegando a la conclusión que el estudio demuestra que la integración de LDA, BERT y el método de etiquetado y filtrado de tópicos propuesto produce una herramienta sólida para el análisis de investigación preliminar con alta precisión y relevancia.

Kherwa y Bansal (2020) en su investigación tuvieron como objetivo proponer modelado de temas semántico N-Gram propuesto se compara con la colocación de asignación de Dirichlet latente (coll-LDA) y la técnica de modelado de temas de última generación más apropiada. Los autores arribaron a la conclusión que la perplejidad mejora drásticamente y se encontró una mejora significativa en la puntuación de coherencia específicamente para conjuntos de datos de texto breves, como reseñas de películas y blogs políticos.

Qiang et al. (2022) en su investigación tuvieron como objetivo realizar una revisión exhaustiva de varias técnicas de modelado de tópicos de textos breves propuestas en la literatura, llegando a las conclusiones que los modelos de tópicos se clasifican en cuatro categorías según sus técnicas de modelado. Las técnicas de modelado son algebraicas, difusas, probabilísticas y neuronales. Los diferentes

modelos y categorías de modelos tienen diferentes características y coexisten para servir en diferentes contextos y en diferentes características de corpus.

Vayansky y Kumar (2020) tuvieron como objetivo presentar diferentes enfoques de modelado de tópicos capaces de abordar la correlación entre tópicos, los cambios de temas a lo largo del tiempo, así como la capacidad de manejar textos breves como los que se encuentran en las redes sociales o datos de texto dispersos. También revisan brevemente los algoritmos que se utilizan para optimizar e inferir parámetros en el modelado de tópicos. Los autores llegaron a la conclusión de que el modelado de tópicos es una herramienta analítica popular para evaluar datos. Se han desarrollado numerosos métodos de modelado de temas que consideran muchos tipos de relaciones y restricciones dentro de conjuntos de datos; sin embargo, estos métodos no se emplean con frecuencia. En cambio, muchos investigadores gravitan hacia el análisis latente de Dirichlet, que, aunque flexible y adaptable, no siempre es adecuado para modelar relaciones de datos más complejas.

Jelodar et al. (2019) tuvieron como objetivo descubrir el desarrollo de la investigación, las tendencias actuales y la estructura intelectual del modelado de temas. En su investigación presentan artículos altamente académicos (entre 2003 y 2016) relacionados con el modelado de temas basado en LDA. Los autores arribaron a la conclusión que los modelos de tópicos tienen un papel importante en la informática para la minería de textos. Los modelos de tópicos no pueden comprender los medios y conceptos de las palabras en documentos de texto para el modelado de temas. En cambio, suponen que cualquier parte del texto se combina seleccionando palabras de probables cestas de palabras donde cada cesta corresponde a un tema.

Abdelrazek et al. (2023) en su artículo tiene como objetivo analizar la amplia variedad de modelos disponibles de cada categoría, destacamos las diferencias y similitudes entre modelos y categorías de modelos utilizando una perspectiva unificada. Los autores llegaron a la conclusión que ningún modelo logra los mejores resultados en todos los criterios y se debe considerar cuatro aspectos clave de la naturaleza de una aplicación para ayudar a determinar qué

modelo de tema utilizar. Evaluar un modelo temático es un desafío.

Por su parte Chauhan y Shah (2022) tuvieron como objetivo analizar los antecedentes y avances de las técnicas de modelado de tópicos así como las técnicas de implementación y evaluación de modelado de tópicos. Los autores llegaron a la conclusión que el modelado de tópicos es una técnica poderosa en la minería de texto para descubrir relaciones entre datos y documentos de texto. Se aplica en campos como la ingeniería de software, la ciencia política, la medicina y la lingüística.

Lange et al. (2021) tuvieron como objetivo analizar un corpus representativo de obras de derecho sustantivo islámico desde los inicios de la jurisprudencia jurídica islámica hasta el período moderno temprano utilizando modelado de tópicos. De los resultados obtenidos llegaron a la conclusión de que se enfatizan en el derecho público y el comercio, y se muestran un gran interés en el derecho de herencia.

Rawat et al. (2022) en su investigación tuvieron como objetivo comprender el corpus de sentencias legales relacionadas con casos bajo la Ley de Matrimonio Hindú de la India. El estudio examinó varios métodos para generar incrustaciones de oraciones a partir de la sentencia. En una de las conclusiones se sostiene el poder del algoritmo BERTopic para generar tópicos importantes en documentos legales.

Silveira et al. (2021) se plantearon como objetivo proporcionar una visión general de las oportunidades y limitaciones del modelado de tópicos dentro de un contexto de ciencias sociales, a través de un ejemplo de investigación concreto, que aplica el modelado temático a un corpus que consta de leyes húngaras de 1990 a 2018. La investigación concluye en el uso de BERTopic para construir modelos temáticos de documentos legales para proporcionar información sobre las leyes mencionadas en el documento. A partir de una valoración cualitativa, el enfoque revela temas coherentes con la temática del documento.

Martino et al. (2022) tuvieron como objetivo proponer un método novedoso de modelado de tópico llamado PRILJ, que identifica regularidades de párrafos en sentencias de casos legales. Los autores llegaron a la conclusión que

el método propuesto permite apoyar a los expertos legales durante la redacción de documentos legales.

Dyevre et al. (2021) tuvo como objetivo analizar una variedad de técnicas novedosas en aprendizaje automático y procesamiento del lenguaje natural como modelado de tópicos a gran escala de textos legales. Se llegó a la conclusión que la minería de textos y el procesamiento del lenguaje natural han avanzado mucho en los últimos años. Algunas de estas técnicas están en el centro de la tan publicitada "revolución de la IA" y están impulsando el desarrollo de la tecnología legal.

1.2.2 Nacionales

Halgekar et al. (2023) en su investigación tuvieron como objetivo proponer un enfoque único para agrupar estos documentos utilizando el algoritmo k-means en incrustaciones de oraciones dimensionalmente reducidas generadas con el uso de DistilBERT y UMAP. Los autores llegaron a la conclusión que el enfoque propuesto al compararlos con enfoques de agrupación y modelado de tópicos de última generación, los ha superado.

Gamboa Unsihuay (2019) en su investigación tuvo como objetivo determinar temas abordados por los distintos grupos de la clase política peruana a través de los análisis de contenidos por sus miembros en sus cuentas de Twitter. El autor llegó a la conclusión que las tres cuartas partes de los contenidos textuales se refieren a la gestión del Poder Ejecutivo y Legislativo.

1.2.3 Locales

Rodríguez Urquiaga (2018) en su investigación tuvo como objetivo visualizar la evolución temática de corpus de documentos usando LDA. El autor llegó a la conclusión de que usando la técnica LDA permitió visualizar de temas organizados por similitud de contenido y temporal, la comparación entre tópicos.

CAPÍTULO II

PLANTEAMIENTO DEL PROBLEMA

2.1 Identificación del problema

En la era de la información, el acceso transparente y comprensible a los datos se ha vuelto esencial para fomentar la participación ciudadana y promover la rendición de cuentas en la gestión pública. Aunque existen numerosas fuentes de información, la falta de herramientas efectivas de análisis de datos puede obstaculizar la capacidad de la ciudadanía para comprender y aprovechar plenamente la información disponible (Dyevre, 2021).

En el ámbito de la gestión y análisis de documentos normativos, se ha evidenciado la necesidad de desarrollar herramientas y metodologías eficientes para el modelado de tópicos (Rawat et al., 2022). Los documentos normativos, que incluyen leyes, reglamentos y políticas, contienen información crucial para diversos sectores, pero su complejidad y extensión dificultan la identificación y comprensión de los tópicos específicos abordados en ellos (Silveira et al., 2021).

La información gubernamental y pública del Gobierno Regional Puno se encuentra en su portal institucional (Gobierno Regional Puno, 2024), a menudo se encuentra formatos escaneados, dificultando su interpretación y análisis por parte de la ciudadanía. La falta de acceso fácil y visualización efectiva de esta información puede conducir a una brecha en la participación ciudadana y limitar la capacidad de la sociedad para tomar decisiones informadas.

Por otro lado, el modelado de tópicos es una de las técnicas más poderosas de procesamiento de lenguaje natural, el descubrimiento de datos latentes y la búsqueda de relaciones entre datos y documentos de texto. Los investigadores han publicado muchos artículos en el campo del modelado de tópicos y los han aplicado en diversos campos, como la ingeniería de software, las ciencias políticas, las ciencias médicas y lingüísticas, etc. (Jelodar et al., 2019).

En el ámbito de la gestión y análisis de documentos normativos, se ha evidenciado la necesidad de desarrollar herramientas y metodologías eficientes para el modelado de tópicos.

En este contexto, la falta de aplicación de técnicas de modelado de tópicos puede generar obstáculos en la interpretación de la información contenida en los documentos normativos, limitando la capacidad de la ciudadanía a acceder de forma precisa a los contenidos relevantes, así como la diversidad de temas tratados en estos documentos, junto con la frecuente actualización y modificación de las normativas, plantea la necesidad para la propuesta de modelos que sean robustos y capaces de adaptarse a los cambios en el tiempo.

2.2 Enunciados del problema

2.2.1 Problema general

- ¿Qué modelo de procesamiento de lenguaje natural obtiene una mayor coherencia en el análisis de tópicos en el corpus de ordenanzas regionales de 2010 a 2024 del Gobierno Regional Puno?

2.2.2 Problemas específicos

- ¿Cómo se construye un corpus y cómo se realiza el pre procesamiento de las ordenanzas regionales del 2010 al 2024 del Gobierno Regional Puno?
- ¿Cuáles son los algoritmos más relevantes del procesamiento de lenguaje natural utilizados para el análisis de tópicos?
- ¿Cuáles son los pasos clave en el proceso de diseño de un modelo de análisis de tópicos en ordenanzas regionales?
- ¿Cuál es la coherencia de modelo de tópicos aplicado a ordenanzas regionales utilizando métricas de análisis de tópicos?

2.3 Justificación

En la actualidad, el gobierno regional Puno ha generado una gran cantidad de ordenanzas regionales que son utilizados para informar a la población a través de su sección transparencia del portal institucional Gobierno Regional Puno (2024), permitiendo que los ciudadanos, la sociedad civil y otros actores interesados tengan acceso a la información necesaria para evaluar su desempeño; sin embargo esta información accesible y disponible, en específico de las ordenanzas regionales, que está en documentos formales no permite la comprensión y análisis de la ciudadanía y conocer

en qué temas las autoridades rinden cuentas por sus acciones y decisiones para el beneficio de la región Puno.

La presente investigación tiene el propósito de encontrar un modelo de análisis de tópicos utilizando técnicas de Procesamiento de Lenguaje Natural con el fin de proporcionar una comprensión profunda y estructurada de los temas abordados en la legislación del Gobierno Regional Puno.

El modelado de tópicos en ordenanzas regionales puede contribuir significativamente a la optimización de los procesos legales. Al identificar y categorizar eficientemente los temas, se puede mejorar la velocidad y precisión en la búsqueda de información relevante, tanto para los profesionales del derecho como para los ciudadanos interesados. La organización de información contenida en las ordenanzas regionales de manera clara y accesible, fomenta la transparencia y la participación ciudadana. Los ciudadanos y las organizaciones pueden comprender mejor las leyes regionales, lo que fortalece la confianza en el sistema legal y promueve la participación activa en la vida democrática.

2.4 Objetivos

2.4.1 Objetivo general

- Determinar un modelo de análisis de tópicos en las ordenanzas regionales emitidas por el Gobierno Regional Puno que obtenga los mejores resultados en la coherencia de tópicos durante el periodo 2010-2024.

2.4.2 Objetivos específicos

- Construir un corpus de las ordenanzas regionales emitidas por el Gobierno Regional Puno desde 2010 hasta 2024.
- Seleccionar algoritmos de análisis de tópicos adecuados que permitan analizar tópicos en las ordenanzas regionales.
- Diseñar el modelo de análisis de tópicos utilizando los algoritmos seleccionados y el corpus de ordenanzas regionales.
- Evaluar la efectividad del modelo mediante métricas de análisis de tópicos.

2.5 Hipótesis

2.5.1 Hipótesis general

- El modelo de análisis de tópicos seleccionado permite analizar con mayor coherencia los tópicos en ordenanzas regionales del 2010 al 2024 del Gobierno Regional Puno.

2.5.2 Hipótesis específicas

- Las técnicas y herramientas NLP permiten construir un corpus de ordenanzas regionales del 2010 al 2024 del Gobierno Regional Puno.
- Existen algoritmos de modelado de tópicos relevantes que analizan los tópicos en ordenanzas regionales.
- Los pasos seguidos de la metodología KDD permiten diseñar el modelado de análisis de tópicos en ordenanzas regionales.
- El modelo propuesto tiene una efectividad aceptable en el análisis de tópicos en ordenanzas regionales.

CAPÍTULO III

MATERIALES Y MÉTODOS

3.1 Lugar de estudio

La investigación se realizó en la ciudad Puno, ubicada a 3,820 metros sobre el nivel del mar, donde se generó el corpus de ordenanzas regionales del periodo 2010 al 2024 del Gobierno Regional Puno.

3.2 Población

La población estuvo compuesta por todas las ordenanzas regionales emitidas por el Gobierno Regional de Puno en el periodo del mes de febrero del 2010 al mes de julio del 2024, el cual se detalla en la Tabla 2.

Tabla 2

Cantidad de ordenanzas regionales emitidas en el periodo 2010 al 2024 por el gobierno regional puno

Año	Cantidad de ordenanzas regionales
2010	15
2011	18
2012	15
2013	23
2014	14
2015	9
2016	33
2017	11
2018	13
2019	12
2020	5
2021	13
2022	25
2023	32
2024	11
Total	**249**

Nota. Gobierno Regional Puno (2024).

3.3 Muestra

El tipo de selección de la muestra es no probabilístico, debido a que la población es relativamente pequeña y homogénea, es decir, los elementos son similares en términos de

características relevantes para la investigación (Hernández-Sampieri y Mendoza, 2020), por lo que la muestra estuvo constituida por toda la población, es decir por un corpus de 249 documentos de ordenanzas regionales.

3.4 Método de investigación

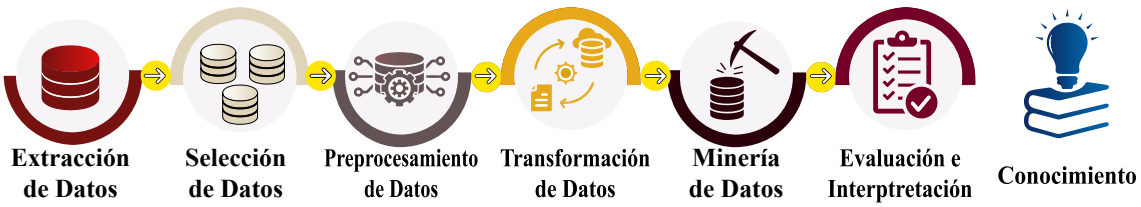
La presente investigación corresponde a un enfoque cuantitativo, con un diseño no experimental transversal, ya que se analiza las ordenanzas regionales emitidas por el Gobierno Regional Puno desde el 2010 al año 2024.

3.5 Descripción detallada de métodos por objetivos específicos

El proceso para determinar el modelo para la identificación de tópicos de ordenanzas regionales emitidas por el Gobierno Regional Puno es presentado en la Figura 5. Este proceso es una adaptación del proceso de descubrimiento de conocimiento en base de datos (KDD) y utilizado ampliamente por autores diversos. Esta metodología abarca cinco fases esenciales que permiten extraer información valiosa de grandes volúmenes de datos, a los cuales se agregó una fase previa, la fase de extracción de datos, que es un paso crucial que implica la recolección de datos de fuentes confiables; seguida de la segunda fase que consiste el seleccionar los datos relevantes para garantizar su calidad. El pre procesamiento de datos permite preparar y limpiar los datos obtenidos en la fase anterior. La fase de transformación de datos, permite transformar y representar los datos en un formato adecuado para su posterior análisis en la fase siguiente de minería de datos; posteriormente, en la evaluación e interpretación, se validan y entienden patrones. Finalmente se obtiene un conocimiento para su aplicación en la toma de decisiones.

Figura 5

Pasos del proceso KDD



3.5.1 Construcción del corpus para la identificación de tópicos

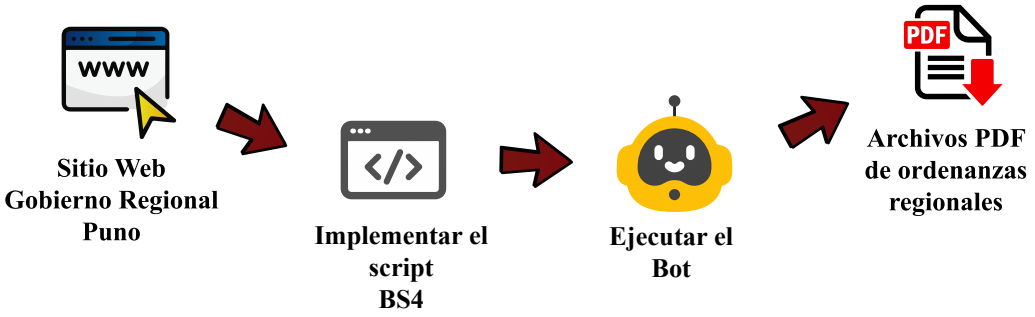
En la construcción del corpus se realizó la recolección de datos mediante técnicas de web scraping y OCR, que permitieron extraer los datos de ordenanzas

regionales del sitio Web del Gobierno Regional Puno.

