Análisis de Tópicos Dinámicos en Noticias de Chubut: Enfoque de Machine Learning y NLP

Tesina de grado

Autor

Markel Jaureguibehere

Tutores

Dr. Ing. Leonardo Ordinez

Esp. Lic. Demián Barry

Universidad Nacional de la Patagonia San Juan Bosco

Facultad de Ingeniería

Sede Puerto Madryn

Agradecimientos

Quiero brindar este espacio para expresar mi más sincero agradecimiento a todas las personas que han contribuido a la realización de esta tesina.

En primer lugar, a mis tutores, el Dr. Ing. Leonardo Ordinez y el Esp. Lic. Demián Barry, por orientarme y motivarme a lo largo de todo el proceso, así como a todos los profesores que formaron parte de mi etapa de profesionalización en la Lic. en Informática.

A mis compañeros, que me apoyaron y vivieron esta experiencia junto a mí, que jamás voy a olvidar. En especial, a Martin, Paula y Lautaro que eligen día a día ser mis amigos y compartir su tiempo conmigo.

A mi familia, que supo estar junto a mí en los buenos y malos momentos, brindándome su amor incondicional y consejos de vida externos a la tesina.

A la Universidad Nacional de la Patagonia San Juan Bosco, a la Facultad de Ingeniería y a todo el ambiente universitario, incluido el deporte y el área de investigación, que creen en mí y me ayudan a superarme en cada aspecto de la vida.

A mí mismo, por demostrarme que soy suficiente, que poseo la disciplina, resiliencia y personalidad necesarias para superar las etapas difíciles e inevitables de la educación y de la vida, las cuales ratifican la pasión por la profesión que elegí. Este trabajo es un broche de oro a mi capacitación desde nivel inicial hasta profesional, que sienta las bases de lo que voy a ser a partir de ahora.

Resumen

El presente informe contiene toda la información sobre el proyecto de tesina de grado del alumno Markel Jaureguibehere, sobre el desarrollo de un software para el **Análisis de Tópicos Dinámicos en Noticias de Chubut**, utilizando técnicas avanzadas de *Machine Learning* y NLP. El documento está organizado en capítulos que fundamentan y explican de manera detallada, desde la introducción hasta las conclusiones del proyecto.

En el primer capítulo, referido a la *Introducción*, se expone la fundamentación y la idea principal que motivaron la realización de esta tesina de grado. Asimismo, en el segundo capítulo, se desglosa el *Glosario*, donde se presentan los términos comunes utilizados a lo largo del documento, acompañados de sus respectivas explicaciones.

El tercer capítulo está dedicado al *Marco Teórico*. En él, se proporciona el respaldo para todos los conceptos utilizados en el proyecto, partiendo desde los más generales hasta los más específicos sobre modelado de tópicos, así como las bases metodológicas para la ejecución del proyecto.

El cuarto capítulo describe la *Metodología* empleada. Se encuentra subdividido en la etapa de *Exploración* y la etapa de definición del *Procedimiento* a seguir durante la ejecución del proyecto.

En el quinto capítulo se presenta la *Solución Propuesta*. Aquí se detalla el uso de todas las herramientas empleadas, la arquitectura general del software desarrollado, las funcionalidades de modelado de tópicos y el análisis tanto estático como

dinámico para estudiar los temas latentes de las noticias obtenidas de Chubut entre los años 2019 y 2022.

El sexto capítulo se centra en el *Análisis de Resultados*. En este apartado, se explican, interpretan y se comparan los resultados obtenidos; también se presentan los experimentos y modelos generados, además de realizar una comparación y discusión sobre ellos. A continuación, el séptimo capítulo expone las *Conclusiones* del proyecto y los *Próximos pasos*, con posibles mejoras y avances que quedaron fuera del alcance de la tesina.

Finalmente, se incluye la *Bibliografía* utilizada y todos los *Anexos* generados.

Índice general

CAPÍTULO 1

INTRODUCCIÓN

Los medios de comunicación juegan un rol crucial al informar sobre las inquietudes, necesidades y prioridades de la sociedad, ofreciendo un registro de los acontecimientos más relevantes. Durante la crisis sanitaria del COVID-19, los medios se convirtieron en una fuente crucial de información, capturando como la crisis sanitaria modificaba la evolución de los temas principales en las noticias entre los años 2020 y 2022.

Este trabajo se centra en el análisis de noticias de medios chubutenses, con el objetivo de identificar y seguir la evolución de los temas principales abordados en ese periodo crítico. Para lograrlo, se propone el uso de técnicas avanzadas de *Machine Learning* 3.2 y *Procesamiento del Lenguaje Natural* 3.4, con el fin de realizar un análisis dinámico de tópicos que permita no solo identificar los temas predominantes en las noticias, sino también observar cómo estos han cambiado con el tiempo.

CAPÍTULO 2

GLOSARIO

Tópico Se refiere a un tema, asunto o área de interés específico que es discutido, investigado o tratado en un contexto particular.

Latente Propiedad que no es directamente observable pero que se infiere a partir de los datos, como los tópicos ocultos en el texto.

Semántica El significado de las palabras.

Corpus Conjunto de documentos sobre los cuales se aplica el modelo LDA.

Diccionario Estructura de datos para mapear las palabras del corpus a identificadores únicos.

Modelo Algoritmo entrenado con datos que realiza tareas específicas, como clasificación o predicción, basándose en patrones aprendidos.

Algoritmo Conjunto de pasos o reglas definidas que se siguen para realizar cálculos o resolver problemas.

Entrenamiento Proceso en el que un modelo aprende de los datos ajustando sus parámetros para mejorar su capacidad de hacer predicciones o clasificaciones.

Métrica Estándar de medida de un grado en el que un sistema o proceso de software posee alguna propiedad.

Coherencia Medida en que las palabras dentro de un tópico están relacionadas y forman un conjunto semántico comprensible.

Perplejidad Medida utilizada en modelado de lenguaje que evalúa qué tan bien un modelo predice una muestra.

Co-ocurrencia Ocurrencia simultánea de dos elementos en un mismo contexto o en relación cercana.

Insight Verdad revelada, algo que es cierto y ya existe pero que no habíamos detectado antes.

CAPÍTULO 3

MARCO TEÓRICO

El análisis del lenguaje es un campo de estudio de gran interés dentro de las ciencias de la computación, ya que permite abordar problemas relacionados con la comprensión y el procesamiento de grandes volúmenes de datos textuales, algo impracticable de manejar de forma manual. En este contexto, el Procesamiento del Lenguaje Natural (**NLP**) 3.4 se consolida como una disciplina clave, enfocada en la interacción entre computadoras y lenguajes humanos. A través del desarrollo de técnicas avanzadas, el NLP permite sintetizar información y extraer conocimiento a partir de grandes conjuntos de datos textuales.

Una técnica central dentro del NLP es el *modelado de tópicos* (Topic Modeling) 3.5, que permite descomponer grandes colecciones de textos en temas latentes o tópicos. Existen diversos métodos para realizar este análisis, como el *Latent Semantic Analysis* (**LSA**) [1], *Non-negative Matrix Factorization* (**NMF**) [2] y el *Biterm Topic Model* (**BTM**) [3], entre otros. En particular, en este informe, se abordará el uso de *Latent Dirichlet Allocation* (**LDA**) 3.6 debido a su naturaleza probabilística y a su capacidad comprobada para identificar tópicos de manera eficiente en conjuntos de datos textuales. A lo largo del trabajo, también se explorarán los desafíos asociados al preprocesamiento de datos, la evolución temporal de los

tópicos mediante *Dynamic Topic Models* (**DTM**) 3.7, y la importancia del *Data cleaning* 3.8 para asegurar la precisión y coherencia en los resultados obtenidos.

A continuación, se presentan los conceptos teóricos fundamentales que constituyen la base del desarrollo de esta tesina. Estos se exponen de manera general para facilitar la comprensión de los principios clave que deben considerarse en la elaboración de la solución propuesta.

3.1. Inteligencia Artificial (IA)

La inteligencia artificial (**IA**) es un campo de estudio que se centra en la creación de sistemas capaces de realizar tareas que normalmente requieren inteligencia humana. Estas tareas incluyen el razonamiento, el aprendizaje, la percepción, la comprensión del lenguaje natural, y la toma de decisiones. Dentro del ámbito de la IA, se destacan dos enfoques principales: el enfoque humano y el enfoque racional.

Sistemas que piensan como humanos	**Sistemas que piensan racionalmente**
«El nuevo y excitante esfuerzo de hacer que los computadores piensen… máquinas con mentes, en el más amplio sentido literal». (Haugeland, 1985) «[La automatización de] actividades que vinculamos con procesos de pensamiento humano, actividades como la toma de decisiones, resolución de problemas, aprendizaje…» (Bellman, 1978)	«El estudio de las facultades mentales mediante el uso de modelos computacionales». (Charniak y McDermott, 1985) «El estudio de los cálculos que hacen posible percibir, razonar y actuar». (Winston, 1992)
Sistemas que actúan como humanos	**Sistemas que actúan racionalmente**
«El arte de desarrollar máquinas con capacidad para realizar funciones que cuando son realizadas por personas requieren de inteligencia». (Kurzweil, 1990) «El estudio de cómo lograr que los computadores realicen tareas que, por el momento, los humanos hacen mejor». (Rich y Knight, 1991)	«La Inteligencia Computacional es el estudio del diseño de agentes inteligentes». (Poole *et al.*, 1998) «IA… está relacionada con conductas inteligentes en artefactos». (Nilsson, 1998)
Figura 1.1 Algunas definiciones de inteligencia artificial, organizadas en cuatro categorías.	

Figura 3.1: Definiciones de inteligencia artificial. Fuente: [4]

Como se describe en el libro *Inteligencia Artificial: Un Enfoque Moderno* de Russell [4], estos dos enfoques se representan en la figura 3.1 a través de 4 categorías que diferencian cómo piensa y cómo actúa el sistema.

Enfoque Humano : Este enfoque se centra en emular la forma en que los seres humanos piensan y actúan. La idea es construir sistemas que no solo logren resultados similares a los de los humanos, sino que también utilicen procesos mentales similares. Este enfoque busca modelar la cognición humana, replicando el razonamiento, la toma de decisiones, y el aprendizaje tal como lo haría una persona. Las técnicas desarrolladas bajo este enfoque están inspiradas en estudios de psicología cognitiva y neurociencia.

Enfoque Racional : Este enfoque se basa en la idea de que un sistema de IA debe actuar de manera racional, es decir, debe realizar acciones que maximicen las posibilidades de éxito en la consecución de sus objetivos. A diferencia del enfoque humano, el enfoque racional no necesariamente intenta imitar los procesos cognitivos humanos. En cambio, se centra en la lógica y las matemáticas para crear algoritmos y modelos que tomen decisiones óptimas en función de la información disponible. Este enfoque se inspira en teorías de la racionalidad y la toma de decisiones en condiciones de incertidumbre.

3.2. Machine Learning

El aprendizaje automático, o *Machine Learning*, es un campo de la inteligencia artificial 3.1 que se centra en el desarrollo de algoritmos y técnicas que permiten a las máquinas mejorar su desempeño en tareas específicas a través de la experiencia. Según Sunila Gollapudi en su libro *Practical Machine Learning*, se define como "un mecanismo para buscar patrones y desarrollar inteligencia en una máquina, permitiéndole aprender y, en consecuencia, mejorar en el futuro a partir

de su propia experiencia" [5]. Esta capacidad de aprendizaje autónomo es lo que distingue al *Machine Learning* de otros enfoques tradicionales de programación.

La importancia del *Machine Learning* radica en su capacidad para permitir que las máquinas perciban su entorno y tomen decisiones informadas sobre cómo categorizar y ajustar su comportamiento. Este proceso, que tradicionalmente se asocia con la cognición humana, es fundamental para desarrollar sistemas que puedan adaptarse dinámicamente a nuevas situaciones sin intervención humana directa.

3.3. Aprendizaje no supervisado

El *aprendizaje no supervisado* es una metodología dentro del campo del *Machine Learning* 3.2 que se caracteriza por la ausencia de conocimiento previo sobre el conjunto de datos de entrada. A diferencia de otros enfoques donde el modelo se entrena con datos etiquetados, en el *aprendizaje no supervisado* el modelo explora los datos sin orientación explícita, ajustándose a las observaciones de manera autónoma. Según IBM, "estos algoritmos descubren agrupaciones de datos o patrones ocultos sin necesidad de ninguna intervención humana" [6].

3.4. Natural Language Processing (NLP)

El *Procesamiento del Lenguaje Natural* (**NLP**, por sus siglas en inglés) es un área de la inteligencia artificial 3.1 que se enfoca en el estudio y la mejora de las interacciones entre las máquinas y los seres humanos a través del uso de lenguajes naturales como el inglés o el español. El objetivo principal del NLP es permitir que las máquinas comprendan, interpreten y generen texto y habla de manera que resulte natural para los humanos [7].

Las técnicas de NLP pueden clasificarse en dos enfoques principales:

Modelos Lógicos : Estos modelos utilizan conjuntos de reglas gramaticales para interpretar el lenguaje. Basándose en estructuras sintácticas y semánticas predefinidas, los modelos lógicos buscan comprender el significado de las oraciones a partir de la aplicación de estas reglas.

Modelos Probabilísticos : Estos modelos se basan en el análisis estadístico del lenguaje utilizando grandes conjuntos de datos (corpus). A través de estos análisis, los modelos probabilísticos identifican patrones y relaciones dentro del lenguaje, lo que les permite realizar tareas como la clasificación de texto, la generación de lenguaje y la identificación de tópicos.

3.5. Topic Modeling

Topic Modeling es una técnica dentro del campo del *Procesamiento del lenguaje natural* (**NLP**) 3.4 y el *aprendizaje automático* 3.2 que permite identificar automáticamente los temas principales de un conjunto de documentos. Estos temas, conocidos como tópicos, son patrones recurrentes de palabras que aparecen juntas en los textos, lo que facilita la comprensión del contenido subyacente de grandes colecciones de documentos. Como señala Blei, "los modelos de tópicos son un conjunto de algoritmos que descubren la estructura temática oculta en las colecciones de documentos. Estos algoritmos ayudan a desarrollar nuevas formas de buscar, explorar y resumir grandes archivos de textos" [8].

Una de las características clave del *Topic Modeling* es que no requiere anotaciones previas ni etiquetado de los documentos; los tópicos emergen directamente del análisis de los textos originales. Esto significa que los modelos de tópicos organizan y resumen archivos electrónicos a una escala que sería imposible de manejar mediante anotación humana. En la Figura 3.2 se muestra un esquema gráfico de la asignación de tópicos a documentos, donde se puede observar cómo los algoritmos analizan la co-ocurrencia de palabras para generar los temas

subyacentes en un conjunto de documentos.

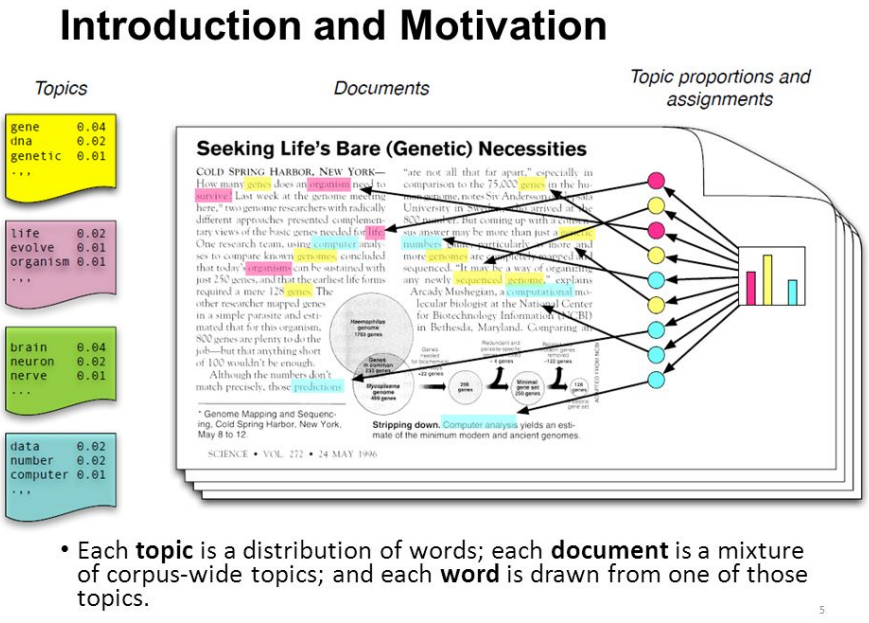

Figura 3.2: Asignación de tópicos a documentos. Fuente: [8]

3.6. Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation (**LDA**) es "un modelo probabilístico generativo de un corpus". La idea central es que los documentos pueden ser representados como una mezcla de tópicos latentes, donde cada tópico está compuesto por una distribución de palabras que describe la probabilidad de que una palabra específica pertenezca a ese tópico. En otras palabras, LDA permite descomponer el corpus en un conjunto de tópicos, facilitando la identificación automática de los temas principales tratados en los documentos [9]. En la Figura 3.3 se presenta un esquema gráfico que ilustra este modelo y cómo los tópicos son inferidos y asignados con probabilidades a los documentos del corpus.

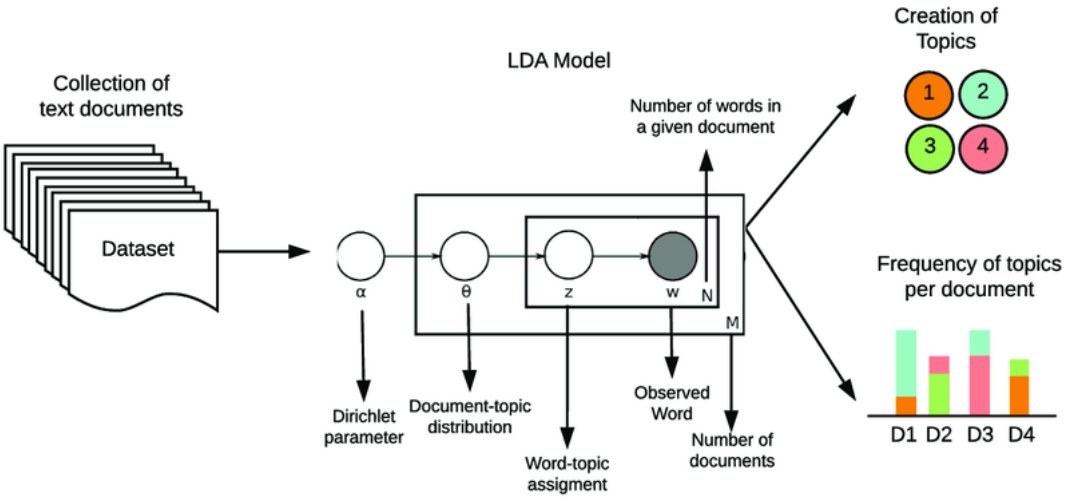

Figura 3.3: Esquema del modelo LDA. Fuente:
https://www.markovml.com/blog/lda-topic-modelling

3.6.1. Parámetros

Las librerías para entrenamiento de modelos LDA permiten modificar diversos parámetros para ajustar el modelo según las características del corpus y los objetivos del análisis. A continuación, se detallan los parámetros más importantes y cómo influyen en el rendimiento del modelo:

1. **Número de tópicos**: Este parámetro define cuántos tópicos se generarán en el modelo. Si el número de tópicos es muy bajo, el modelo podría no captar la diversidad temática del corpus. Si es demasiado alto, los tópicos generados pueden ser redundantes o carecer de coherencia. Por ejemplo, en un corpus de noticias de una región, elegir un número muy pequeño de tópicos podría generar temas muy amplios (como ”política” o ”economía”), mientras que un número alto podría crear tópicos más específicos y detallados (como ”elecciones locales” o ”crisis económica”). La selección del número de tópicos óptimo suele ser un proceso iterativo, evaluando la coherencia de los temas y su relevancia en función del corpus.

2. **Alpha**: Este parámetro controla la distribución de los tópicos dentro de cada documento. Este puede ajustarse numéricamente, donde un valor alto indica

que los documentos tienen una distribución más uniforme de los tópicos, mientras que un valor bajo indica que un documento está conformado por pocos tópicos. LDA también permite establecer el alpha con dos valores predefinidos:

- `"symmetric"`: Establece una probabilidad uniforme para todos los tópicos, es decir, se asume que todos los temas tienen la misma posibilidad de aparecer en cualquier documento. Esto puede ser útil cuando no se tiene una hipótesis previa sobre la distribución de tópicos en los documentos. Por ejemplo, en un conjunto de artículos donde se espera que cada uno cubra una amplia variedad de temas, esta configuración puede ser adecuada.
- `"asymmetric"`: Permite que algunos tópicos tengan una mayor probabilidad de aparecer que otros, lo cual puede ser más realista en muchos casos. Por ejemplo, en un corpus donde ciertos temas son mucho más frecuentes que otros (como "economía" en comparación con "deporte"), una distribución asimétrica puede capturar esta tendencia de manera más precisa.

3.6.2. Métricas

El ser un método de *aprendizaje no supervisado* 3.3, no hay manera de dar anotaciones previas al dataset para indicar resultados esperados del modelado de tópicos 3.5, pero si hay ciertas métricas que ayudan a diferenciar modelos e intuir qué parámetros son los adecuados para un modelo LDA 3.6 de calidad.

3.6.2.1. Coherencia

La coherencia de un tópico, a nivel de métrica, se refiere a la capacidad de las palabras que lo componen para co-ocurrir en el mismo contexto. En otras

palabras, las palabras dentro de un tópico deben ser semánticamente similares entre sí y distintas de las palabras de otros tópicos, reflejando una agrupación lógica y significativa.

Existen varias medidas para evaluar la coherencia de los tópicos, cada una de las cuales utiliza diferentes parámetros y enfoques para determinar qué tan coherente es un modelo `LDA` en su representación de la realidad. Las métricas más comunes incluyen:

- **Coherencia `CV`**: Es una de las métricas más avanzadas y comúnmente utilizadas, basada en la combinación de varios enfoques que incluyen medidas de probabilidad condicional y validación externa usando conjuntos de datos de referencia. Evalúa cómo las palabras de un tópico tienden a aparecer juntas en corpus externos y considera el contexto semántico, lo que la hace una de las más robustas [10].

- **Coherencia `Umass`**: Esta medida se basa en la probabilidad logarítmica de co-ocurrencia de palabras dentro del mismo documento. A pesar de ser ampliamente utilizada, tiende a estar más sesgada por el corpus específico y no utiliza información externa, lo que puede reducir su capacidad para generalizar [11].

- **Coherencia `UCI`**: Evalúa la coherencia calculando las co-ocurrencias de las palabras dentro de un tópico a lo largo de ventanas de palabras en el corpus, lo que permite un análisis más localizado y detallado de las relaciones entre las palabras [11].

- **Coherencia `NPMI`** (Pointwise Mutual Information Normalizada): Utiliza la información mutua entre las palabras de un tópico, normalizada para corregir la frecuencia con la que ocurren las palabras por sí solas. Esta métrica se basa en la suposición de que si dos palabras tienden a aparecer juntas más

de lo que se esperaría por azar, entonces están estrechamente relacionadas [12].

Por ejemplo, en el caso de la coherencia `CV`, si un tópico contiene las palabras "virus", ”pandemia”, ”vacuna” y ”contagio”, la métrica evaluará no solo si estas palabras co-ocurren en el corpus, sino también si lo hacen en corpus externos, sugiriendo que representan un tema coherente.

Cada una de estas métricas ofrece una perspectiva distinta sobre la calidad de los tópicos, y en esta tesis se utilizará principalmente la coherencia `CV`, dada su capacidad para combinar diferentes enfoques y su fiabilidad en estudios anteriores sobre la evaluación de modelos de tópicos.

3.6.2.2. Perplejidad

La **perplejidad** es una medida utilizada para evaluar la eficacia de un modelo de lenguaje, reflejando su capacidad para predecir palabras en un corpus de prueba. En términos simples, la perplejidad se puede entender como una forma de medir qué tan ”confundido” está un modelo al intentar predecir el siguiente término en una secuencia de texto.

Para una explicación más detallada sobre cómo se calcula y se interpreta la perplejidad en modelos LDA, se puede consultar el trabajo de Griffiths y Steyvers (2004), titulado *”Finding scientific topics”* [13], que presenta un enfoque completo sobre el uso de LDA y métricas relacionadas.

3.7. Dynamic Topic Models (DTM)

Los *Dynamic Topic Models* (**DTM**) extienden los modelos de tópicos estáticos, como *Latent Dirichlet Allocation* (**LDA**) 3.11.1, para capturar la evolución de los temas en una colección de documentos a lo largo del tiempo. A diferencia de los modelos tradicionales que asumen que los tópicos son fijos, DTM permite que la

distribución de los temas cambien en función de los periodos temporales. Esto es esencial para analizar cómo los temas emergen, se desarrollan o declinan a lo largo del tiempo, proporcionando una visión más dinámica del contenido textual [14]. Esto es especialmente útil en colecciones de documentos históricamente distribuidos o publicaciones periódicas.

3.8. Data Cleaning

El *Data Cleaning*, o limpieza de datos, es un paso esencial en la preparación de datos para el análisis y modelado. Consiste en una serie de técnicas y procedimientos aplicados a los datos para transformarlos, limpiarlos y estructurarlos adecuadamente, con el fin de que sean útiles y fiables para el análisis posterior [15]. Este proceso es crucial para garantizar la calidad y precisión de los resultados en cualquier proyecto de análisis de datos.

El objetivo principal del *Data Cleaning* es eliminar el "ruido" en los datos y corregir errores, inconsistencias o anomalías, lo que resulta en datos más precisos y coherentes. Un *Data Cleaning* efectivo mejora la calidad de los datos y reduce la complejidad de los modelos, lo que a su vez puede conducir a un análisis más eficiente y preciso. La importancia del *Data Cleaning* es evidente en una amplia gama de aplicaciones, desde el análisis de datos hasta el *aprendizaje automático* 3.2, donde datos de alta calidad son fundamentales para obtener resultados confiables y útiles [15].

