Escuela Superior
de Ingeniería y Tecnología
Universidad de La Laguna

Trabajo de Fin de Grado

Comprendiendo a los Parlamentos usando Inteligencia Artificial

Understanding Parliaments using Artificial Intelligence

Carlos Pío Reyes

La Laguna, 10 de julio de 2024

Dña. **Maria Elena Sanchez Nielsen**, profesora Titular de Universidad adscrita al Departamento de Ingeniería Informática y de Sistemas de la Universidad de La Laguna, como tutora

C E R T I F I C A (N)

Que la presente memoria titulada:

”*Comprendiendo a los Parlamentos usando Inteligencia Artificial*”

ha sido realizada bajo su dirección por D. **Carlos Pío Reyes**.

Y para que así conste, en cumplimiento de la legislación vigente y a los efectos oportunos firman la presente en La Laguna a 10 de julio de 2024

Agradecimientos

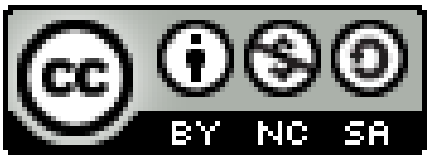
En primer lugar, quisiera agradecerle a mi tutora la profesora María Elena Sánchez Nielsen por ayudarme, aconsejarme y guiarme en todo momento durante la realización del proyecto hasta lograr la realización actual del mismo, ha sido en mi más sincera opinión una tutora ejemplar.

Agradecer a mi familia el apoyo y los ánimos que me han dado durante toda esta etapa de mi vida.

Quiero también agradecer a los amigos que he hecho durante la carrera y a los que han llegado en estos últimos dos años por haber sido un gran apoyo, aquellos que me han ayudado con los obstáculos que me han aparecido y que se han alegrado conmigo por los logros obtenidos.

Y por último quiero agradecer, especialmente, a mi pareja por haber sido mi mayor apoyo y un pilar fundamental para mi en todo este tiempo.

Licencia



Resumen

Tradicionalmente, el trabajo realizado en los Parlamentos a lo largo de su historia abarca un elevado volumen de información que no es fácilmente comprensible. La cantidad ingente de documentos legislativos producidos en cada legislatura hace difícil comprender de forma rápida y eficiente los temas tratados por cada partido político.

El objetivo de este Trabajo Fin de Grado es desarrollar un sistema software que permita visualizar de forma atractiva los temas/tópicos más propuestos por cada partido político, permitiendo comprender de manera temporal cuales han sido los temas que cada partido político ha considerado más importantes.

Se utilizará el Parlamento de Canarias, como caso de uso.

Palabras clave: Aprendizaje automático no supervisado, Iniciativa parlamentaria, Temas, Análisis

Abstract

Traditionally, the work carried out in parliaments throughout their history includes a large volume of information that is not easily understood. a large volume of information that is not easily comprehensible. The sheer volume The sheer volume of legislative documents produced in each legislature makes it difficult to understand the issues dealt with the issues dealt with by each political party quickly and efficiently. The objective of this Final Degree Thesis is to develop a software system that allows to to visualize in an attractive way the topics/topics most proposed by each political party, allowing to understand in a temporal way which have been the topics that each political party has considered most important. political party has considered most important. The Parliament of the Canary Islands will be used as a use case.

Keywords: Unsupervised machine learning, Parliamentary iniciative, Topics, Analysis

Índice general

Índice de Figuras

Índice de Tablas

Capítulo 1

Introducción

En la actualidad la inteligencia artificial es una herramienta en gran auge usada para resolver tareas complejas en menos tiempo del que le llevaría a un humano. En este campo hay una rama que está teniendo especial relevancia: el aprendizaje automático. Esta es una subdisciplina de la inteligencia artificial que se enfoca en la habilidad de las máquinas para aprender de los datos y mejorar su rendimiento de manera autónoma a lo largo del tiempo, sin necesidad de ser programadas para cada tarea específica. Utiliza algoritmos y modelos estadísticos para permitir que los sistemas informáticos realicen tareas basándose en patrones y deducciones.

Una de las características más fascinantes del aprendizaje automático es su capacidad para adaptarse y mejorar a medida que se expone a más datos. Esto lo hace invaluable en campos como la medicina, donde puede ayudar a predecir enfermedades, o en el sector financiero, para detectar fraudes.