El proceso de extracción de datos, se ilustra en la Figura 6, este proceso se realizó utilizando la técnica de Web Scraping a través de la biblioteca BS4, que facilitó la extracción de datos de documentos Web, permitiendo navegar por la estructura de un documento HTML y descargar automáticamente todos los archivos PDF de las ordenanzas regionales escaneadas publicadas entre 2010 y 2024 del sitio Web del Gobierno Regional Puno a través de un bot.

Figura 6

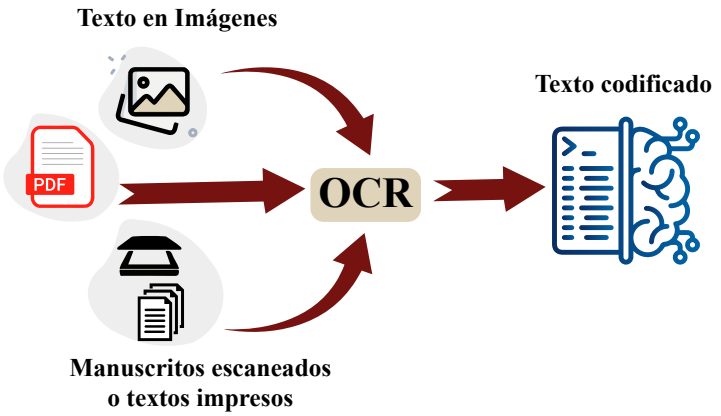
Pasos de extracción de datos mediante Web Scraping



En la Figura 7, se muestra la tarea de conversión de texto renderizado, los archivos PDF obtenidos en el proceso anterior, en texto codificado, en formato texto, para ello se utilizó Tesseract, que es un motor del reconocimiento óptico de caracteres (OCR).

Figura 7

Conversión de texto renderizado en texto codificado con OCR



3.5.2 Procesamiento de datos

Los pasos cruciales para el procesamiento de texto son la tokenización, la eliminación de stop words y la lematización son pasos cruciales en el procesamiento de texto.

a) **Tokenización:** Es el proceso de dividir el texto en unidades más pequeñas llamadas tokens. Estos tokens pueden ser palabras, frases, oraciones u otros elementos significativos. En el contexto de NLP, la tokenización generalmente se refiere a dividir el texto en palabras.

b) **Eliminación de stop words:** Los stop words son palabras comunes en un idioma que generalmente no aportan mucho valor semántico y pueden eliminarse para simplificar el procesamiento de texto.

c) **Lematización:** La lematización es el proceso de reducir una palabra a su forma base o lema. A diferencia de la stemming, que simplemente corta los sufijos, la lematización tiene en cuenta el contexto y transforma las palabras en su forma raíz gramaticalmente correcta.

d) **Vectorización** Es el paso de convertir los tokens procesados en una representación numérica que pueda ser utilizada por algoritmos de machine learning. Los métodos utilizados fueron:

- Bolsa de palabras (Bag of Words): Representa el texto como una colección de palabras ignorando el orden.
- TF-IDF (Term Frequency-Inverse Document Frequency): Pondera las palabras según su frecuencia en el documento y su rareza en el corpus.
- Embeddings de palabras: Utiliza técnicas como Word2Vec o GloVe para representar palabras en un espacio vectorial continuo.

3.5.3 Modelo de identificación de tópicos

Para la selección del modelo más adecuado para la identificación de tópicos se realizó la revisión de la literatura de los algoritmos NLP utilizados en la identificación de tópicos.

3.5.4 Evaluación del modelo

La métrica seleccionada y ampliamente utilizada como, es la coherencia de tópicos basadas en palabras, cuyo objetivo es la interpretabilidad semántica de los temas descubiertos (Hosseiny Marani y Baumer, 2023). Esta métrica de coherencia como la media de las medidas de co-ocurrencia para todos los pares de palabras en todos los tópicos. Cuanto mayor sea la métrica de coherencia, más coherentes se considerarán los tópicos (Rüdiger et al., 2022).

La coherencia se calcula como:

$$Coherence = \sum_{k=2}^{N} \sum_{l=1}^{k-1} sim(word_k, word_l)$$

Donde:

- $word_k$ y $word_l$: Son la k-ésima y l-ésima palabra temática generadas por un modelo de tópicos.
- N: Es el número de palabras temáticas de salida.
- $sim(.,.)$: Es el coseno de similaridad de dos palabras.

CAPÍTULO IV

RESULTADOS Y DISCUSIÓN

4.1 Resultados

Los resultados se presentan en relación al desarrollo de los objetivos de la investigación y su respectiva discusión. En la sección 4.1.1 se presenta el corpus construido de ordenanzas regionales emitidas por el Gobierno Regional Puno desde 2010 a 2024. En la sección 4.1.2 se presenta la determinación de algoritmos de identificación de tópicos adecuados para el análisis de ordenanzas regionales. En la sección 4.1.3 se presenta el diseño e implementación del modelo de identificación de tópicos utilizando técnicas NLP. En la sección 4.1.4 se muestra el resultado de la evaluación de la efectividad del modelo mediante métricas.

4.1.1 Resultado conforme al primer objetivo específico

La construcción del corpus de ordenanzas regionales emitidas por el Gobierno Regional Puno abarcó un período de 14 años, desde 2010 hasta 2024. A continuación, se detallan los resultados obtenidos:

A. Recopilación y selección de documentos

Se recolectaron y digitalizaron un total de 249 documentos de ordenanzas regionales emitidas durante el período indicado. Estos documentos fueron obtenidos a través de fuentes oficiales, el Portal Institucional del Gobierno Regional Puno.

Se recuperó un total de 249 documentos escaneados en formato PDF de acuerdo a los pasos de extracción de datos mediante web scraping. La herramienta BeautifulSoup (bs4) fue fundamental para llevar a cabo el Web scraping de las ordenanzas regionales desde el sitio web del Gobierno Regional Puno.

El proceso que facilitó fue en la extracción de contenido HTML, que permitió la extracción de contenido HTML de las páginas Web que alojaban las ordenanzas regionales. Al cargar el código HTML en bs4, se pudo navegar y manipular fácilmente la estructura de la página,

accediendo a elementos específicos como tablas, enlaces y bloques de texto donde se encontraban las ordenanzas. Así mismo, facilitó la navegación y filtrado de etiquetas, con bs4, fue posible seleccionar y filtrar las etiquetas HTML que contenían la información relevante, como , <a>, <p>, y otras etiquetas de estructura, que fue crucial para localizar y extraer los enlaces de descarga de las ordenanzas. También permitió el manejo de datos anidados dentro de múltiples capas de etiquetas a través de métodos como find() y find_all() para navegar por estas estructuras anidadas y extraer datos de manera precisa y eficiente. Finalmente se logró automatizar la extracción masiva de ordenanzas desde diferentes secciones y años del sitio web. Esto ahorró tiempo y redujo errores en comparación con la extracción manual, permitiendo descargar en forma masiva y automática las 249 ordenanzas regionales de manera efectiva, como se ilustra en la Figura 8.

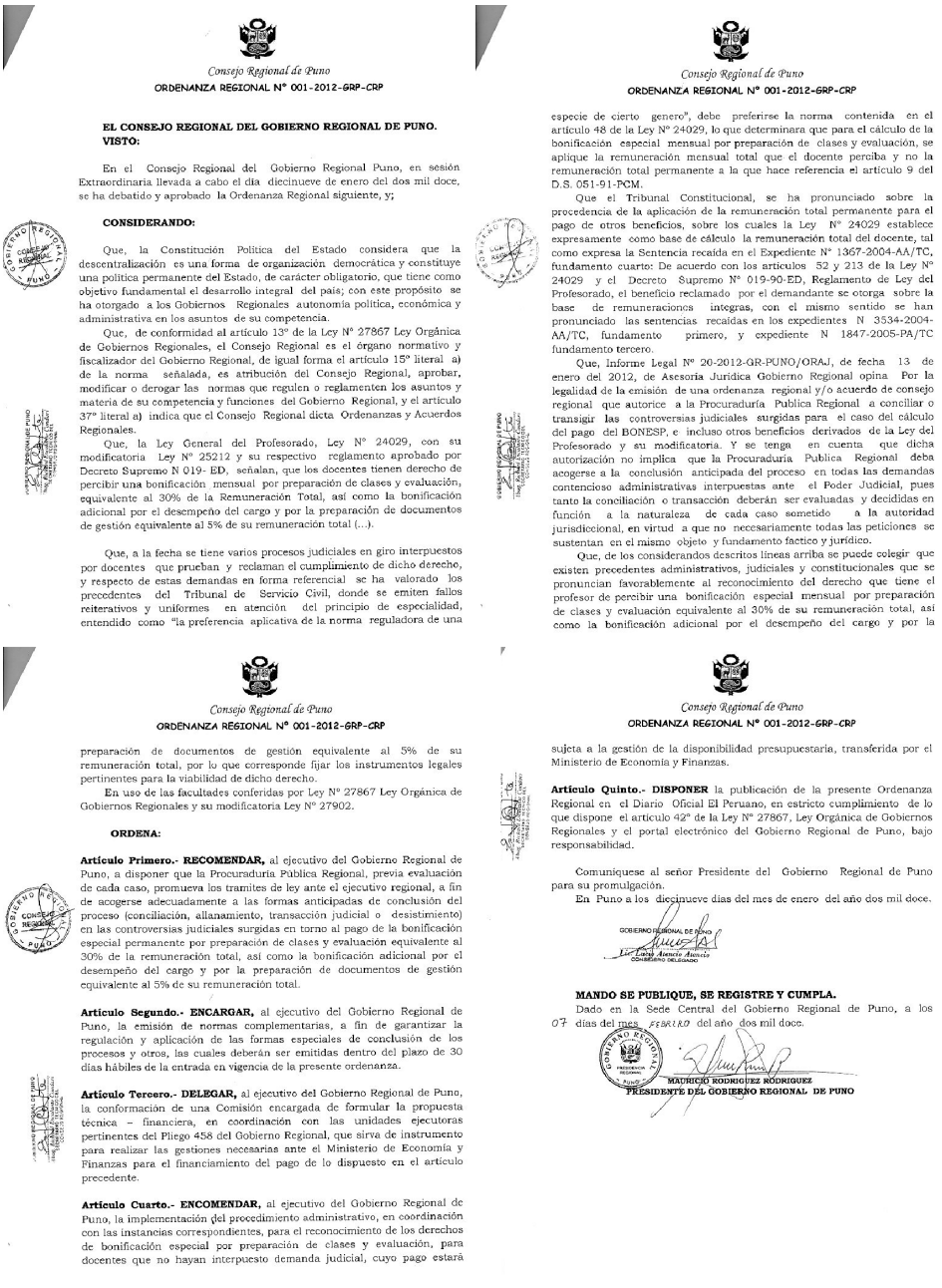
Figura 8

Documentos de ordenanza regional extraída en formato PDF

En la Figura 9, se ilustra un ejemplo de documento PDF que contiene la información escaneada de una ordenanza regional. En promedio cada ordenanza regional contiene 5 páginas.

Figura 9

Documento de ordenanza regional extraída en formato PDF



Consejo Regional de Puno
ORDENANZA REGIONAL N° 001-2012-GRP-CRP

EL CONSEJO REGIONAL DEL GOBIERNO REGIONAL DE PUNO.
VISTO:

En el Consejo Regional del Gobierno Regional Puno, en sesión Extraordinaria llevada a cabo el día diecinueve de enero del dos mil doce, se ha debatido y aprobado la Ordenanza Regional siguiente, y;

CONSIDERANDO:

Que, la Constitución Política del Estado considera que la descentralización es una forma de organización democrática y constituye una política permanente del Estado, de carácter obligatorio, que tiene como objetivo fundamental el desarrollo integral del país; con este propósito se ha otorgado a los Gobiernos Regionales autonomía política, económica y administrativa en los asuntos de su competencia.

Que, de conformidad al artículo 13° de la Ley N° 27867 Ley Orgánica de Gobiernos Regionales, el Consejo Regional es el órgano normativo y fiscalizador del Gobierno Regional, de igual forma el artículo 15° literal a) de la norma señalada, es atribución del Consejo Regional, aprobar, modificar o derogar las normas que regulen o reglamenten los asuntos y materia de su competencia y funciones del Gobierno Regional, y el artículo 37° literal a) indica que el Consejo Regional dicta Ordenanzas y Acuerdos Regionales.

Que, la Ley General del Profesorado, Ley N° 24029, con su modificatoria Ley N° 25212 y su respectivo reglamento aprobado por Decreto Supremo N 019- ED, señalan, que los docentes tienen derecho de percibir una bonificación mensual por preparación de clases y evaluación, equivalente al 30% de la Remuneración Total, así como la bonificación adicional por el desempeño del cargo y por la preparación de documentos de gestión equivalente al 5% de su remuneración total (...).

Que, a la fecha se tiene varios procesos judiciales en giro interpuestos por docentes que prueban y reclaman el cumplimiento de dicho derecho, y respecto de estas demandas en forma referencial se ha valorado los precedentes del Tribunal de Servicio Civil, donde se emiten fallos reiterativos y uniformes en atención del principio de especialidad, entendido como "la preferencia aplicativa de la norma reguladora de una

Consejo Regional de Puno
ORDENANZA REGIONAL N° 001-2012-GRP-CRP

especie de cierto genero", debe preferirse la norma contenida en el artículo 48 de la Ley N° 24029, lo que determinara que para el cálculo de la bonificación especial mensual por preparación de clases y evaluación, se aplique la remuneración mensual total que el docente perciba y no la remuneración total permanente a la que hace referencia el artículo 9 del D.S. 051-91-PCM.

Que el Tribunal Constitucional, se ha pronunciado sobre la procedencia de la aplicación de la remuneración total permanente para el pago de otros beneficios, sobre los cuales la Ley N° 24029 establece expresamente como base de cálculo la remuneración total del docente, tal como expresa la Sentencia recaída en el Expediente N° 1367-2004-AA/TC, fundamento cuarto: De acuerdo con los artículos 52 y 213 de la Ley N° 24029 y el Decreto Supremo N° 019-90-ED, Reglamento de Ley del Profesorado, el beneficio reclamado por el demandante se otorga sobre la base de remuneraciones integras, con el mismo sentido se han pronunciado las sentencias recaídas en los expedientes N 3534-2004-AA/TC, fundamento primero, y expediente N 1847-2005-PA/TC fundamento tercero.

Que, Informe Legal N° 20-2012-GR-PUNO/ORAJ, de fecha 13 de enero del 2012, de Asesoría Jurídica Gobierno Regional opina Por la legalidad de la emisión de una ordenanza regional y/o acuerdo de consejo regional que autorice a la Procuraduría Publica Regional a conciliar o transigir las controversias judiciales surgidas para el caso del cálculo del pago del BONESP, e incluso otros beneficios derivados de la Ley del Profesorado y su modificatoria. Y se tenga en cuenta que dicha autorización no implica que la Procuraduría Publica Regional deba acogerse a la conclusión anticipada del proceso en todas las demandas contencioso administrativas interpuestas ante el Poder Judicial, pues tanto la conciliación o transacción deberán ser evaluadas y decididas en función a la naturaleza de cada caso sometido a la autoridad jurisdiccional, en virtud a que no necesariamente todas las peticiones se sustentan en el mismo objeto y fundamento factico y jurídico.

Que, de los considerandos descritos líneas arriba se puede colegir que existen precedentes administrativos, judiciales y constitucionales que se pronuncian favorablemente al reconocimiento del derecho que tiene el profesor de percibir una bonificación especial mensual por preparación de clases y evaluación equivalente al 30% de su remuneración total, así como la bonificación adicional por el desempeño del cargo y por la

Consejo Regional de Puno
ORDENANZA REGIONAL N° 001-2012-GRP-CRP

preparación de documentos de gestión equivalente al 5% de su remuneración total, por lo que corresponde fijar los instrumentos legales pertinentes para la viabilidad de dicho derecho.

En uso de las facultades conferidas por Ley N° 27867 Ley Orgánica de Gobiernos Regionales y su modificatoria Ley N° 27902.