3.9. Web scraping

El *Web Scraping* es una técnica utilizada para extraer información de sitios web de manera automatizada. Consiste en utilizar programas o scripts que navegan por páginas web y extraen datos específicos que luego pueden ser almacenados y analizados. Esta práctica es comúnmente utilizada para recopilar grandes vo-

lúmenes de datos que están disponibles públicamente en la web, pero que no se presentan en un formato fácilmente descargable [16].

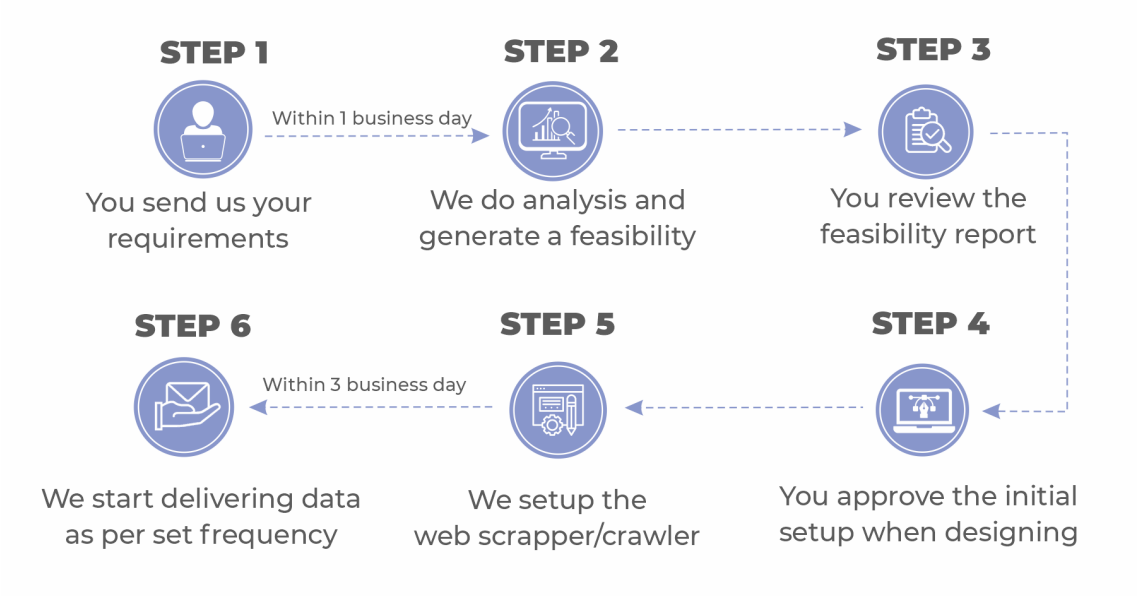

Figura 3.4: Ejemplo de proceso de *Web scraping*. Fuente: https://www.mindbowser.com/real-estate-web-scraping/

En la Figura 3.4 se ilustra un ejemplo de cómo funciona un proceso típico de *Web scraping*, desde la navegación por las páginas hasta la extracción y almacenamiento de los datos.

3.10. Knowledge Discovery in Databases (KDD)

El *descubrimiento de conocimiento en bases de datos* (**KDD**, por sus siglas en inglés) es un proceso integral que busca transformar grandes volúmenes de datos en conocimiento útil y comprensible. A diferencia de la simple minería de datos, que se centra en la aplicación de técnicas para encontrar patrones dentro de los datos, KDD abarca un enfoque más amplio y sistemático que incluye varias etapas desde la preparación de los datos hasta la interpretación de los resultados. Según Fayyad "el *descubrimiento de conocimiento en bases de datos* es el proceso de encontrar patrones válidos, novedosos, útiles y comprensibles en grandes bases de datos"

[17]. Este proceso es fundamental para extraer información relevante que pueda ayudar en la toma de decisiones estratégicas.

3.10.1. Etapas del KDD

1. **Selección de Datos**: Determinación de las fuentes de datos relevantes y su recopilación.

2. **Limpieza de Datos**: Corrección de errores y eliminación de ruido en los datos para asegurar su calidad.

3. **Transformación de Datos**: Conversión de los datos en un formato adecuado para el análisis, lo que puede incluir la normalización y la reducción de dimensionalidad.

4. **Minería de Datos**: Aplicación de técnicas de *aprendizaje automático* 3.2 y estadística para descubrir patrones o modelos dentro de los datos.

5. **Evaluación e Interpretación**: Análisis de los patrones descubiertos para validar su utilidad y comprensión, y su integración en el conocimiento práctico.

3.11. Metodología Ágil

La metodología ágil es un enfoque para la gestión y desarrollo de proyectos que enfatiza la entrega rápida y continua de valor, a través de iteraciones cortas y ciclos de retroalimentación frecuentes. Este enfoque surgió como una alternativa a los métodos tradicionales de desarrollo de software, que a menudo eran considerados rígidos y propensos a retrasos. La agilidad se centra en la flexibilidad, la colaboración cercana entre los equipos y los clientes, y la adaptación constante a los cambios en los requisitos y en el entorno del proyecto.

3.11.1. Scrum

Scrum es una de las metodologías ágiles más populares y se caracteriza por su enfoque estructurado en ciclos iterativos denominados *sprints*, que suelen tener una duración de entre dos a cuatro semanas. Según Ken Schwaber y Jeff Sutherland, creadores de *Scrum*, esta metodología proporciona un marco de trabajo para abordar problemas complejos mediante la colaboración y la autoorganización del equipo [18]. La Figura 3.5 muestra un ejemplo del ciclo de *Scrum*, destacando las fases clave que componen cada *sprint*.

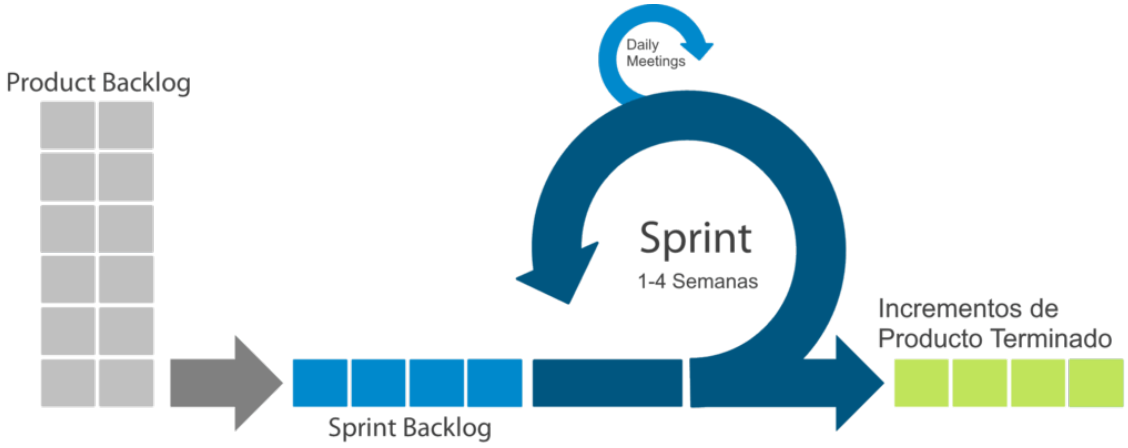

Figura 3.5: Ejemplo de ciclo de *Scrum*. Fuente:
https://programmingbabel.blogspot.com/2017/08/principios-de-scrum.html

CAPÍTULO 4

METODOLOGÍA

En esta sección se detalla el proceso de resolución del problema y cómo se llegó a la solución propuesta. Se describen las estrategias empleadas y los resultados obtenidos. Dado que se enfrenta un problema orientado a *Data Analytics*, se decidió abordarlo siguiendo las etapas de *Knowledge Discovery in Databases* (**KDD**) 3.10. Durante el desarrollo, esta metodología se complementó con prácticas ágiles 3.11, específicamente *Scrum* 3.11.1, implementando iteraciones semanales y organizando reuniones *weekly* con los tutores para asegurar una constante revisión y mejora del trabajo.

Previo a la definición de la solución final y del procedimiento a seguir, se realizaron reuniones con los tutores, investigaciones teóricas, análisis de problemas similares y sus soluciones, así como pruebas de concepto. Estas actividades permitieron identificar tecnologías y enfoques reconocidos en los campos de *Machine Learning* 3.2 y *Topic Modeling* 3.5, además de definir el *pipeline* del *procesamiento del lenguaje natural* (**NLP**) 3.4.

Es importante resaltar que el proceso se mantuvo iterativo e incremental, alineado con los principios de *Scrum* 3.11.1. Este enfoque permitió que, durante las iteraciones, nuevos problemas o ajustes fueran identificados y resueltos de manera

ágil y oportuna, asegurando una retroalimentación constante y una adaptación continua a las necesidades del proyecto.

4.1. Exploración

Para el inicio del proyecto, se llevaron a cabo pruebas de concepto tanto con tecnologías conocidas para problemas del tipo *Machine Learning* 3.2 como con soluciones aplicadas a problemas similares de *Topic Modeling* 3.5. En esta etapa se exploraron varias herramientas:

Python

En *Python*, se realizaron pruebas básicas para entender y familiarizarse con la sintaxis del lenguaje de programación. Además, se llevaron a cabo algunas pruebas siguiendo tutoriales de *Machine Learning* 3.2 y *Natural Language Processing* (**NLP**) 3.4, con el objetivo de investigar las librerías más comunes para la resolución de problemas de *Inteligencia artificial* y *Procesamiento del lenguaje natural*.

Wordcloud

La librería *Wordcloud* permite la visualización de nubes de palabras, facilitando la identificación rápida de las palabras más frecuentes en un conjunto de datos. Esta herramienta resultó particularmente útil durante la fase de experimentación, ya que permitió detectar la necesidad de realizar un procesamiento adicional del texto. Posteriormente, tras aplicar las técnicas de preprocesamiento necesarias, la nube de palabras reflejó de manera más precisa y clara la relevancia de los términos en el texto analizado.

La Figura 4.1 muestra un ejemplo de una nube de palabras generada como parte del trabajo de Carlos Emanuel Balcazar [19], donde se visualizan los términos más frecuentes de las noticias de la región.

Figura 4.1: Ejemplo de *Wordcloud*. Fuente: [19]

InfraNodus

InfraNodus es una herramienta avanzada de análisis de texto y visualización de datos que destaca por su capacidad para detectar y relacionar tópicos dentro de los documentos. Utiliza gráficos de redes para visualizar las ideas subyacentes, proporcionando una comprensión clara y visual de cómo se interconectan los conceptos. Además, *InfraNodus* ofrece una funcionalidad dinámica para analizar la evolución de los tópicos a lo largo del tiempo mediante la inserción de textos con marcas temporales.

Aunque se realizaron pruebas conceptuales con algunas noticias, las limitaciones espaciales del plan *premium* de *InfraNodus*, que solo permite datasets menores a 3 MB, llevaron a la decisión de desarrollar una herramienta propia para abordar este problema de manera más eficiente.

A pesar de estas limitaciones, *InfraNodus* demostró ser una herramienta extremadamente útil y sirvió de inspiración y ejemplo para crear un sistema capaz de analizar tópicos y su variación temporal de manera efectiva.

4.2. Procedimiento

Después de la etapa de prueba e investigación teórica, se buscó dividir el problema principal en varios subproblemas clave más abordables para llegar al objetivo. Con esto en mente, se adoptó la metodología *Knowledge Discovery in Databases* (**KDD**) 3.10.

KDD es un procedimiento que permite generar conocimiento a partir de grandes conjuntos de datos, sistematizando el proceso e identificando claramente los pasos necesarios para resolver el problema e interpretar los resultados. Esta metodología incluye técnicas de *minería de datos*, cruciales para descubrir patrones y relaciones ocultas en los datos. Este enfoque se adapta perfectamente al problema que se busca abordar, como se ilustra en la Figura 4.2, que muestra un ejemplo del ciclo de *Data Mining*. Este ciclo sistematiza el proceso de descubrimiento de patrones ocultos en grandes conjuntos de datos.

Con esto en cuenta, el problema principal se dividió en cinco subproblemas clave, que se detallan a continuación:

1. **Obtención del contenido**: Este subproblema aborda cómo se obtiene el contenido del dataset de noticias para su análisis. Incluye la recopilación y almacenamiento de las noticias en un formato adecuado para el procesamiento posterior.

2. **Preprocesamiento de texto**: Se enfoca en cómo realizar una correcta normalización del texto para garantizar una mejor entrada en el posterior modelado de tópicos 3.5.

3. **Modelado de tópicos**: Este subproblema trata sobre cómo identificar y modelar los tópicos latentes en las noticias utilizando el modelo de *LDA* 3.6. Incluye la configuración del modelo, el entrenamiento con el dataset preprocesado y la evaluación de los resultados para asegurar la coherencia y relevancia de los tópicos.

4. **Análisis de tópicos estático**: Se refiere a cómo realizar un análisis y visualización de los tópicos de manera estática, es decir, considerando la totalidad del dataset sin tener en cuenta las marcas temporales. Esto incluye la generación de gráficos y métricas que muestran la relación entre los diferentes tópicos, su similitud y su importancia relativa.

5. **Análisis de tópicos dinámico**: Este subproblema aborda cómo realizar un análisis y visualización de los tópicos de manera dinámica, observando cómo varían a través del tiempo. Involucra la segmentación del dataset según marcas temporales y la creación de visualizaciones que muestran la evolución de los tópicos a lo largo de diferentes períodos.

Figura 4.2: Ejemplo de ciclo de *Data Mining*. Fuente:
https://www.masterdatascienceucm.com/que-es-el-data-mining/

CAPÍTULO 5

SOLUCIÓN PROPUESTA

Basado en el *Procedimiento* 4.2 planteado en la Metodología, se planteó un sistema basado en módulos independientes para abordar cada subproblema del análisis de tópicos, asegurando un proceso estructurado. Cada módulo se encargó de una fase específica del flujo de trabajo, lo que facilitó la integración de los resultados en el análisis final. En esta sección se presenta en detalle el desarrollo de todas las herramientas creadas y utilizadas para realizar un correcto y robusto análisis dinámico de tópicos a partir de un enfoque de *Machine Learning* y NLP.

Python fue seleccionado como lenguaje de programación por varias razones fundamentales: su versatilidad para manejar grandes volúmenes de datos textuales, su capacidad para paralelizar tareas mediante multiprocesamiento, y su vasto ecosistema de librerías dedicadas al análisis de datos y NLP, tales como `gensim`, `nltk`, y `spaCy`, entre otras. De acuerdo con el índice TIOBE, Python se posiciona como uno de los lenguajes más utilizados en programación general [20] y, según un estudio de Rice University, se encuentra entre los más recomendados para ciencia de datos [21], manteniéndose entre los primeros lugares de los índices de popularidad debido a su simplicidad y potencia.

El proceso comienza con el módulo de preprocesamiento (Figura 5.1), encar-

gado de normalizar el texto y eliminar ruido para mejorar la calidad de los datos. A continuación, se emplea el módulo de modelado de tópicos (Figura 5.2), que utiliza LDA junto con técnicas de optimización y refinamiento para identificar los tópicos más relevantes en el conjunto de datos. El análisis se complementa con el módulo de análisis estático (Figura 5.3), que extrae métricas y genera gráficos basados en la totalidad del dataset, proporcionando una visión general de los tópicos identificados. Finalmente, el módulo de análisis dinámico (Figura 5.4) permite observar la evolución de estos tópicos a lo largo del tiempo, obteniendo métricas y visualizaciones que reflejan cómo cambian en diferentes periodos.

Durante cada ejecución, todos los resultados, que incluyen modelos entrenados, visualizaciones, matrices de análisis y logs del sistema, se almacenan sistemáticamente en una carpeta generada. Esta práctica asegura la accesibilidad y gestión eficiente de los datos generados durante todo el proceso. Además, permite acceder fácilmente a los datos y gráficos posteriormente, facilitando la revisión y el análisis continuo de los resultados obtenidos. Se puede observar la estructura general de los resultados en el Anexo B.I

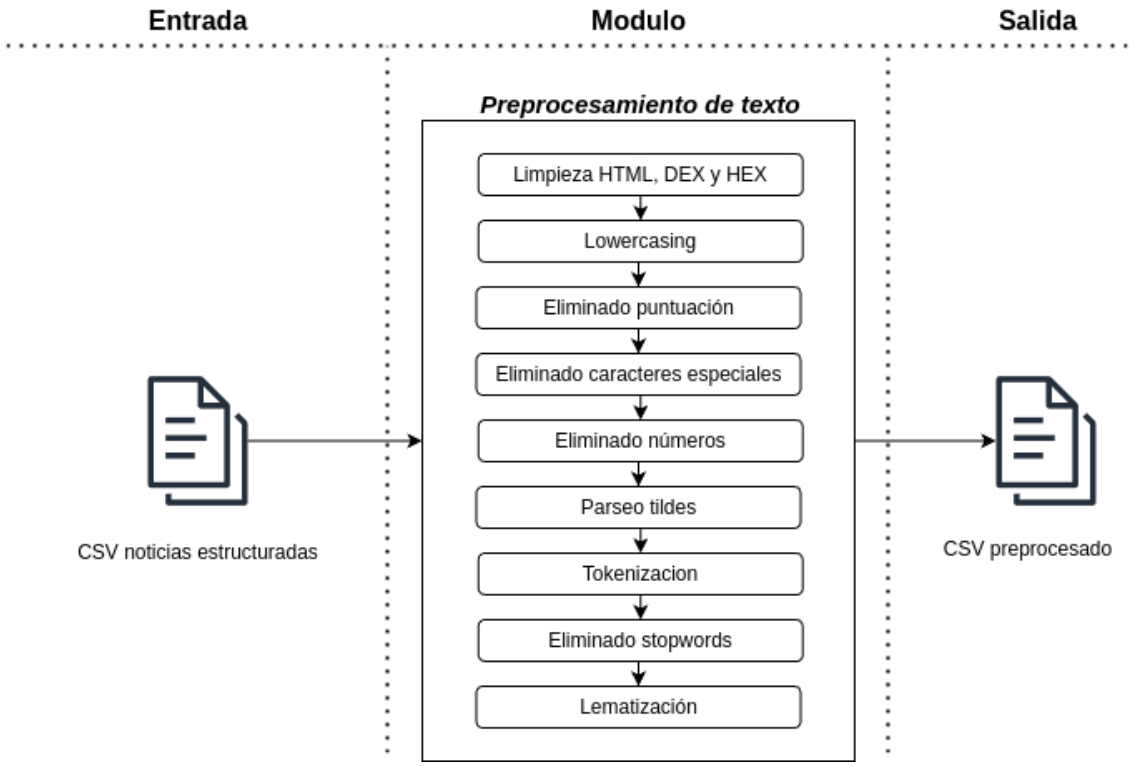

Figura 5.1: Modulo de preprocesado de texto

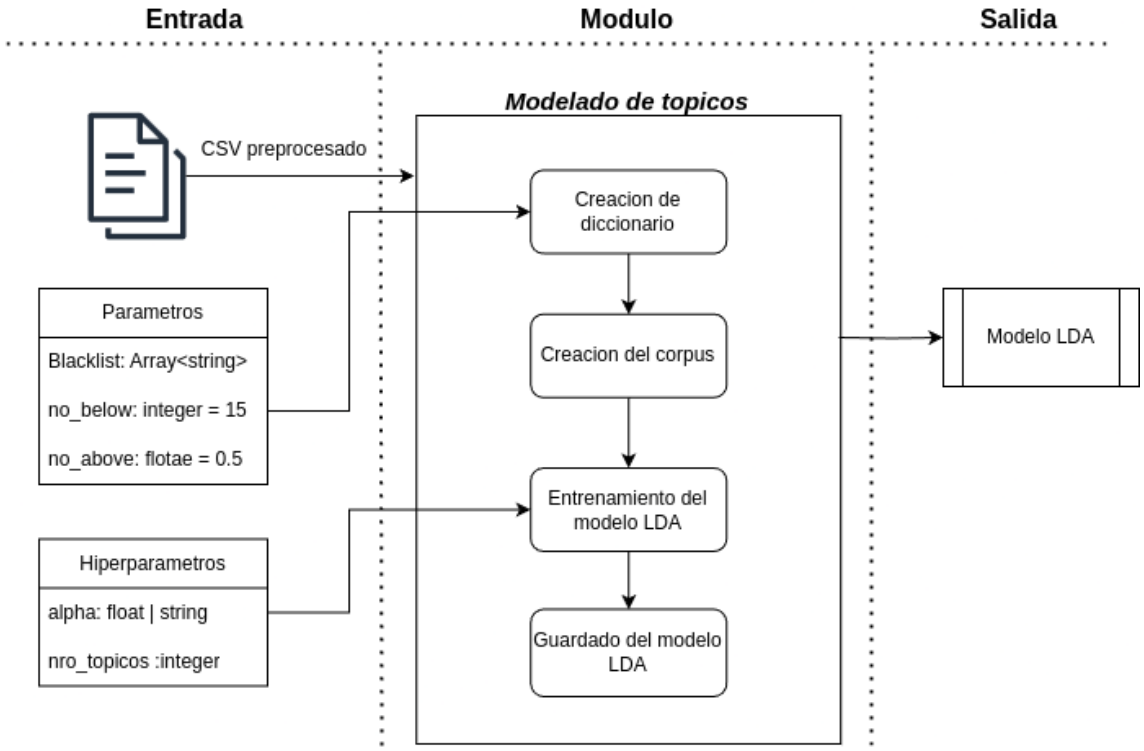

Figura 5.2: Modulo de modelado de tópicos

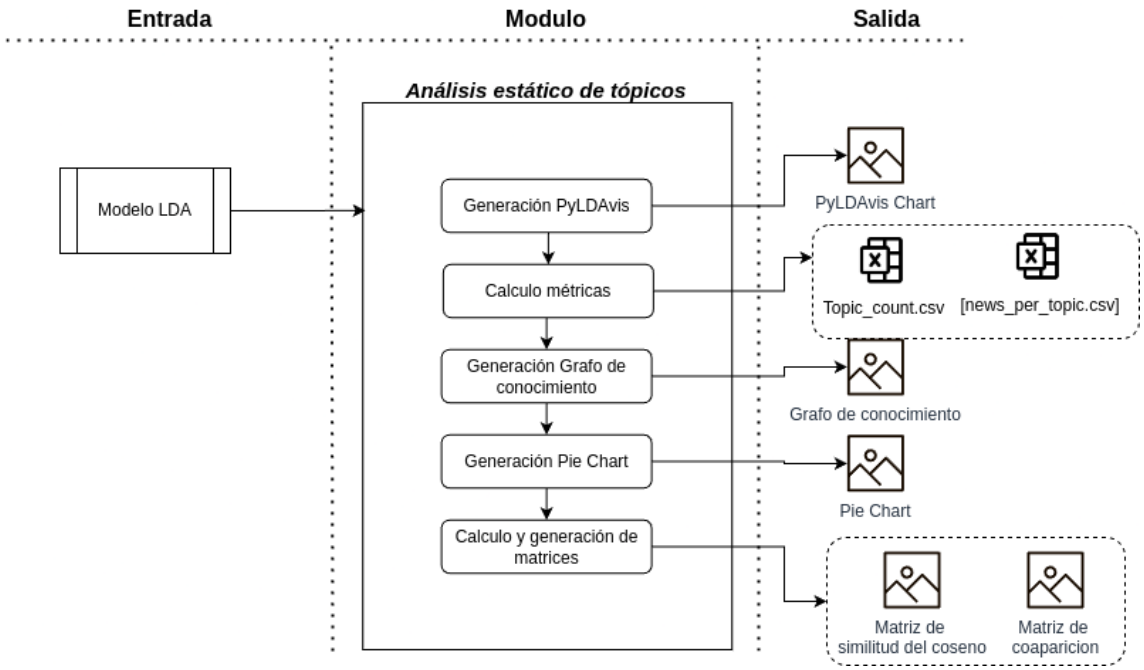

Figura 5.3: Modulo de análisis estático de tópicos

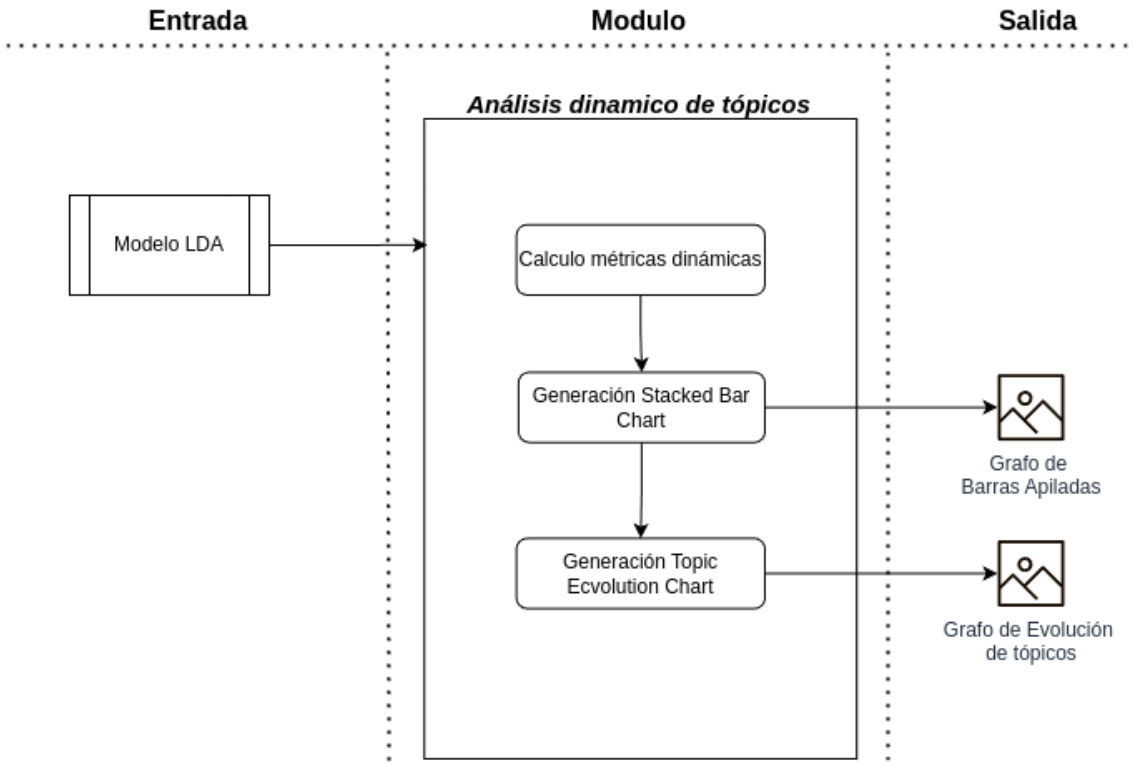

Figura 5.4: Modulo de análisis dinámico de tópicos

5.1. Herramientas utilizadas

```
Python: 3.10.12
    pip: 22.0.2
    numpy: 1.22.4
    numexpr: 2.10.0
    pandas: 2.2.2
    scipy: 1.7.3
    sklearn: 1.4.2
    gensim: 4.3.2
    nltk: 3.8.1
    spacy: 3.7.4
    stanza: 1.8.2
    selenium: 4.20.2
    py2neo: 2021.2.4
    wordcloud: 1.9.3
    pyLDAvis: 3.4.0
    seaborn: 0.13.2
    matplotlib: 3.8.4
    plotly: 5.22.0
Neo4j: 4.1.13
PostgreSQL: 16.3
PGAdmin4: 8.7
Git: 2.34.1
```

5.2. Obtención de datos

Para el análisis de los datos de noticias, se utilizará como base la tesina de grado de Emmanuel Balcazar, titulada *"Extracción, Análisis y Procesamiento Automático de Información Periodística relacionada al COVID-19 en Chubut"* [19]. En su investigación, mediante una metodología de *Web scraping* 3.9, se obtuvo un extenso conjunto de datos compuesto por más de 100 mil noticias provenientes de los medios de comunicación más importantes de la región.