El aprendizaje automático se divide en varias categorías: el aprendizaje supervisado, no supervisado, semi-supervisado y por refuerzo, cada uno con sus propias técnicas y aplicaciones específicas. A medida que esta tecnología avanza, se espera que desempeñe un papel aún más importante en nuestra vida cotidiana, haciendo que estos sistemas sean más inteligentes y eficientes.

1.1. Contexto

A lo largo de la historia, los Parlamentos han sido centros neurálgicos de actividad legislativa, acumulando una vasta cantidad de información que, debido a su volumen y complejidad, resulta difícil de procesar y comprender. La producción masiva de documentos en cada legislatura presenta un desafío significativo para analizar de manera rápida y eficiente las prioridades y temas abordados por los distintos partidos políticos, consiguiendo así que una gran cantidad de los datos legislativos puede llevar a que temas cruciales pasen desapercibidos o que se subestimen tendencias importantes en la agenda política.

En este Trabajo Fin de Grado se propone abordar esta problemática mediante el desarrollo de un sistema de software que estará diseñado para ofrecer una visualización atractiva y comprensible de los tópicos más recurrentes propuestos por cada partido político en el Parlamento de Canarias. El objetivo es facilitar una comprensión temporal de los asuntos que han sido considerados de mayor importancia por cada grupo político, proporcionando una herramienta que mejore la transparencia y el acceso a la información legislativa. Permitiendo así identificar rápidamente los temas predominantes de cada legislatura, ayudando a ciudadanos, investigadores y periodistas entender mejor las dinámicas políticas y las prioridades de estos, contribuyendo a un proceso legislativo más transparente y accesible para todos.

1.2. Objetivos

El presente trabajo pretende hacer frente a un caso real basado en automatizar el análisis de las propuestas parlamentarias del Parlamento de Canarias mediante el análisis de sus extractos. Estas iniciativas serán clasificadas usando dos modelos de aprendizaje automáticos, utilizando el vocabulario de cada iniciativa para indicar el tema que refleja.

Una vez finalizado el desarrollo del software, es importante realizar una evaluación de los resultados obtenidos por ambos modelos para verificar su correcto funcionamiento y asegurar que los temas se creen correctamente.

1.3. Organización

Para realizar el trabajo de manera adecuada es necesario identificar cada punto a desarrollar siguiendo el siguiente orden:

- **Estado del arte:** Información y situación actual sobre el tema que se trata.
- **Análisis de requisitos:** Estudio sobre las necesidades más relevantes del proyecto y las posibles soluciones para satisfacerlas.
- **Desarrollo de software:** Diseño y codificación del software con ayuda de las librerías requeridas para satisfacer los objetivos.
- **Evaluación de herramientas de algoritmos:** Estudio de los algoritmos para especificar cual obtiene resultados más fiables.
- **Memoria del proyecto:** Realización de la memoria del proyecto.

Capítulo 2

Estado del arte

En el capítulo anterior se ha hablado sobre el ámbito del proyecto y sus principales objetivos. En este capítulo se procederá a estudiar la situación acerca de la inteligencia artificial, el aprendizaje automático, algunos de sus usos más recientes e información relevante acerca de esta rama.

2.1. Situación actual

La inteligencia artificial es una tecnología que está cada vez más presente en nuestro día a día, generando un gran revuelo por las implicaciones que esta teniendo en todos los ámbitos y todas las herramientas que se han creado en estos días.

Muchos de estos sistemas software son capaces de hacer tareas de manera infinitamente más eficiente que los humanos, y están llevando a cabo actividades que hasta hace poco se creían que eran imposibles de realizar por máquinas. Un ejemplo destacado que ha generado mucha discusión es la creación de arte, lo que ha suscitado un intenso debate sobre todas las cuestiones tanto éticas como morales que presentan el uso de la inteligencia artificial en este tema.

Estos avances provocan que la opinión pública se divida en dos posturas ante la inteligencia artificial. Por un lado ha generado un gran temor por la posibilidad de que se puedan sustituir a personas en sus respectivos trabajos y cause otros problemas sociales más complejos[1]. En el otro lado, existe una postura más optimista que cree que la inteligencia artificial acelerará el progreso en muchos campos, ayudando también a mejorar la eficiencia en varios sectores y mejorar la calidad de vida de los seres humanos.

El desarrollo de estas capacidades de la inteligencia artificial es fruto del aprendizaje automático, el cual permite aprender a las máquinas de manera autónoma a partir de un conjunto de datos inicial, reconocer patrones y, con suficiente entrenamiento, predecir patrones futuros.