ORDENA:

Artículo Primero.- RECOMENDAR, al ejecutivo del Gobierno Regional de Puno, a disponer que la Procuraduría Pública Regional, previa evaluación de cada caso, promueva los tramites de ley ante el ejecutivo regional, a fin de acogerse adecuadamente a las formas anticipadas de conclusión del proceso (conciliación, allanamiento, transacción judicial o desistimiento) en las controversias judiciales surgidas en torno al pago de la bonificación especial permanente por preparación de clases y evaluación equivalente al 30% de la remuneración total, así como la bonificación adicional por el desempeño del cargo y por la preparación de documentos de gestión equivalente al 5% de su remuneración total.

Artículo Segundo.- ENCARGAR, al ejecutivo del Gobierno Regional de Puno, la emisión de normas complementarias, a fin de garantizar la regulación y aplicación de las formas especiales de conclusión de los procesos y otros, las cuales deberán ser emitidas dentro del plazo de 30 días hábiles de la entrada en vigencia de la presente ordenanza.

Artículo Tercero.- DELEGAR, al ejecutivo del Gobierno Regional de Puno, la conformación de una Comisión encargada de formular la propuesta técnica – financiera, en coordinación con las unidades ejecutoras pertinentes del Pliego 458 del Gobierno Regional, que sirva de instrumento para realizar las gestiones necesarias ante el Ministerio de Economía y Finanzas para el financiamiento del pago de lo dispuesto en el artículo precedente.

Artículo Cuarto.- ENCOMENDAR, al ejecutivo del Gobierno Regional de Puno, la implementación del procedimiento administrativo, en coordinación con las instancias correspondientes, para el reconocimiento de los derechos de bonificación especial por preparación de clases y evaluación, para docentes que no hayan interpuesto demanda judicial, cuyo pago estará

Consejo Regional de Puno
ORDENANZA REGIONAL N° 001-2012-GRP-CRP

sujeta a la gestión de la disponibilidad presupuestaria, transferida por el Ministerio de Economía y Finanzas.

Artículo Quinto.- DISPONER la publicación de la presente Ordenanza Regional en el Diario Oficial El Peruano, en estricto cumplimiento de lo que dispone el artículo 42° de la Ley N° 27867, Ley Orgánica de Gobiernos Regionales y el portal electrónico del Gobierno Regional de Puno, bajo responsabilidad.

Comuníquese al señor Presidente del Gobierno Regional de Puno para su promulgación.

En Puno a los diecinueve días del mes de enero del año dos mil doce.

GOBIERNO REGIONAL DE PUNO
Lic. Lenin Antonio Aranda
CONSEJERO DELEGADO

MANDO SE PUBLIQUE, SE REGISTRE Y CUMPLA.

Dado en la Sede Central del Gobierno Regional de Puno, a los 07 días del mes FEBRERO del año dos mil doce.

MAURICIO RODRIGUEZ RODRIGUEZ
PRESIDENTE DEL GOBIERNO REGIONAL DE PUNO

B. Proceso de extracción de texto de documentos escaneados

En esta etapa se llevó a cabo la extracción de información textual de los documentos PDF escaneados para convertirlos a texto plano utilizando la herramienta Tesseract OCR (Reconocimiento Óptico de Caracteres) en el que se transformó los documentos escaneados en

formato digital editable y procesable para su posterior análisis en tareas de procesamiento de texto. Este proceso es esencial para la digitalización de documentos en formato físico, facilitando la preservación, análisis y manejo de información en el ámbito de la administración pública. Este proceso incluyó las siguientes actividades:

B.1 Descripción de los documentos

Los documentos de origen consisten en ordenanzas regionales que fueron digitalizados mediante escaneo, almacenándose en formato PDF. Dado que estos documentos no contienen texto seleccionable, sino imágenes que representan el texto, se requirió una herramienta de OCR para extraer el contenido textual. Tesseract OCR fue seleccionada debido a su capacidad para manejar una variedad de fuentes y su soporte para múltiples idiomas, incluyendo el español, que es el idioma de los documentos analizados.

B.2 Configuración y preparación del entorno

El proceso de extracción comenzó con la configuración del entorno de trabajo, donde se instaló Tesseract OCR junto con sus librerías asociadas, Además, se descargaron los datos de entrenamiento del idioma español para optimizar la precisión del OCR en documentos escritos en este idioma, como se muestra en el siguiente código:

Código de instalación y configuración de Tesseract

```
pip install PyMuPDF pytesseract Pillow
!sudo apt-get update
!sudo apt-get install -y
 tesseract-ocr
 tesseract-ocr-spa
!pip install pytesseract
os.environ['TESSDATA_PREFIX'] =
    '/usr/share/tesseract-ocr/4.00/tessdata/'
```

B.3 Carga y conversión de PDF a imágenes

Cada documento PDF fue procesado para extraer las imágenes de cada una de sus páginas. Esto se realizó utilizando la librería PyMuPDF, como se muestra en el código siguiente, que permitió la conversión de cada página del PDF en una imagen individual, preservando la calidad original del escaneo.

Código de importar las librerías necesarias en especial PyMuPDF

```
import fitz # PyMuPDF
import pytesseract
from PIL import Image
import io
```

B.4 Aplicación de OCR a las imágenes

Una vez obtenidas las imágenes, se aplicó Tesseract OCR a cada una para convertir el contenido visual en texto plano. Tesseract fue configurado para identificar y procesar texto en español, utilizando su modelo de idioma específico para mejorar la precisión en la detección de caracteres, especialmente en textos que contienen acentos y caracteres especiales propios del idioma español.

Código de la función que convierte imagen a texto con Tesseract

```
def pdf_to_text(pdf_path, output_txt_path):

    # Abre el archivo PDF
    pdf_document = fitz.open(pdf_path)

    # Variable para almacenar el texto extraído
    full_text = ""

    for page_num in range(pdf_document.page_count):
        # Extrae la página
        page = pdf_document.load_page(page_num)

        #Extrae la imagen de la página
```

```
        pix = page.get_pixmap()

        #Convierte la imagen a formato PIL
        img = Image.open(io.BytesIO(pix.tobytes()))

        #Usa pytesseract para realizar OCR en la
            imagen
        text = pytesseract.image_to_string(img,
            lang='spa')
        # Agrega el texto extraído al resultado total
          full_text += text + "\n"

        # Escribe el texto extraído en un archivo de
            texto
          with open(output_txt_path, "w",
              encoding="utf-8") as text_file:
            text_file.write(full_text)

    print(f"Texto extraído guardado en
        {output_txt_path}")
```

Código de la función que procesa la conversión en forma múltiple

```
def process_multiple_pdfs(pdf_directory,
    output_directory):
    # Crear el directorio de salida si no existe
    if not os.path.exists(output_directory):
        os.makedirs(output_directory)

    # Iterar sobre todos los archivos PDF en el
        directorio dado
    for pdf_filename in
        os.listdir(pdf_directory):
        if pdf_filename.endswith(".pdf"):
            pdf_path = os.path.join(pdf_directory,
                pdf_filename)
            output_txt_path =
                os.path.join(output_directory,
```

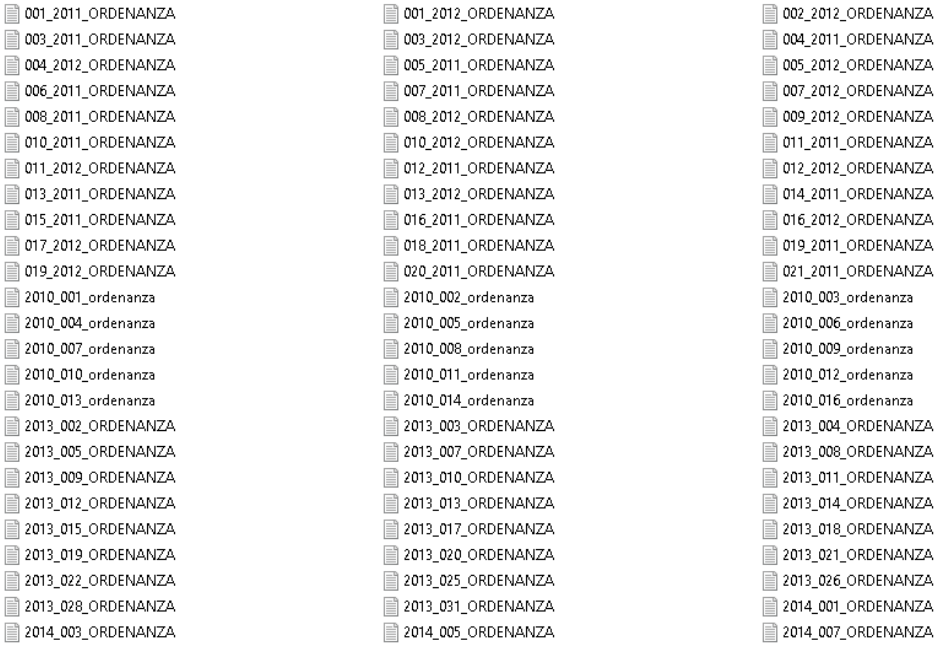
```
                pdf_filename.replace(".pdf",
                ".txt"))
11          pdf_to_text(pdf_path, output_txt_path)
```

B.5 Resultados de la extracción

El proceso de extracción logró convertir exitosamente el contenido de los documentos PDF escaneados a texto plano. Los textos resultantes fueron adecuados para su uso en análisis de NLP, como la identificación de tópicos. A continuación, en la Figura 10 se ilustra el resultado de la extracción, es decir las ordenanzas regionales en formato texto (.txt).

Figura 10

Documentos de ordenanzas regionales convertidos en formato texto



En la Figura 11, se ilustra un ejemplo de documento texto que contiene la información en texto plano de una ordenanza regional.

Figura 11

Documento de ordenanza regional extraída en formato texto

Consejo Regional de Puno
ORDENANZA REGIONAL N* 001-2012-GRP-CRP

EL CONSEJO REGIONAL DEL GOBIERNO REGIONAL DE PUNO.
VISTO:

En el Consejo Regional del Gobierno Regional Puno, en sesión Extraordinaria llevada a cabo el día: diecinueve de enero del dos mil doce, se ha debatido y aprobado la Ordenanza Regional siguiente, y;

CONSIDERANDO:

Que, la Constitución Política del Estado considera que la descentralización es una forma de organización democrática y constituye una política permanente del Estado, de carácter obligatorio, que tiene como objetivo fundamental el desarrollo integral del país; con este propósito se ha otorgado a los Gobiernos Regionales autonomía política, económica y administrativa en los asuntos de su competencia,

Que, de conformidad al artículo 13? de la Ley N* 27867 Ley Orgánica de Gobiernos Regionales, el Consejo Regional es el órgano normativo y fiscalizador del Gobierno Regional, de igual forma el artículo 15* literal a) de la norma señalada, es atribución del Consejo Regional, aprobar, modificar o derogar las normas que regulen o reglamenten los asuntos y materia de su competencia y funciones del Gobierno Regional, y el artículo 37 literal a) indica que el Consejo Regional dicta Ordenanzas y Acuerdos

Regionales.

Que, la Ley General del Profesorado, Ley N" 24029, con su modificatoria Ley N* 25212 y su respectivo reglamento aprobado por Decreto Supremo N 019- ED, señalan, que los docentes tienen derecho de percibir una bonificación mensual por preparación de clases y evaluación, equivalente al 30% de la Remuneración Total, así como la bonificación adicional por el desempeño del cargo y por la preparación de documentos de gestión equivalente al 5% de su remuneración total (...

Que, a la fecha se tiene varios procesos judiciales en giro interpuestos por docentes que prueban y reclaman el cumplimiento de dicho derecho, y respecto de estas demandas en forma referencial se ha valorado los precedentes del Tribunal de Servicio Civil, donde se emiten fallos reiterativos y uniformes en atención del principio de especialidad, entendido como "la preferencia aplicativa de la norma reguladora de una

Consejo Regional de Puno
ORDENANZA REGIONAL N* 001-2012-GRP-CRP.

especie de cierto genero", debe preferirse la norma contenida en el artículo 48 de la Ley N* 24029, lo que determinara que para el cálculo de la bonificación especial mensual por preparación de clases y evaluación, se aplique la remuneración mensual total que el docente perciba y no la remuneración total permanente a la que hace referencia el artículo 9 del D.S. 051-91-PCM.

Que el Tribunal Constitucional, se ha pronunciado sobre la procedencia de la aplicación de la remuneración total permanente para el pago de otros beneficios, sobre los cuales la Ley N" 24029 establece expresamente como base de cálculo la remuneración total del docente, tal como expresa la Sentencia recaída en el Expediente N* 1367-2004-AA/TC, fundamento cuarto: De acuerdo con los artículos 52 y 213 de la Ley N* 24029 y el Decreto Supremo N* 019-90-ED, Reglamento de Ley del Profesorado, el beneficio reclamado por el demandante sc otorga sobre la base de remuneraciones integras, con el mismo sentido se han pronunciado las sentencias recaídas en los expedientes N 3534-2004-AA/TC, fundamento – primero, y expediente N 1847-2005-PA/TC fundamento tercero.

Que, Informe Legal N* 20-2012-GR-PUNO/ORAJ, de fecha 13 de enero del 2012, de Asesoria Jurídica Gobierno Regional opina – Por la legalidad de la emisión de una ordenanza regional y/o acuerdo de consejo regional que autorice a la Procuraduría Publica Regional a conciliar o transigir las controversias judiciales surgidas para el caso del cálculo del pago del BONESP, e incluso otros beneficios derivados de la Ley del Profesorado y su modificatoria. Y se tenga en cuenta que dicha autorización no implica que la Procuraduria Publica Regional deba acogerse a la conclusión anticipada del proceso en todas las demandas contencioso administrativas interpuestas ante el Poder Judicial, pues tanto la conciliación o transacción deberán ser evaluadas y decididas en función a la naturaleza de cada caso sometido a la autoridad jurisdiccional, en virtud a que no necesariamente todas las peticiones se sustentan en el mismo objeto y fundamento factico y jurídico.

Que, de los considerandos descritos líncas arriba se puede colegir que existen precedentes administrativos, judiciales y constitucionales que se pronuncian favorablemente al reconocimiento del derecho que tiene el profesor de percibir una bonificación especial mensual por preparación de clases y evaluación equivalente al 30% de su remuneración total, así como la bonificación adicional por el desempeño del cargo y por la

Consejo Regional de Puno
ORDENANZA REGIONAL N* 001-2012-GRP-CRP

preparación de documentos de gestión equivalente al 5% de su remuneración total, por lo que corresponde fijar los instrumentos legales pertinentes para la viabilidad de dicho derecho.

:n uso de las facultades conferidas por Ley N* 27867 Ley Orgánica de Gobiernos Regionales y su modificatoria Ley N* 27902,

ORDENA:

Artículo Primero.- RECOMENDAR, al ejecutivo del Gobierno Regional de Puno, a disponer que la Procuraduría Pública Regional, previa evaluación de cada caso, promueva los tramites de ley ante el ejecutivo regional, a fin de acogerse adecuadamente a las formas anticipadas de conclusión del proceso (conciliación, allanamiento, transacción judicial o desistimiento] en las controversias judiciales surgidas en torno al pago de la bonificación especial permanente por preparación de clases y evaluación equivalente al 30% de la remuneración total, así como la bonificación adicional por el desempeño del cargo y por la preparación de documentos de gestión equivalente al 5% de su remuneración total.

Artículo Segundo.- ENCARGAR, al ejecutivo del Gobierno Regional de Puno, la emisión de normas complementarias, a fin de garantizar la regulación y aplicación de las formas especiales de conclusión de los procesos y otros, las cuales deberán ser emitidas dentro del plazo de 30 días hábiles de la entrada en vigencia de la presente ordenanza.

Artículo Tercero.- DELEGAR, al ejecutivo del Gobierno Regional de Puno, la conformación de una Comisión encargada de formular la propuesta técnica - financiera, en coordinación con las unidades ejecutoras pertinentes del Pliego 458 del Gobierno Regional, que sirva de instrumento para realizar las gestiones necesarias ante el Ministerio de Economía y Finanzas para el financiamiento del pago de lo dispuesto en el artículo precedente.

Artículo Cuarto.- ENCOMENDAR, al ejecutivo del Gobierno Regional de Puno, la implementación gel procedimiento administrativo, en coordinación con las instancias correspondientes, para el reconocimiento de los derechos de bonificación especial por preparación de clases y evaluación, para docentes que no hayan interpuesto demanda judicial, cuyo pago estará

Consejo Regional de Puno
ORDENANZA REGIONAL N* 001-2012-SRP-CRP

sujeta a la gestión de la disponibilidad presupuestaria, transferida por el

Ministerio de Economía y Finanzas.

Artículo Quinto. DISPONER la publicación de la presente Ordenanza

Regional en el Diario Oficial El Peruano, en estricto cumplimiento de lo

que dispone el artículo 42? de la Ley N* 27867, Ley Orgánica de Gobiernos

Regionales y el portal electrónico del Gobierno Regional de Puno, bajo lad.

responsa!