Las noticias se encuentran encapsuladas en un archivo CSV con el siguiente formato:

- **displayLink:** Contiene el enlace al medio de comunicación al que pertenece la noticia.

- **body:** Contiene el texto plano de la noticia, extraído directamente del HTML de la página web.

- **expected_date:** Contiene la fecha de publicación de la noticia, formateada como YYYY-MM-DD.

Aunque el resultado del *scraping* de las noticias incluye más columnas, como el título, el enlace directo a la noticia y el "snippet", las mencionadas anteriormente son las más relevantes para el análisis de tópicos que se llevará a cabo en este desarrollo.

En total, el *dataset* contiene 109,000 noticias, distribuidas de acuerdo a lo expresado en el Cuadro 5.1, según el medio de comunicación analizado.

Título	**Link**	**Cantidad noticias**
El Chubut	www.elchubut.com.ar	57,597
Diario Jornada	www.diariojornada.com.ar	19,601
Diario La Portada	www.diariolaportada.com.ar	9,442
Diario Crónica	www.diariocronica.com.ar	6,269
El Patagónico	www.elpatagonico.com	5,074
El Diario Web	www.eldiarioweb.com	4,032
Red 43	www.red43.com.ar	4,017
Radio 3 Cadena Patagonia	www.radio3cadenapatagonia.com.ar	3,652

Cuadro 5.1: Distribución de noticias por medio de comunicación.

La cantidad de noticias obtenidas por año se muestra en el Cuadro 5.2.

Año	**Cantidad noticias**
2019	8,377
2020	37,192
2021	42,612
2022	21,503

Cuadro 5.2: Distribución de noticias por año.

Finalmente, en la Figura 5.5, se muestra la cantidad de noticias por mes, de acuerdo al medio de comunicación de origen.

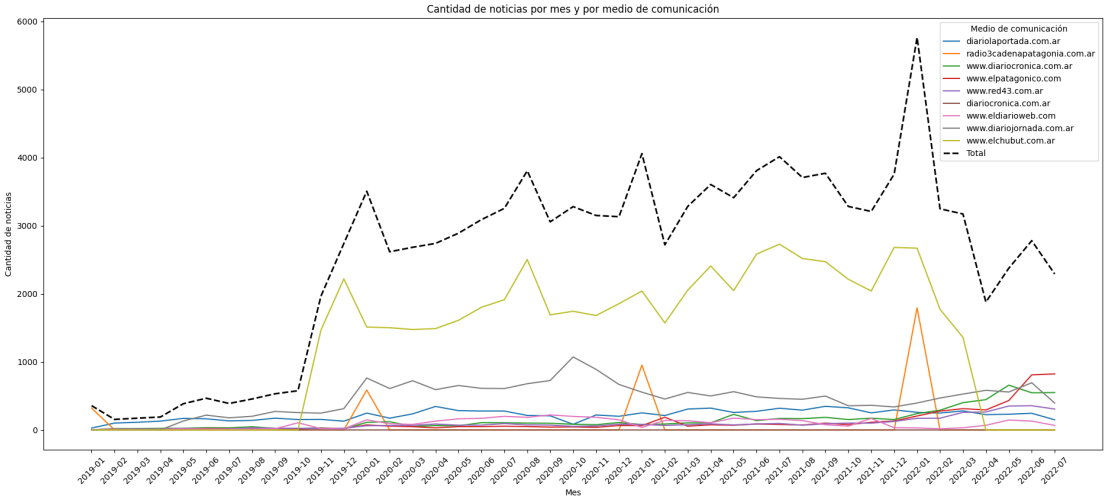

Figura 5.5: Cantidad de noticias por mes y medio de comunicación

5.3. Preprocesamiento de texto

El preprocesamiento de datos, comúnmente denominado como *Data Cleaning* 3.8, es una etapa clave en la minería de datos. Esta fase afecta significativamente a la calidad de los resultados que se obtienen en los análisis siguientes, y es por eso que requiere una atención cuidadosa.

Debido a que el objetivo es realizar un modelado de tópicos, es necesario identificar las palabras claves de cada noticia. Por lo tanto, es importante normalizar el texto y eliminar todas aquellas palabras que no contribuyen al análisis. Para esto se desarrollaron los siguientes pasos:

5.3.1. Normalización

En el contexto de este preprocesamiento de texto, la normalización es el proceso mediante el cual se limpia el texto de cada noticia sin realizar modificaciones a las palabras. El objetivo es obtener el texto plano de la noticia, libre de caracteres especiales y otras anomalías. La normalización se lleva a cabo mediante los siguientes seis pasos:

1. **Limpieza de caracteres HTML, decimales y hexadecimales** En esta etapa del preprocesamiento de datos, se utilizaron funciones específicas de *parsing* para decodificar y reemplazar cualquier carácter HTML, decimal o hexadecimal encontrado en el texto por su equivalente en caracteres normales. Este paso es fundamental debido a la naturaleza de las noticias, las cuales son extraídas directamente de páginas web (ver Cuadro 5.3).

ENTRADA	**SALIDA**
`El ni&ntilde;o juega al f&#xFA;tbol`	El niño juega al fútbol

Cuadro 5.3: Ejemplo de reemplazo de caracteres especiales.

2. **Conversión de texto a minúsculas** Fase de conversión de todas las letras a minúsculas para asegurar la consistencia y reducir la complejidad del texto.

Para esto, Python provee su propia funcionalidad de "*lower casing*" (ver Cuadro 5.4).

ENTRADA	SALIDA
La Luna brilla en la noche oscura.	la luna brilla en la noche oscura.

Cuadro 5.4: Ejemplo de conversión de texto a minúsculas.

3. **Eliminación de puntuación** Eliminar signos de puntuación como comas, puntos, etc. Para mantener solo el contenido textual. Para esto, se utilizaron detecciones de caracteres y filtrado de los mismos (ver Cuadro 5.5).

ENTRADA	SALIDA
la luna brilla, el niño duerme.	la luna brilla el niño duerme

Cuadro 5.5: Ejemplo de eliminación de puntuación.

4. **Eliminación de caracteres especiales** Eliminar caracteres especiales como signos de exclamación, almohadillas, ampersands, etc. Para esto, se utiliza detección de caracteres especiales con expresiones regulares y eliminación de los mismos (ver Cuadro 5.6).

ENTRADA	SALIDA
Hola! ¿Cómo estás? #FelizDía	Hola Como estas FelizDía

Cuadro 5.6: Ejemplo de eliminación de caracteres especiales.

5. **Eliminación de números** Eliminar números del texto, ya que no contribuyen al significado de la noticia. Para esto, se utiliza detección de números con expresiones regulares y eliminación de los mismos (ver Cuadro 5.7).

ENTRADA	SALIDA
Hoy es 25 de junio de 2024	Hoy es de junio de

Cuadro 5.7: Ejemplo de eliminación de números.

6. **Conversión de tildes** Se utilizan funciones específicas para la detección y reemplazo de caracteres con tildes por sus letras normalizadas correspon-

dientes. Este paso es esencial para estandarizar el texto y facilitar el análisis posterior (ver Cuadro 5.8).

Es crucial tener en cuenta que, dado que se trabaja con el idioma español, es fundamental no reemplazar la ”ñ” por la ”n”, ya que esto podría alterar o perder el significado semántico de las palabras. Por ejemplo:

- Peña → Pena
- Campaña → Campana
- Moño → Mono
- Uña → Una
- Año → Ano

ENTRADA	**SALIDA**
Mañana será un día soleado	Mañana sera un dia soleado

Cuadro 5.8: Ejemplo de conversión de tildes.

5.3.2. Tokenización

En esta etapa del pre-procesamiento se divide el texto en unidades más pequeñas llamadas ”tokens”, que suelen ser palabras o frases. Para esto, se utiliza la librería gensim que provee funcionalidades de tokenización (ver Cuadro 5.9).

ENTRADA	**SALIDA**
mañana sera un dia soleado	[“mañana”,”sera”,”un”,”dia”,”soleado”]

Cuadro 5.9: Ejemplo de tokenización.

5.3.3. Eliminación de stopwords

La eliminación de stopword es el paso por el cual se eliminan las palabras comunes y frecuentes que no aportan un significado sustancial al análisis de texto. Las stopwords incluyen artículos, preposiciones, pronombres y otras palabras de

enlace, como ”el”, ”la”, ”de”, ”y”, ”en”, entre otras. Para esto, se utiliza un filtrado del texto basado en la lista de stopwords que provee la librería `Spacy` (ver Cuadro 5.10).

ENTRADA	**SALIDA**
[’El’, ’perro’, ’y’, ’la’, ’mesa’, ’en’, ’el’, ’parque’]	[’perro’, ’mesa’,’parque’]

Cuadro 5.10: Ejemplo de eliminación de stopwords.

5.3.4. Lematización

La lematización es el proceso de reemplazar una palabra flexionada por su “lemma”. Consiste en reducir las palabras a su forma base o raíz, lo que ayuda a normalizar el texto y reducir la variabilidad. Básicamente, se basa en remover el plural, el tiempo, o los atributos finales de la palabra. “El lema es la forma que por convenio se acepta como representante de todas las formas flexionadas de una misma palabra.” Wikipedia. La lematización se aplicó a través de la librería `Stanza` (ver Cuadro 5.11).

ENTRADA	**SALIDA**
[’perro’, ‘esta’,’corriendo’, ’rapidamente’]	[’perro’, ‘estar’,’correr’, ’rapido’]

Cuadro 5.11: Ejemplo de lematización de tokens.

5.4. Modelado de tópicos

Para la funcionalidad del modelado de tópicos se utilizó la librería gensim. Gensim está especializada en el *Procesamiento del lenguaje natural* (**NLP**) 3.4, particularmente en el *Topic modeling* 3.5. Posee también capacidad para manejar grandes volúmenes de datos textuales y funciones de procesamiento de los mismos.

Gensim además tiene funcionalidades para el uso de múltiples procesadores e implementaciones eficientes para LDA 3.6 lo que la hace una librería ideal además en términos de optimización para este proyecto.

5.4.1. Latent Dirichlet Allocation (LDA)

Para el modelado de tópicos a través de LDA 3.6, se utilizó la funcionalidad `LdaMulticore` de Gensim. Esto permite hacer uso eficiente de varios núcleos de procesamiento para calcular los tópicos de los modelos, optimizando significativamente el tiempo de cálculo.

El proceso de creación de un modelo requiere de 3 pasos:

1. **Creación de un diccionario**: El diccionario es una estructura que asigna un identificador único a cada palabra en nuestro conjunto de documentos, permitiendo así una representación coherente y manejable de las palabras en nuestro análisis. Se utiliza la función `Dictionary` de Gensim.

2. **Creación del corpus**: El corpus es una representación de nuestros documentos basada en el diccionario previamente creado, donde cada documento es transformado en una lista de tuplas que contienen el identificador de la palabra y su frecuencia en ese documento, utilizando el método `doc2bow` del diccionario.

3. **Creación del modelo LDA**: Con el corpus, el modelo y los parámetros seteados, se procede a entrenar al modelo. Durante el entrenamiento, el

modelo analiza el corpus iterativamente, ajustando las probabilidades de cada palabra de pertenecer a los distintos tópicos.

El resultado de ejecutar el entrenamiento del modelo de LDA es un conjunto de tópicos que representan temáticas latentes dentro del corpus de documentos analizados. Cada tópico está definido por un conjunto de sus palabras clave más relevantes, junto con las probabilidades de cada palabra de pertenecer a ese tópico. Además, el modelo proporciona las distribuciones de tópicos para cada noticia del corpus.

Con un entrenamiento adecuado, ya es posible obtener resultados interesantes a partir del corpus heredado. Por ejemplo, es posible clasificar y agrupar las noticias de la región dentro de los tópicos encontrados en el aprendizaje. Esto se logra identificando el tópico predominante en cada noticia y agrupándolas en función de dicho tópico.

La Figura 5.6 muestra los cinco tópicos generados en un modelo LDA de prueba, con las palabras clave más relevantes para cada tópico y su correspondiente peso.

```
(0, '0.007*"policia" + 0.006*"policial" + 0.006*"hecho" + 0.005*"personal"
     + 0.005*"mujer" + 0.005*"momento" + 0.005*"encontrar" + 0.004*"fiscal
    " + 0.004*"persona" + 0.004*"noticia"')
(1, '0.007*"equipo" + 0.007*"club" + 0.006*"partido" + 0.005*"final" +
    0.005*"fecha" + 0.005*"punto" + 0.005*"deportivo" + 0.005*"jugar" +
    0.004*"torneo" + 0.004*"deporte"')
(2, '0.006*"trabajo" + 0.006*"actividad" + 0.005*"municipal" + 0.005*"obra
    " + 0.005*"trabajar" + 0.005*"social" + 0.005*"desarrollo" + 0.004*"
    proyecto" + 0.004*"intendente" + 0.004*"area"')
(3, '0.008*"gobierno" + 0.005*"trabajador" + 0.004*"millon" + 0.004*"
    frente" + 0.004*"parte" + 0.004*"servicio" + 0.004*"politico" +
    0.004*"ley" + 0.004*"pago" + 0.004*"diputado"')
(4, '0.007*"salud" + 0.007*"persona" + 0.007*"escuela" + 0.006*"caso" +
    0.004*"trabajar" + 0.004*"hospital" + 0.004*"docente" + 0.003*"chico"
    + 0.003*"tiempo" + 0.003*"familia"')
```

Figura 5.6: Resultados de un modelo LDA de 5 tópicos.

5.4.2. Optimización y refinamiento

Durante el proceso de modelado de LDA, aun con el preprocesamiento de palabras, existe mucho ruido en los documentos que ralentiza el proceso y altera de forma negativa los resultados.

Para detectar mejor este ruido, se desarrolló un proceso en Python que permite, dado un dataset de documentos (en este caso, las noticias ya preprocesadas), generar un documento de frecuencias de palabras únicas. Al resultado se le agrega las siguientes columnas en un archivo CSV:

- **Word**: La palabra única.
- **Count**: Cantidad de apariciones.

- **DocCount**: Cantidad de documentos en donde aparece la palabra al menos una vez.
- **POS**: Tipo de palabra, detectado a través de Spacy.
- **POS_ES**: Tipo de palabra, representado en español.

Además, se ordenaron las filas por la columna **Count**. La Figura 5.12 muestra un ejemplo de las palabras más comunes del diccionario, limitado a las primeras 10 filas:

Word	Count	DocCount	POS	POS_ES
año	100712	45364	NOUN	Sustantivo
chubut	57364	28114	NOUN	Sustantivo
nacional	55143	28091	ADJ	Adjetivo
ciudad	54543	29808	PROPN	Nombre propio
argentino	53436	24705	ADJ	Adjetivo
tener	48584	30722	VERB	Verbo
provincia	47588	23852	NOUN	Sustantivo
trabajo	45223	25108	NOUN	Sustantivo
persona	45206	24153	NOUN	Sustantivo
poder	42035	28683	AUX	Verbo auxiliar

Cuadro 5.12: Frecuencias de palabras únicas en el dataset de noticias.

Este proceso ayudó a dividir el ruido del dataset en 2 grupos:

- **Palabras irrelevantes:** Palabras que, aun con el preprocesamiento previo, no representan nada a la hora de intentar inferir de qué habla una noticia.
- **Extremos:** Palabras muy comunes en el idioma o muy raras que ralentizan el proceso y no producen modificaciones de valor al modelo LDA.

Estos dos grupos de palabras se tratan como una limpieza adicional del *dataset*, para lo cual se desarrollaron procesos específicos de filtrado previo al armado del diccionario. Esta etapa se diferencia del preprocesamiento de texto 5.3 inicial, ya que está específicamente orientada al *Topic Modeling* 3.5. En este punto, las

palabras ya están normalizadas, por lo que el objetivo es filtrar aquellas que no aportan significado relevante a las noticias.

En el código 5.4.2 se presenta el código en Python que muestra el proceso de creación y filtrado del diccionario aplicado a un ejemplo de documentos:

```
from gensim import corpora

exampleDocuments = [
["economia", "crecer", "poder", "gobierno", "medida"],
["gobierno", "cambio", "social", "poder"],
["poblacion", "medida", "haber", "beneficiar", "crecer"],
["inflacion", "cambio", "consecuencia", "global"],
["gobierno", "crecer", "haber", "inflacion", "medida"],
["social", "poblacion", "crecer", "economia"]
]

blacklist = ['poder', 'haber']
no_below= 0.3
# 30% de los documentos. La palabra debe aparecer en 2 o mas
   documentos
no_above = 0.5
# 50% de los documentos. La palabra no debe aparecer 4 o mas
   documentos

dictionary = corpora.Dictionary(exampleDocuments)
# Diccionario inicial:
# { 'economia': 0, 'crecer': 1, 'poder': 2, 'gobierno': 3, '
   medida': 4, 'cambio': 5, 'social': 6, 'poblacion': 7, 'haber
   ': 8, 'beneficiar': 9, 'inflacion': 10, 'consecuencia': 11, '
```

```
        global': 12}

    dictionary.filter_tokens(blacklist)
    # Se elimina 'poder' y 'haber'

    dictionary.filter_extremes(no_below, no_above)
    # Se elimina 'consecuencia', 'global', 'beneficiar' por no_below
    # Se elimina 'crecer' por no_above

    # Diccionario final despues del filtrado:
    # { 'economia': 0, 'gobierno': 1, 'medida': 2, 'cambio': 3, '
        social': 4, 'poblacion': 5, 'inflacion': 6}
```

En el Anexo, la sección A detalla el comportamiento de los modelos después de aplicar los procesos explicados a continuación.

Los dos procesos de optimización y refinamiento son los siguientes:

5.4.2.1. Blacklist

En varias de las pruebas iniciales se pudieron detectar muchas palabras irrelevantes que se colaban en los principales tópicos, dificultando la capacidad de inferir cuál era la semántica del tema. Algunos ejemplos de palabras irrelevantes son:

- **Verbos Genéricos**: Se trata de verbos que no aportan un significado específico al contexto del tópico analizado. Su uso es frecuente en el lenguaje general. Ejemplos incluyen: “hacer”, “ir”, “realizar”, “dar”, “llevar”, “haber”, entre otros.
- **Palabras espaciales**: Palabras que indican ubicaciones geográficas recurrentes en el corpus, las cuales no son esenciales para el análisis temático es-

pecífico y pueden sesgar la interpretación. Ejemplos incluyen: “argentina”, “chubut”, “patagonia”, “provincia”, “pais”, “ciudad”, entre otros.

- **Palabras temporales**: Similar al punto anterior. Palabras que denotan temporalidad y son comunes en las noticias, pero no agregan valor analítico al tema de interés: “día”, “hora”, “mes”, “año”, “semana”, etc.

- **Nombres propios**: Nombres de personas comunes en el idioma que no son identificativos y, por tanto, no contribuyen significativamente a la diferenciación temática en el análisis: “Juan”, “José”, “Luis”, “Mario”, “Martín”, etc.

Estas palabras son comunes en el lenguaje y en la escritura de una noticia, pero para el análisis realizado ocupan espacio y tiempo de procesamiento que podrían tener palabras con significados más relevantes para el tópico. Es por eso que se desarrolló una llamada blacklist antes de crear el diccionario de LDA, que permite agregar estas palabras para que no sean tenidas en cuenta a la hora de realizar el modelado de tópicos.

La asignación de palabras a la blacklist debe ser cuidadosa para no perder información importante por agregar palabras que sí son relevantes para el modelado de tópicos, es por esto que además se realizaron múltiples pruebas de modelados de tópicos para saber qué palabras no relevantes son las que terminan apareciendo en los múltiples tópicos diferentes con frecuencia y dificultan el análisis de los mismos.

5.4.2.2. Filtrado de extremos

Apoyado por el desarrollo del diccionario del dataset en CSV, se pudieron detectar que existen dos extremos de palabras que no aportan significado a los tópicos, estas son:

- **Extremo Superior**: Palabras tan comunes que aparecen en más del 50 % de las noticias, suelen ser verbos o conectores comunes en el idioma.
- **Extremo Inferior**: Palabras tan raras que aparecen en muy pocos documentos y solo ralentizan el modelado de tópicos ya que no terminan siendo relevantes. Estas se pueden dividir en varios tipos:
  - **Palabras muy específicas**: Estas son palabras que son muy particulares de una noticia, como por ejemplo: “bellamente”, “arqueológicamente”, “subóptima”.
  - **Nombres propios**: Estos son apellidos, nombres raros o nombres de ciudades/países que aparecen en muy pocas noticias, como por ejemplo: “Bielorrusia”, “Olszanowski”, “Franceschelli”.
  - **Errores ortográficos**: Debido a la naturaleza de las noticias que suelen ser escritos por personas, es normal que existan errores ortográficos en la escritura que no se pueden detectar en el preprocesamiento, como por ejemplo: “psdchwmvqm”, “notifficaddo”, “phyton”.
  - **Palabras combinadas**: Similar a los errores ortográficos, las palabras combinadas son dos o más palabras a las que no se les agregó el espacio correspondiente y no pueden ser preprocesadas correctamente, como por ejemplo: “alavesmañana”, “subocupacionnumero”, “turisticospor”.

De las 170.000 palabras únicas detectadas en el dataset, solo 32.000 palabras aparecen en más del 0,01 % de los documentos. Esto quiere decir que el 82 % de las palabras únicas sólo aparecen en menos del 11 noticias.

En el Gráfico 5.7 se muestra la distribución de palabras únicas en el *corpus*, donde se puede apreciar que el pequeño porcentaje de palabras mas comunes es utilizado en la mayoría de documentos. El gráfico incluso se muestra en escala logarítmica por la diferencia de apariciones entre las palabras.

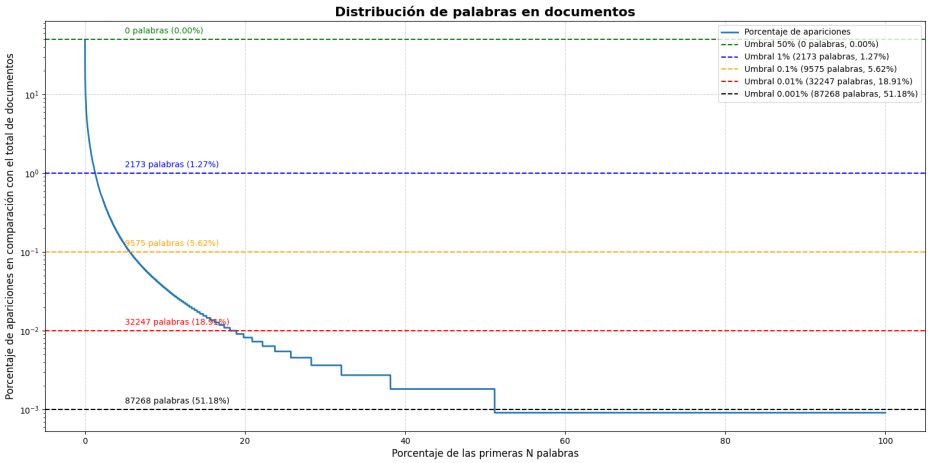

Figura 5.7: Gráfico de aparición de palabras únicas sobre el total de documentos

Es por eso que se realizó un filtrado para que solo se tengan en cuenta las palabras que aparecen en más del 0,01 % del total de las noticias. Con el dataset presentado de 109.684 noticias, una palabra tiene que aparecer en al menos 11 noticias para ser tenida en cuenta en el modelado de tópicos. Este proceso ayuda tanto a limpiar el dataset de palabras no relevantes, como a acelerar el proceso de modelado significativamente al tener menos palabras a las cuales calcular las probabilidades de aparecer en un tópico.

5.4.3. Interpretación de un modelo de tópicos

Una forma propuesta (ademas de las métricas 3.6.2) para analizar un modelo de tópicos con LDA es mediante la interpretación humana. Esta consiste en una serie de pasos que permiten identificar si una bolsa de palabras representativa de un tópico tiene sentido o no. Este proceso se utiliza para reconocer e identificar anomalías en el modelo, permitiendo validar o descartar resultados a partir del sentido común. Si bien estas medidas son completamente subjetivas y dependientes del observador, ayudan a dar un primer vistazo o complementar las métricas de comparación entre modelos que permitan elegir uno u otro.