En este sentido, el aprendizaje automático se aplica en varias tareas de nuestra vida cotidiana como los siguientes ejemplos[2]:

- **Automatización de procesos:** El aprendizaje automático juega un papel crucial en la automatización y optimización de procesos en diversas industrias como la fabricación, logística y atención al cliente, permitiendo a las empresas mejorar la eficiencia y reducir costos.
- **Medicina y salud:** el aprendizaje automático ha transformado la medicina, permitiendo diagnósticos precisos y tratamientos personalizados.
- **Innovación en transporte:** El aprendizaje automático ha revolucionado el transporte, con vehículos autónomos y sistemas de tráfico inteligentes que mejoran la eficiencia y reducen la congestión.
- **Educación y formación:** El aprendizaje automático ha mostrado su potencial en educación. Los tutores inteligentes personalizan el aprendizaje, las herramientas de análisis de texto asisten la evaluación y las plataformas en línea democratizan el acceso a la educación a nivel global.
- **Asistentes virtuales:** Con el lanzamiento de varias inteligencias artificiales generativas, estas se integrarán comúnmente en el trabajo, permitiendo escribir, investigar y resumir correos electrónicos, cambiando la vida laboral.

Además de estos ejemplos, en el caso más concreto del uso del *machine learning* en los parlamentos, hay estudios que exploran la implementación de esta tecnología para realizar diversas tareas con el objetivo de mejorar la eficiencia y transparencia.

Por ejemplo, un artículo analiza esta posibilidad para el parlamento de Malasia[3] donde la investigación reveló que la implementación de Inteligencia artificial en este parlamento podría mejorar la eficiencia y la productividad. plantea desafíos éticos y de responsabilidad que aún necesitan resolverse, y sugiere que futuros estudios deberían enfocarse en la precisión de las respuestas y en el análisis de contenido.

O también otro estudio de cómo se podría hacer esta tarea usando cómo caso de uso el parlamento helénico[4], un estudio donde se recopilan propuestas para mejorar la eficiencia parlamentaria usando Inteligencia artificial. Este estudio destaca por un lado la necesidad de que se desarrolle una colaboración internacional y por otro la necesidad de un diálogo extenso sobre el uso de la Inteligencia artificial en este ámbito. Finalmente se insta a las partes interesadas del parlamento griego post-elecciones 2023 a discutir las soluciones adaptadas al propio parlamento usando un enfoque colaborativo.

dos estudios que concluyen con que el uso del *machine learning* podría conseguir mejoras significativas en estos parlamentos, y que en el futuro podría implementarse de manera más amplia.

Estos ejemplos son solo una muestra de los esfuerzos por investigar cómo esta tecnología podría mejorar la vida parlamentaria y su calidad cuando lleguen a implementarse en los parlamentos de todo el mundo. Un cambio al que el parlamento europeo se ha estado adelantando creando un reglamento para regular la inteligencia artificial, el primero de su tipo en el mundo.

2.2. Tipos de aprendizaje

En la rama del aprendizaje automático encontramos varios tipos que explicaremos brevemente a continuación[5]:

Tenemos dos de los más importantes actualmente que son el **aprendizaje supervisado** y el **no supervisado**.

Por el lado del aprendizaje supervisado, se trata de un tipo de aprendizaje que requiere un conjunto de datos etiquetados para predecir un resultado y un conjunto de resultados deseados. Encontramos un ejemplo de su funcionamiento en la Figura 2.1:

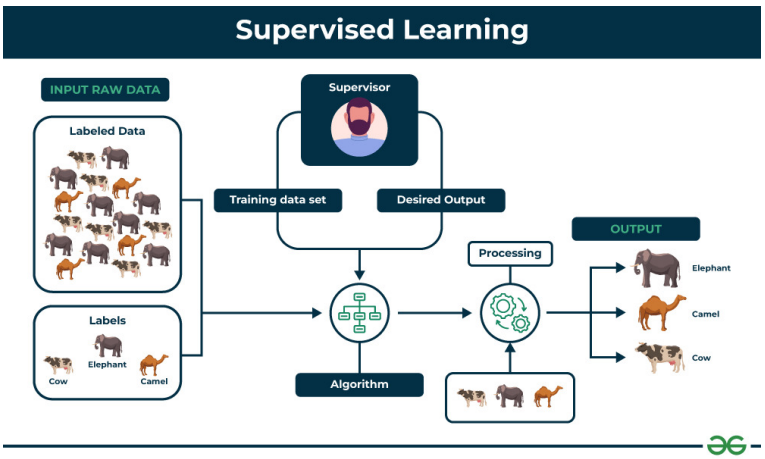


Figura 2.1: Ejemplo de aprendizaje supervisado, extraído de geeksforgeeks.org

Esta técnica requiere de una gran supervisión humana, tal y como su nombre indica, ya que depende de la provisión y los resultados deseados. Este tipo de aprendizaje es muy útil para las empresas ya que se puede implementar en una serie de aplicaciones, como por ejemplo[6]:

- Regresión Lineal: Predice valores numéricos basándose en relaciones lineales.
- Regresión Logística: Predice probabilidades de categorías binarias.
- Máquinas de vectores de soporte (SVM): Clasifica datos separándolos con un hiperplano óptimo.
- Bosques Aleatorios: Combinan múltiples árboles de decisión para mejorar la precisión.
- Redes Neuronales: Simulan la estructura del cerebro para reconocer patrones complejos.
- k-Vecinos más Cercanos (k-NN): Clasifica basándose en la mayoría de sus vecinos cercanos.
- Naive Bayes: Clasifica usando la probabilidad condicional basada en el teorema de Bayes.

Mientras que el aprendizaje no supervisado, es una técnica que no requiere de que se etiquete el conjunto de entrada. Se basa en la capacidad de identificar patrones y determinar aquellos con gran similitud sin tener ningún conocimiento previo. El aprendizaje no supervisado es muy útil para encontrar patrones en grandes volúmenes de datos rápidamente.La Figura 2.2 ilustra cómo opera este tipo de aprendizaje:

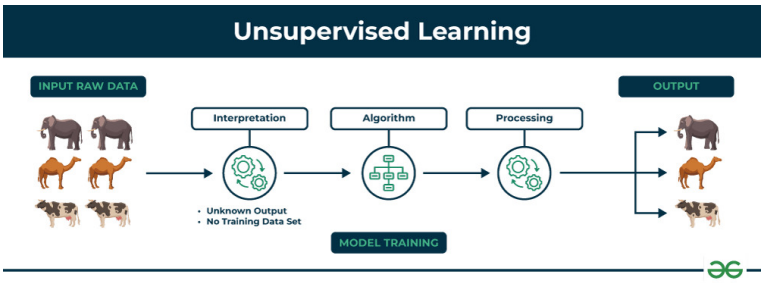


Figura 2.2: Ejemplo de aprendizaje no supervisado, extraído de geeksforgeeks.org

Los modelos de aprendizaje no supervisado se utilizan para tres tareas principales[7]:

- Agrupación en clústeres, asociación y reducción de dimensionalidad: Organiza datos no etiquetados en grupos basados en similitudes, utilizando métodos como k-medias, agrupación jerárquica y modelos de mezcla gaussiana.

- Reglas de asociación: identifica relaciones entre variables en datos, destacándose en análisis de cesta de la compra con algoritmos como A priori para recomendar productos basados en patrones de consumo.

- Reducción de dimensionalidad: Simplifica conjuntos de datos de alta dimensión para mejorar el rendimiento y facilitar la visualización, usando métodos como PCA, SVD y codificadores automáticos.

Además de estos dos tipos de aprendizajes ya mencionados, también tenemos el **aprendizaje semi-supervisado**[8], que combina los dos anteriores para poder clasificar de manera adecuada. Este enfoque considera tanto los datos etiquetados como los no etiquetados.

Por otro lado tenemos el **aprendizaje por refuerzo**[9] que basa su sistema de aprendizaje en el ensayo y error, recompensando los comportamientos adecuados y penalizando los inadecuados. Este tipo de aprendizaje está pensado para que la máquina cometa varios errores al inicio y mediante un refuerzo positivo, termine realizando correctamente las acciones.

2.3. Empresas que utilizan el aprendizaje automático

En estos últimos años, cada vez más empresas reconocidas a nivel mundial, tanto del sector tecnológico como en otros sectores, están realizando grandes inversiones en el campo de la inteligencia artificial con la idea de mejorar tanto sus plataformas como productos, aprovechando todas las ventajas que esta ofrece para mantenerse líderes en sus respectivos mercados. Dentro de este campo, una de las ramas en la que más se está invirtiendo es precisamente el aprendizaje automático. Las empresas no solo están integrando esta tecnología en sus productos, sino que también la aplican en los servicios que prestan y la utilizan para fomentar la innovación dentro de sus organizaciones.