Comuniquese al señor Presidente del Gobierno Regional de Puno
para su promulgación.
En Puno alos dlc; opere dias del mes de enero del año dos mil doce

MANDO SE PUBLIQUE, SE REGISTRE Y CUMPLA.
Dado en la Sede Central del Gobierno Regional de Puno, a los
07 días del

C. Pre procesamiento del texto

El pre procesamiento de los documentos en formato texto (.txt) correspondientes a las ordenanzas regionales emitidas por el Gobierno Regional se realizó con el objetivo de preparar los datos para un análisis posterior, como el modelado de tópicos. A continuación, se describen las tareas clave del pre procesamiento y los resultados obtenidos.

C.1 Eliminación de caracteres, números y signos de puntuación

En esta tarea que eliminaron los caracteres especiales detallados en la Tabla 3 haciendo uso de Expresiones Regulares.

Tabla 3

Eliminación de caracteres y signos de puntuación

Categoría 1	Categoría 2
Caracteres especiales	Todos los caracteres especiales, tales como símbolos (@, #, $, etc.)
Dígitos	Los dígitos [0-9] y su secuencia.
Signos de puntuación	Punto, puntos suspensivos, coma, punto y coma, dos puntos), signos de exclamación y apertura de exclamación en español, signos de interrogación y apertura de interrogación en español, comillas dobles, comillas simples, paréntesis de apertura y cierre, corchetes de apertura y cierre, llaves de apertura y cierre, guión), guión largo, barra inclinada, etc.

Una vez que el texto fue normalizado, se procedió a la tokenización, es decir, la segmentación del texto en unidades más pequeñas llamadas tokens. Este proceso permitió la conversión de cada documento en una lista de palabras, facilitando así el análisis detallado y la manipulación del texto.

C.2 Eliminación de texto irrelevante

En el contexto de la identificación de tópicos mediante técnicas de procesamiento de lenguaje natural (NLP), la eliminación de texto irrelevante es un paso fundamental para mejorar la calidad y precisión de los resultados. Durante el pre procesamiento, se elimina el texto que no aporta valor semántico significativo, como encabezados, numeraciones, fechas y cualquier otra información redundante que pueda distraer o distorsionar el análisis. Este proceso permite depurar el corpus, centrándose en las palabras y frases que realmente contribuyen a la formación de tópicos coherentes y representativos. Como resultado, se

observó una mejora en la coherencia de los tópicos generados, ya que se redujo el ruido y se potenció la identificación de patrones temáticos más claros y relevantes dentro del conjunto de datos.

Se aplicó expresiones regulares para eliminar segmentos específicos del texto, como párrafos introductorios repetitivos o cláusulas legales estándar, contribuyó significativamente a la limpieza del corpus. Esto no solo optimizó el tiempo de procesamiento y análisis, sino que también permitió un enfoque más directo en el contenido sustantivo de los documentos. La eliminación de este tipo de texto irrelevante resultó en tópicos más precisos y detallados, lo que mejoró la interpretación y la utilidad práctica de los modelos de identificación de tópicos, especialmente en documentos oficiales o legales, donde la relevancia semántica es clave para un análisis efectivo.

En documentos legales, como es el caso de las ordenanzas regionales, se evidencia que el cuerpo relevante del texto siga a encabezados, introducciones, o consideraciones preliminares. La palabra “Ordena:” suele marcar el inicio de las disposiciones principales del documento, y por lo tanto, se considera el punto de interés a partir del cual se desea conservar el texto, que se ilustra en la Figura 12:

Figura 12

Sección de interés en el documento de ordenanza regional

Consejo Regional de Puno
ORDENANZA REGIONAL N° 001-2012-GRP-CRP

preparación de documentos de gestión equivalente al 5% de su remuneración total, por lo que corresponde fijar los instrumentos legales pertinentes para la viabilidad de dicho derecho.

:n uso de las facultades conferidas por Ley N° 27867 Ley Orgánica de Gobiernos Regionales y su modificatoria Ley N° 27902,

ORDENA:

Artículo Primero.- RECOMENDAR, al ejecutivo del Gobierno Regional de Puno, a disponer que la Procuraduría Pública Regional, previa evaluación de cada caso, promueva los tramites de ley ante el ejecutivo regional, a fin de acogerse adecuadamente a las formas anticipadas de conclusión del proceso (conciliación, allanamiento, transacción judicial o desistimiento] en las controversias judiciales surgidas en torno al pago de la bonificación especial permanente por preparación de clases y evaluación equivalente al 30% de la remuneración total, así como la bonificación adicional por el desempeño del cargo y por la preparación de documentos de gestión equivalente al 5% de su remuneración total.

Artículo Segundo.- ENCARGAR, al ejecutivo del Gobierno Regional de Puno, la emisión de nonmas complementarias, a fin de garantizar la regulación y aplicación de las formas especiales de conclusión de los procesos y otros, las cuales deberán ser emitidas dentro del plazo de 30 días hábiles de la entrada en vigencia de la presente ordenanza.

Artículo Tercero.- DELEGAR, al ejecutivo del Gobierno Regional de Puno, la conformación de una Comisión encargada de formular la propuesta técnica - financiera, en coordinación con las unidades ejecutoras pertinentes del Pliego 458 del Gobierno Regional, que sirva de instrumento para realizar las gestiones necesarias ante el Ministerio de Economía y Finanzas para el financiamiento del pago de lo dispuesto en el artículo precedente.

C.3 Eliminación de stopwords y personalización de stopwords

En el proceso de pre procesamiento de texto para el análisis de tópicos, la eliminación de stopwords juega un papel crucial al reducir el ruido en los datos textuales. Se aplicó un filtro para eliminar las stopwords estándar en español, tales como preposiciones, conjunciones, entre otras, que son comúnmente utilizadas, pero no aportan significado contextual en el análisis, en la Tabla 4 se presenta éstos.

Tabla 4

Stopword considerados para su eliminación

Stopwords	Palabras
Preposiciones en español	a, ante, bajo, con, contra, de, desde, en, entre, hacia, hasta, para, por, según, sin, sobre, tras
Conjunciones en español	y, e, ni, o, u, pero, sino, mas, aunque

Al eliminar estas palabras, se logra que el corpus se concentre en términos más significativos que realmente contribuyen a la identificación de patrones y tópicos dentro del texto. Como resultado de este proceso, el texto pre procesado se vuelve más limpio y enfocado, lo que facilita un análisis más preciso y eficiente en etapas posteriores del modelado de tópicos.

Además, la eliminación de stopwords personalizadas, adaptadas al contexto específico del corpus, mejora aún más la calidad del texto para su análisis. En algunos casos, términos que no son generalmente considerados stopwords pueden repetirse frecuentemente en un corpus específico y no aportar valor al análisis; por ejemplo, nombres de lugares o términos técnicos que no son relevantes para la extracción de tópicos. Al eliminar estas stopwords personalizadas, se logra una mayor precisión en la identificación de los temas más relevantes, lo que contribuye a obtener resultados más coherentes y representativos en el modelado de tópicos.

Además, se incluyeron stopwords personalizadas específicas para el contexto de las ordenanzas regionales, que se obtuvieron en retroalimentación al modelo, las cuales se consideran irrelevantes para el objetivo del análisis de tópicos, estos se presentan en la Tabla 5.

Tabla 5

Stopword personalizados a ordenanzas regional

Stopwords	Palabras
Stopwords personalizados	Ordenanza, comisión, gobierno, regional, ley, puno, puno, comisión, artículo, capítulo sesión, ordinaria, extraordinaria, constitución, reglamento, norma, código, conforme, ejecutar, jurídico, vigente,

La eliminación de estas palabras permitió una mayor precisión en la identificación de los temas relevantes.

C.4 Lematización

Finalmente, se realizó la lematización del texto, que consistió en reducir cada palabra a su forma base o lema. Por ejemplo, “implementación” se redujo a “implementar”,“cumpliendo” a “cumplir”, entre otros. Este paso fue crucial para normalizar las variaciones morfológicas de las palabras, asegurando que términos con significados similares fueran tratados como equivalentes durante el análisis.

En la Figura 13 se muestra un ejemplo obtenido del resultado de este pre procesamiento:

Figura 13

Documento de ordenanza regional pre procesado

```
visto sesión extraordinario llevada cabo diecinueve enero mil
debatir aprobar considerar constitución político descentralización
forma organización democrático constituir política permanente
carácter obligatorio objetivo fundamental desarrollo integral país
propósito otorgar gobierno regional autonomía político económico
administrativo asunto competencia conformidad   n   orgánico
gobierno regional órgano normativo fiscalizador formar   literal
norma señalado atribución aprobar modificar derogar norma regular
reglamenten asunto materia competencia función   literal indicar
dictar ordenanza acuerdo regional general profesorado n
modificatoria n   respectivo reglamento aprobado decreto supremo n
ed señalar docente derecho percibir bonificación mensual preparación
clase evaluación equivalente   remuneración bonificación adicional
desempeño cargo preparación documento gestión equivalente
remuneración fecha proceso judicial giro interpuesto docente probar
reclamar cumplimiento derecho demanda forma referencial valorar
precedente tribunal servicio civil emitir fallo reiterativo uniforme
atención principio especialidad entendido preferencia aplicativo
norma regulador n    grp crp especie genero preferir él norma
contenido articulo determinar cálculo bonificación especial mensual
preparación clase evaluación aplicar remuneración mensual docente
percir remuneración permanente referencia   d s    pcm tribunal
constitucional pronunciar procedencia aplicación remuneración
permanente pago beneficio n   establecer expresamente base cálculo
remuneración docente expresar sentencia recaído expediente n    aa
tc fundamento cuarto artículo   decreto supremo n    ed reglamento
profesorado beneficio reclamado demandante sc otorgar base
remuneración integra sentido pronunciar sentencia recaída expediente
fundamento expediente n    pa tc fundamento informir legal fecha
enero   asesoria jurídico opinar legalidad emisión autorizar
procuraduría publicar conciliar transigir controversia judicial
surgido caso cálculo pago bonesp beneficio derivado profesorado
modificatoria dicho autorización implicar procuraduria publicar
```

4.1.2 Resultado conforme al segundo objetivo específico

La revisión de la literatura sobre algoritmos de procesamiento de lenguaje natural para la identificación de tópicos en ordenanzas regionales, permitió determinar los modelos más adecuados para la presente investigación, considerando para ello los resultados de evaluación obtenidos a nivel de desempeño y frecuencia de uso. La Tabla 6 presenta los autores con el detalle de las técnicas utilizadas y las métricas aplicadas para su evaluación.

Según la Tabla 6, los modelos de identificación de tópicos más empleados son LDA y BERTopic, debido a sus enfoques efectivos para descomponer grandes volúmenes de texto en temas significativos. LDA, uno de los métodos

más tradicionales, utiliza un enfoque probabilístico para inferir la estructura subyacente de los documentos, lo que permite identificar y analizar los tópicos con una base sólida en teoría estadística. Por otro lado, BERTopic representa una evolución moderna al aprovechar embeddings contextuales generados por modelos de lenguaje pre entrenados, como BERT, para capturar significados semánticos más ricos y complejos. Por lo tanto, en la investigación se consideró a estos dos modelos.

Así mismo la métrica de evaluación de modelos de identificación de tópicos es la métrica de la Coherencia, que fue considerada para evaluar el modelo propuesto.

Tabla 6

Técnicas de NLP y métricas aplicadas para identificación de tópicos en documentos normativos

Nº	Referencia	Fuente	Técnica NLP	Métricas		
				Coherencia de Tópicos	Perplejidad	Dif. de Tópicos
1	Lange et al. (2021)	SJR	LDA	X		
2	Dyevre (2021)	IEEE Xplore	LDA		X	
3	Lee et al. (2023)	Science Direct	BERTopic	X		
4	Mazumder y Barui (2021)	Science Direct	LDA	X		
5	Mifrah y Benlahmar (2022)	IEEE Xplore	LSA			X
6	Zankadi et al. (2023)	IEEE Xplore	HDP	X		

Nº	Referencia	Fuente	Técnica NLP	Coherencia de Tópicos	Métricas	
					Perplejidad	Dif. de Tópicos
7	Dillan y Fudholi (2023)	SJR	BERTopic	X		
8	Pedregosa et al. (2011)	SJR	BERTopic	X		
9	Jelodar et al. (2019)	Science Direct	LDA		X	
10	Fahlevvi y Azhari (2022)	Science Direct	LDA	X		
11	Wang et al. (2023)	SJR	BERTopic	X		
12	Lee et al. (2023)	SJR	LDA	X		X

4.1.3 Resultado conforme al tercer objetivo específico

Al diseñar e implementar un modelo de análisis de tópicos utilizando los algoritmos seleccionados (LDA y BERTopic) y el corpus de ordenanzas regionales, los resultados esperados incluirían la identificación de tópicos relevantes, la distribución de tópicos en documentos, la visualización de los tópicos y la comparación de los modelos.

A. Diseño e implementación del modelo de tópicos LDA

Se eligió Latent Dirichlet Allocation (LDA) como el modelo para la identificación de tópicos debido a su capacidad para descubrir patrones latentes en grandes volúmenes de texto.

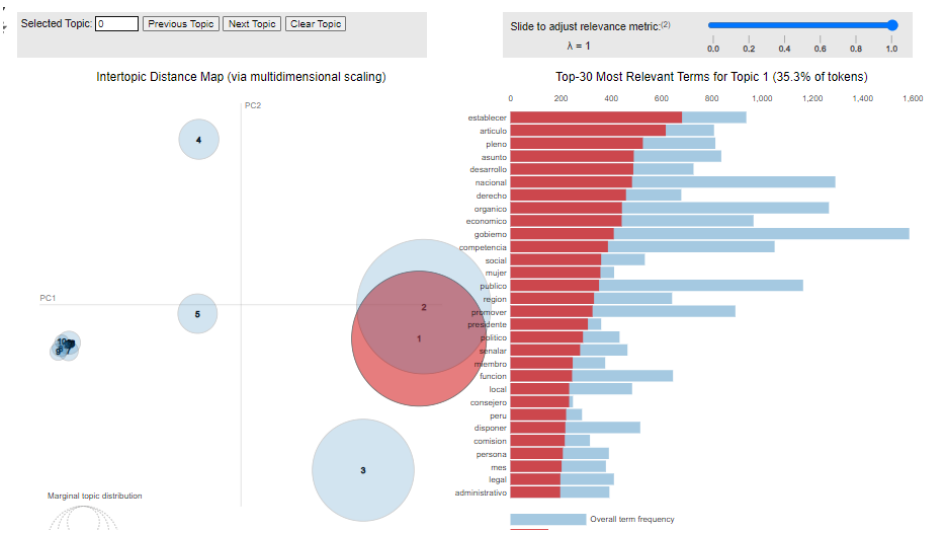
A.1 Entrenamiento del modelo

El modelo LDA fue entrenado utilizando las ordenanzas pre procesadas. Para esto, se empleó la biblioteca Gensim en Python, que es ampliamente utilizada para la implementación de modelos LDA.

Como resultado del modelo se ilustra cada tópico representado por un conjunto de palabras con alta probabilidad de aparición conjunta en los documentos. Para facilitar la interpretación, se utilizaron herramientas como pyLDAvis para visualizar la distribución de tópicos y palabras clave en los documentos, ayudando a identificar las áreas de enfoque del gobierno regional.

Figura 14

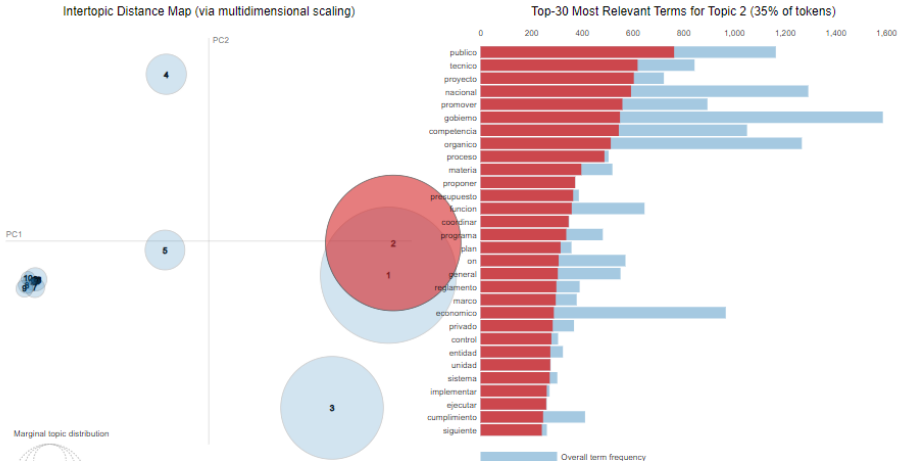
Tópico 1 del resultado del modelo LDA



En la Figura 14, se puede apreciar 5 tópicos visibles y 3 tópicos dominantes. En el tópico 1, se visualiza que un conjunto específico de palabras aparece con alta frecuencia dentro de las ordenanzas regionales, algunas de estas palabras asociadas son: desarrollo, orgánico, económico, social, mujer; que refleja ordenanzas regionales que abordan aspectos relacionado con el desarrollo social.