Coherencia e intrusión de palabras

Al observar una bolsa de palabras, se intenta inferir un título representativo para cada tópico a partir de las mismas. Es decir, que en caso que haya palabras que rompan con esa coherencia, eso generaría ruido.

Ejemplo 1:

- Perro
- Gato
- Caballo

Se observa que esta bolsa de palabras tiene coherencia, ya que las palabras son semánticamente similares y un título adecuado podría ser "Animales".

Ejemplo 2:

- Flor
- Césped
- Árbol
- Ventilador
- Pasto

En este ejemplo, no se respeta una coherencia, ya que algunas palabras rompen la similitud semántica en el tópico.

Esta noción de coherencia se relaciona con la tarea conocida como **Word Intrusion**, que fue presentada por Chang et al. (2009) en su libro *Reading Tea Leaves: How Humans Interpret Topic Models* [22].

Contextualización

Utilizando la **noticia más representativa de un tópico** 9, es posible analizar si existe correspondencia entre la noticia y la bolsa de palabras que caracteriza dicho tópico. Por ejemplo, si el tópico está compuesto por las palabras:

- Tecnología
- Innovación
- Software
- Empresa

Debería observarse que las noticias más representativas de este tópico traten sobre avances tecnológicos, lanzamientos de software o noticias sobre empresas del sector tecnológico.vos.

Conocimientos previos

Debido a que el dataset está conformado por noticias de diarios de Chubut entre los años 2019 y 2022 5.2, hay ciertos conocimientos sobre las noticias ocurridas y temáticas más populares que se espera que aparezcan dentro de los N tópicos más importantes de la región, algunos ejemplos de estos son:

- **Coronavirus**: Inició en Argentina en marzo de 2020.
- **Turismo**: Mayor relevancia en época de temporada (verano/invierno).
- **Alcoholemia**: Mayor relevancia en fiestas.
- **Eventos deportivos relevantes**: Copa América (Jun-2021), Mundial (Dic-2022), Olimpiadas Tokio (Jul-2021), Automovilismo, etc.
- **Otros**: Política, Economía, Noticias locales, Producción, Educación, Transporte, Seguridad, Social, Guerras, etc.

Varios estudios han incorporado el concepto de ”conocimientos previos” al analizar los resultados de modelos de tópicos aplicados a noticias. Por ejemplo, el trabajo titulado *Temporal topics in online news articles: Migration crisis in Venezuela* [23] destaca la relevancia de entender el contexto histórico y social de los eventos para interpretar adecuadamente los temas emergentes en el análisis de noticias.

5.5. Análisis estático de los tópicos

Para el análisis de los tópicos de manera estática, se desarrollaron métricas, gráficos y matrices que ayudan tanto a visualizar el modelo de tópicos generado por LDA como a validarlo y evaluar su coherencia y reflejo con la realidad.

Para este análisis es importante dar definición de algunos conceptos que se presentan a continuación:

Definición 1 (Peso de un tópico en una noticia)**.** *Valor entre 0 y 1 que representa la relevancia del tópico en una noticia. Valores cercanos a 1 indican que la noticia trata casi en su totalidad sobre el tópico, mientras que valores cercanos a 0 indican que la noticia casi no contiene el tópico. Gensim provee una función que dado un documento y un modelo LDA entrenado, se obtiene una lista con los tópicos que componen al documento. Se representa con:*

$$Peso(i, tópico)$$

Definición 2 (Peso total de un tópico)**.** *Es la sumatoria del peso de un tópico en cada noticia del dataset analizado. Se representa con la fórmula:*

$$Peso\ total\ tópico = \sum_{i=1}^{N} Peso(i, topico)$$

Definición 3 (Peso promedio de un tópico)**.** *Es la división entre el peso total de un tópico sobre la cantidad de noticias analizadas.*

$$Avg(i) = \frac{\sum_{i=1}^{N} Peso(i, tópico)}{N}$$

Definición 4 (Aparición de un tópico en una noticia)**.** *Es una función indicadora que*

retorna 1 si el tópico aparece en la noticia analizada (valor de peso mayor a 0) y 0 en caso contrario:

$$\delta(tópico \in Doc_i) = \begin{cases} 1 & si\ Peso(i, tópico) > 0 \\ 0 & si\ Peso(i, tópico) = 0 \end{cases}$$

Definición 5 (Apariciones totales de un tópico)**.** *Es la sumatoria del cálculo de la función indicadora de aparición en todas las noticias, representa en cuántas noticias del dataset aparece el tópico:*

$$A(i) = \sum_{i=1}^{N} \delta(tópico \in Doc_i)$$

Definición 6 (Predominancia de un tópico en una noticia)**.** *Un tópico es predominante en una noticia si su peso es mayor al 50 % de la noticia, es decir, mayor a 0.5. La función devuelve 1 en caso de ser predominante y 0 en caso de no serlo:*

$$Predominancia(tópico, i) = \begin{cases} 1 & si\ Peso(i, tópico) \geq 0{,}5 \\ 0 & si\ Peso(i, tópico) < 0{,}5 \end{cases}$$

Definición 7 (Predominancia total de noticias de un tópico)**.** *Es la sumatoria de la función de predominancia en todas las noticias, y representa en cuántos documentos el tópico es predominante:*

$$Pt = \sum_{i=1}^{N} Predominancia(tópico, i)$$

Definición 8 (Proporción de predominancia de un tópico)**.** *Es un valor entre 0 y 1 que se calcula dividiendo la predominancia total de un tópico sobre la cantidad de noticias en las que aparece. Mientras más alto el valor, indica que el tópico suele ser más*

predominante en las noticias en las que aparece; valores más bajos indican que el tópico suele aparecer pero no suele ser el tema principal de las noticias en donde se lo menciona.

$$\textit{Proporción de predominancia}(\textit{tópico}) = \frac{\textit{Predominancia total}}{A(\textit{tópico})}$$

Definición 9 (Noticia más representativa de un tópico)**.** *Se define como la noticia dentro del conjunto de documentos analizados que más peso tiene de un tópico específico.*

En la definición 6, se puede observar cómo se determina si un tópico predomina en una noticia.

A continuación se presentan las herramientas desarrolladas y utilizadas para el análisis estático de tópicos:

5.5.1. Pie Chart

A través de la librería `pyplot`, y dado un modelo LDA, se genera un gráfico de torta que permite visualizar la relevancia relativa de cada tópico en el conjunto de datos. El cálculo de la relevancia total de cada tópico se realizó a partir del promedio de aparición del tópico en el dataset 3. Rápidamente se puede visualizar qué tópicos son más relevantes en el conjunto de datos e incluso se observa el porcentaje que no pudo ser asignado a ningún tópico (Ver Figura 5.8).

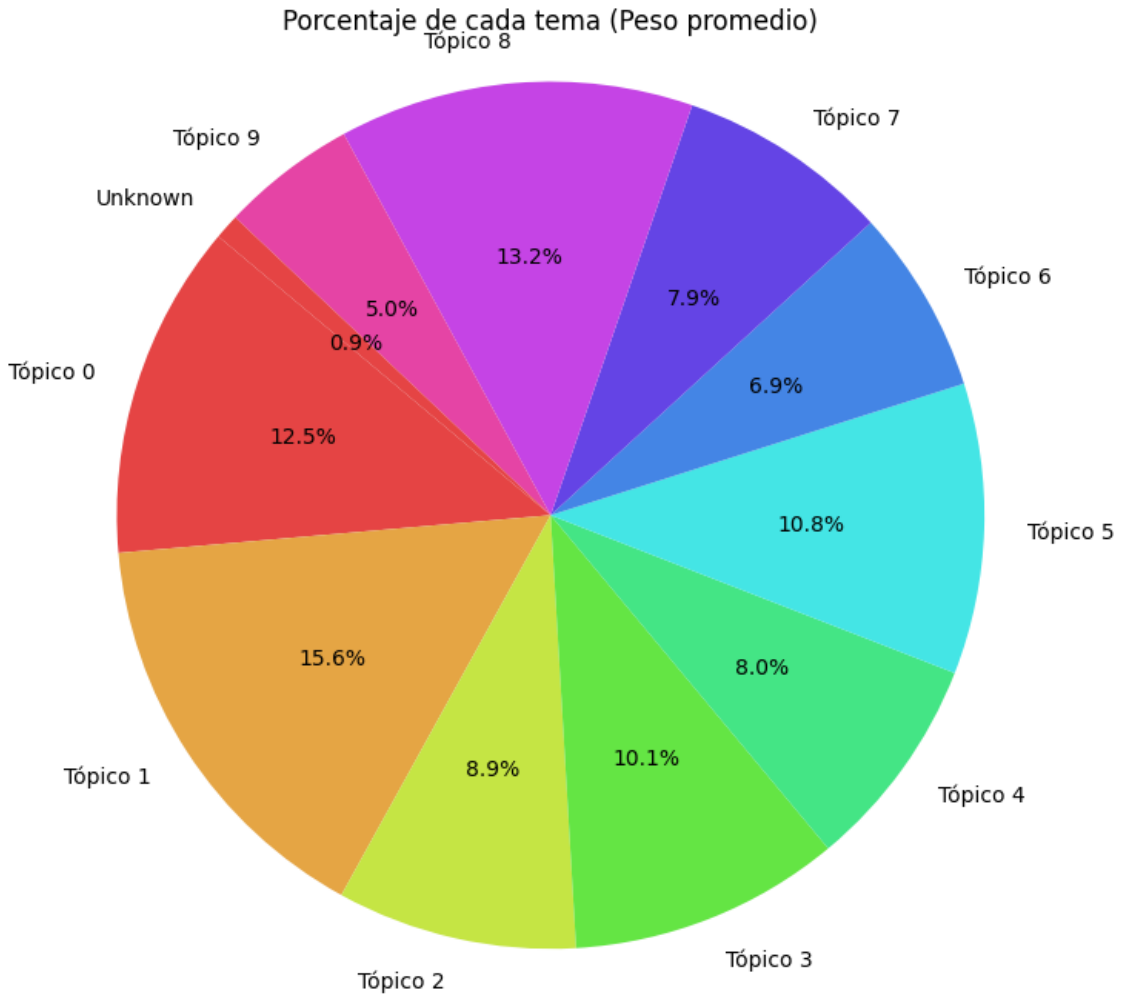

Figura 5.8: Ejemplo de Gráfico torta de un modelo LDA

5.5.2. PyLDAvis

`PyLDAvis` es una herramienta de visualización interactiva para *Topic Modeling* 3.5. Permite visualizar un modelo *Latent Dirichlet Allocation* (**LDA**) 3.6 en un plano bidimensional mediante círculos, donde el tamaño de cada uno indica su relevancia, la distancia entre ellos su similitud, y el contenido muestra las palabras más relevantes de cada tópico. [24].

`PyLDAvis` facilita la evaluación rápida de un modelo LDA, ya que al entrenar el modelo se busca que las palabras más relevantes de cada tópico sean suficientemente diferentes entre sí. En la visualización, esto se representa con círculos de cada tópico que están lo suficientemente separados entre sí.

Se presenta un ejemplo de un gráfico *PyLDAvis* en la Figura 5.9, donde se destacan los puntos mencionados (distancia, magnitud y contenido).

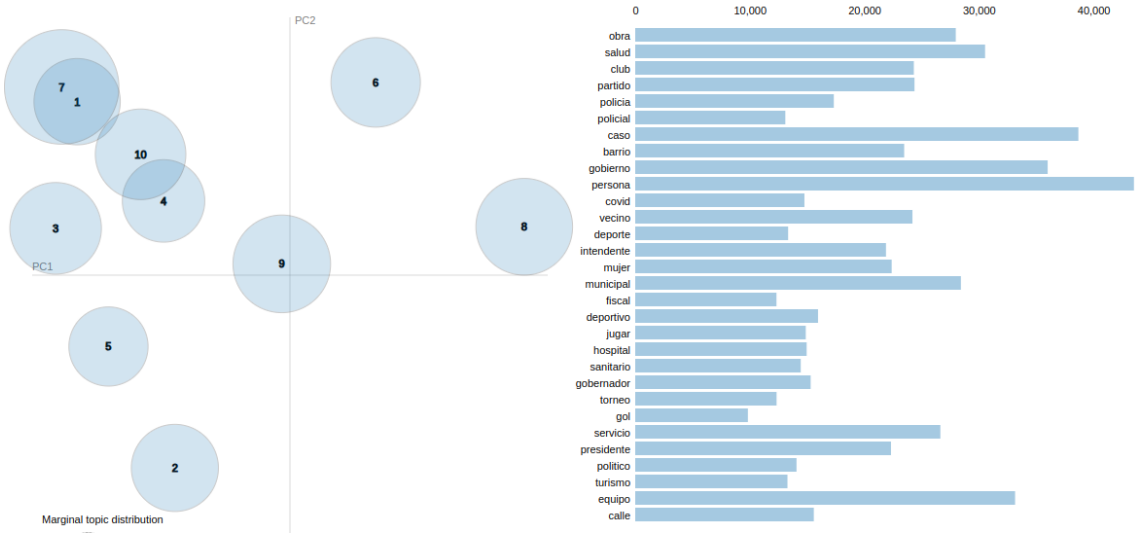

Figura 5.9: Ejemplo de la visualización PyLDAvis de un modelo LDA

5.5.3. Grafo de conocimiento

Se realizó un modelado de grafo de conocimiento (*Knowledge Graph*) con `Neo4J`, ejemplificado en la figura 5.10, que muestra la relación de un tópico con sus N palabras más relevantes. Esto permite visualizar rápidamente qué palabras están relacionadas a múltiples tópicos, así como qué tópicos son más cercanos según sus palabras en común.

Para su construcción, primero se obtuvieron los resultados del modelo LDA, estas palabras y sus respectivas relaciones fueron almacenadas como nodos y aristas en Neo4J, permitiendo su visualización en forma de grafo.

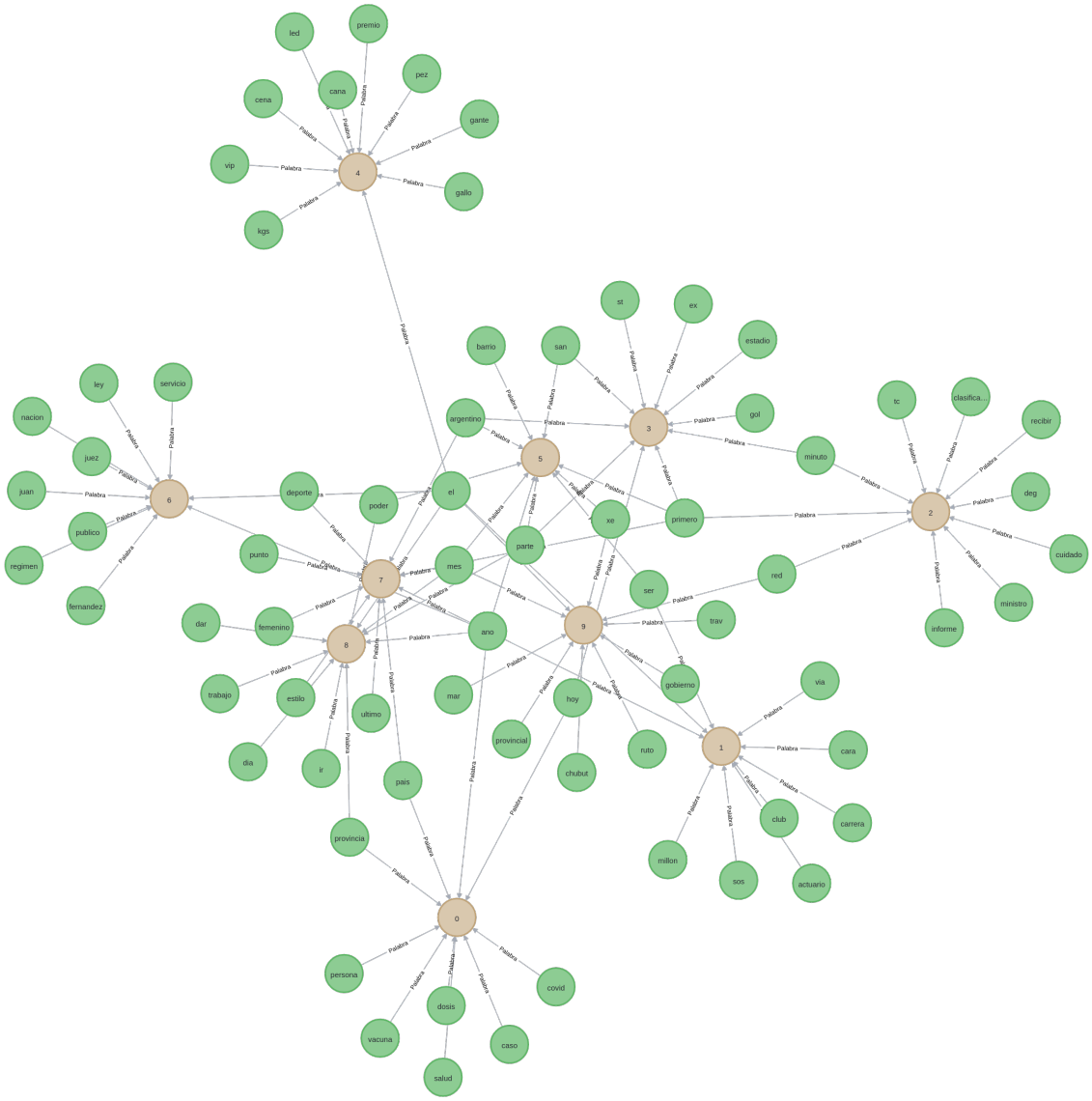

Figura 5.10: Ejemplo de grafo de conocimiento de un modelo LDA, visualizado a través de **Neo4J Bloom**

5.5.4. Matrices

Se desarrollaron dos tipos de matrices que permiten comparar los tópicos entre sí desde dos enfoques diferentes. Estas matrices proporcionan información más analítica en comparación con los gráficos y permiten una evaluación detallada y cuantitativa de la relación entre los tópicos, proporcionando una base sólida para el análisis y la interpretación del modelo.

Las matrices son cuadradas con dimensiones de N x N, siendo N la cantidad

de tópicos del modelo. Esto significa que, a medida que el número de tópicos del modelo aumenta, el tamaño de la matriz también crece.

5.5.4.1. Matriz de similitud del coseno

La similitud del coseno es una medida que evalúa la similitud entre dos vectores calculando el coseno del ángulo entre ellos [25]. Esta medida varía entre -1 y 1, donde 1 significa que los vectores son idénticos, 0 indica que no tienen relación (son ortogonales) y -1 indica que son diametralmente opuestos (ver Figura 5.11).

1. $cos(0°) = 1$
2. $cos(90°) = 0$
3. $cos(180°) = -1$

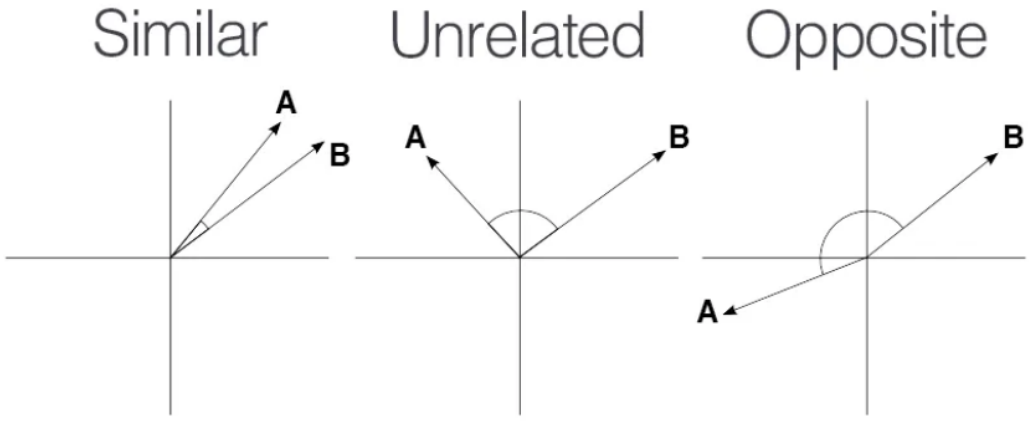

Figura 5.11: Ejemplo de similitud del coseno de vectores. Fuente: https://medium.com/geekculture/cosine-similarity-and-cosine-distance-48eed889a5c4

En **LDA**, calcular la *similitud del coseno* entre dos tópicos permite medir su similitud basada en sus palabras. Los tópicos son interpretados como **vectores** en un espacio multidimensional conformados por las *probabilidades* de sus palabras más relevantes. Es importante destacar que, como los vectores de tópicos están formados por probabilidades de palabras (que van de valores de 0 a 1), estos no

pueden ser negativos y, por consecuencia, no puede haber tópicos "**opuestos**" entre sí. La mayor distancia entre dos tópicos es que sean **ortogonales**, lo que significa que no tienen relación alguna y, por lo tanto, su similitud es 0.

La **matriz de similitud del coseno** desarrollada representa de forma analítica la similitud entre cada par de tópicos. Lo que en *PyLDAvis* se muestra con una distancia entre ellos, en esta sección se puede ver de forma exacta. Vale destacar que la matriz es **simétrica**, la similitud entre el Tópico "i" con el Tópico "j" es igual que la del Tópico "j" con el Tópico "i". La similitud de un tópico con consigo mismo es el valor máximo, el cual es 1. La figura 5.12 muestra una Matriz de similitud del coseno de un modelo de ejemplo.

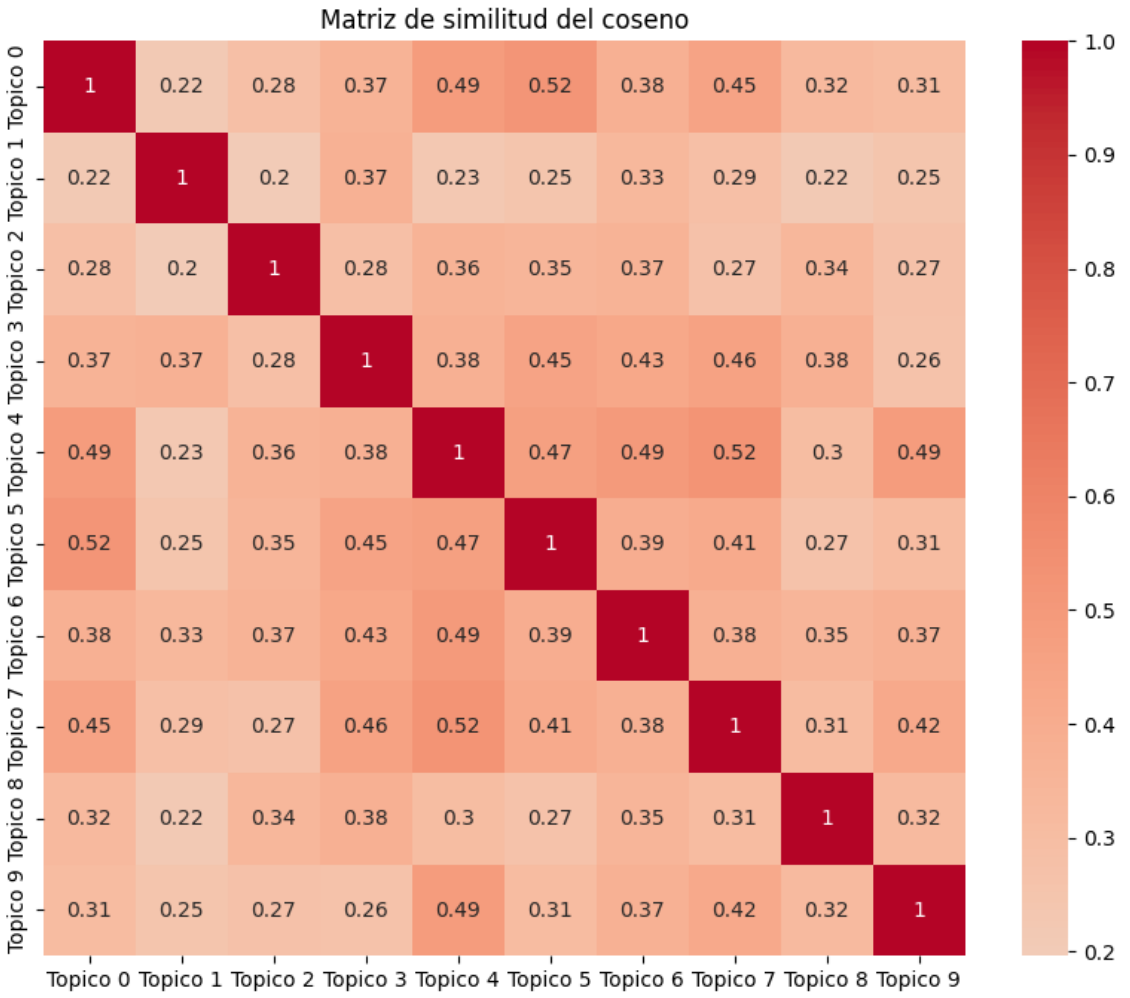

Figura 5.12: Ejemplo de Matriz de similitud del coseno de un modelo LDA

5.5.4.2. Matriz de coaparición

Se desarrolló una matriz que representa la *coaparición* entre los tópicos en el conjunto de documentos.

La *coaparición* se define como la cantidad de veces que dos tópicos aparecen juntos en los documentos. Para dar una representación porcentual, cada celda (i, j) de la matriz indica la proporción de veces que el tópico "i" y el tópico "j" aparecen juntos en relación con la frecuencia del tópico "i" en los documentos (ver Figura de ejemplo 5.13).