Figura 15

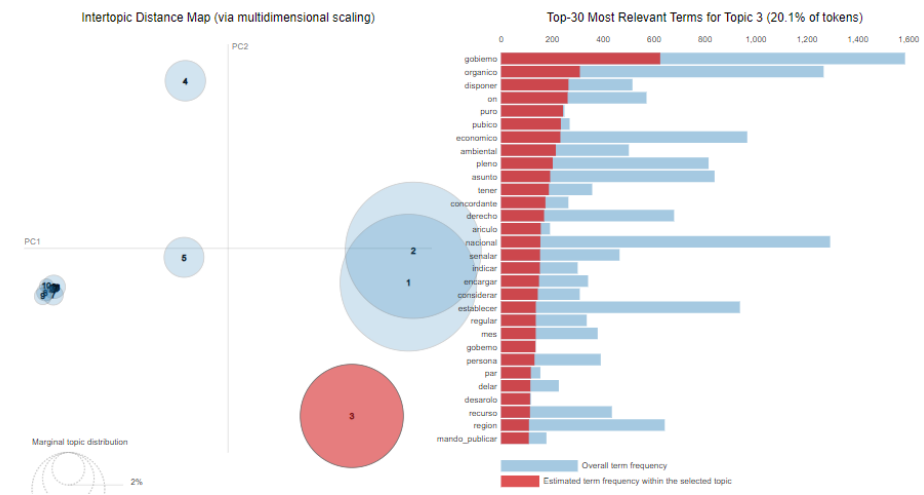
Tópico 2 del resultado del modelo LDA



En la Figura 15, se visualiza el tópico 2, caracterizado por un conjunto específico de palabras que aparecen con alta frecuencia dentro de las ordenanzas regionales, algunas de estas palabras asociadas son: proyecto, promover, presupuesto, programa y económico; que refleja ordenanzas regionales que abordan aspectos relacionado con el desarrollo económico, que es un tema central en las ordenanzas regionales de Puno, reflejando las políticas y estrategias que guían el crecimiento económico regional.

Figura 16

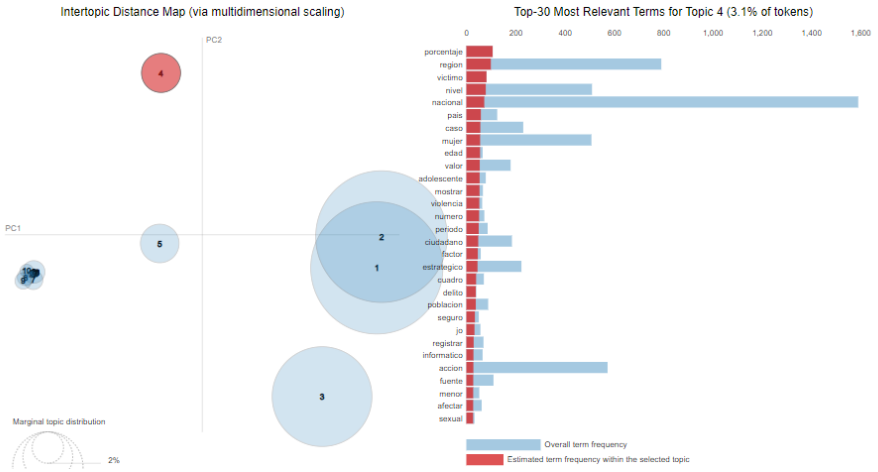
Tópico 3 del resultado del modelo LDA



En la Figura 16, se visualiza el tópico 3, conformado por un conjunto específico de palabras que aparecen con alta frecuencia dentro de las ordenanzas regionales, algunas de estas palabras asociadas son: público, ambiental, derecho, recurso y persona; que refleja ordenanzas regionales que abordan aspectos relacionado con el medio ambiente, que es un tema relevante en las ordenanzas regionales de Puno, se centra en las políticas y regulaciones relacionadas con la protección del medio ambiente y la gestión sostenible de los recursos naturales en la región de Puno.

Figura 17

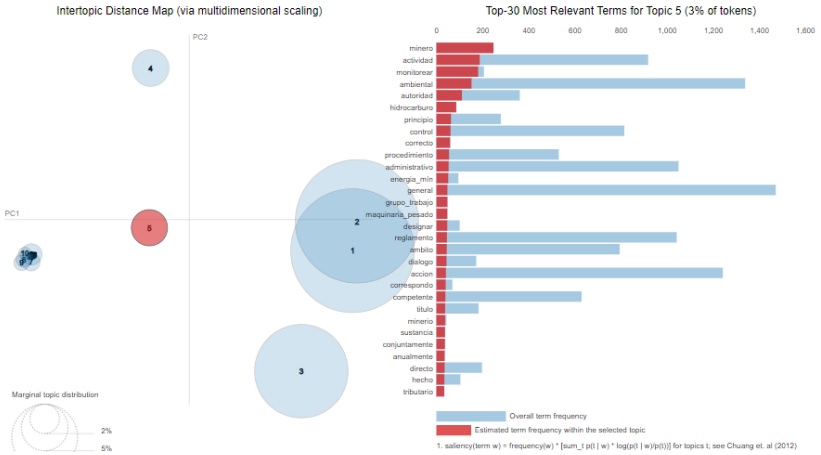
Tópico 4 del resultado del modelo LDA



En la Figura 17, se ilustra el tópico 4, representado por un conjunto específico de palabras que aparecen con mayor frecuencia dentro de las ordenanzas regionales, algunas de estas palabras asociadas son: caso, mujer, adolescente, violencia y delito. Este tópico refleja el creciente enfoque de las políticas regionales en la protección de los derechos humanos, la igualdad de género y la atención a casos de violencia doméstica, agresiones sexuales y delitos relacionados.

Figura 18

Tópico 5 del resultado del modelo LDA



En la Figura 18, se visualiza el tópico 5, caracterizado por un conjunto específico de palabras que aparecen con frecuencia dentro de las ordenanzas regionales, algunas de estas palabras asociadas son: hidrocarburo, minero, control, ambiental y maquinaria pesada. Este tópico se centra en las ordenanzas regionales que abordan la regulación y el control de las actividades extractivas de hidrocarburos y minerales, buscando regular las actividades extractivas de hidrocarburos y minerales, enfocándose especialmente en su impacto ambiental y en el uso de maquinaria pesada durante los procesos de extracción.

A.2 Distribución de tópicos en documentos

Se obtuvo las distribuciones de los tópicos en cada documento del corpus. Esto permitió entender la prevalencia de cada tópico en diferentes ordenanzas y cómo se distribuyen los tópicos a lo largo del corpus.

Tabla 8

Tópicos identificados con el modelo LDA

N° de tópico	Tópicos dominantes	Palabras clave del tópico	Número de documentos
0	14	“Desarrollo”, “orgánico”, “económico”, “social”, "mujer"	106
1	25	“Proyecto”, “promover”, “presupuesto”, “programa”, .ecoonómico"	114
2	23	“Público”, “ambiental”, "derecho", recurso", "persona"	16
3	5	”Caso”, “mujer”, .adolescente", "violencia", "delito"	9
4	3	”Hidrocarburo”, “minero”, çontrol", .ambiental", "maquinaria pesada"	4

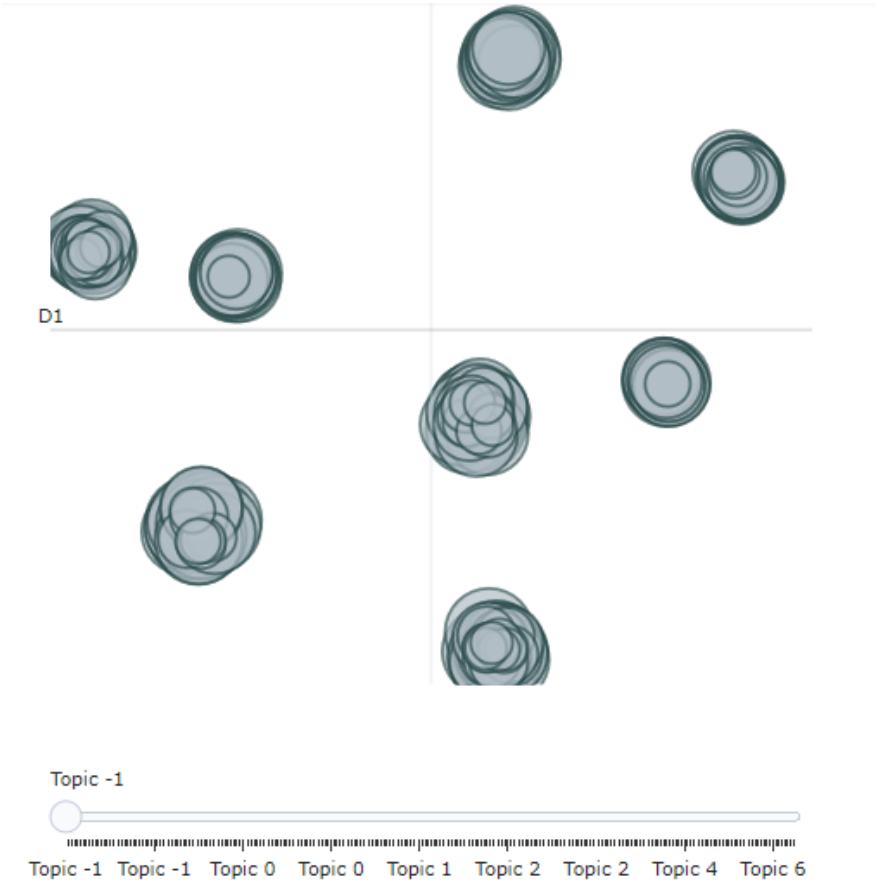
B. Diseño e implementación del modelo de tópicos BERTopic

El diseño e implementación del modelo BERTopic para la identificación de tópicos en ordenanzas regionales de la región Puno comenzó con los textos pre procesados, que fueron vectorizados utilizando modelos de transformadores, como BERT, que capturan las relaciones contextuales y semánticas entre las palabras. Estos embeddings contextuales fueron luego reducidos dimensionalmente usando técnicas como UMAP para facilitar la agrupación de temas.

Una vez obtenidos los embeddings, se utilizó HDBSCAN para agrupar los documentos en tópicos basados en la similitud de sus representaciones vectoriales. Cada clúster representaba un tópico específico, y los términos más representativos de cada clúster fueron extraídos para identificar y nombrar los tópicos.

Figura 19

Visualización de tópicos como resultado del modelo BERTopic



El análisis de las ordenanzas regionales de la región Puno mediante el modelo BERTopic reveló una serie de tópicos clave, que se presentan en la Tabla 9, que reflejan las principales áreas de enfoque y preocupación del gobierno regional durante el período analizado.

Tabla 9

Tópicos identificados con el modelo BERTopic

N° de tópico	Cantidad	Tópico	Representación
0	76	Transporte	"carretera", "vialidad", "construcción" y "mantenimiento"
1	13	Cultura	"escuela", "capacitación", "proyecto educativo" y "cultura"
2	36	Salud	"salud", "hospital", "atención médica" y "seguridad social"
3	34	Presupuesto	"producción", "desarrollo", "economía"
4	14	Ambiental	"riesgo", "desastre", "prevención", "seguridad", "forestación"

A continuación, se detallan los principales tópicos identificados con ambos modelos de tópicos LDA y BERTopic:

1. **Transporte:** Este tópico abarcó un conjunto significativo de documentos relacionados con la planificación, construcción y

mantenimiento de infraestructuras viales, puentes, y redes de transporte público. Las ordenanzas dentro de este tópico frecuentemente mencionaron términos como "carretera", "vialidad", "construcción" y "mantenimiento", indicando un enfoque continuo en mejorar la conectividad y la infraestructura física en la región.

2. **Cultura:** Otro tópico destacado estuvo relacionado con la promoción de la educación y el fortalecimiento de la cultura local. Los documentos agrupados bajo este tema discutieron la implementación de programas educativos, la construcción de centros educativos, y la preservación de tradiciones culturales. Palabras clave como "escuela", "capacitación", "proyecto educativo" y "cultura" fueron recurrentes, reflejando un compromiso por parte de las autoridades regionales en mejorar la calidad educativa y preservar la identidad cultural de Puno.

3. **Salud:** Un tópico crucial identificado por BERTopic estuvo centrado en la salud pública y el bienestar social. Las ordenanzas en este grupo trataron temas como la construcción y equipamiento de centros de salud, campañas de vacunación, y programas de asistencia social para poblaciones vulnerables. Este tópico incluyó términos como "salud", "hospital", "atención médica" y "seguridad social", subrayando las prioridades en la provisión de servicios de salud y el apoyo a las comunidades más necesitadas.

4. **Medio Ambiente:** Las preocupaciones ambientales y la sostenibilidad también surgieron como un tópico relevante. Documentos en este grupo abordaron la protección de recursos naturales, la gestión de residuos, y la promoción de prácticas sostenibles. Palabras como "medio ambiente", "gestión de residuos", "recursos naturales" y "sostenibilidad" fueron prominentes, indicando esfuerzos legislativos dirigidos a la preservación del entorno natural de la región Puno.

5. **Desarrollo:** Finalmente, el análisis identificó un tópico centrado en

el desarrollo económico y la promoción de actividades productivas. Las ordenanzas relacionadas con este tópico incluyeron incentivos para el desarrollo de la agricultura, la promoción del turismo, y el apoyo a pequeñas y medianas empresas. Términos clave como "economía", "turismo", "agricultura" y "emprendimiento" destacaron la orientación del gobierno regional hacia el fortalecimiento de la economía local y la diversificación de las fuentes de ingresos. Estos resultados proporcionan una visión integral de las áreas prioritarias en las ordenanzas regionales de la región Puno, permitiendo a los investigadores y tomadores de decisiones identificar tendencias y áreas de mejora en la gestión regional. El uso de BERTopic y LDA ha facilitado la extracción de estos tópicos de manera precisa y eficiente, ofreciendo un recurso valioso para futuros análisis y planificaciones estratégicas en la región.

4.1.4 Resultado conforme al cuarto objetivo específico

Se evaluó el modelo utilizando métricas de coherencia, perplejidad y diferenciación de tópico, esto de acuerdo a la revisión de la literatura en la Tabla 7: Técnicas de NLP y métricas aplicadas para identificación de tópicos en documentos normativos.

Los tópicos fueron analizados para asegurar que tuvieran sentido en el contexto de las ordenanzas regionales de la región

En la Tabla 10, se presenta el cuadro comparativo entre los resultados obtenidos por LDA y BERTopic, evaluando la coherencia, la perplejidad de los tópcios identificados por cada modelo.

Tabla 10

Comparación de los modelos LDA y BERTopic

Modelo	Métrica de coherencia	Diferenciación de tópicos	Métrica de perplejidad
LDA	0,57	0,43	250
BERTopic	0,67	0,61	120

La evaluación de los modelos LDA y BERTopic en la identificación de tópicos del Gobierno Regional Puno muestra diferencias significativas en su rendimiento. Para LDA, la métrica de coherencia se sitúa en 0,57, lo que indica una moderada relación entre las palabras dentro de los tópicos generados. La diferenciación de tópicos es de 0.43, sugiriendo que los temas presentan una considerable superposición, lo que dificulta su interpretación. Además, una perplejidad de 250 sugiere que el modelo tiene dificultades para predecir los datos, lo que indica que no captura adecuadamente la estructura de los documentos.

Por otro lado, BERTopic muestra un mejor desempeño en todas las métricas evaluadas. Con una coherencia de 0,67 los tópicos generados son más significativos y fáciles de interpretar. La diferenciación de tópicos alcanza un valor de 0,61 lo que significa que los temas son más claros y bien definidos. Finalmente, la perplejidad de 120 indica que BERTopic logra un mejor ajuste a los datos, capturando de manera más efectiva la complejidad de los documentos. En conjunto, estos resultados sugieren que BERTopic es la opción preferible para la identificación de tópicos en este contexto.

Por tanto, la evaluación comparativa de los modelos LDA y BERTopic en la identificación de tópicos del Gobierno Regional Puno revela que BERTopic supera a LDA en términos de coherencia, diferenciación de tópicos y perplejidad. Mientras que LDA presenta métricas que indican un ajuste deficiente y una alta superposición de temas, BERTopic demuestra una mayor claridad y significado en los tópicos generados, lo que lo convierte en la opción más efectiva para este tipo de análisis. Estos hallazgos sugieren que, para la identificación de tópicos en documentos complejos, BERTopic es preferible, proporcionando resultados más interpretables y relevantes.

Al evaluar el modelo de identificación de tópicos BERTopic, la métrica de coherencia se elige como criterio principal debido a su enfoque en la calidad semántica de los tópicos generados. La coherencia mide la relación entre las palabras que componen un tópico, lo que permite asegurar que estos reflejan conceptos claros y significativos. Esta métrica es crucial en contextos donde la

interpretabilidad y la relevancia de los tópicos son fundamentales, ya que tópicos coherentes facilitan la comprensión y aplicación práctica de los resultados.