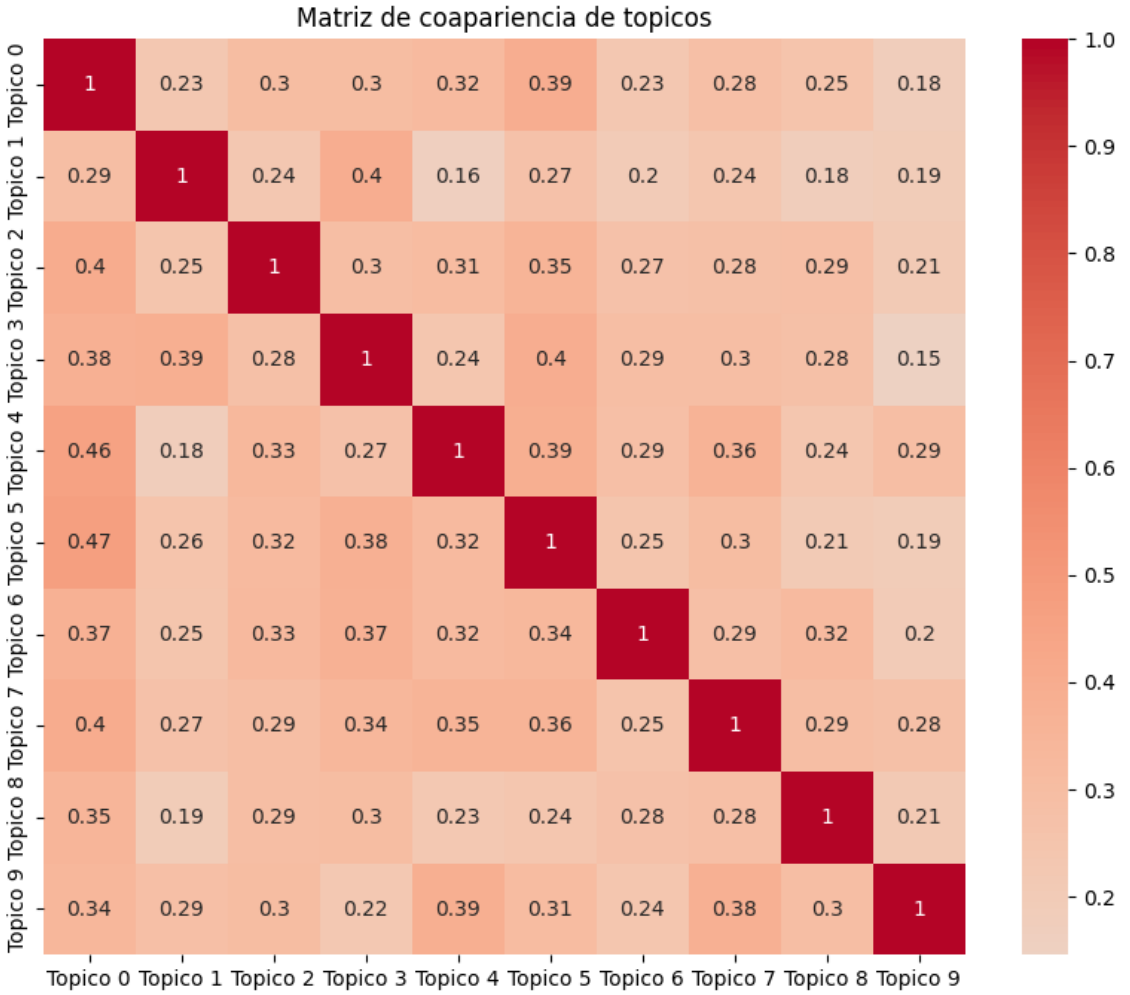

Figura 5.13: Ejemplo de Matriz de coaparición de tópicos de un modelo LDA

La fórmula que se utiliza para el cálculo de cada celda es la siguiente:

$$Ca_{ij} = \frac{\text{Número de veces que aparecen que el tópico i y el tópico j aparecen juntos}}{\text{Numero total de veces que aparece el tópico i}}$$

En la matriz de comparación, se compara cada par de tópicos y se representa de manera relativa la cantidad de veces que aparecen juntos en el dataset. Es decir, mientras más frecuentemente aparezcan juntos en relación a la aparición individual del tópico "i", el valor se acercará más a 1. Si no suelen aparecer juntos, el valor se acercará a 0.

Es importante mencionar que la matriz de coaparición **no es simétrica**, ya que la cantidad de veces que un tópico "i" y un tópico "j" aparecen juntos en relación a la frecuencia de "i" no es la misma que en relación a la frecuencia de "j". Esto se ilustra en el diagrama de Venn de la Figura 5.14.

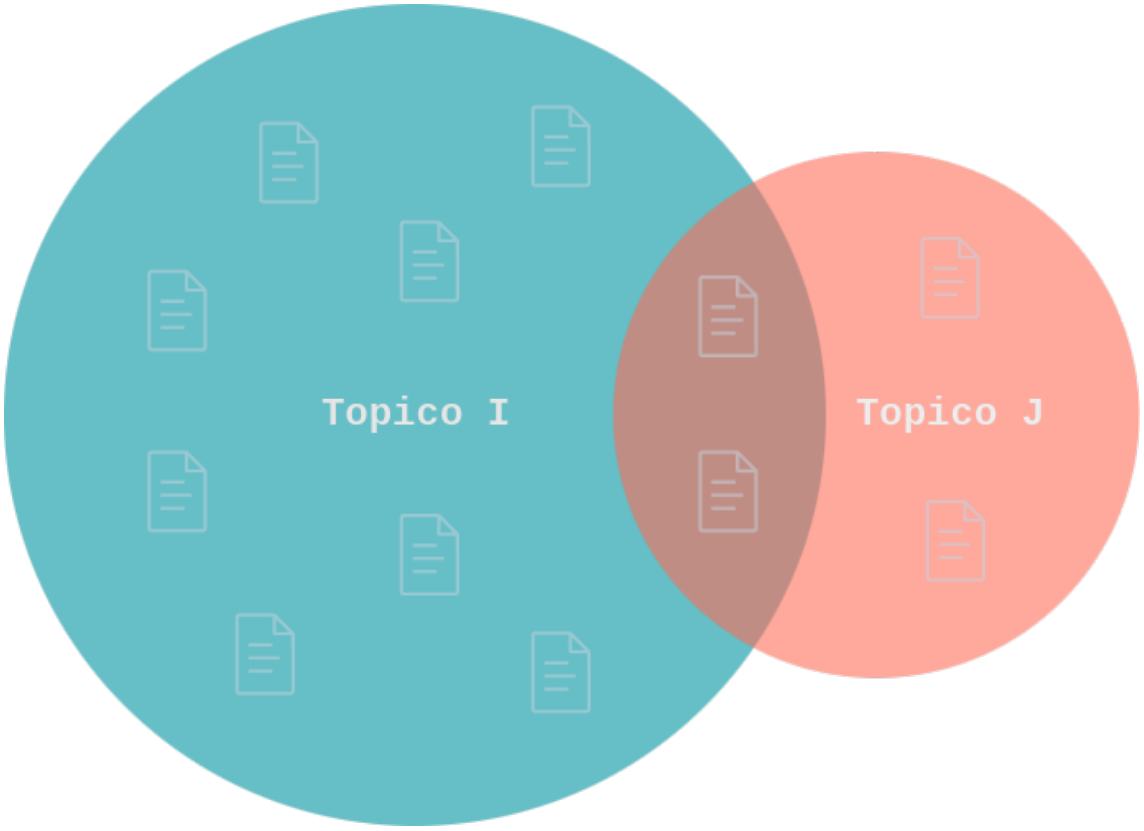

Figura 5.14: Ejemplo de Diagrama de Venn sobre la coaparición entre dos tópicos

En el diagrama, el tópico "i" aparece en 10 documentos, y el tópico "j" aparece

en 4 documentos. Ambos aparecen de manera conjunta en 2 documentos, por lo tanto:

$$Ca_{ij} = 2/10 = 0{,}2$$

$$Ca_{ji} = 2/4 = 0{,}5$$

Esta matriz permite analizar la relación entre los tópicos desde otra perspectiva, ya que muestra qué tan relacionados están dos tópicos según su *coaparición* en los documentos. Así, se puede identificar qué tópicos tienen mayor tendencia a aparecer juntos y cuáles están mas distantes según esta formula.

5.6. Análisis dinámico de los tópicos

En la siguiente sección, se procederá a la solución para un análisis dinámico de los tópicos, es decir, se examinará cómo varían las noticias a lo largo del tiempo dentro del periodo analizado. Para llevar a cabo este análisis, es fundamental determinar y definir la **granularidad mínima** del sistema. Dado que el enfoque del estudio se centra en noticias, el conjunto de datos proporciona la fecha de publicación de cada una de ellas. Esto implica que la granularidad estará restringida a días completos, ya que no es posible detectar variaciones en horas, minutos o segundos. De igual manera, es relevante analizar las noticias en periodos determinados, tales como: diario, semanal, mensual, trimestral y anual. Estos análisis son útiles para identificar tanto temas virales de corto plazo como cambios de tendencia en las publicaciones de los medios de comunicación.

Para esta sección también es importante calcular la **relevancia de un tópico en un lapso de tiempo**. Este se define como la relevancia relativa del tópico repartida en las noticias que pertenecen a ese lapso de tiempo. Es decir, el promedio de la relevancia en ese subconjunto de noticias analizadas. Se puede definir con la siguiente fórmula:

$$R(d,t) = \frac{\sum_{i=1}^{N(d)} \text{Peso}(topico, i)}{N(d)}$$

Donde:

- $R(d,t)$: Relevancia del tópico “t” en el lapso de tiempo “d”.
- $N(d)$: Cantidad de noticias en el lapso de tiempo “d”.
- $\sum_{i=1}^{N(d)} \text{Peso}(topico, i)$: Peso total del tópico “t” en el lapso de tiempo “d”.
- d: Lapso de tiempo
- t: Tópico

Para analizar la evolución temporal de los tópicos, se desarrollaron gráficos que permiten visualizar y comparar la relevancia de cada uno de los temas a lo largo del tiempo dentro del intervalo del dataset.

5.6.1. Stacked Bar Chart

El gráfico de barras apiladas fue una herramienta utilizada que permite conocer la **relevancia relativa** de todos los tópicos agrupados en lapsos de tiempo. Cada barra vertical indica un lapso de tiempo, cada color dentro de una barra indica un tópico y la amplitud del color su relevancia dentro del lapso de tiempo. Se puede observar que la suma de amplitudes de un lapso de tiempo siempre es cercana a 1, lo que indica que la casi totalidad del contenido de los documentos se puede representar con los tópicos detectados por LDA (ver Figura 5.15).

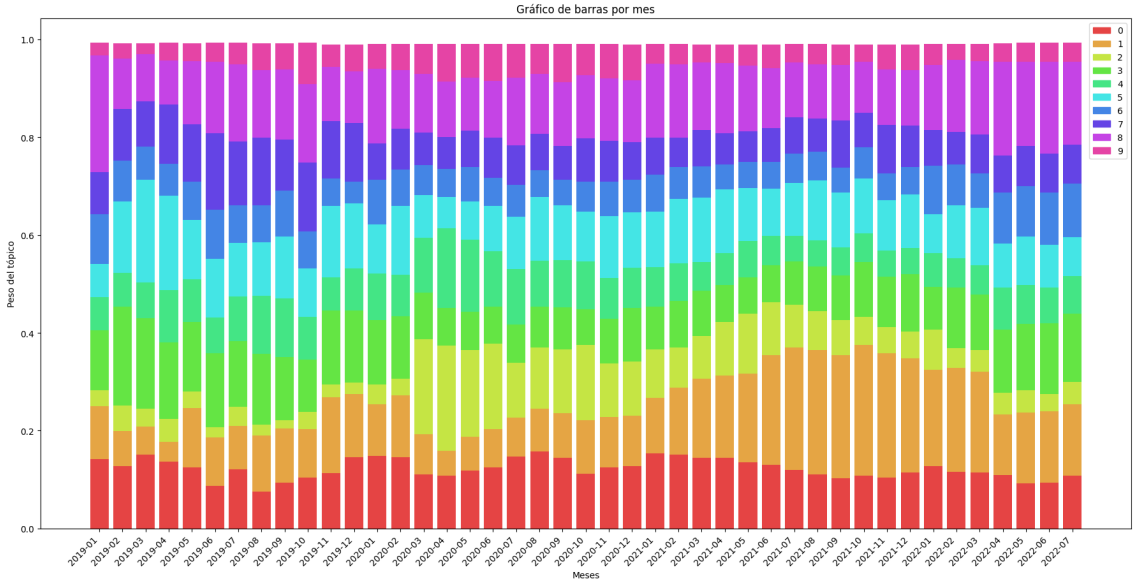

Figura 5.15: Ejemplo de *Stacked Bar chart* segmentado por meses de un modelo LDA

Este gráfico desarrollado utilizando `pyplot` resulta muy útil dado que es fácil visualizar los tópicos más relevantes y su evolución en los lapsos de tiempo marcados, permitiendo comparar incluso varios tópicos a la vez.

Sin embargo, cuando aumenta la cantidad de tópicos se vuelve más difícil visualizar las amplitudes y comparar la relevancia de un tópico contra otro, a su

vez que se vuelve más complejo evaluar los cambios de tendencia de un tópico al desplazarse verticalmente en cada barra. Una desventaja más es que al ser fijos los lapsos de tiempo del gráfico pueden perderse momentos virales del tópico dentro del *dataset* de noticias, ya que son suavizados por el resto de noticias dentro del lapso de tiempo seteado.

5.6.2. Topic Evolution Chart

Teniendo en cuenta las limitaciones del gráfico anterior, se desarrolló un **gráfico interactivo** que permite visualizar más claramente la **evolución** de los tópicos temporalmente y su comparación unos a otros. Para esta sección, utilizando la herramienta `plotly`, se generó un gráfico temporal que representa la relevancia de tópico utilizando la granularidad mínima del *dataset*, la cual es diaria. Esto permite visualizar su peso durante todo el dominio temporal, detectando incluso los días donde el tópico no aparece o donde es tendencia (ver Figura 5.16).

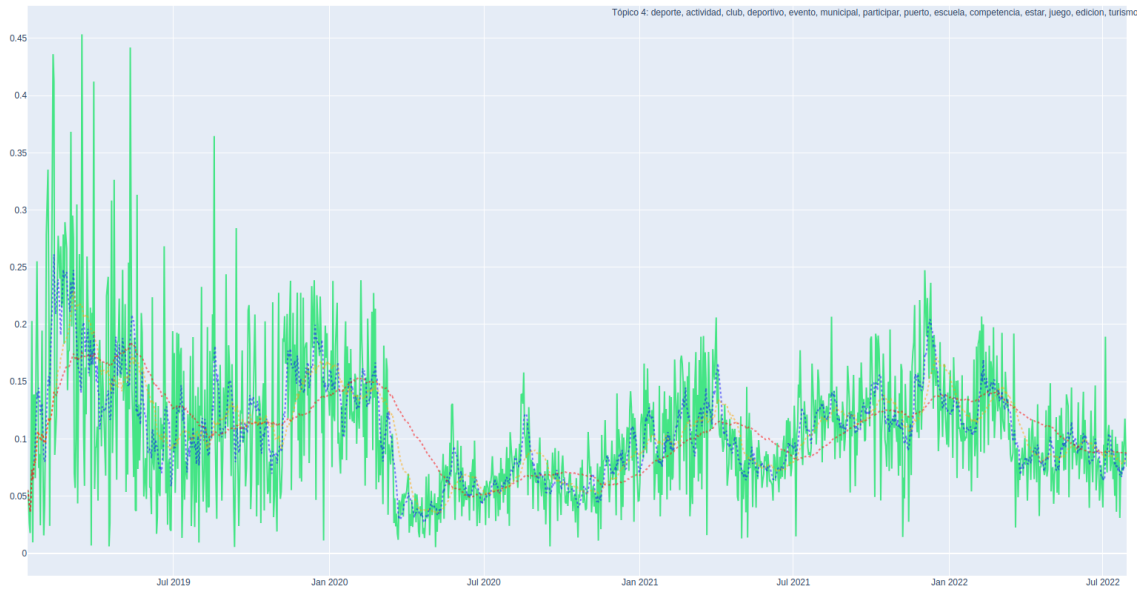

Figura 5.16: Ejemplo de *Topic evolution chart* de un tópico de un modelo LDA

El gráfico permite elegir entre visualizar un tópico específico o varios tópicos a la vez para compararlos. Además, posee la funcionalidad de **suavizado**. Esta permite visualizar la relevancia ponderada en lapsos de tiempo más largos para

detectar mejor los cambios de tendencia en la relevancia de un tópico, seteadas inicialmente en semanal, mensual y trimestral.

Como clara desventaja al gráfico de barras, este resulta más complejo y no permite comparar visualmente más de dos tópicos a la vez, por lo que ambos resultan ser complementarios y no excluyentes uno de otro.

CAPÍTULO 6

ANÁLISIS DE RESULTADOS

En esta sección se analizarán, a través de las métricas 3.6.2 y la interpretabilidad 5.4.3, la cantidad óptima de tópicos para el modelado basado en el *corpus* de noticias de Chubut entre los años 2019 y 2022 . Además, se presentarán los experimentos realizados con modelos tanto estáticos como dinámicos, junto con un análisis detallado y una discusión sobre sus resultados.

6.1. Comparación de métricas

En este trabajo, se entrenaron múltiples modelos LDA con diferentes cantidades de tópicos para evaluar las métricas descritas. En los gráficos que se presentan a continuación, se comparan los puntajes de coherencia 3.6.2.1 (CV y UMass) y perplejidad 3.6.2.2 para diferentes cantidades de tópicos.

En el eje x se muestra la cantidad de tópicos, mientras que en el eje y se presenta el valor de la métrica correspondiente. Cada punto en la gráfica representa el promedio de cinco entrenamientos realizados con el mismo número de tópicos.

En la Figura 6.1, se muestra la coherencia CV para modelos LDA con diferentes números de tópicos. La Figura 6.2 presenta la coherencia UMass, mientras que la Figura 6.3 ilustra la perplejidad para los modelos LDA.

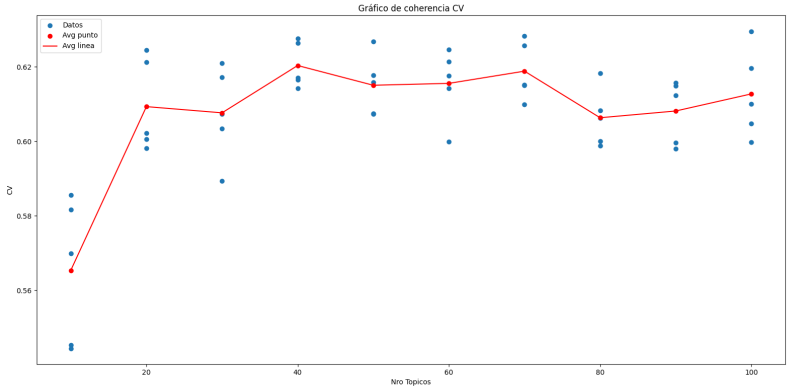

Figura 6.1: Coherencia CV para modelos LDA con diferentes números de tópicos. Valores más altos indican mayor coherencia.

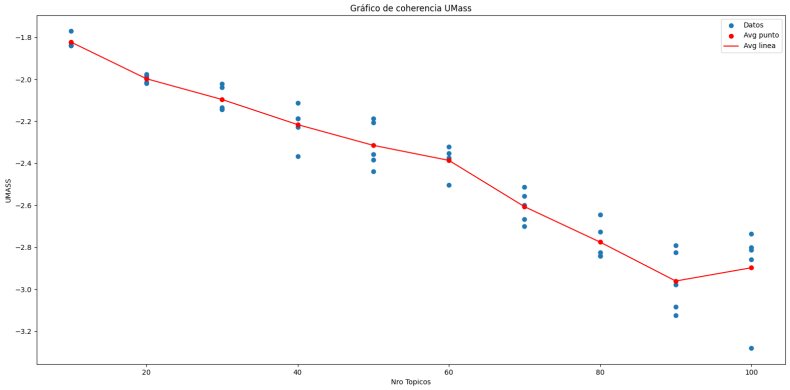

Figura 6.2: Coherencia UMass para modelos LDA con diferentes números de tópicos. Valores más altos indican mayor coherencia.

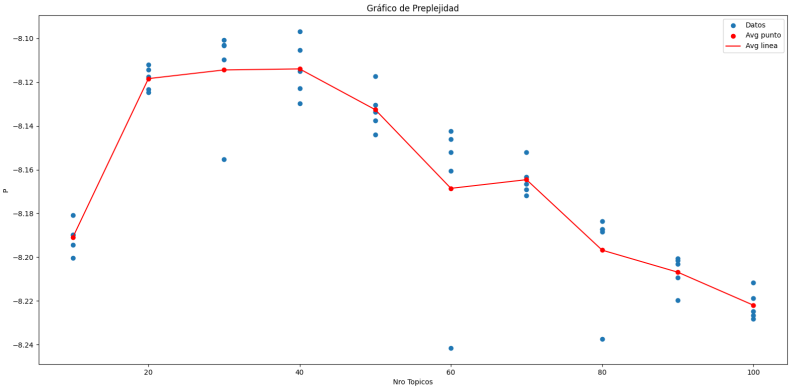

Figura 6.3: Perplejidad para modelos LDA con diferentes números de tópicos. Valores más negativos indican menor confusión del modelo.

6.1.1. Discusión

Basado en las comparaciones de métricas 6.1, se observa que la elección de un número de tópicos para el modelo no es sencilla. Por un lado, la perplejidad tiende a disminuir a medida que se incrementa el número de tópicos, lo que sugiere que el modelo es más efectivo prediciendo las palabras asociadas a cada tópico. Sin embargo, un aumento excesivo en el número de tópicos puede llevar a una sobre-segmentación, donde los temas se vuelven demasiado específicos y pierden sentido interpretativo.

Por otro lado, las métricas de coherencia suelen favorecer un número menor de tópicos, ya que buscan maximizar la consistencia semántica entre las palabras dentro de cada tópico. De hecho, mientras más tópicos se agregan, es común observar una disminución en la coherencia, lo que sugiere que los temas resultantes son menos interpretables 5.4.3.

En este trabajo, se ha elegido utilizar la coherencia `CV` como métrica principal, dado que es una de las más robustas y validadas. No obstante, es importante destacar que esta métrica es complementaria a la interpretación humana y debe ser contrastada con el contexto específico del *dataset* utilizado, en este caso, noticias.

Para ilustrar esto, se entrenaron varios modelos con una cantidad excesiva de tópicos, para comparar con las métricas obtenidas anteriormente (ver Tabla 6.1).

Modelo	Coherencia CV	Coherencia Umass	Perplejidad
100 Tópicos (media)	0.613	-2.897	-8.222
500 Tópicos	0.538	-7.311	-9.255
1000 Tópicos	0.493	-8.166	-10.894
2000 Tópicos	0.427	-5.555	-17.550

Cuadro 6.1: Comparación de métricas para distintos números de tópicos

Como puede observarse, cuando se excede un umbral (de aproximadamente 100 tópicos), la coherencia disminuye drásticamente. Esto indica que modelos con una cantidad excesiva de tópicos no logran capturar una representación coherente de la realidad, a pesar de que la perplejidad continúe disminuyendo.

6.2. Modelos estáticos

Dado que, basado en la discusión 6.1.1, no existe un número exacto de tópicos que represente de manera perfecta la cantidad de temas latentes en las noticias de Chubut, se identificaron ciertos umbrales que permiten mantener la coherencia del modelo. En particular, se determinó que no se deben utilizar menos de 10 tópicos ni exceder los 100, y que, según la comparación de las métricas de coherencia `CV`, el modelo con 40 tópicos ofrece la mejor media con menor desviación.

Para esta experimentación, se entrenaron tres modelos *Latent Dirichlet Allocation* (**LDA**) 3.6 con diferentes cantidades de tópicos:

- Modelo con **10** tópicos: Umbral inferior.
- Modelo con **40** tópicos: Modelo recomendado por la coherencia `CV`.
- Modelo con **100** tópicos: Umbral superior.

6.2.1. Modelo con 10 tópicos

En el siguiente cuadro se muestra el modelo de 10 tópicos. Con las columnas **T** (numero de tópico), **Palabras** (las palabras mas relevantes del tópico), y **Titulo inferido** (titulo coherente que representa a la bolsa de palabras).

T	Palabras	Título inferido
0	obra, trabajo, intendente, trabajar, vecino, municipal, municipio, servicio, barrio, sector, público, importante, gestión, gobierno, proyecto	Gestión municipal y obras públicas
1	salud, caso, persona, covid, sanitario, hospital, coronavirus, aire, pandemia, vacuna, médico, medida, vacunación, dosis, protocolo	Salud y pandemia
2	zona, ruta, río, agua, vehículo, seguridad, incendio, control, bombero, personal, vial, lago, puerto, localidad, servicio	Emergencias y seguridad vial
3	policía, policial, fiscal, mujer, persona, investigación, momento, hombre, personal, causa, juez, encontrar, noticia, caso, víctima	Investigación policial y criminal
4	deporte, actividad, club, deportivo, evento, municipal, participar, puerto, escuela, competencia, estar, juego, edición, turismo, categoría	Actividades deportivas y eventos
5	escuela, desarrollo, trabajo, educación, social, ministerio, programa, proyecto, capacitación, actividad, educativo, docente, persona, trabajar, objetivo	Educación y desarrollo social
6	gobierno, presidente, político, trabajador, frente, diputado, gobernador, ley, situación, ministro, proyecto, fernández, elección, reunión, arcioni	Política y gobierno
7	equipo, partido, fecha, club, jugar, vs, fútbol, gol, torneo, punto, jugador, tiempo, copa, ganar, liga	Fútbol y competiciones deportivas
8	vida, ver, familia, momento, gente, historia, tiempo, vivir, pasar, mundo, casa, llegar, persona, salir, hijo	Historias de vida y experiencias personales
9	millón, peso, empresa, precio, producto, dólar, pago, mercado, mil, banco, aumento, valor, venta, producción, comercio	Economía y finanzas

Cuadro 6.2: Tópicos identificados con palabras clave y títulos inferidos

En la Figura 6.4 se muestra el gráfico *PyLDAvis* del modelo de 10 tópicos. Se destaca como los tópicos similares se juntan en el plano bidimensional y también como los círculos representativos de los tópicos mas relevantes son mas grandes.

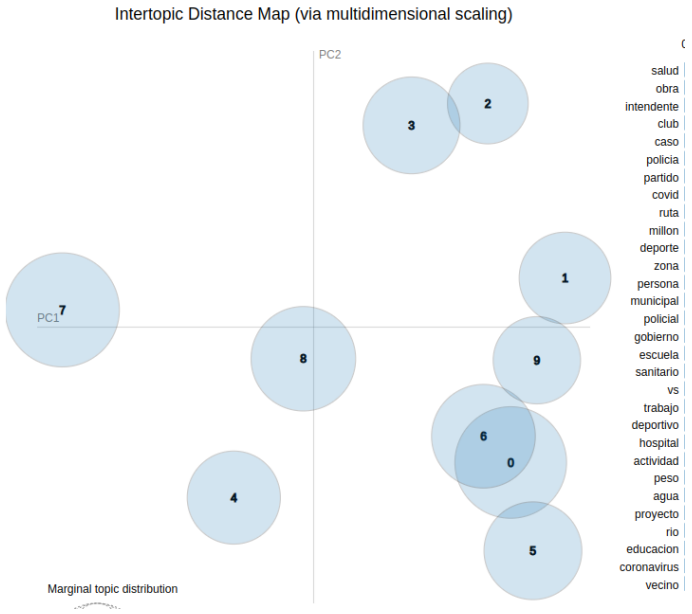

Figura 6.4: Gráfico PyLDAvis 5.5.2 del Modelo LDA de 10 tópicos

En la Figura 6.5 se presenta el gráfico *Pie Chart* 5.5.1 del modelo de 10 tópicos, donde se puede observar el valor exacto de la relevancia de cada tópico.

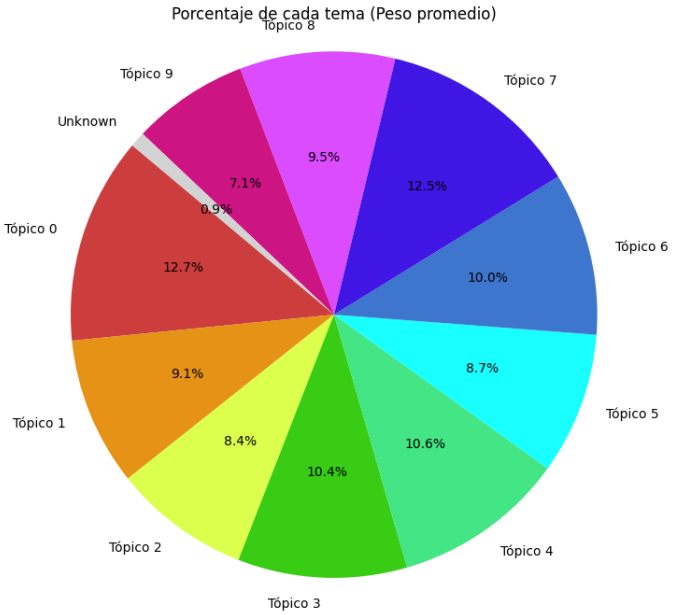

Figura 6.5: Pie Chart 5.5.1 del Modelo LDA de 10 tópicos

En el siguiente cuadro se muestran las estadísticas del modelo de 10 tópicos. Con las columnas **TOPIC** (numero de tópico), **NEWS** 5 (apariciones totales del tópico), **PRED_NEWS** 7 (cantidad de noticias donde el tópico es predominante), **PRED_RATIO** 8 (proporción de predominancia del tópico) y **AVG_WEIGHT** 3 (peso promedio del tópico).

TOPIC	NEWS 5	PRED_NEWS 7	PRED_RATIO 8	AVG_WEIGHT 3
0	48599	9937	0.2045	0.1269
1	37966	7434	0.1958	0.0912
2	36013	6179	0.1716	0.0840
3	31419	10330	0.3288	0.1039
4	36984	9170	0.2479	0.1059
5	35588	6404	0.1799	0.0871
6	36133	8249	0.2283	0.1000
7	29886	13318	0.4456	0.1248
8	38274	6912	0.1806	0.0954
9	32252	4871	0.1510	0.0715

Cuadro 6.3: Estadísticas del modelo de 10 tópicos

6.2.2. Modelo con 40 tópicos

En el siguiente cuadro se muestra el modelo de 40 tópicos. Con las columnas **T** (numero de tópico), **Palabras** (las palabras mas relevantes del tópico) y **Titulo inferido** (titulo coherente que representa a la bolsa de palabras).

T	Palabras	Título inferido
0	gol, partido, dt, equipo, estadio, local, arbitro, fecha, punto	Fútbol y deportes
1	premio, playa, peso, fiesta, edicion, categoria, evento, ganador	Premios y eventos locales
2	vehículo, accidente, conductor, auto, ruta, bombero, camioneta, rodado, hospital, personal	Accidentes de tránsito y seguridad vial
3	peso, precio, pago, banco, dolar, impuesto, venta, ingreso, valor	Economía y finanzas
4	trabajador, gobierno, situación, sector, gremio, sindicato, trabajo, reclamo, deuda, pago	Conflictos laborales y sindicales
5	medida, protocolo, sanitario, persona, actividad, pandemia, situación, deber, covid, social	Medidas sanitarias y pandemia
6	carrera, fecha, categoria, tc, prueba, campeonato, piloto, pista, competencia, puesto	Competencias automovilísticas
7	proyecto, concejo, bombero, sesion, concejal, deliberante, bloque	Política local y proyectos municipales
8	animal, especie, natural, cientifico, area, investigador, ciencia, conicet, investigacion, perro	Ciencias naturales y conservación
9	gente, familia, pasar, ver, trabajar, llegar, tiempo, vivir, salir	Historias de vida
10	transporte, vuelo, viaje, aeropuerto, pasajero, empresa, avion, aereo, viajar, aire	Transporte aéreo y viajes
11	club, deporte, deportivo, actividad, juego, municipal, torneo, futbol, equipo, evento	Deportes y actividades comunitarias
12	vs, deportivo, torneo, partido, equipo, fecha, jugar, cancha, club	Torneos deportivos
13	fiscal, juez, juicio, imputado, prision, causa, delito, tribunal, penal	Justicia y sistema penal
14	santo, aire, rio, cruz, negro, mendoza, fe, cordoba, tierra, fuego	Provincias Argentinas

Continúa en la siguiente página

T	Palabras	Título inferido
15	turismo, turistico, temporada, lago, incendio, turista, zona, comarca, parque, bosque	Turismo y actividades en la naturaleza
16	escuela, educacion, docente, alumno, clase, nivel, secundario	Educación y escuelas
17	social, trabajo, desarrollo, programa, capacitacion, encuentro, area, familia, objetivo	Desarrollo social y capacitación
18	ley, consejo, judicial, justicia, comision, cargo, federal, tribunal, resolucion, presidente	Sistema judicial y leyes
19	mundo, historia, vida, libro, tiempo, pelicula, forma, ver, serie	Historias y cultura popular
20	obra, servicio, agua, vivienda, publico, construccion, infraestructura, gobierno, energia	Obras públicas e infraestructura
21	red, social, video, twitter, noticia, imagen, mensaje, pedir, medio	Medios y redes sociales
22	cultura, artista, actividad, taller, musica, evento, feria, arte, grupo	Eventos culturales y arte
23	caso, covid, positivo, coronavirus, salud, persona, contagio, reportar, fallecido, contacto	Casos de COVID y salud pública
24	madre, niño, padre, hijo, cuerpo, mujer, joven, menor, hija, niña	Familia y relaciones personales
25	seguridad, policia, ministro, jefe, gobernador, massoni, policial, fuerza, gobierno, arcioni	Seguridad y fuerzas policiales
26	vacuna, unido, dosis, millon, chino, agencia, persona, ruso, mundial, mundo	Vacunas y distribución global
27	intendente, municipal, vecino, secretario, municipalidad, gestion, sastre, residuo, trabajar	Gestión municipal y gobierno local
28	ruta, vial, control, seguridad, transito, vehiculo, personal, operativo, zona, rio	Seguridad vial y controles de tránsito
29	mundial, partido, messi, campeon, madrid, españa, equipo, ganar	Fútbol internacional
30	club, boca, jugador, equipo, partido, futbol, copa, river, plantel	Fútbol local y clubes
31	salud, hospital, medico, vacunacion, persona, paciente, sanitario, atencion, enfermedad	Atención médica y hospitales
32	produccion, proyecto, desarrollo, empresa, industria, productor, sector, producto, generar	Industria y desarrollo económico

Continúa en la siguiente página

T	Palabras	Título inferido
33	puerto, pesca, chile, mar, brasil, buque, uruguay, pesquero, portuario, barco	Comercio marítimo y pesca
34	federal, universidad, petrolero, ypf, facultad, malvino, petroleo, secretario, junio, sur	Energía y recursos naturales
35	acto, pueblo, guerra, malvino, militar, aniversario, comunidad, bandera, historia, plaza	Historia y actos conmemorativos
36	mujer, genero, violencia, derecho, persona, ley, diversidad, sexual	Derechos de género y diversidad
37	politico, presidente, diputado, gobierno, eleccion, fernandez, partido, candidato, electoral	Política y elecciones
38	barrio, calle, rc, vecinal, vecino, zona, ubicado, sede, avenida	Barrios y vida comunitaria
39	policia, policial, personal, persona, calle, hombre, arma, encontrar, detenido, vivienda	Seguridad pública y delincuencia

En la Figura 6.6 se muestra el gráfico *PyLDAvis* del modelo de 40 tópicos. Se destaca como los tópicos similares se juntan en el plano bidimensional y también como los círculos representativos de los tópicos mas relevantes son mas grandes.

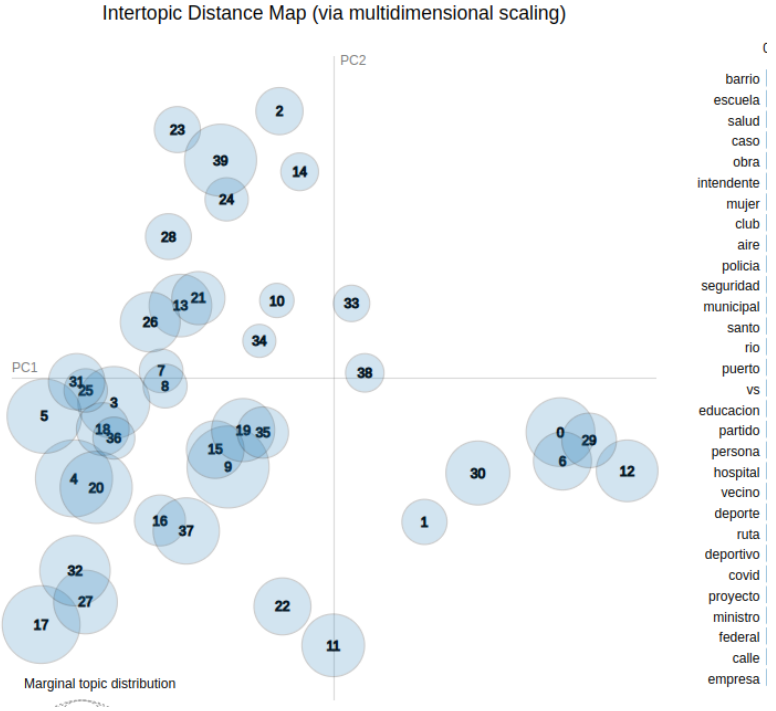

Figura 6.6: Gráfico PyLDAvis 5.5.2 del Modelo LDA de 40 tópicos

En la Figura 6.7 se presenta el gráfico *Pie Chart* 5.5.1 del modelo de 40 tópicos, donde se puede observar el valor exacto de la relevancia de cada tópico.

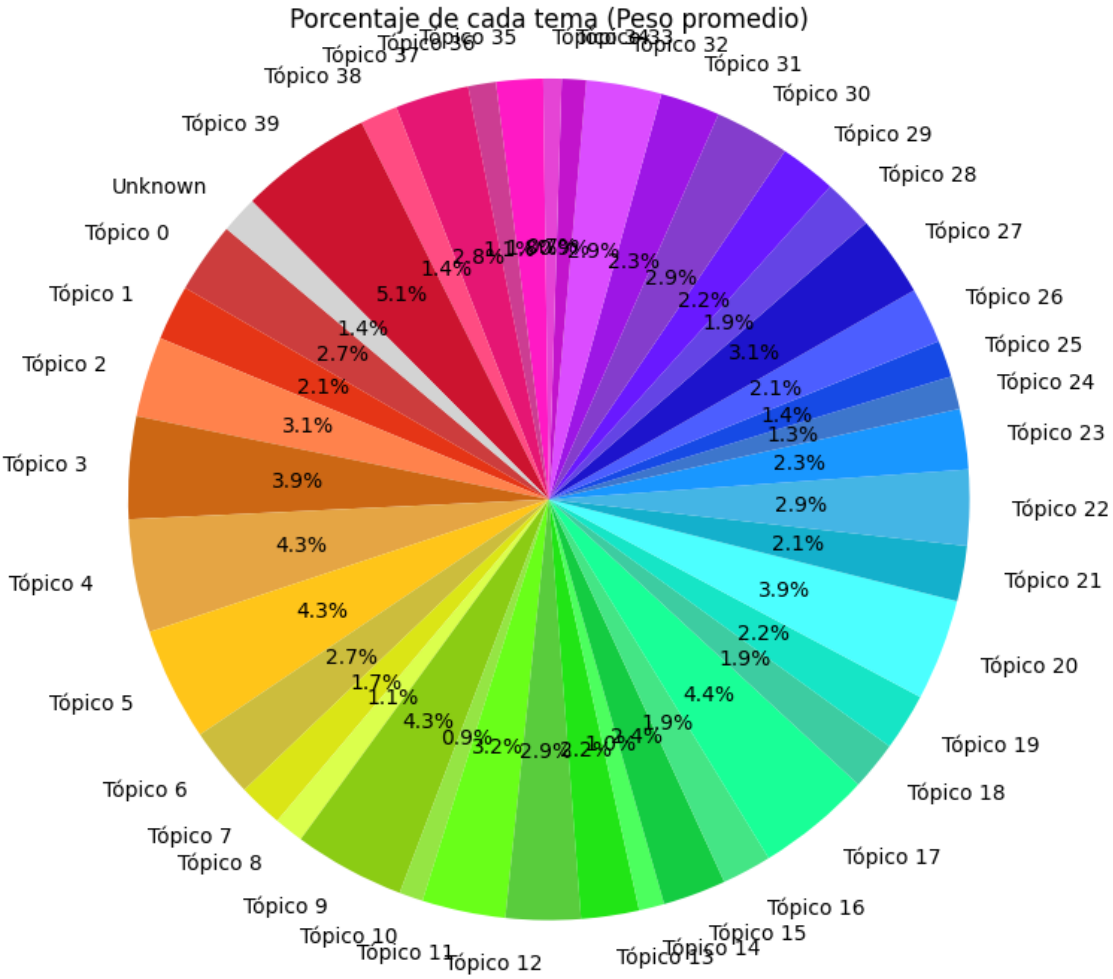

Figura 6.7: Pie Chart 5.5.1 del Modelo LDA de 40 tópicos

En el siguiente cuadro se muestran las estadísticas del modelo de 40 tópicos. Con las columnas **TOPIC** (numero de tópico), **NEWS** 5 (apariciones totales del tópico), **PRED_NEWS** 7 (cantidad de noticias donde el tópico es predominante), **PRED_RATIO** 8 (proporción de predominancia del tópico) y **AVG_WEIGHT** 3 (peso promedio del tópico).

TOPIC	NEWS 5	PRED_NEWS 7	PRED_RATIO 8	AVG_WEIGHT 3
0	15991	2128	0.1331	0.0273
1	16019	1139	0.0711	0.0213
2	15732	2710	0.1723	0.0308
3	21577	2417	0.1120	0.0393
4	21298	2894	0.1359	0.0435
5	25997	1940	0.0746	0.0434
6	12703	2215	0.1744	0.0268
7	14116	621	0.0440	0.0171
8	10668	352	0.0330	0.0113
9	29625	909	0.0307	0.0427
10	11577	177	0.0153	0.0091
11	16464	1892	0.1149	0.0321
12	14265	2192	0.1537	0.0287
13	11786	1648	0.1398	0.0223
14	14545	355	0.0244	0.0097
15	16905	971	0.0574	0.0243
16	14683	601	0.0409	0.0190
17	24306	1956	0.0805	0.0440
18	16063	438	0.0273	0.0186
19	16671	757	0.0454	0.0220
20	18844	2786	0.1478	0.0389
21	19915	448	0.0225	0.0212
22	15137	1935	0.1278	0.0290
23	12364	1742	0.1409	0.0232
24	12429	164	0.0132	0.0127
25	13592	384	0.0283	0.0141
26	12691	1251	0.0986	0.0214
27	20973	843	0.0402	0.0313
28	14163	899	0.0635	0.0194
29	11234	1669	0.1486	0.0218
30	12524	2156	0.1721	0.0287
31	14480	1159	0.0800	0.0229
32	17406	1330	0.0764	0.0289
33	11593	165	0.0142	0.0094
34	8191	262	0.0320	0.0070
35	14596	555	0.0380	0.0180
36	11476	338	0.0295	0.0106
37	15941	1678	0.1053	0.0283
38	15798	326	0.0206	0.0144
39	17595	4807	0.2732	0.0512

Cuadro 6.5: Estadísticas del modelo de 40 tópicos

6.2.3. Modelo con 100 tópicos

En el siguiente cuadro se muestra el modelo de 100 tópicos. Con las columnas **T** (numero de tópico), **Palabras** (las palabras mas relevantes del tópico) y **Titulo inferido** (titulo coherente que representa a la bolsa de palabras).

T	Palabras	Título inferido
0	paso, jardin, indio, cargo, sapo, salado, inicial, maestro, infante, turno, pichiñan, sala, maternal, docente, lagunita	Educación infantil
1	trabajador, gremio, gobierno, sindicato, salarial, salario, reclamo, sueldo, pago, empleado, secretario, situacion, trabajo, sector, paro	Sindicatos y reclamos laborales
2	intendente, municipal, trabajo, trabajar, desarrollo, reunion, gestion, secretario, encuentro, funcionario, gobierno, conjunto, social	Gestión municipal
3	messi, pari, lionel, frances, psg, city, saint, pared, manchester, francia, germain, neymar, astro, league, rosarino	Fútbol internacional (Lionel Messi)
4	velazquez, candela, regata, delegacion, palista, canotaje, kayak, selectivo, escobar, bote, blues, junior, velocidad, alvarez	Deportes acuáticos
5	real, madrid, barcelona, español, españa, volver, sevilla, futuro, valencia, catalan, junio, atletico, sociedad, bilbao, anuncio	Fútbol español
6	vecino, familia, barrio, vivienda, social, necesitar, situacion, gente, entregar, casa, ayuda, hogar, trabajo, trabajar, gas	Vivienda social y ayuda comunitaria
7	medico, caso, paciente, hospital, persona, covid, positivo, sintoma, contacto, salud, enfermedad, intensivo, coronavirus, detectar	Salud y COVID-19
8	iglesia, papa, navidad, celebrar, comunidad, familia, celebracion, francisco, padre, santo, religioso, navideño, medicinal, parroquia	Religión y celebraciones
9	fecha, partido, arbitro, gol, zona, tiempo, local, punto, brown, deportivo, estudiante, estadio, victoria, torneo, tiro	Competencias locales
10	punto, rebote, asistencia, triple, golf, doble, tanto, muzio, temporada, base, nba, golpe, vega, equipo, juego	Baloncesto
11	solidario, jubilado, donacion, aporte, social, colaborar, persona, donar, campaña, recibir, asignacion, anse, hijo, peña, jubilacion	Jubilados
12	politico, eleccion, partido, frente, candidato, electoral, voto, lista, dirigente, junto, senador, diputado, cambio, gobierno, presidente	Política y elecciones

Continúa en la siguiente página

T	Palabras	Título inferido
13	vuelo, transporte, aeropuerto, viaje, pasajero, empresa, avion, aereo, estacion, viajar, patagonico, aire, tren, colectivo, aerolinea	Transporte y viajes
14	fiesta, popular, sierra, caballo, gaucho, clandestino, reina, jineteado, campo, evento, modulo, asado, desarrollar, baile, predio	Festividades populares
15	agua, bombero, corte, zona, voluntario, servicio, sector, problema, personal, situacion, calle, vecino, tarea, cuartel, electrico	Servicios públicos y emergencias
16	presidente, fernandez, acto, alberto, jefe, malvino, mandatario, gobierno, bandera, guerra, ex, kirchner, ceremonia, general	Gobierno y presidencia
17	ruso, ucrania, unido, guerra, militar, fuerza, agencia, ataque, nuclear, estadounidense, gobierno, ucraniano, presidente	Conflicto ruso-ucraniano
18	turismo, temporada, turista, actividad, destino, puerto, visitante, area, ballena, protegido, sector, atractivo, natural, verano	Turismo y actividades recreativas
19	tc, piloto, carrera, fecha, vuelta, pista, campeonato, ford, serie, clasificacion, autodromo, chevrolet, auto, posicion, equipo	Automovilismo y carreras (TC)
20	pandemia, medida, sanitario, salud, covid, coronavirus, caso, cuarentena, emergencia, aislamiento, protocolo, contagio	Pandemia y medidas sanitarias
21	trabajar, trabajo, gente, chico, idea, deporte, importante, tiempo, equipo, llegar, estar, esperar, momento, entrenamiento, empezar	Trabajo y deportes
22	rugby, puma, toma, seleccionado, pumo, sudafrica, ledesma, try, australia, frente, capitan, equipo, conversion, encuentro, zelando	Rugby
23	gonzalez, diaz, garcia, lopez, martinez, luciano, fernandez, hernandez, gabriel, perez, alvarez, sanchez, gomez, federico, ariel	Nombres
24	incendio, fuego, bosque, zona, comarca, bombero, hoyo, andino, lago, civil, forestal, afectado, manejo, defensa, llama	Incendios forestales
25	barrio, espacio, vecino, municipal, plaza, luque, intendente, centro, vecinal, sector, calle, trabajo, ubicado, distinto, trabajar	Espacios públicos y barrios
26	pueblo, derecho, comunidad, chile, memoria, humano, originario, chileno, mapuche, justicia, social, libertad, organizacion	Comunidades indigenas
27	vehiculo, accidente, conductor, auto, rodado, camioneta, ruta, hospital, trasladar, policial, personal, camion, sufrir, lesion	Accidentes de tránsito

Continúa en la siguiente página

T	Palabras	Título inferido
28	ley, derecho, deber, resolucion, decision, establecer, articulo, medida, caso, decreto, considerar, norma, forma, legal, accion	Legislación y derechos
29	niño, niña, adolescente, edad, menor, adulto, persona, vida, enfermedad, fisico, madre, padre, infantil, pequeño, riesgo	Salud infantil
30	minero, meseta, mineria, gan, proyecto, zonificacion, gastre, actividad, desarrollo, mina, marcha, telsen, vecino, ley, pueblo	Minería y proyectos de zonificación
31	pesca, petrolero, ypf, pesquero, mar, empresa, petroleo, pesquera, langostino, barco, actividad, puerto, buque, gas, flota	Pesca y petróleo
32	team, circulo, rayo, luz, fuerza, rw, loma, rosas, kick, central, sarmiento, vela, cantero, chueco, gym	Equipos deportivos
33	viento, temperatura, lluvia, meteorologico, frio, grado, mañana, alerta, fuerte, servicio, maximo, zona, calor, esperar, km	Clima
34	camaron, esteban, perez, navarro, busto, rossi, estrella, ponce, baez, camara, autonomo, celebro, zarate, rocio, amarillo	Nombres
35	gales, jon, whatsapp, gal, asociacion, polo, lewis, david, colono, valle, julio, colectividad, idioma, villagra, inmigrante	Comunidad galesa
36	comercio, precio, producto, supermercado, empleado, compra, venta, local, consumidor, descuento, cliente, cadena, programa	Comercio y precios
37	playa, union, costa, verano, actividad, censo, deporte, balneario, municipal, mar, colonia, temporada, guardavido, enero, parador	Actividades en la playa
38	santo, aire, cruz, mendoza, fe, cordoba, rio, tierra, fuego, negro, pampa, corriente, chaco, tucuman, jujuy	Provincias argentinas
39	vino, cross, maldonado, guemes, bodega, resultar, mañana, salteño, amadeo, mantener, kms, tn, vitivinicola, corral, colega	Vitivinicola
40	evento, feria, entrada, espectaculo, teatro, cultura, artista, cine, local, edicion, disfrutar, cultural, presentar, publico, artesano	Eventos culturales
41	pozo, inclusive, version, peso, linea, caja, gama, faro, obsequio, puerta, bolilla, ofrecer, motor, especial, manual	Productos y obsequios
42	pago, banco, deber, credito, inscripcion, cuota, tarjeta, impuesto, acceder, interesado, beneficio, abonar, requisito, estar, bancario	Bancos y pagos

Continúa en la siguiente página

T	Palabras	Título inferido
43	partido, equipo, boca, copa, river, gol, jugar, estadio, dt, fecha, liga, jugador, tiempo, ganar, delantero	Fútbol argentino
44	estudio, laboratorio, tratamiento, enfermedad, medicamento, virus, cientifico, investigacion, prueba, salud, analisis, oms	Investigación en salud
45	policia, policial, mujer, hombre, joven, encontrar, persona, casa, comisaria, detenido, victima, momento, calle, personal, vivienda	Crímenes y policía
46	informacion, digital, dato, plataforma, sistema, pagina, web, electronico, correo, telefono, usuario, aplicacion, registro, internet	Información digital
47	rn, caravana, avila, silvano, cm, concientizacion, organo, carpa, cancer, cirujano, quevedo, ansv, cendrar, concientizar, corporal	Concientización de salud
48	animal, perro, mascota, veterinario, medio, gato, campaña, castracion, fundacion, cuidado, raza, maltrato, intervencion, zoonosis	Bienestar animal
49	control, seguridad, vial, transito, vehiculo, operativo, infraccion, alcoholemia, policia, personal, agencia, positivo, distinto, conductor, test	Alcoholemias vehiculares
50	millon, dolar, precio, aumento, mercado, incremento, subir, promedio, nivel, inflacion, us, valor, ciento, venta, ultimo	Economía e inflación
51	fiscal, imputado, juicio, prision, juez, audiencia, penal, tribunal, delito, pena, fiscalia, acusado, victima, publico, homicidio	Juicios y delitos
52	justicia, causa, juez, judicial, fiscal, abogado, federal, ex, tribunal, publico, funcionario, denuncia, juzgado, investigar, cargo	Justicia y causas politicas
53	torneo, equipo, club, jugar, femenino, futbol, partido, liga, jugador, certamen, deportivo, sub, masculino, municipal, disputar	Futbol valorado
54	vs, racing, germinal, moreno, zona, cancha, deportivo, fecha, independiente, roca, valle, huracan, brown, fc, dep	Fútbol local
55	peso, mil, premio, sorteo, valor, ronda, ganador, loteria, entregar, juego, bingo, agencia, efectivo, dinero, telebingo	Premios y sorteos
56	rc, concejo, deliberante, bigornia, concejal, patoruzu, goch, draig, cargo, consejo, ferrari, dr, puerto, magistratura, williams	Consejo deliberante
57	empresa, sector, generar, trabajo, inversion, economico, empleo, desarrollo, produccion, productivo, economia, industrial, empresario, industria, proyecto	Desarrollo industrial