Por otro lado, aunque la perplejidad y la diferenciación de tópicos son métricas útiles, no capturan de manera efectiva la calidad semántica de los temas. La perplejidad se centra en el ajuste del modelo a los datos, lo que puede no reflejar la interpretabilidad de los tópicos. Asimismo, la diferenciación de tópicos mide la separación entre ellos, pero no garantiza que cada tópico sea cohesivo y significativo. En resumen, la métrica de coherencia proporciona una evaluación más directa de la utilidad y claridad de los tópicos generados por BERTopic, lo que la convierte en la opción más adecuada para este análisis.

4.1.5 Prueba de hipótesis

La prueba de hipótesis del modelo de identificación de tópicos en ordenanzas regionales del gobierno regional puno de 2010 a 2024 utilizando técnicas de procesamiento de lenguaje natural, se planteó de la siguiente manera:

A. Estadístico t

La fórmula para el valor t en una prueba para muestras emparejadas es:

$$t = \frac{\bar{x} - \mu}{S/\sqrt{n}}$$

B. Nivel de significancia

Para comprobar la prueba de hipótesis, se consideró un nivel de significancia de 0,05 o 5 % de error:

$$\alpha = 0{,}05 = 5\,\%$$

Se utilizó la distribución t-Student para el área de una cola en relación al grado de libertad GL, $n - 1 = 2$, donde n es el número de especies, por lo tanto, el valor de t_α es:

$$t_{\alpha} = t_{0,05} = 2,92$$

C. Respecto al modelo LDA

H_0: El modelo de identificación de tópicos basados en LDA no obtiene los mejores resultados de coherencia en la identificación de tópicos de las ordenanzas regionales emitidas del 2010 al 2024 por el Gobierno Regional Puno.

H_1: El modelo de identificación de tópicos basados en LDA obtiene los mejores resultados de coherencia en la identificación de tópicos de las ordenanzas regionales emitidas del 2010 al 2024 por el Gobierno Regional Puno.

Tabla 11

Prueba de muestra única en función del modelo LDA

Modelo	**Prueba de muestra única**			
	Nivel de Coherencia			
LDA	N	X	S	μ
	especies	coherencia		
	3	0,57	0,2	0,50

En el reemplazo de la ecuación, se usaron los datos de la Tabla 11, rechazamos la hipótesis alterna, pues el valor de t = 0,61 está en la zona de aceptación. Entonces se acepta la hipótesis nula, que señala que el modelo de identificación de tópicos basados en LDA no obtiene los mejores resultados de coherencia en la identificación de tópicos de las ordenanzas regionales emitidas del 2010 al 2024 por el Gobierno Regional Puno.

D. Respecto al modelo BERTopic

H_0: El modelo de identificación de tópicos basados en BERTopic no obtiene los mejores resultados de coherencia en la identificación de tópicos de las ordenanzas regionales emitidas del 2010 al 2024 por el Gobierno Regional Puno.

H_1: El modelo de identificación de tópicos basados en BERTopic obtiene los mejores resultados de coherencia en la identificación de tópicos de las ordenanzas regionales emitidas del 2010 al 2024 por el Gobierno Regional Puno.

Tabla 12

Prueba de muestra única en función del modelo BERTopic

	Prueba de muestra única			
Modelo	**Nivel de Coherencia**			
BERTopic	N	X	S	μ
	especies	coherencia		
	3	0,67	0,09	0,50

En el reemplazo de la ecuación, se usaron los datos de la Tabla 11, rechazamos la hipótesis nula, pues el valor de t = 3,28 está en la zona de rechazo. Entonces se acepta la hipótesis alterna, que señala que el modelo de identificación de tópicos basados en BERTopic obtiene los mejores resultados de coherencia en la identificación de tópicos de las ordenanzas regionales emitidas del 2010 al 2024 por el Gobierno Regional Puno.

Finalmente, de las Tablas 11 y 12, se observa que el modelo de BERTopic obtiene los mejores resultados de coherencia en la identificación de tópicos de las ordenanzas regionales emitidas del 2010 al 2024 por el Gobierno Regional Puno.

4.2 Discusión

La presente investigación busca avanzar en el análisis temático de documentos legales a través del modelado de tópicos, abordando las ordenanzas regionales emitidas por el Gobierno Regional de Puno desde 2010 hasta 2024.

El corpus es esencial para capturar la evolución y temas recurrentes en las políticas regionales. La metodología empleada en la construcción de corpus, aborda la extracción, el pre procesamiento y análisis de la estructura de documentos legales y gubernamentales, garantizando una compilación representativa y manejable para un análisis de modelado de tópicos como Sánchez (2019) y Dyevre (2021) destacan. Los autores como McInnes et al.

(2018), con su trabajo para reducción de dimensionalidad, y Mazumder y Barui (2021), con su investigación sobre LDA, coinciden en que el éxito de los modelos depende en gran parte del pre procesamiento de datos. Este proceso de preparación se vuelve crucial en contextos legislativos y normativos, donde el lenguaje especializado y formal puede presentar desafíos en el análisis.

La creación de un corpus estructurado y adecuado es un paso fundamental en cualquier análisis de modelado de tópicos. Hosseiny Marani y Baumer (2023) destacan la importancia de la estabilidad y coherencia del corpus para garantizar que el modelo sea capaz de identificar temas significativos y estables. En esta investigación se permite a los modelos captar mejor las estructuras temáticas, especialmente relevante en contextos de análisis histórico y legislativo, como sugiere el estudio de Grajzl y Murrell (2022) en el campo de la historia legal. Estos autores enfatizan cómo la preparación del corpus influye en la calidad de los resultados de modelo de tópicos, subrayando la necesidad de datos textuales bien estructurados en estudios normativos.

En la identificación de los algoritmos más adecuados en concordancia con los estudios de Chauhan y Shah (2022) y Jelodar et al. (2019) se emplearon los modelos de tópicos como LDA (Latent Dirichlet Allocation) y BERTopic. La comparación de algoritmos en esta investigación puede orientar la selección de técnicas que se adapten mejor a la estructura de las ordenanzas de Puno. La discusión en torno a BERTopic (Grootendorst, 2022) y el uso de BERT para mejorar la detección de temas en textos legales también es valiosa, dada la complejidad y especificidad del lenguaje legal que contienen las ordenanzas.

En el proceso de identificación de tópicos en las ordenanzas regionales de la región de Puno, se utilizaron dos enfoques diferentes: Latent Dirichlet Allocation (LDA) y BERTopic. Ambos modelos ofrecen métodos potentes para el análisis temático, pero presentan diferencias clave en su diseño, implementación, y resultados, lo que impacta en la interpretación y utilidad de los tópicos identificados.

LDA es un modelo probabilístico generativo tradicional que ha sido ampliamente utilizado para la identificación de tópicos en grandes colecciones de textos. Su fortaleza radica en su simplicidad y en su capacidad para descomponer documentos en una mezcla de tópicos, donde cada tópico está representado por una distribución de palabras. Sin

embargo, LDA tiene limitaciones en su capacidad para capturar la semántica contextual compleja, especialmente en textos normativos que suelen contener un lenguaje técnico y formal. Además, la interpretación de los tópicos generados por LDA puede ser menos precisa, ya que depende en gran medida de la selección manual de palabras clave y del número de tópicos, que debe ser definido a priori. Aunque LDA es un modelo robusto para la identificación de tópicos, presenta ciertas limitaciones. Asuncion et al. (2012) señalan que LDA puede tener dificultades para capturar relaciones complejas entre tópicos debido a su naturaleza bag-of-words, y su rendimiento puede verse afectado por la necesidad de predefinir el número de tópicos. Además, Wallach et al. (2009) discuten cómo la calidad de los resultados de LDA depende en gran medida de la elección de los hiperparámetros y del tamaño del corpus.

Por otro lado, BERTopic aprovecha los modelos de transformadores como BERT para capturar relaciones semánticas más profundas y contextuales en los textos. Esto es particularmente beneficioso en documentos normativos, donde el significado de una palabra puede depender fuertemente de su contexto. BERTopic, al combinar embeddings contextuales con técnicas de reducción dimensional y agrupamiento no supervisado, como UMAP y HDBSCAN, permite una identificación de tópicos más flexible y dinámica. Los tópicos generados por BERTopic suelen ser más coherentes y contextualmente relevantes, lo que facilita una interpretación más precisa y detallada de las políticas públicas presentes en las ordenanzas. Sin embargo, es importante destacar que BERTopic, aunque más potente en términos de análisis contextual, también requiere mayor poder computacional y puede ser más complejo de ajustar y personalizar en comparación con LDA. Además, la naturaleza no supervisada de HDBSCAN en BERTopic puede llevar a la identificación de tópicos que, aunque coherentes, podrían no ser inmediatamente interpretables sin una adecuada revisión por expertos.

Los resultados obtenidos al aplicar BERTopic para analizar las ordenanzas regionales en Puno revelan un enfoque avanzado en modelado de tópicos, cuya efectividad puede contextualizarse a partir del trabajo de Blei et al. (2010), quiénes establecieron las bases de los modelos probabilísticos en NLP. En su investigación, destacan cómo el modelado de tópicos permite comprender patrones en grandes volúmenes de texto, siendo relevante para este tipo de análisis. Sin embargo, BERTopic logra una mejora significativa en coherencia y adaptabilidad a contextos específicos

como el de ordenanzas regionales. Este hallazgo se alinea con el trabajo de Grootendorst (2022), quien desarrolló BERTopic y probó su eficacia para identificar temas en texto, optimizando la detección de tópicos menos evidentes, pero coherentes en corpus especializados.

En la evaluación de la efectividad del modelo, se aplicaron las métricas de coherencia y perplejidad, sin embargo, Hosseiny Hosseiny Marani y Baumer (2023) revisan métodos de estabilidad, útiles para verificar la consistencia de los temas a través de múltiples ejecuciones. Por otro lado, Wallach et al. (2009) resaltan cómo las métricas de coherencia, aplicadas a modelos como LDA, ayudan a evaluar la calidad de los temas generados, por su parte, Rüdiger et al. (2022) también sugieren métricas de interpretabilidad y ajuste en el contexto legal, ofreciendo un marco robusto para medir la precisión en temas jurídicos y gubernamentales.

La transparencia y equidad en el análisis de políticas públicas también son cuestiones clave. Binns (2018) explora cómo los algoritmos deben ser interpretados con cuidado en contextos de relevancia social y política, ya que cualquier sesgo inherente en los modelos de análisis de datos podría reflejarse en interpretaciones sesgadas de las políticas públicas. En esta investigación, es esencial entender que los tópicos identificados en las ordenanzas regionales reflejan las prioridades y enfoques de los legisladores, pero también las limitaciones del corpus.

La aplicabilidad de BERTopic en contextos locales sugiere que este modelo podría escalarse para el análisis de normativas en otras regiones. Según el análisis de Lange et al. (2021) en textos islámicos, el uso de NLP en análisis legales ayuda a descubrir patrones no evidentes y a crear perfiles legislativos regionales. Asimismo, la capacidad de BERTopic para adaptarse a contextos particulares, como leyes y regulaciones locales, confirma que estos modelos pueden servir de apoyo en el diseño de políticas que respondan de manera más adecuada a las necesidades de cada región.

La investigación motiva a la aplicación prácticas de los resultados del análisis de tópicos obtenidos, como la clasificación de nuevas ordenanzas, la identificación de tendencias legislativas regionales, y la asistencia en la redacción y revisión de nuevas ordenanzas basadas en temas emergentes.

CONCLUSIONES

- Se determinó un modelo para la identificación de tópicos en las ordenanzas regionales emitidas por el Gobierno Regional Puno durante el periodo 2010-2024, este modelo BERTopic identifica correctamente en un 67 %, es decir una coherencia en la categorización de los temas tratados en las ordenanzas facilitando un mayor entendimiento del impacto y la evolución de las normativas en la región Puno.
- La construcción de un corpus completo de las ordenanzas regionales emitidas por el Gobierno Regional de Puno desde 2010 hasta 2024 ha sido un logro significativo en el análisis legislativo. Se elaboró un conjunto de datos, la tarea con el mayor trabajo involucrado corresponde al extraer a información de los documentos escaneados del portal institucional del Gobierno Regional, con los cuáles se logró recabar un corpus con un total de 249 documentos.
- La selección de algoritmos de identificación de tópicos adecuados para la identificación de tópicos de las ordenanzas regionales ha revelado las ventajas y limitaciones de dos enfoques prominentes: LDA (Latent Dirichlet Allocation) y BERTopic, siendo los dos modelos ampliamente usado en la revisión de la literatura.
- El diseño e implementación del modelo de análisis de tópicos utilizando los algoritmos seleccionados: LDA y BERTopic y el corpus de ordenanzas regionales ha sido exitoso. El proceso se dio desde la extracción de los documentos, pre procesamiento de la información, el entrenamiento y evaluación del modelo.
- Se evaluó el rendimiento del modelo de tópicos LDA, haciendo uso de las métricas de coherencia y perplejidad, obteniéndose en la coherencia 0,57 por otro lado, en la métrica de perplejidad se obtuvo 250. Respecto al rendimiento del modelo BERTopic se evaluó con la métrica de coherencia obteniéndose 0,67 y en la métrica de perplejidad se obtuvo 120, lo que indica que el modelo es aceptable para la identificación de tópicos en ordenanzas regionales.

RECOMENDACIONES

- Se recomienda adoptar el modelo BERTopic como el modelo principal para el análisis de ordenanzas regionales. Su capacidad para captar temas contextuales y su alta precisión hacen de este modelo una herramienta valiosa para el estudio de las normativas regionales.
- Se recomienda el uso de técnicas de Web Scraping para la extracción de documentos de sitios Web públicos, así mismo estandarizar el formato de los documentos recopilados para facilitar su procesamiento y análisis posterior y el uso de OCR para la extracción de información en textos escaneados.
- Se recomienda para la selección de los algoritmos de identificación de tópicos más adecuados para el análisis de ordenanzas regionales, realizar una comparación exhaustiva entre diferentes algoritmos de identificación de tópicos, considerando tanto enfoques tradicionales (LDA) como métodos avanzados (BERTopic).
- Se recomienda en el diseño del modelo, realizar un pre procesamiento personalizado del contenido textual del conjunto de datos, empleando una lista específica de stopwords adaptada al contexto del estudio y utilizando técnicas avanzadas como la lematización.
- Se recomienda validar los resultados del modelo con expertos en el área de legislación regional para garantizar que los tópicos identificados sean relevantes y precisos.

BIBLIOGRAFÍA

Abdelrazek, A., Eid, Y., Gawish, E., Medhat, W., y Hassan, A. (2023). Topic modeling algorithms and applications: A survey. https://doi.org/10.1016/j.is.2022.102131

Aráoz, M. (2015). Desarrollo regional y políticas de descentralización en Perú. *Editorial San Marcos.*

Ashihara, K., Vaigh, C. B. E., Chu, C., Renoust, B., Okubo, N., Takemura, N., Nakashima, Y., y Nagahara, H. (2020). Improving topic modeling through homophily for legal documents. *Applied Network Science*, *5*. https://doi.org/10.1007/s41109-020-00321-y

Asuncion, A., Welling, M., Smyth, P., y Teh, Y. W. (2012). On smoothing and inference for topic models. *arXiv preprint arXiv:1205.2662*.

Binns, R. (2018). Fairness in machine learning: Lessons from political philosophy. *Conference on Fairness, Accountability and Transparency*.

Blei, D., Carin, L., y Dunson, D. (2010). Probabilistic topic models. *IEEE Signal Processing Magazine*, *27*, 55-65. https://doi.org/10.1109/MSP.2010.938079

Brynjolfsson, E., y McAfee, A. (2014). The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies. *W. W. Norton y Company*.

Buenaño-Fernandez, D., González, M., Gil, D., y Luján-Mora, S. (2020). Text Mining of Open-Ended Questions in Self-Assessment of University Teachers: An LDA Topic Modeling Approach. *IEEE Access*, *8*, 35318-35330. https://doi.org/10.1109/ACCESS.2020.2974983

Chauhan, U., y Shah, A. (2022). Topic Modeling Using Latent Dirichlet allocation: A Survey. *ACM Computing Surveys*, *54*. https://doi.org/10.1145/3462478

Devlin, J., Chang, M. W., Lee, K., y Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Dillan, T., y Fudholi, D. H. (2023). LDAViewer: An Automatic Language-Agnostic System for Discovering State-of-the-Art Topics in Research Using Topic Modeling, Bidirectional Encoder Representations From Transformers, and Entity Linking. *IEEE Access*, *11*. https://doi.org/10.1109/ACCESS.2023.3285116

Dyevre, A. (2021). Text-mining for Lawyers: How Machine Learning Techniques Can Advance our Understanding of Legal Discourse. *Erasmus Law Review*, *14*. https://doi.org/10.5553/elr.000191

Dyevre, A., Glavina, M., y Ovádek, M. (2021). The voices of european law: Legislators, judges and law professors. *German Law Journal*, *22*, 956-982. https://doi.org/10.1017/glj.2021.47

Fahlevvi, M. R., y Azhari, S. (2022). Topic Modeling on Online News. Portal Using Latent Dirichlet Allocation (LDA). *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, *16*(4), 335-344. https://doi.org/https://doi.org/10.22146/ijccs.74383

Gamboa Unsihuay, J. E. (2019). Topic modeling en twitter: determinación de la agenda política peruana en el periodo de enero a setiembre del 2018. *Anales Científicos*, *80*(2), 308-327. https://doi.org/https://doi.org/10.21704/ac.v80i2.1446

Gobierno Regional Puno. (2024). *Ordenanzas Regionales Gobierno Regional Puno: Portal Institucional del Gobierno Regional de Puno*. https://www.regionpuno.gob.pe/ordenanzas-regionales/

Gonzales de Olarte, E. (2003). El sistema educativo en Perú: avances y desafíos. Instituto de Estudios Peruanos.