Continúa en la siguiente página

T	Palabras	Título inferido
58	rural, localidad, comuna, aniversario, aldea, cushamen, pluma, comunal, pueblo, altar, atilio, tecka, poblador, festejo, dragon	Fiestas y tradicion
69	policia, policial, robo, personal, allanamiento, arma, investigacion, calle, division, elemento, droga, secuestro, barrio, domicilio, detenido	Narcotrafico
60	escuela, educacion, docente, educativo, alumno, clase, estudiante, escolar, nivel, secundario, institucion, chico, aula, ciclo, primario	Educación escolar
61	brasil, mundial, unido, mundo, internacional, americo, chile, chino, colombia, uruguay, mexico, millon, peru, europeo, agencia	Noticias internacionales
62	protocolo, persona, actividad, horario, medida, deber, sanitario, respetar, municipal, cumplir, local, ingreso, social, circulacion, gente	Protocolos sociales
63	noticia, seccion, bajo, recibir, momento, importante, tweet, red, denuncia, social, video, caso, persona, denunciar, foto	Noticias en redes sociales
64	mundial, italiano, italia, partido, ganar, grupo, campeon, jugar, español, titulo, victoria, ranking, torneo, aleman, ronda	Futbol internacional
65	caso, reportar, fallecido, epidemiologico, salud, covid, nexo, confirmado, positivo, contagio, coronavirus, persona, ministerio, puerto, fecha	Epidemiología y virus
66	pelea, boxeo, titulo, combate, kg, campeon, mundial, ko, profesional, boxeador, ganar, amateur, velada, festival, gimnasio	Boxeo
67	color, azul, moto, blanco, casino, negro, super, prenda, rojo, ropa, futsal, moda, estilo, vestido, lanus	Moda y estilo
68	casa, llegar, campo, pequeño, vender, resto, encontrar, masters, ver, noche, llamar, dueño, tierra, quedo, antiguo	Vida rural
69	universidad, proyecto, desarrollo, programa, ciencia, tecnologia, trabajo, capacitacion, objetivo, formacion, profesional, tecnico, social, conocimiento, tecnologico	Educacion profesional
70	seguridad, policia, jefe, ministro, massoni, policial, fuerza, comisario, federico, gomez, unidad, regional, personal, cargo, director	Seguridad en Chubut
71	romero, gutierrez, omar, acevedo, jaramillo, sismo, esquivel, gareca, edison, salomon, borja, caro, magnitud, ram, kilometro	Sismos y terremotos

Continúa en la siguiente página

T	Palabras	Título inferido
72	puesto, do, categoria, er, to, ii, juvenil, championship, negro, green, iii, iv, capon, humphreys, bichito	Ediciones y competencias
73	salud, hospital, atencion, centro, medico, ministerio, area, equipo, sanitario, servicio, trabajo, social, personal, profesional, puratich	Ministerio de salud
74	vacuna, vacunacion, dosis, persona, salud, covid, vacunar, poblacion, campaña, esquema, variante, mayor, aplicar, recibir, sputnik	Vacunas de Covid-19
75	club, futbol, deportivo, institucion, presidente, jugador, deporte, cancha, dirigente, entidad, comision, socio, contrato, directivo, social	Contratos deportivos
76	comision, vocal, presidente, titular, directivo, asociacion, rojo, vicepresidente, yrigoyen, suplente, hugo, filial, moyano, afiliado, asamblea	Gremios y sindicatos
77	consejo, discapacidad, persona, comision, representante, municipal, alquiler, miembro, tema, asamblea, reunion, ordenanza, actividad, arco, accesibilidad	Discapacidad y Accessibilidad
78	servicio, cooperativo, energia, electrico, empresa, publico, usuario, tarifa, gas, cooperativa, costo, luz, factura, lote, subsidio	Tarifas y subsidios
79	puerto, portuario, muelle, madrynense, prefectura, naval, buque, terminal, crucero, embarcacion, golfo, administracion, tripulante, barco, omnibus	Cruceros y Actividad naval
80	edificio, ministerio, general, instalacion, personal, oficina, superior, gobierno, infraestructura, tarea, trabajo, publico, servicio, reparacion, administrativo	Infraestructura
81	competencia, deporte, evento, prueba, kilometro, edicion, carrera, km, distancia, organizacion, deportivo, fecha, inscripcion	Running y competencias locales
82	obra, construccion, trabajo, publico, calle, infraestructura, ejecucion, proyecto, licitacion, sector, metro, ejecutar, acceso, planificacion, avanzar	Proyectos publicos y obras
83	laguna, metro, reserva, cerro, parque, natural, piedra, recorrido, zona, ubicado, sendero, chiquichano, kilometro, cacique	Parques y naturaleza
84	bahia, the, mayo, of, cipolletti, presentar, blanco, mitre, rincon, olimpo, presidente, convocar, ro, serie, sportivo	Deportes Rio negro

Continúa en la siguiente página

T	Palabras	Título inferido
85	libro, musica, artista, banda, arte, obra, historia, cancion, danza, grupo, musico, musical, presentar, disco, voz	Arte y Cultura
86	arcioni, gobierno, millon, ministro, deuda, peso, fondo, economia, economico, financiero, mandatario, nacion, pagar, situacion	Gobierno de Chubut
87	especie, natural, marino, agua, cientifico, investigador, area, zona, conservacion, cambio, ambiente, animal, mar, climatico, estudio	Medio Ambiente
88	ruta, nieve, tramo, vialidad, vial, precaucion, camino, rp, transitable, acceso, calzada, kilometro, zona, hoya, sector	Rutas y vialidad
89	productor, produccion, agricultura, ganaderia, industria, comercio, cavaco, ministerio, producto, emprendedor, zona, emprendimiento, agropecuario, campo	Sector agropecuario
90	categoria, dama, master, caballero, categorio, fecha, pesca, ganador, promocional, pareja, concurso, pto, ta, elite, premio	Deportes de pesca
91	federal, malvino, isla, junio, sur, departamento, atlantico, secretario, mono, fm, puerto, aire, unidad, antartido, pedido	Islas malvidas y Antartida
92	proyecto, diputado, sesion, bloque, ley, legislatura, concejal, frente, ejecutivo, concejo, comision, legislativo, deliberante	Proyectos de ley
93	olimpico, juego, medalla, tokio, atleta, muñoz, metro, oro, maraton, tiempo, deportista, joaquin, coco, arbe, logro	Juego olimpicos y atletismo
94	residuo, ambiental, ambiente, planta, limpieza, recoleccion, urbano, material, vecino, basura, sustentable, plastico, tratamiento, limpio, secretaria	Residuos y gestion ambiental
95	actividad, taller, capacitacion, encuentro, participar, cultural, municipalidad, espacio, centro, secretaria, cargo, cultura, curso	Talleres y capacitaciones
96	producto, alimento, consumo, bebida, carne, dulce, consumir, marca, alcoholico, leche, alcohol, venta, agua, chocolate, recibio	Consumo y productos alimenticios
97	rio, lago, negro, localidad, caleta, pico, sarmiento, zona, mayo, olivio, costa, gallego, gobernador, corcovado, bariloche	Localidades de la region
98	ver, vida, pasar, momento, gente, vivir, salir, tiempo, hablar, familia, hijo, volver, sentir, amigo, empezar	Experiencias de vida

Continúa en la siguiente página

T	Palabras	Título inferido
99	mujer, genero, violencia, derecho, persona, sexual, diversidad, ley, social, situacion, perspectiva, politico, integral, femicidio, publico	Derechos de la Mujer

En la Figura 6.8 se muestra el gráfico *PyLDAvis* del modelo de 100 tópicos. Se destaca como los tópicos similares se juntan en el plano bidimensional y también como los círculos representativos de los tópicos mas relevantes son mas grandes.

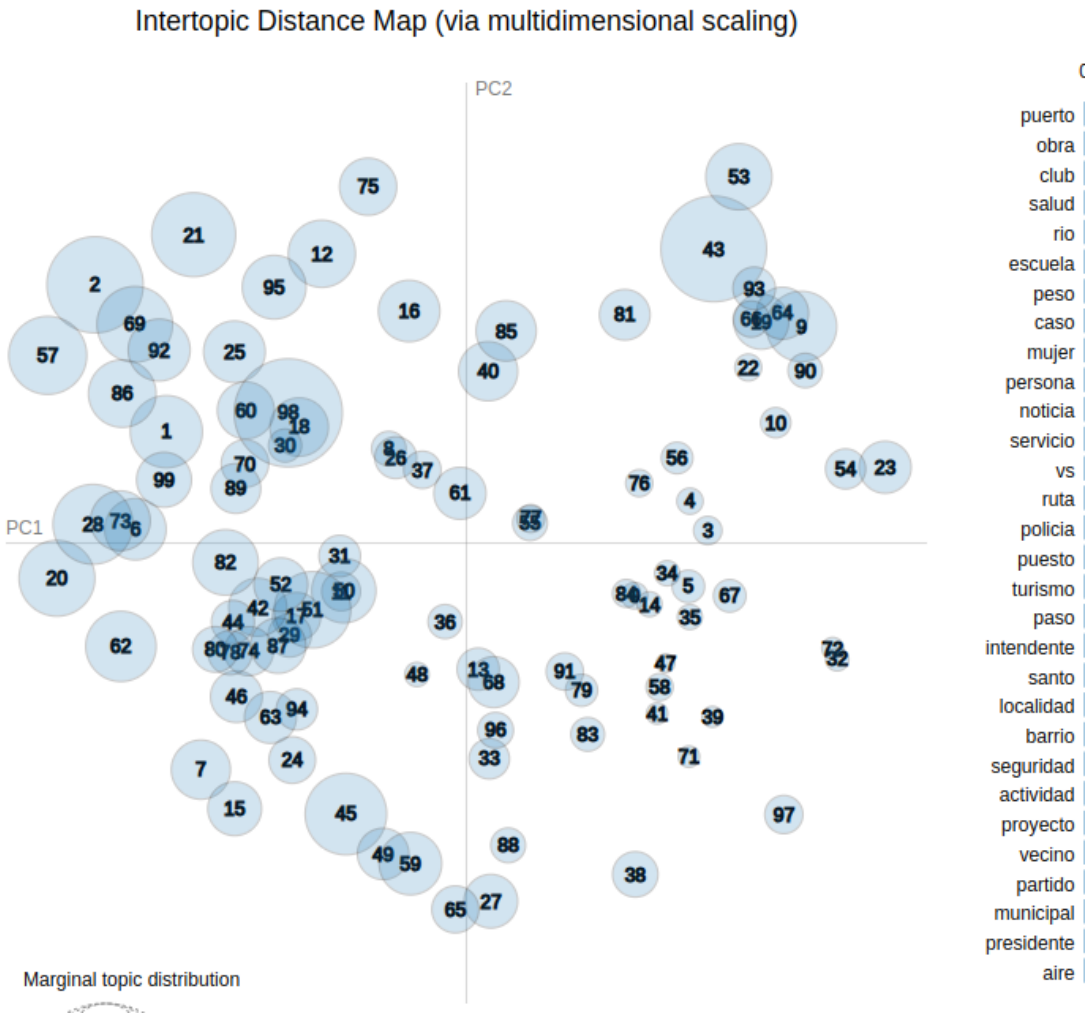

Figura 6.8: Gráfico PyLDAvis 5.5.2 del Modelo LDA de 100 tópicos

En la Figura 6.9 se presenta el gráfico *Pie Chart* 5.5.1 del modelo de 100 tópicos, donde se puede observar el valor exacto de la relevancia de cada tópico.

Figura 6.9: Pie Chart 5.5.1 del Modelo LDA de 100 tópicos

En el siguiente cuadro se muestran las estadísticas del modelo de 100 tópicos. Con las columnas **TOPIC** (numero de tópico), **NEWS** 5 (apariciones totales del tópico), **PRED_NEWS** 7 (cantidad de noticias donde el tópico es predominante), **PRED_RATIO** 8 (proporción de predominancia del tópico) y **AVG_WEIGHT** 3 (peso promedio del tópico).

TOPIC	NEWS 5	PRED_NEWS 7	PRED_RATIO 8	AVG_WEIGHT 3
0	6974	17	0.0024	0.0025
1	12922	1893	0.1465	0.0237
2	21263	767	0.0361	0.0341
3	4583	28	0.0061	0.0026
4	5151	42	0.0082	0.0023
5	5329	60	0.0113	0.0034
6	13964	317	0.0227	0.0146
7	10525	535	0.0508	0.0136
8	7302	68	0.0093	0.0046
9	9875	774	0.0784	0.0147
10	5518	131	0.0237	0.0034

11	8566	110	0.0128	0.0065
12	9798	763	0.0779	0.0144
13	9554	101	0.0106	0.0069
14	5910	8	0.0014	0.0023
15	11369	478	0.0420	0.0135
16	12714	281	0.0221	0.0128
17	7116	221	0.0311	0.0064
18	9997	290	0.0290	0.0112
19	6309	1057	0.1675	0.0117
20	16918	346	0.0205	0.0203
21	19928	181	0.0091	0.0232
22	4175	112	0.0268	0.0024
23	15421	55	0.0036	0.0096
24	8137	354	0.0435	0.0091
25	12149	180	0.0148	0.0131
26	8220	57	0.0069	0.0053
27	10574	1613	0.1525	0.0188
28	18025	117	0.0065	0.0178
29	9799	39	0.0040	0.0058
30	4764	94	0.0197	0.0036
31	6643	75	0.0113	0.0054
32	4556	4	0.0009	0.0016
33	8478	422	0.0498	0.0079
34	4798	52	0.0108	0.0020
35	6292	5	0.0008	0.0023
36	7722	65	0.0084	0.0043
37	8718	56	0.0064	0.0062
38	10718	336	0.0313	0.0069
39	4077	10	0.0025	0.0014
40	11280	587	0.0520	0.0155
41	4510	56	0.0124	0.0018
42	11962	599	0.0501	0.0151

43	12840	2889	0.225	0.0345
44	6303	108	0.0171	0.0049
45	13690	1812	0.1324	0.0291
46	13682	106	0.0077	0.0103
47	3959	3	0.0008	0.0011
48	4731	19	0.0040	0.0022
49	8796	511	0.0581	0.0112
50	11582	809	0.0698	0.0137
51	9350	1042	0.1114	0.0164
52	9354	141	0.0151	0.0087
53	11285	727	0.0644	0.0187
54	6168	346	0.0561	0.0064
55	9039	229	0.0253	0.0055
56	5785	52	0.0090	0.0039
57	14240	233	0.0164	0.0172
58	7552	7	0.0009	0.0030
59	9638	1527	0.1584	0.0205
60	11845	213	0.0180	0.0117
61	12090	20	0.0017	0.0086
62	14580	686	0.0471	0.0199
63	15527	133	0.0086	0.0117
64	8889	536	0.0603	0.0100
65	7843	1235	0.1575	0.0139
66	4878	320	0.0656	0.0046
67	6624	139	0.0210	0.0035
68	12300	13	0.0010	0.0074
69	12737	461	0.0361	0.0168
70	9564	210	0.0219	0.0083
71	4561	4	0.0008	0.0015
72	5329	5	0.0009	0.0016
73	10758	254	0.0236	0.0125
74	5878	485	0.0825	0.0086

75	11377	106	0.0093	0.0122
76	6680	12	0.0018	0.0027
77	6234	78	0.0125	0.0031
78	9648	48	0.0049	0.0072
79	9615	36	0.0037	0.0040
80	10137	191	0.0188	0.0083
81	8719	594	0.0681	0.0119
82	11322	658	0.0581	0.0151
83	8748	52	0.0059	0.0046
84	4286	47	0.0109	0.0022
85	10160	249	0.0245	0.0119
86	12612	342	0.0271	0.0142
87	8133	206	0.0253	0.0078
88	7377	192	0.0260	0.0055
89	7911	209	0.0264	0.0078
90	6415	169	0.0263	0.0052
91	5908	193	0.0326	0.0043
92	10756	324	0.0301	0.0131
93	6531	279	0.0427	0.0070
94	7782	237	0.0304	0.0069
95	13973	323	0.0231	0.0182
96	7261	49	0.0067	0.0042
97	11504	9	0.0007	0.0058
98	24135	810	0.0335	0.0335
99	9670	234	0.0242	0.0087

6.2.4. Análisis y discusión de modelos estáticos

Al comparar los tres modelos estáticos con distintas cantidades de tópicos (10, 40 y 100), se puede ver que a medida que se aumenta la cantidad de tópicos del modelo, los temas se subdividen y se vuelven mas específicos. Los modelos con menor cantidad de tópicos tienden a generalizar y presentar temáticas más amplias y abstractas, mientras que aquellos con mayor cantidad permiten descubrir subtemas más específicos.

Las métricas sugieren que un modelo de 40 tópicos es un equilibrio adecuado entre sobresegmentar los temas y generalizarlos demasiado, lo que podría resultar en la pérdida de información relevante.

Por ejemplo, el tópico inicial ”Salud y pandemia” en el modelo de 10 tópicos se descompone progresivamente en temas más detallados al aumentar la cantidad de tópicos. A continuación, en la Tabla 6.8 se muestra cómo este tópico se subdivide en cada modelo:

Cantidad de Tópicos	**Nº de Tópico**	**Subtemas del tópico "Salud y pandemia"**
10 tópicos (6.2.1)	1	Salud y pandemia
40 tópicos (6.2.2)	5	Medidas sanitarias y pandemia
	23	Casos de COVID y salud pública
	26	Vacunas y distribución global
	31	Atención médica y hospitales
100 tópicos (6.2.3)	20	Pandemia y medidas sanitarias
	29	Salud infantil
	44	Investigación en salud
	47	Concientización de salud
	63	Protocolos sociales
	65	Epidemiología y virus
	73	Ministerio de salud
	74	Vacunas de COVID-19

Cuadro 6.8: Subtemas del tópico ”Salud y pandemia” en diferentes modelos estáticos

La inferencia de títulos apoyada por *Inteligencia Artificial* 3.1 ayuda a etiquetar los conjuntos de palabras que representan cada subtema. Esto contribuye a una interpretación más clara de la estructura temática a medida que los modelos se vuelven más específicos.

Ademas, las herramientas desarrolladas en esta tesina para el *Análisis estático de los tópicos* 5.5 contribuyen a entender mejor cada modelo y como se representa el corpus de noticias de Chubut con los parámetros seteados, inclusive las relaciones entre cada uno de los temas.

6.3. Modelos Dinámicos

En esta sección se presenta el análisis de la evolución de los tópicos a lo largo del tiempo, con el fin de visualizar de manera efectiva cómo cambiaron las temáticas principales a medida que las noticias evolucionaban. Para ello, se detallan dos experimentos: el primero se basa en la metodología de los *Modelos de Tópicos Dinámicos* (**DTM**) 3.7 propuesta por Blei [14], mientras que el segundo emplea las herramientas de visualización dinámica desarrolladas en esta tesina.

El tópico ”Salud y Pandemia” fue seleccionado para el análisis debido a su notable relevancia e impacto en las noticias de la región entre 2019 y 2022, periodo en el que la pandemia de COVID-19 fue un tema central en el discurso público y mediático.

6.3.1. Experimento Blei

Para el análisis dinámico de los tópicos, basado en los términos más representativos de cada uno, Blei propone los *Dynamic Topic Models* (**DTM**) 3.7. A través de esta metodología se puede visualizar la evolución de cada tópico y cómo sus palabras más relevantes aumentan o disminuyen su relevancia dentro del tópico. A través de los DTM se puede observar el cambio en cada temática individualmente.

Para este experimento, se utilizó la biblioteca `gensim` y su implementación del modelo `LdaSeqModel` para aplicar los DTM. Se seleccionó el tópico número 1 del modelo de 10 tópicos (ver figura 6.2), inferido como ”Salud y pandemia”. El análisis se realizó utilizando un *time slice* (ventana temporal) de 1 año.

En la tabla 6.9 se muestra una comparación de los términos más relevantes de este tópico durante cada año. Se puede observar que términos como ”persona”, ”hospital” y ”vacuna” mantienen una relevancia constante a lo largo de los años, reflejando su importancia en el discurso sobre la pandemia. Sin embargo, otros términos surgen y desaparecen en función de las distintas etapas de la pandemia. Por ejemplo, en 2020 se destacan términos como ”covid”, ”coronavirus” y ”aislamiento”, vinculados al inicio de la crisis sanitaria y las medidas de confinamiento. En 2021 y 2022, términos relacionados con las ”vacunas” y ”dosis” ganan protagonismo, mientras que expresiones como ”pandemia” y ”aislamiento” disminuyen, señalando el avance de las campañas de vacunación y la transición hacia un nuevo enfoque de la salud pública. En contraste, 2019 refleja una etapa previa a la pandemia.

2019	2020	2021	2022
hospital	persona	persona	persona
persona	covid	vacuna	enfermedad
paciente	coronavirus	covid	covid
enfermedad	medida	dosis	hospital
medico	aislamiento	vacunacion	sanitario
tratamiento	hospital	coronavirus	vacunacion
ministerio	social	pandemia	vacuna
sanitario	situacion	actividad	medico
vuelo	pandemia	hospital	dosis
ministro	sanitario	paciente	paciente
importante	paciente	sanitario	contagio
area	ministerio	medico	indicar
centro	cumplir	medida	pandemia
explicar	deber	enfermedad	ministerio
alto	area	ministerio	medida
sistema	positivo	positivo	alto
vacuna	personal	reportar	niño
personal	obligatorio	deber	trabajar
indicar	ministro	gobierno	centro
trabajar	gobierno	fallecido	area

Cuadro 6.9: Tabla de términos mas relevantes por año

6.3.2. Experimento Markel

En este experimento de análisis dinámico, se utilizó la herramienta desarrollada en esta tesina denominada *Topic Evolution Chart* 5.6.2, aplicada al tópico número 1 del modelo de 10 tópicos, inferido como ”Salud y pandemia” (ver figura 6.2).

El gráfico de la figura 6.10 ilustra la evolución de la relevancia del tópico a lo largo del tiempo. Se observa un aumento significativo en la relevancia del tópico a partir de marzo de 2020, indicando una gran preocupación por la pandemia inminente, seguido de un descenso medido y picos de aumento en abril de 2021 y enero de 2022.

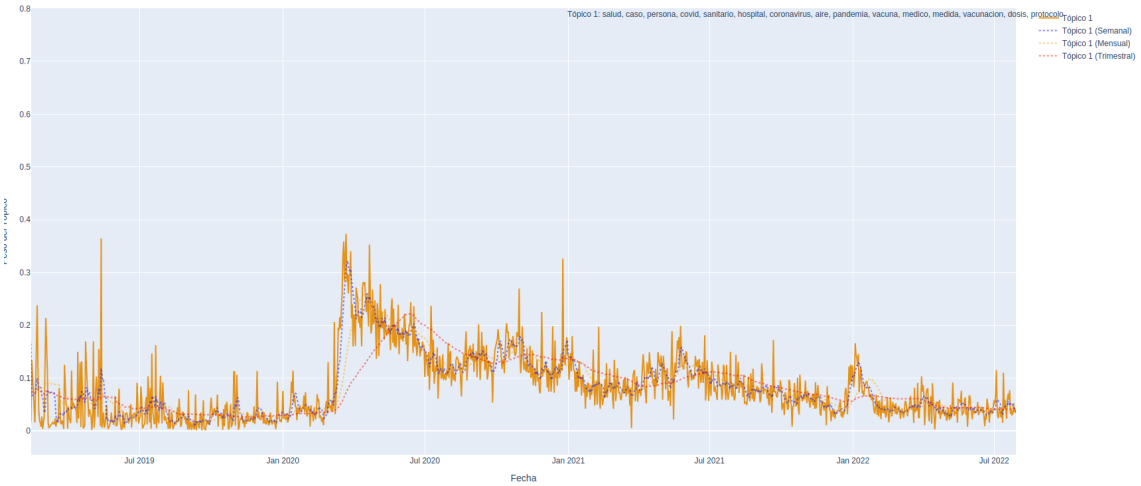

Figura 6.10: Gráfico de relevancia del tópico a través del tiempo

Ademas, se puede consultar la figura 6.11, la cual muestra el gráfico *Stacked Bar Chart* 5.6.1 del modelo de 10 tópicos completo. Este gráfico proporciona una visión integral del cambio en la relevancia relativa mensual de todos los tópicos analizados.

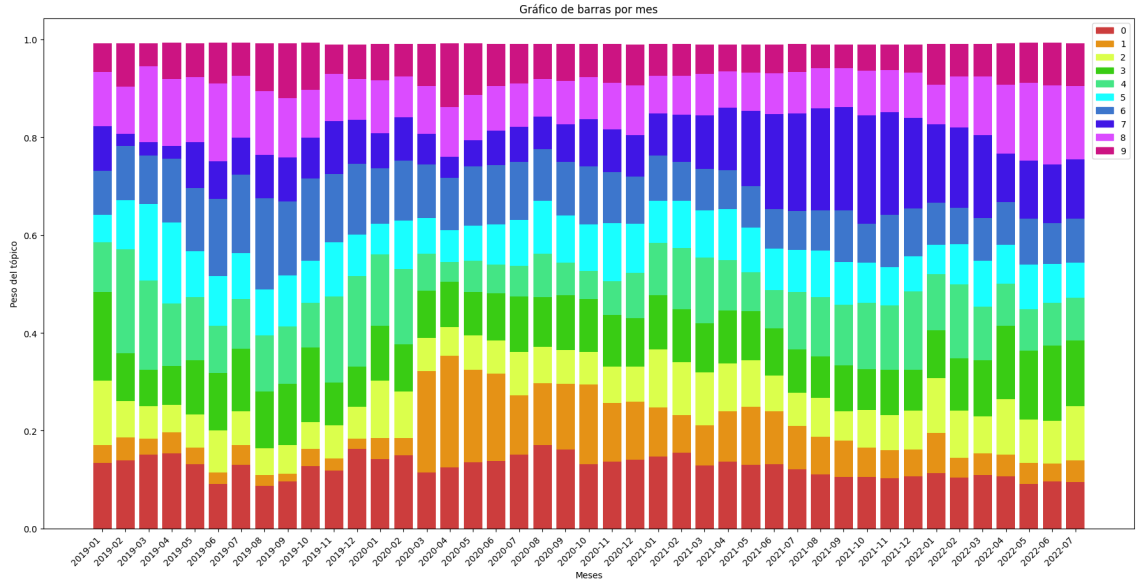

Figura 6.11: Stacked Bar Chart 5.6.1 del Modelo LDA de 10 tópicos

6.3.3. Análisis y discusión de modelos dinámicos

Los análisis de los dos experimentos se complementan entre sí, proporcionando una visión más completa sobre la dinámica de los tópicos. El experimento de Blei permite observar cómo, a través de los DTM 3.7, se pueden identificar los cambios en cada temática individualmente, destacando la evolución de términos clave a lo largo de los años.

Por otro lado, el uso de herramientas visuales desarrolladas en la tesina, como el *Topic Evolution Chart* 5.6.2 y el *Stacked Bar Chart* 5.6.1, permite representar de manera clara y efectiva la variación en la relevancia de los tópicos en cada lapso de tiempo. Estas visualizaciones enriquecen el análisis, facilitando la identificación de tendencias y patrones, así como temáticas virales en la evolución de los tópicos de las noticias recolectadas.

En conjunto, estos modelos dinámicos ofrecen un enfoque robusto para analizar fenómenos complejos y en constante cambio, comprender mejor cómo los discursos y preocupaciones sociales evolucionan con el tiempo.

CAPÍTULO 7

CONCLUSIONES

El análisis de temas mediante *Latent Dirichlet Allocation* (**LDA**) 3.6 aplicado al extenso corpus de noticias de Chubut entre 2019 y 2022 5.2 demostró ser una herramienta poderosa para determinar y clasificar los principales tópicos presentes en el contenido noticioso. El análisis identificó de manera efectiva los temas más relevantes presentes en el corpus, incluso permitiendo saber de entre los temas, cuáles fueron los más presentes y predominantes en las noticias analizadas.

Las técnicas de análisis dinámico de tópicos 3.7, propuestas por Blei, brindaron una visión detallada sobre la evolución de los términos asociados a cada tópico a lo largo del tiempo. Además, los gráficos desarrollados en esta investigación, como el *Stacked Bar Chart* 5.6.1 y el *Topic Evolution Chart* 5.6.2, facilitaron la visualización clara de la relevancia y fluctuación de los tópicos a lo largo del tiempo, permitiendo observar cómo los medios incrementaron o disminuyeron la frecuencia de publicación de noticias relacionadas con ciertos temas en respuesta a las preocupaciones y prioridades cambiantes de la sociedad.

Los resultados son muy positivos y podrían ayudar como apoyo en la toma de decisiones y políticas públicas; podrían ofrecer insights valiosos sobre las áreas de interés y preocupación. Además, la segmentación de noticias a través de tópicos puede ser útil en el filtrado de noticias de los portales de comunicación, automatizando la selección y difusión de noticias relevantes según el interés del público. Esto mejoraría tanto a priorizar el contenido de mayor importancia como a mejorar la experiencia del usuario.

7.1. Próximos pasos

1. **Análisis de sentimiento**: El **análisis de sentimiento** es una técnica que permite identificar y clasificar las emociones o actitudes presentes en un texto, asignándoles categorías como positiva, negativa o neutra. Este proceso se aplica comúnmente en grandes volúmenes de datos textuales, como noticias, reseñas o publicaciones en redes sociales. Esta herramienta puede ayudar a identificar cómo los periodistas o medios abordan ciertos temas, revelando su postura o intención comunicativa. Además, permite evaluar el impacto emocional que se intenta generar en los lectores, detectando patrones de sesgo o inclinación y proporcionando una visión más profunda sobre la influencia de los medios en la opinión pública.

2. **Modelado completo únicamente con DTM**: Durante el proceso de modelado de tópicos, se utilizó *Latent Dirichlet Allocation* (**LDA**) 3.6 para agrupar y diferenciar las noticias. Posteriormente, se implementó *Dynamic Topic Models* (**DTM**) 3.7 de manera independiente para cada tópico, con el fin de analizar la evolución de los mismos a lo largo del tiempo. Debido a limitaciones computacionales, no fue posible aplicar DTM sobre el corpus completo de noticias de Chubut. Realizar un experimento utilizando únicamente DTM en todo el corpus y ademas variando los *time slice* (el experimento de la tesina fue con un *time slice* anual) podría ofrecer una perspectiva más global sobre la evolución temporal de los temas, permitiendo identificar subdivisiones más detalladas y descubrir dinámicas distintas de las obtenidas con LDA en esta tesis. Este enfoque podría arrojar nuevos resultados y enriquecer el análisis temático presentado.

3. **Variaciones de corpus**: Debido a que el sistema desarrollado recibe un CSV estructurado con noticias. Obtener resultados de modelos con noticias de otras regiones u otro dataset de noticias mas amplio también modificando los medios de comunicaciones de donde se obtienen podría arrojar resultados diferentes y enriquecedores para analizar.

4. **Mejora y búsqueda de otros tipos de gráficos**: Muchos de los gráficos presentados en esta tesis son útiles para modelos LDA con una cantidad reducida de tópicos, ya que permiten visualizar claramente las relaciones y la evolución de estos temas. Sin embargo, a medida que se aumenta la cantidad de tópicos entrenados, los gráficos tienden a volverse confusos e ilegibles debido a la superposición de información y la complejidad creciente del modelo. Un futuro paso importante sería explorar otras formas de visualización que no dependan tanto de la cantidad de tópicos. Esto podría incluir el uso de gráficos interactivos o el desarrollo de representaciones más especializadas, que permitan analizar la evolución de múltiples tópicos de manera clara y accesible.

Bibliografía

[1] Pooja Kherwa y Poonam Bansal. «Latent Semantic Analysis: An Approach to Understand Semantic of Text». En: *2017 International Conference on Current Trends in Computer, Electrical, Electronics and Communication (CTCEEC)*. 2017, págs. 870-874. DOI: 10.1109/CTCEEC.2017.8455018.

[2] Tian Shi et al. «Short-Text Topic Modeling via Non-negative Matrix Factorization Enriched with Local Word-Context Correlations». En: *Proceedings of the 2018 World Wide Web Conference*. WWW '18. Lyon, France: International World Wide Web Conferences Steering Committee, 2018, págs. 1105-1114. ISBN: 9781450356398. DOI: 10.1145/3178876.3186009.

[3] Xiaohui Yan et al. «A biterm topic model for short texts». En: *Proceedings of the 22nd International Conference on World Wide Web*. WWW '13. Rio de Janeiro, Brazil: Association for Computing Machinery, 2013, págs. 1445-1456. ISBN: 9781450320351. DOI: 10.1145/2488388.2488514.

[4] Stuart J. Russell y Peter Norvig. *Inteligencia artificial: Un enfoque moderno*. Accedido: 2024-09-05. Pearson, 2017.

[5] Sunila Gollapudi. *Practical Machine Learning*. Accedido: 2024-09-05. Apress, 2016.

[6] IBM. *Aprendizaje no supervisado*. Accedido: 2024-09-05. 2024.

[7] Antonio Moreno. *¿Qué es el Procesamiento de Lenguaje Natural?* Accedido: 2024-09-05. 2024.

[8] David M. Blei. *Probabilistic Topic Models*. Vol. 55 | No. 4 Accedido: 2024-09-05. Abr. de 2012.

[9] David M. Blei, Andrew Y. Ng y Michael I. Jordan. «Latent Dirichlet Allocation». En: *Journal of Machine Learning Research* 3 (2003). Accedido: 2024-09-05, págs. 993-1022.

[10] Michael Röder, Andreas Both y Alexander Hinneburg. «Exploring the Space of Topic Coherence Measures». En: *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*. WSDM '15. Accedido: 2024-10-16. Shanghai, China: Association for Computing Machinery, 2015, págs. 399-408. ISBN: 9781450333177. DOI: 10.1145/2684822.2685324.

[11] Frank Rosner et al. *Evaluating topic coherence measures*. Accedido: 2024-10-16. 2014. arXiv: 1403.6397 [cs.LG].

[12] Gerlof J. Bouma. «Normalized (pointwise) mutual information in collocation extraction». En: *From Form to Meaning: Processing Texts Automatically, Proceedings of the Biennial GSCL Conference 2009*. Accedido: 2024-10-16. 2009.

[13] Thomas L. Griffiths y Mark Steyvers. «Finding scientific topics». En: *Proceedings of the National Academy of Sciences* 101.suppl_1 (2004). Accedido: 2024-10-21, págs. 5228-5235. DOI: 10.1073/pnas.0307752101.

[14] David M. Blei y John D. Lafferty. «Dynamic Topic Models». En: *Proceedings of the 23rd International Conference on Machine Learning (ICML)* (2006). Accedido: 2024-10-16, págs. 113-120.

[15] Wikipedia. *Limpieza de datos — Wikipedia, La enciclopedia libre*. [Internet; descargado 4-agosto-2024]. 2024.

[16] Sitelabs. *Web Scraping: Introducción y Herramientas*. Accedido: 2024-09-05. 2017.

[17] Usama Fayyad, Gregory Piatetsky-Shapiro y Padhraic Smyth. «From Data Mining to Knowledge Discovery in Databases». En: *AI Magazine* 17.3 (mar. de 1996). Accedido: 2024-10-16, pág. 37.

[18] Ken Schwaber. *Agile Project Management with Scrum*. Accedido: 2024-10-16. Redmond, WA: Microsoft Press, 2004. ISBN: 9780735619937.

[19] Carlos Emanuel Balcazar. «Extracción, Análisis y Procesamiento Automático de Información Periodística relacionada al COVID-19 en Chubut». Accedido: 2024-10-16. Tesina de grado. Puerto Madryn, Argentina: Universidad Nacional de la Patagonia San Juan Bosco Sede Puerto Madryn, En proceso.

[20] Paul Jansen. *TIOBE Index for October 2024*. Accedido: 2024-10-09. 2024.

[21] *12 Best Programming Languages for Data Science and Analytics*. Accedido: 2024-10-09. Ene. de 2024.

[22] Jonathan Chang et al. «Reading Tea Leaves: How Humans Interpret Topic Models». En: *Neural Information Processing Systems*. Vol. 32. Accedido: 2024-10-16. Ene. de 2009, págs. 288-296.

[23] Paulette Vázquez et al. «Temporal topics in online news articles: Migration crisis in Venezuela». En: *2020 Seventh International Conference on eDemocracy eGovernment (ICEDEG)*. Accedido: 2024-10-16. 2020, págs. 106-113. DOI: 10.1109/ICEDEG48599.2020.9096804.

[24] Carson Sievert y Kenneth Shirley. «LDAvis: A method for visualizing and interpreting topics». En: *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*. Accedido: 2024-10-28. 2014, págs. 63-70.

[25] Christopher D. Manning, Prabhakar Raghavan e Hinrich Schütze. *Introduction to Information Retrieval*. Accedido: 2024-10-16. Cambridge University Press, 2008, págs. 121-125.

ANEXOS

En esta sección se incluye información complementaria que, aunque no forma parte del cuerpo principal, proporciona detalles adicionales relevantes para una mejor comprensión del trabajo realizado.

A. Comparación de coherencia entre distintos corpus

Es interesante analizar la coherencia `CV` del mismo modelo manteniendo exactamente los mismos parámetros, pero variando el *dataset* de entrada. De este modo, es posible comparar cómo cambia el rendimiento del modelo según el corpus utilizado: noticias sin procesar, noticias preprocesadas, y las noticias con preprocesamiento, lista negra (*blacklist*) y el filtrado de extremos explicado en esta tesina.

Para el ejemplo se entrenó un modelo con 10 tópicos utilizando el 100 % de las noticias (Ver cuadro de comparación 1).

Corpus	Coherencia CV	Tiempo de ejecución
Noticias crudas	0.426 `CV`	550.43 Segundos
Noticias preprocesadas	0.586 `CV`	451.05 Segundos
Noticias preprocesadas, con blacklist y filtrado de extremos	0.629 `CV`	383.45 Segundos

Cuadro 1: Comparación de corpus y su impacto en la coherencia y tiempo de ejecución

Se observa que las técnicas utilizadas de preprocesamiento de texto 5.3, *blacklist* 5.4.2.1 y filtrado de extremos 5.4.2.2 mejoran notablemente la coherencia del modelo, además de optimizar significativamente el tiempo de ejecución.

B. Procedimiento

En esta sección, se vuelcan todos los anexos referentes al desarrollo del proyecto de tesina

B.I. Carpeta de resultados

Se presenta la estructura general del resultado de ejecución de un experimento (ver Figura 1).

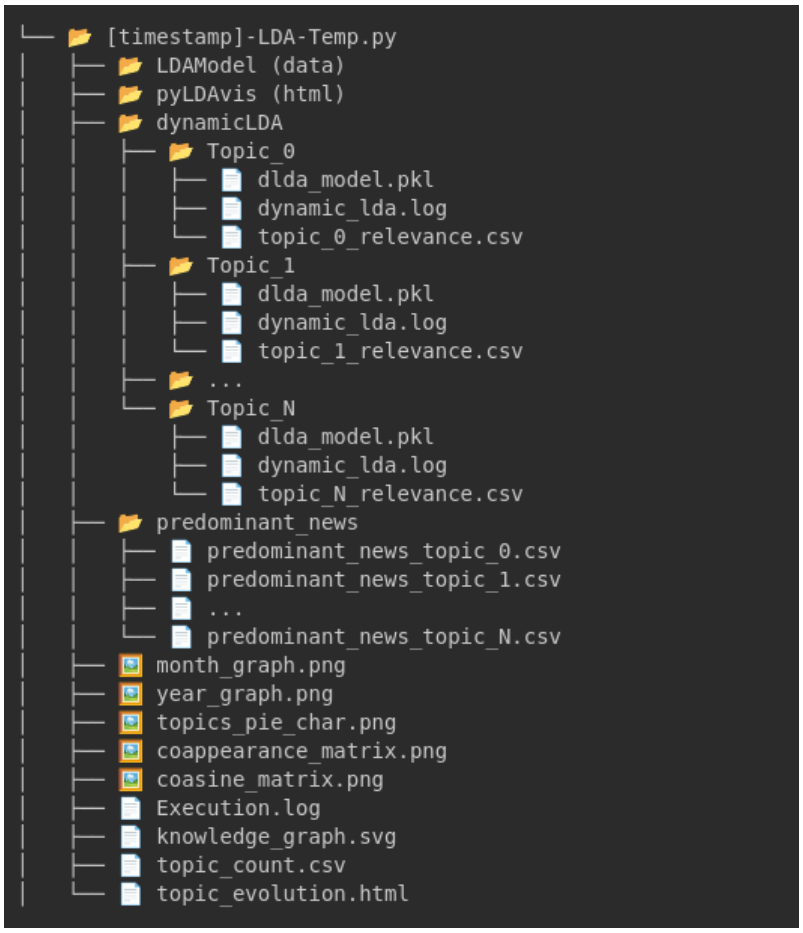

Figura 1: Carpeta de resultados: Estructura general

1. **LDAModel** [carpeta]: Modelo LDA resultante de la ejecución 5.4.
2. **pyLDAvis** [carpeta]: Contiene la información y el archivo HTML de la visualización interactiva de pyLDAvis 5.5.2 del modelo.
3. **dynamicLDA** [carpeta]: Directorios con los modelos dinámicos separados por tópico.
    - **Topic_0**, **Topic_1**, ..., **Topic_N** [carpetas]: Cada una contiene los archivos relacionados con el modelo dinámico para cada tópico específico.
        - **dlda_model.pkl**: Archivo binario del modelo dinámico del tópico.

- **dynamic_lda.log**: Registro de logs específicos del modelo dinámico.
- **topic_X_relevance.csv**: Archivo CSV con los términos más relevantes de cada tópico segmentado por *time slice*.

4. **predominant_news** [carpeta]: Almacena archivos CSV con las noticias donde el tópico es mas relevante 7.
5. **month_graph.png**: Stacked Bar Chart con la relevancia mensual de tópicos 5.6.1.
6. **year_graph.png**: Stacked Bar Chart con la relevancia anual de tópicos 5.6.1..
7. **topic_pie_chart.png**: Gráfico circular que representa la proporción de cada tópico 5.5.1.
8. **coasine_matrix.png**: Matriz de similitud del coseno entre los tópicos del modelo 5.5.4.1.
9. **coappearance_matrix.png**: Matriz de coaparición entre los tópicos 5.5.4.2.
10. **Execution.log**: Registro completo de la ejecución del experimento.
11. **knowledge_graph.svg**: Grafo con la relación entre palabras y tópicos del modelo 5.5.3.
12. **topic_count.csv**: Archivo CSV con las estadísticas de cada tópico.
13. **topic_evolution.html**: HTML con la evolución temporal de los tópicos 5.6.2.

B.II. Tablero Trello

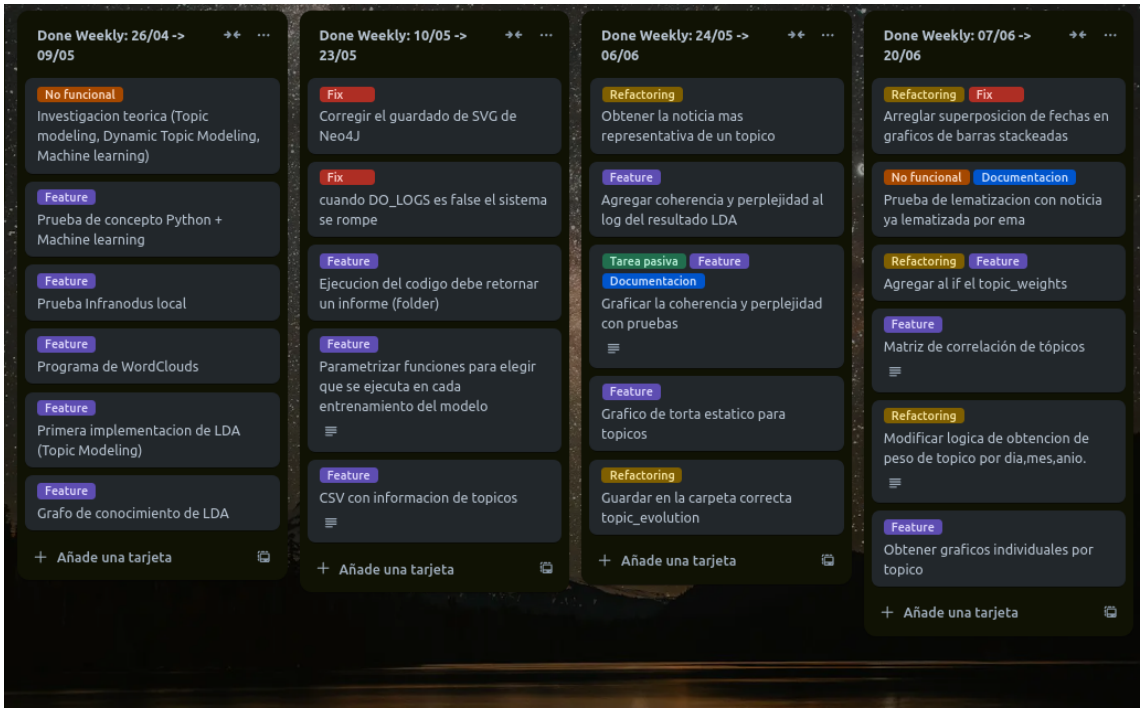

Figura 2: Tablero Trello: Tarjetas realizadas las primeras 4 semanas

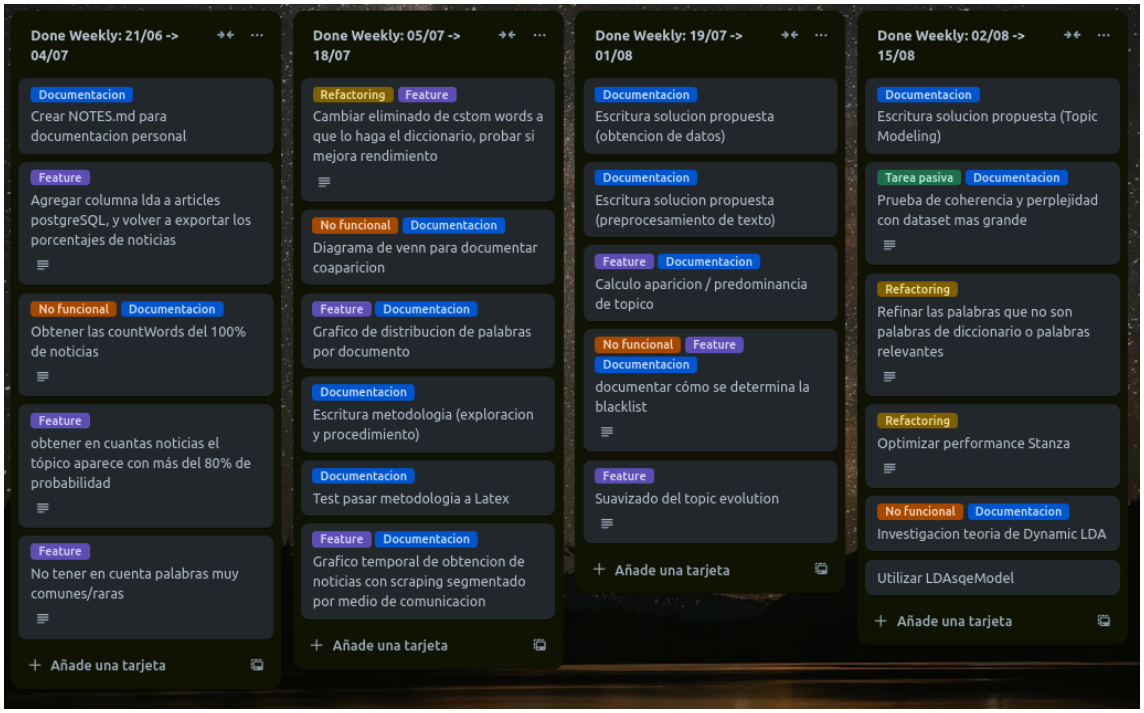

Figura 3: Tablero Trello: Tarjetas realizadas las segundas 4 semanas

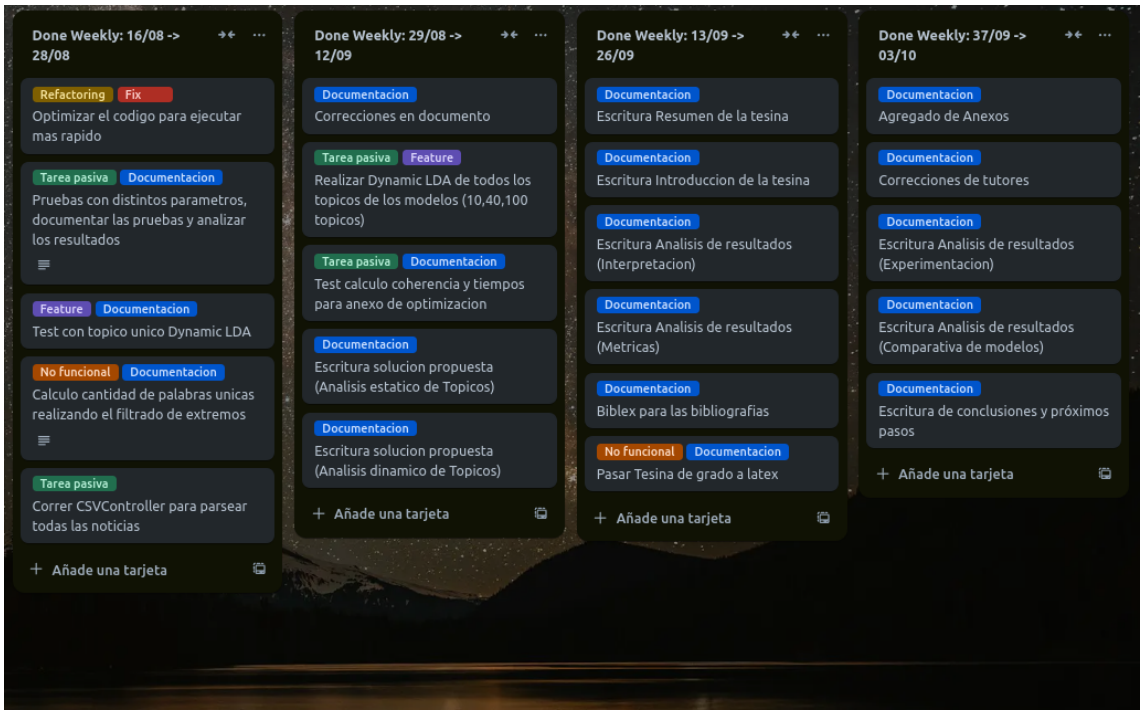

Figura 4: Tablero Trello: Tarjetas realizadas las terceras 4 semanas