Grajzl, P., y Murrell, P. (2022). Using Topic-Modeling in Legal History, with an Application to Pre-Industrial English Case Law on Finance. *Law and History Review*, *40*. https://doi.org/10.1017/S0738248022000153

Grisales-Aguirre, A. M., y Figueroa-Vallejo, C. J. (2022). Modelado de tópicos aplicado al análisis del papel del aprendizaje automático en revisiones sistemáticas. *Revista de Investigación, Desarrollo e Innovación*, *12*, 279-292. https://doi.org/10.19053/20278306.v12.n2.2022.15271

Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*.

Halgekar, A., Rao, A., Khankhoje, D., Khetan, I., y Bhowmick, K. (2023). Topic Modelling-Based Approach for Clustering Legal Documents. *Lecture Notes in Networks and Systems*, *400*. https://doi.org/10.1007/978-981-19-0095-2_17

Hamilton, L. M., y Lahne, J. (2022). Natural Language Processing. https://doi.org/10.1016/B978-0-12-821936-2.00004-2

Hernández-Sampieri, R., y Mendoza, C. (2020). Metodología de la investigación: las rutas cuantitativa, cualitativa y mixta.

Honnibal, M. (2015). SpaCy Documentation. *SpaCy.io*.

Hosseiny Marani, A., y Baumer, E. P. (2023). A review of stability in topic modeling: Metrics for assessing and techniques for improving stability. *ACM Computing Surveys*, *56*(5), 1-32. https://doi.org/https://doi.org/10.1145/3623269

Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., y Zhao, L. (2019). Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. *Multimedia tools and applications*, *78*, 15169-15211. https://doi.org/https://doi.org/10.1007/s11042-018-6894-4

Kherwa, P., y Bansal, P. (2020). Semantic N-Gram Topic Modeling. *EAI Endorsed Transactions on Scalable Information Systems*, *7*. https://doi.org/10.4108/eai.13-7-2018.163131

Khurana, D., Koli, A., Khatter, K., y Singh, S. (2023). Natural language processing: state of the art, current trends and challenges. *Multimedia Tools and Applications*, *82*. https://doi.org/10.1007/s11042-022-13428-4

Lange, C., Latief, M. A., Çelik, Y., Lyklema, A. M., Kuppevelt, D. E. V., y Zwaan, J. V. D. (2021). Text mining Islamic law. https://doi.org/10.1163/15685195-bja10009

Lee, H., Lee, S. H., Lee, K. R., y Kim, J. H. (2023). ESG Discourse Analysis Through BERTopic: Comparing News Articles and Academic Papers. *Computers, Materials and Continua*, *75*, 6023-6037. https://doi.org/10.32604/cmc.2023.039104

Martino, G. D., Pio, G., y Ceci, M. (2022). PRILJ: an efficient two-step method based on embedding and clustering for the identification of regularities in legal case judgments. *Artificial Intelligence and Law*, *30*. https://doi.org/10.1007/s10506-021-09297-1

Mazumder, S., y Barui, T. (2021). Discovering topics from the titles of the Indian LIS theses. *Library Philosophy and Practice (e-journal)*, 1-23.

McInnes, L., Healy, J., y Melville, J. (2018). Uniform manifold approximation and projection for dimension reduction. arXiv.

Mendoza, R., y García, J. (2010). Reformas en el sector salud en Perú: Una evaluación. *Revista Peruana de Salud Pública*.

Mifrah, S., y Benlahmar, E. H. (2022). Topic Modeling with Transformers for Sentence-Level Using Coronavirus Corpus. *International Journal of Interactive Mobile Technologies*, *16*. https://doi.org/10.3991/ijim.v16i17.33281

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., y Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, *26*.

Murakami, R., y Chakraborty, B. (2022). Investigating the Efficient Use of Word Embedding with Neural-Topic Models for Interpretable Topics from Short Texts. *Sensors*, *22*. https://doi.org/10.3390/s22030852

Murshed, B. A. H., Mallappa, S., Abawajy, J., Saif, M. A. N., Al-ariki, H. D. E., y Abdulwahab, H. M. (2023). Short text topic modelling approaches in the context of big data: taxonomy, survey, and analysis. *Artificial Intelligence Review*, *56*, 5133-5260. https://doi.org/10.1007/s10462-022-10254-w

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, *12*, 2825-2830.

Pennington, J., Socher, R., y Manning, C. D. (2014). GloVe: Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Péter, G., Miklós, S., y Orsolya, R. (2022). The opportunities and constraints of topic modelling – the case of a corpus of laws. *Statisztikai Szemle*, *100*, 783-814. https://doi.org/10.20311/STAT2022.8.HU0783

Qiang, J., Qian, Z., Li, Y., Yuan, Y., y Wu, X. (2022). Short Text Topic Modeling Techniques, Applications, and Performance: A Survey. https://doi.org/10.1109/TKDE.2020.2992485

Rawat, A. J., Ghildiyal, S., y Dixit, A. K. (2022). Topic modelling of legal documents using NLP and bidirectional encoder representations from transformers. *Indonesian Journal of Electrical Engineering and Computer Science*, *28*, 1749-1755. https://doi.org/10.11591/ijeecs.v28.i3.pp1749-1755

Reyes, M. A. (2017). Legislatura fallida e investidura convulsa. anÁlisis y consecuencias. *Revista Espanola de Derecho Constitucional*. https://doi.org/10.18042/cepc/redc.109.01

Richardson, L. (2023). Beautiful Soup 4.12.0 documentation. *Beautiful Soup Documentation*.

Rodríguez Urquiaga, R. J. (2018). *Análisis visual de la evolución de temas en Corpus de documentos usando árboles de Similitud* [Tesis de maestría, Universidad Nacional de San Agustín de Arequipa].

Rüdiger, M., Antons, D., Joshi, A. M., y Salge, T.-O. (2022). Topic modeling revisited: New evidence on algorithm performance and quality metrics. *Plos one*, *17*(4), e0266325. https://doi.org/https://doi.org/10.1371/journal.pone.0266325

Rudolph, M., y Feldman, J. (2018). Gensim: Topic Modeling for Humans. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Russell, S., y Norvig, P. (2020). Artificial Intelligence: A Modern Approach. *Pearson*.

Sánchez, V. (2019). Evaluación de Políticas Públicas en Perú a través de Modelado de Tópicos. *Universidad Nacional del Altiplano*. https://repositorio.unmsm.edu.pe/handle/20.500.12672/12345

Silveira, R., Fernandes, C. G. O., Neto, J. A. M., Furtado, V., y Filho, J. E. P. (2021). Topic modelling of legal documents via LEGAL-BERT. *CEUR Workshop Proceedings*, *2896*. https://doi.org/10.2139/ssrn.4539091

Vayansky, I., y Kumar, S. A. P. (2020). A review of topic modeling methods. *Information Systems*, *94*. https://doi.org/10.1016/j.is.2020.101582

Wallach, H., Mimno, D., y McCallum, A. (2009). Rethinking LDA: Why priors matter. *Advances in neural information processing systems*, *22*.

Wang, Z., Chen, J., Chen, J., y Chen, H. (2023). Identifying interdisciplinary topics and their evolution based on BERTopic. *Scientometrics*, 1-26. https://doi.org/https://doi.org/10.1007/s11192-023-04776-5

Wendel, L., Shadrova, A., y Tischbirek, A. (2022). From Modeled Topics to Areas of Law: A Comparative Analysis of Types of Proceedings in the German Federal Constitutional Court. *German Law Journal*, *23*, 493-531. https://doi.org/10.1017/glj.2022.39

Zankadi, H., Idrissi, A., Daoudi, N., y Hilal, I. (2023). Identifying learners' topical interests from social media content to enrich their course preferences in MOOCs using topic modeling and NLP techniques. *Education and Information*

Technologies, *28*(5), 5567-5584. https://doi.org/https://doi.org/10.1007/s10639-022-11373-1

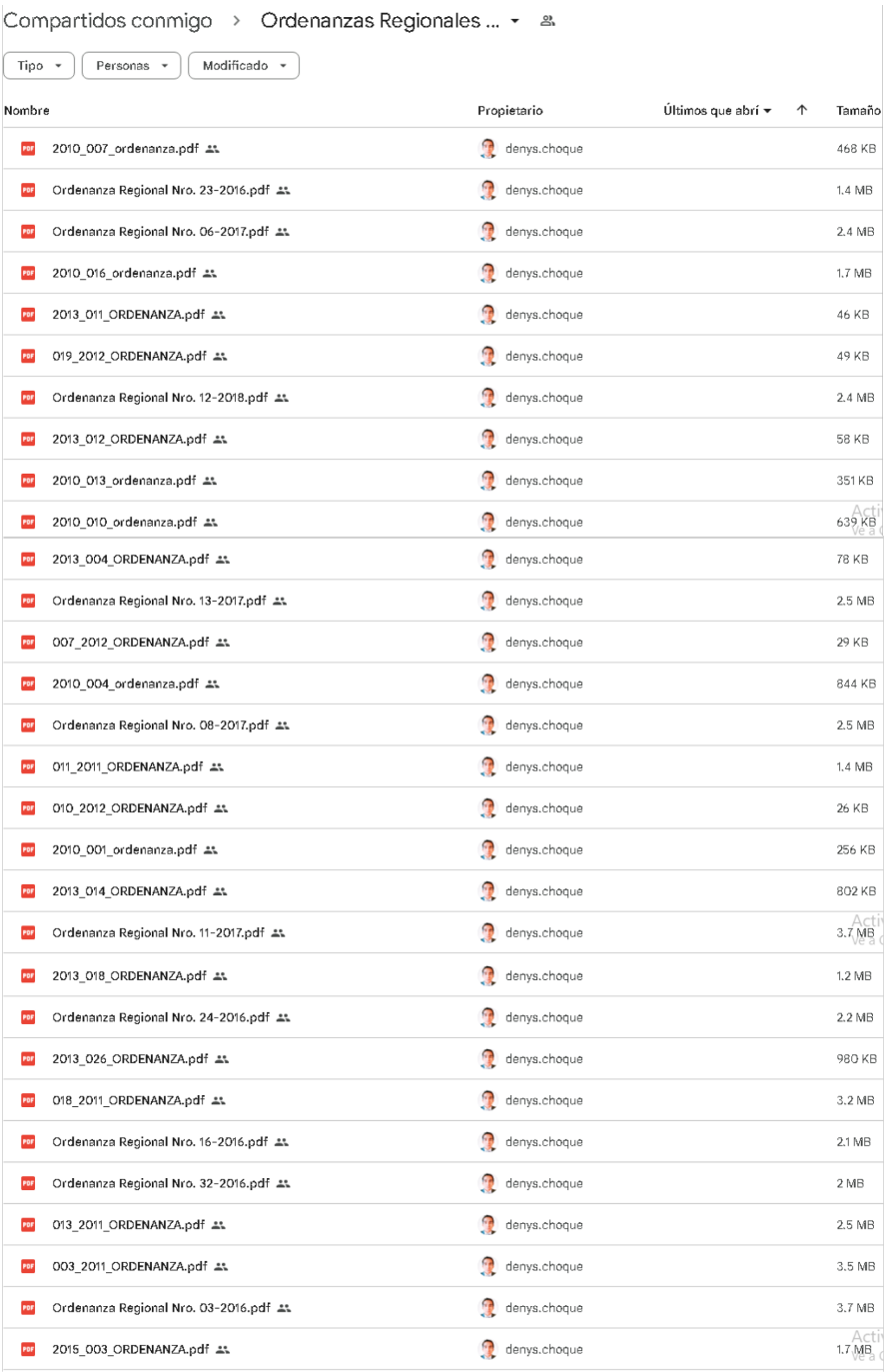
ANEXOS

Anexo 1. Matriz de consistencia

Problema	Hipótesis	Objetivo	Variables	Indicadores	Métodos	Prueba Estadística
Problema General	**Hipótesis General**	**Objetivo General**		**General**	**General**	
PG: ¿Qué modelo de procesamiento de lenguaje natural obtiene una mayor coherencia en el análisis de tópicos en el corpus de ordenanzas regionales de 2010 a 2024 del Gobierno Regional Puno?	HG: El modelo de análisis de tópicos seleccionado permite analizar con mayor coherencia los tópicos en ordenanzas regionales del 2010 al 2024 del Gobierno Regional Puno.	OG: Determinar un modelo de análisis de tópicos en las ordenanzas regionales emitidas por el Gobierno Regional Puno que obtenga los mejores resultados en la coherencia de tópicos durante el periodo 2010 a 2024.	VI: Modelo de tópicos con mayor efectividad. VD: Análisis de tópicos en ordenanzas regionales del Gobierno Regional Puno de 2010 a 2024.	Coherencia: Cuyo valor entre 0 y 1, donde 0 indica falta de coherencia y 1 indica máxima coherencia	Métrica de análisis de tópicos	Para la prueba de hipótesis se realizará con el estadístico de prueba t-student y será aplicado a cada técnica de Procesamiento de Lenguaje Natural utilizado $H_0 : \mu \leq \mu_0$ $H_1 : \mu \leq \mu_0$ La hipótesis nula H_0 señala que μ es menor o igual que μ_0 y la hipótesis alterna nos indica que μ es mayor que μ_0 $t = \frac{\bar{x} - \mu}{\frac{S}{\sqrt{n}}}$
Preguntas Especificas	**Hipótesis Especificas**	**Objetivos Específicos**		**Específico**	**Específico**	
PE 1: ¿Cómo se construye un corpus y cómo se realiza el pre procesamiento de las ordenanzas regionales del 2010 al 2024 del Gobierno Regional Puno?	HG 1: Las técnicas y herramientas NLP permiten construir un corpus de ordenanzas regionales del 2010 al 2024 del Gobierno Regional Puno.	OE 1: Construir un corpus de ordenanzas regionales emitidas por el Gobierno Regional Puno desde 2010 hasta 2024.	VI: Extracción y pre procesamiento de ordenanzas regionales del Gobierno Regional Puno de 2010 a 2024. VD: Corpus construido.	Data set Construido y pre procesado de las ordenanzas regionales del Gobierno Regional Puno de 2010 a 2024.	Extracción - Web Scraping Pre Procesamiento - Stopwords - Tokenización - Lematización	

PE2: ¿Cuáles son los algoritmos más relevantes del procesamiento de lenguaje natural utilizados para el análisis de tópicos?	HG2: Existen algoritmos de modelado de tópicos relevantes que analizan los tópicos en ordenanzas regionales.	OE2: Seleccionar algoritmos de análisis de tópicos adecuados que permitan analizar tópicos en las ordenanzas regionales.	VI: Métrica de coherencia de las técnicas de NLP. VD: Técnicas de NLP para el modelado de tópicos.	Algoritmos seleccionados de Procesamiento de Lenguaje Natural.	Métrica de Coherencia	
PE3: ¿Cuáles son los pasos clave en el proceso de diseño de un modelo de análisis de tópicos en ordenanzas regionales?	HG3: Los pasos seguidos de la metodología KDD permiten diseñar el modelado de análisis de tópicos en ordenanzas regionales.	OE3: Diseñar el modelo de análisis de tópicos utilizando los algoritmos seleccionados y el corpus de ordenanzas regionales.	VI: Metodología KDD VD: Proceso de análisis de tópicos.	Modelo de tópicos para ordenanzas regionales.	Pasos de la metodología KDD.	
PE4: ¿Cuál es la coherencia del modelo de tópicos aplicado a ordenanzas regionales utilizando métricas de análisis de tópicos?	HG4: El modelo propuesto tiene una efectividad aceptable en el análisis de tópicos en ordenanzas regionales.	OE4: Evaluar la efectividad del modelo desarrollado mediante métricas de análisis de tópicos.	VI: Métricas de análisis de tópicos VD: Modelo de análisis de tópicos.	Valor obtenido de las métricas de análisis de tópicos.	Métricas de análisis de tópicos	

Anexo 2. Ordenanzas regionales emitidas por el Gobierno Regional Puno durante el periodo 2010 – 2024



Compartidos conmigo › Ordenanzas Regionales ...

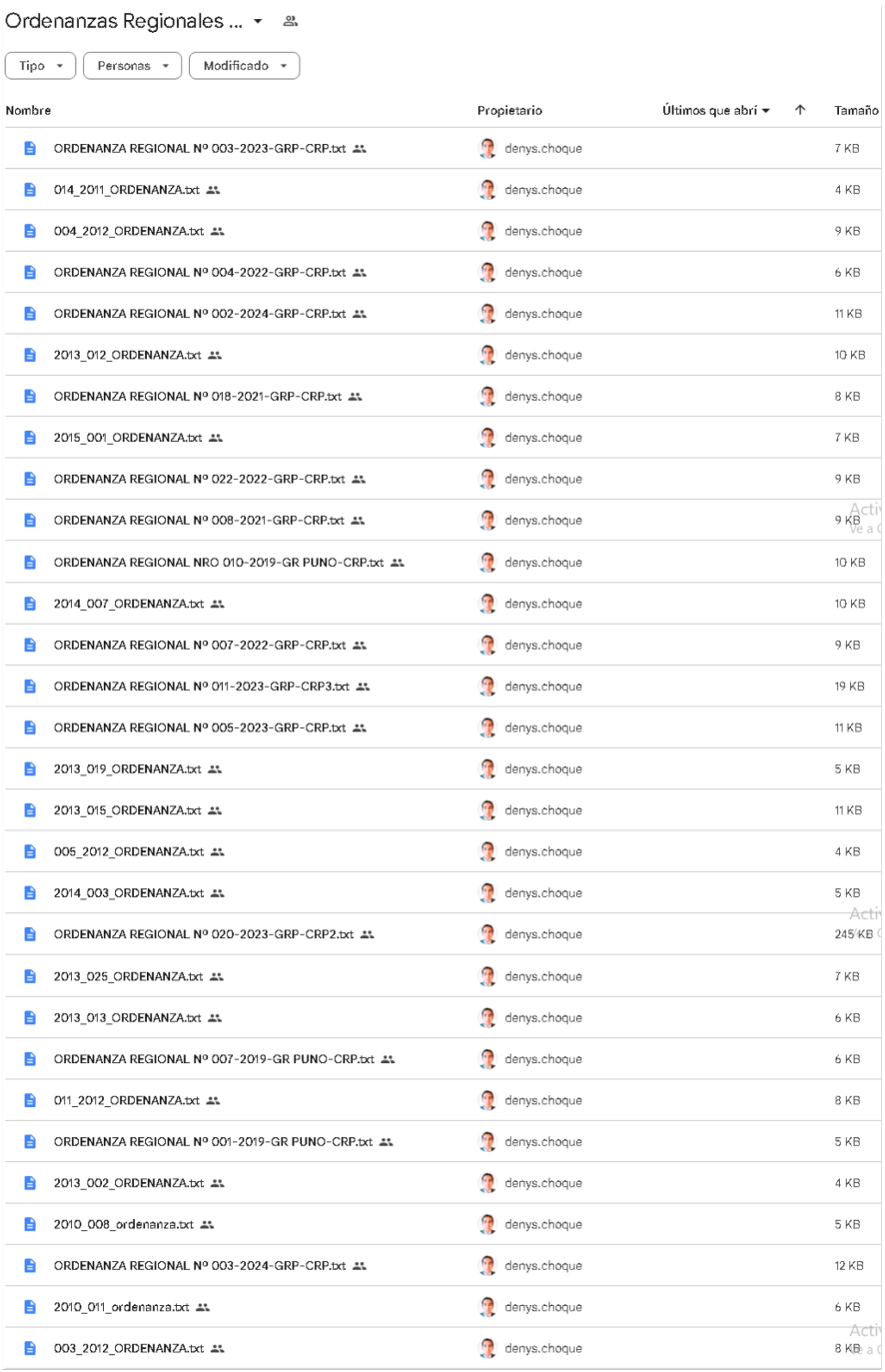
Tipo | Personas | Modificado

Nombre	Propietario	Últimos que abrí	Tamaño
2010_007_ordenanza.pdf	denys.choque		468 KB
Ordenanza Regional Nro. 23-2016.pdf	denys.choque		1.4 MB
Ordenanza Regional Nro. 06-2017.pdf	denys.choque		2.4 MB
2010_016_ordenanza.pdf	denys.choque		1.7 MB
2013_011_ORDENANZA.pdf	denys.choque		46 KB
019_2012_ORDENANZA.pdf	denys.choque		49 KB
Ordenanza Regional Nro. 12-2018.pdf	denys.choque		2.4 MB
2013_012_ORDENANZA.pdf	denys.choque		58 KB
2010_013_ordenanza.pdf	denys.choque		351 KB
2010_010_ordenanza.pdf	denys.choque		639 KB
2013_004_ORDENANZA.pdf	denys.choque		78 KB
Ordenanza Regional Nro. 13-2017.pdf	denys.choque		2.5 MB
007_2012_ORDENANZA.pdf	denys.choque		29 KB
2010_004_ordenanza.pdf	denys.choque		844 KB
Ordenanza Regional Nro. 08-2017.pdf	denys.choque		2.5 MB
011_2011_ORDENANZA.pdf	denys.choque		1.4 MB
010_2012_ORDENANZA.pdf	denys.choque		26 KB
2010_001_ordenanza.pdf	denys.choque		256 KB
2013_014_ORDENANZA.pdf	denys.choque		802 KB
Ordenanza Regional Nro. 11-2017.pdf	denys.choque		3.7 MB
2013_018_ORDENANZA.pdf	denys.choque		1.2 MB
Ordenanza Regional Nro. 24-2016.pdf	denys.choque		2.2 MB
2013_026_ORDENANZA.pdf	denys.choque		980 KB
018_2011_ORDENANZA.pdf	denys.choque		3.2 MB
Ordenanza Regional Nro. 16-2016.pdf	denys.choque		2.1 MB
Ordenanza Regional Nro. 32-2016.pdf	denys.choque		2 MB
013_2011_ORDENANZA.pdf	denys.choque		2.5 MB
003_2011_ORDENANZA.pdf	denys.choque		3.5 MB
Ordenanza Regional Nro. 03-2016.pdf	denys.choque		3.7 MB
2015_003_ORDENANZA.pdf	denys.choque		1.7 MB

Disponible en:

Google Drive - Ordenanzas Regionales en PDF

Anexo 3. Corpus de ordenanzas regionales emitidas por el Gobierno Regional Puno durante el periodo 2010 – 2024



Ordenanzas Regionales ...

Tipo | Personas | Modificado

Nombre	Propietario	Últimos que abrí	Tamaño
ORDENANZA REGIONAL Nº 003-2023-GRP-CRP.txt	denys.choque		7 KB
014_2011_ORDENANZA.txt	denys.choque		4 KB
004_2012_ORDENANZA.txt	denys.choque		9 KB
ORDENANZA REGIONAL Nº 004-2022-GRP-CRP.txt	denys.choque		6 KB
ORDENANZA REGIONAL Nº 002-2024-GRP-CRP.txt	denys.choque		11 KB
2013_012_ORDENANZA.txt	denys.choque		10 KB
ORDENANZA REGIONAL Nº 018-2021-GRP-CRP.txt	denys.choque		8 KB
2015_001_ORDENANZA.txt	denys.choque		7 KB
ORDENANZA REGIONAL Nº 022-2022-GRP-CRP.txt	denys.choque		9 KB
ORDENANZA REGIONAL Nº 008-2021-GRP-CRP.txt	denys.choque		9 KB
ORDENANZA REGIONAL NRO 010-2019-GR PUNO-CRP.txt	denys.choque		10 KB
2014_007_ORDENANZA.txt	denys.choque		10 KB
ORDENANZA REGIONAL Nº 007-2022-GRP-CRP.txt	denys.choque		9 KB
ORDENANZA REGIONAL Nº 011-2023-GRP-CRP3.txt	denys.choque		19 KB
ORDENANZA REGIONAL Nº 005-2023-GRP-CRP.txt	denys.choque		11 KB
2013_019_ORDENANZA.txt	denys.choque		5 KB
2013_015_ORDENANZA.txt	denys.choque		11 KB
005_2012_ORDENANZA.txt	denys.choque		4 KB
2014_003_ORDENANZA.txt	denys.choque		5 KB
ORDENANZA REGIONAL Nº 020-2023-GRP-CRP2.txt	denys.choque		245 KB
2013_025_ORDENANZA.txt	denys.choque		7 KB
2013_013_ORDENANZA.txt	denys.choque		6 KB
ORDENANZA REGIONAL Nº 007-2019-GR PUNO-CRP.txt	denys.choque		6 KB
011_2012_ORDENANZA.txt	denys.choque		8 KB
ORDENANZA REGIONAL Nº 001-2019-GR PUNO-CRP.txt	denys.choque		5 KB
2013_002_ORDENANZA.txt	denys.choque		4 KB
2010_008_ordenanza.txt	denys.choque		5 KB
ORDENANZA REGIONAL Nº 003-2024-GRP-CRP.txt	denys.choque		12 KB
2010_011_ordenanza.txt	denys.choque		6 KB
003_2012_ORDENANZA.txt	denys.choque		8 KB

Disponible en:

Google Drive - Ordenanzas Regionales en TEXTO

Anexo 4. Código usado en el web scraping utilizando la herramienta bs4

```
import requests
from bs4 import BeautifulSoup
import os

def download_pdfs(url, folder):
   response = requests.get(url)
   response.raise_for_status()s
   soup = BeautifulSoup(response.text, 'html.parser')

   pdf_links = [a['href'] for a in soup.find_all('a',
      href=True) if a['href'].endswith('.pdf')]

   for link in pdf_links:
      pdf_url = link if link.startswith('http') else
         f"https://www.regionpuno.gob.pe{link}"
      pdf_name = pdf_url.split('/')[-1]
      pdf_path = ps.path.join(folder, pdf_name)

      response = requests.get(pdf_url)
      response.raise_for_status()

      with open(pdf_path, 'wb') as g:
            f.write(response.content)

   print(f"PDFs descargados en {folder}")
download_pdfs(
"https://www.regionpuno.gob.pe/ordenanzas-regionales/",
   "1_pdfs")
```

Anexo 5. Código para extracción de texto de documento escaneado en formato pdf a formato txt

```
def pdf_to_text(pdf_path, output_txt_path):
    # Abre el archivo PDF
    pdf_document = fitz.open(pdf_path)

    # Variable para almacenar el texto extraido
    full_text = ""

    for page_num in range(pdf_document.page_count):
        # Extrae la pagina
        page = pdf_document.load_page(page_num)

        #Extrae la imagen de la pagina
        pix = page.get_pixmap()

        #Convierte la imagen a formato PIL
        img = Image.open(io.BytesIO(pix.tobytes()))

        #Usa pytesseract para realizar OCR en la imagen
        text = pytesseract.image_to_string(img, lang='spa')
            #Cambia 'spa' por el idioma deseado

        # Agrega el texto extraido al resultado total
        full_text += text + "\n"

    # Escribe el texto extraido en un archivo de texto
    with open(output_txt_path, "w", encoding="utf-8") as
        text_file:
        text_file.write(full_text)

    print(f"Texto extraido guardado en {output_txt_path}")

def process_multiple_pdfs(pdf_directory, output_directory):
    #Crear el directorio de salida si no existe
    if not os.path.exists(output_directory):
        os.makedirs(output_directory)

```

```
35    #Iterar sobre todos los archivos PDF en el directorio dado
36    for pdf_filename.endswith(".pdf"):
37       if pdf_filename.endswith("pdf_directory"):
38          pdf_path = os.path.join(pdf_directory, pdf_filename)
39          output_txt_path = os.path.join(output_directory,
               pdf_filename.replace(".pdf", ".txt"))
40          pdf_to_text(pdf_path, output_txt_path)
```

DECLARACIÓN JURADA DE AUTENTICIDAD DE TESIS

Por el presente documento, Yo **MARCOS DENYS CHOQUE CASTRO** identificado(a) con N° DNI: **46431369** en mi condición de egresado(a) de la:
MAESTRÍA EN INGENIERÍA DE SISTEMAS
con código de matrícula N° 184132, informo que he elaborado la tesis denominada:
MODELO DE ANÁLISIS DE TÓPICOS EN ORDENANZAS REGIONALES DEL GOBIERNO REGIONAL DE PUNO DE 2010 A 2024 UTILIZANDO TÉCNICAS DE PROCESAMIENTO DE LENGUAJE NATURAL

Es un tema original.

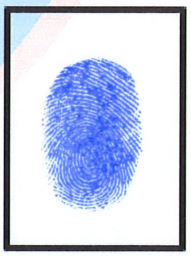
Declaro que el presente trabajo de tesis es elaborado por mi persona y no existe plagio/copia de ninguna naturaleza, en especial de otro documento de investigación (tesis, revista, texto, congreso, o similar) presentado por persona natural o jurídica alguna ante instituciones académicas, profesionales, de investigación o similares, en el país o en el extranjero.

Dejo constancia que las citas de otros autores han sido debidamente identificadas en el trabajo de investigación, por lo que no asumiré como suyas las opiniones vertidas por terceros, ya sea de fuentes encontradas en medios escritos, digitales o Internet.

Asimismo, ratifico que soy plenamente consciente de todo el contenido de la tesis y asumo la responsabilidad de cualquier error u omisión en el documento, así como de las connotaciones éticas y legales involucradas.

En caso de incumplimiento de esta declaración, me someto a las disposiciones legales vigentes y a las sanciones correspondientes de igual forma me someto a las sanciones establecidas en las Directivas y otras normas internas, así como las que me alcancen del Código Civil y Normas Legales conexas por el incumplimiento del presente compromiso

Puno, 25 de Noviembre del 2024.



FIRMA (Obligatorio)

Huella

AUTORIZACIÓN PARA EL DEPÓSITO DE TESIS O TRABAJO DE INVESTIGACIÓN EN EL REPOSITORIO INSTITUCIONAL

Por el presente documento, Yo **MARCOS DENYS CHOQUE CASTRO** identificado(a) con N° DNI: **46431369**, en mi condición de egresado(a) del **Programa de Maestría o Doctorado:**
MAESTRÍA EN INGENIERÍA DE SISTEMAS,
informo que he elaborado la tesis denominada:

MODELO DE ANÁLISIS DE TÓPICOS EN ORDENANZAS REGIONALES DEL GOBIERNO REGIONAL DE PUNO DE 2010 A 2024 UTILIZANDO TÉCNICAS DE PROCESAMIENTO DE LENGUAJE NATURAL

para la obtención de ☒ **Grado.**

Por medio del presente documento, afirmo y garantizo ser el legítimo, único y exclusivo titular de todos los derechos de propiedad intelectual sobre los documentos arriba mencionados, las obras, los contenidos, los productos y/o las creaciones en general (en adelante, los "Contenidos") que serán incluidos en el repositorio institucional de la Universidad Nacional del Altiplano de Puno.

También, doy seguridad de que los contenidos entregados se encuentran libres de toda contraseña, restricción o medida tecnológica de protección, con la finalidad de permitir que se puedan leer, descargar, reproducir, distribuir, imprimir, buscar y enlazar los textos completos, sin limitación alguna.

Autorizo a la Universidad Nacional del Altiplano de Puno a publicar los Contenidos en el Repositorio Institucional y, en consecuencia, en el Repositorio Nacional Digital de Ciencia, Tecnología e Innovación de Acceso Abierto, sobre la base de lo establecido en la Ley N° 30035, sus normas reglamentarias, modificatorias, sustitutorias y conexas, y de acuerdo con las políticas de acceso abierto que la Universidad aplique en relación con sus Repositorios Institucionales. Autorizo expresamente toda consulta y uso de los Contenidos, por parte de cualquier persona, por el tiempo de duración de los derechos patrimoniales de autor y derechos conexos, a título gratuito y a nivel mundial.

En consecuencia, la Universidad tendrá la posibilidad de divulgar y difundir los Contenidos, de manera total o parcial, sin limitación alguna y sin derecho a pago de contraprestación, remuneración ni regalía alguna a favor mío; en los medios, canales y plataformas que la Universidad y/o el Estado de la República del Perú determinen, a nivel mundial, sin restricción geográfica alguna y de manera indefinida, pudiendo crear y/o extraer los metadatos sobre los Contenidos, e incluir los Contenidos en los índices y buscadores que estimen necesarios para promover su difusión.

Autorizo que los Contenidos sean puestos a disposición del público a través de la siguiente licencia:

Creative Commons Reconocimiento-NoComercial-CompartirIgual 4.0 Internacional. Para ver una copia de esta licencia, visita: https://creativecommons.org/licenses/by-nc-sa/4.0/

En señal de conformidad, suscribo el presente documento.

Puno, 25 de Noviembre del 2024.

FIRMA (Obligatorio)

Huella