En esta sección, mencionaremos algunos ejemplos de cómo se utiliza el aprendizaje automático en diversas empresas y para qué se emplea:

Spotify: Debido a la gran cantidad de información que maneja la aplicación de streaming musical Spotify[10], la organización decidió utilizar de toda la información que obtiene de sus usuarios de manera implícita para que, usando la tecnología de machine learning de la compañía Niland, crease y mejorase muchos de los servicios de los que disponemos actualmente en esta plataforma.Estos servicios incluyen mejores búsquedas, descubrimientos semanales, entre otros. Esta integración, junto con la tecnología de *blockchain*, se ha convertido en el núcleo sobre el que se sustenta toda la experiencia de usuario de Spotify permitiéndole convertirse en una de las mayores empresas de su sector.

Netflix: Netflix[11], está expandiendo su uso de *machine learning* para mejorar la experiencia del usuario y optimizar sus operaciones internas. Con el crecimiento en el tamaño de los datos y la complejidad de los modelos de IA, la Machine Learning Platform(MPL), un equipo en el que la compañía está invirtiendo recursos, se ha adaptado para soportar aplicaciones de alto rendimiento en tiempo real y de apoyo a decisiones con las que esperan mejorar varios servicios presentes actualmente en la compañía, como por ejemplo, las recomendaciones de series y películas o la calidad del streaming.

Facebook: *DeepText* de Facebook[12] es una herramienta de la plataforma que le permite entender los textos publicados con una precisión casi humana. Es una de las formas en las que Facebook implementa el aprendizaje automático en forma de *deep learning*, lo que le permite solventar muchos de los desafíos lingüísticos que se presentan en una red social con tantos usuarios como esta. En este contexto, *DeepText* está siendo utilizado principalmente para entender mejor los intereses de los usuarios, unificar el entendimiento de los contenidos visuales y textuales y por último para mejorar la experiencia general del usuario en la plataforma.

Capítulo 3

Análisis de requerimientos

Si se recuerda el primer capítulo, el proyecto propone obtener resultados mediante el análisis de cada iniciativa parlamentaria. Por lo tanto,es fundamental considerar al menos cuatro componentes antes de iniciar el diseño e implementación del mismo: el lenguaje de programación, el formato en el que se trabajará, el procesamiento del lenguaje natural y herramientas de aprendizaje automático.

3.1. Requisitos

En esta sección se determinará qué herramientas se ajustan mejor al proyecto para evitar posibles errores y problemas durante el proceso de desarrollo, e incluso después de este.

3.1.1. Lenguaje de programación

Uno de los elementos más importantes en un proyecto de desarrollo de software es la selección del lenguaje de programación. La selección adecuada puede proporcionar características específicas que son fundamentales para el éxito del proyecto. En el caso de este proyecto, es crucial elegir un lenguaje que ofrezca las siguientes características:

- Uso generalizado y gran facilidad de trabajo en el campo de la inteligencia artificial.
- Facilidad de aprendizaje y uso.
- Gran comunidad.
- Abundancia de librerías, especialmente en el ámbito mencionado.

3.1.2. Formato

Para obtener los extractos y temas de cada iniciativa parlamentaria a través de la API, es necesario decidir el formato de los datos con el que podrá trabajar el programa.

Aunque existen varios formatos disponibles, los más utilizados en este contexto son o JSON o CSV, ya que son los formatos más comunes a la hora de trabajar con datos en el desarrollo web.

Además, la complejidad asociada con cada formato debe considerarse al momento de la elección, tanto a la hora de leer datos como de escribirlos usando dicho formato.

3.1.3. Procesamiento del lenguaje natural

El **procesamiento del lenguaje natural** (**PLN**)[13] es una tecnología de aprendizaje automático que permite a la computadora entender, interpretar y generar texto y voz en lenguaje humano. Utiliza una combinación de modelización del lenguaje humano basado en reglas y modelos estadísticos y de *machine learning*. Para poder realizar la tarea de analizar los extractos parlamentarios de las iniciativas necesitaremos realizar correctamente este procesamiento mediante una serie de métodos como:

- Tokenización.
- Eliminación de palabras vacías.
- Lematización.

3.1.4. Aprendizaje automático

En esta fase del análisis, se debe elegir una herramienta que nos resulte completa para realizar los dos modelos que se han propuesto para realizar el análisis de los datos obtenidos anteriormente con el PLN. La herramienta seleccionada debe ofrecer:

- Creación de los modelos indicados para realizar el análisis de los datos procesados.
- Entrenamiento de los modelos con los datos.
- Obtención de los resultados del entrenamiento de los modelos.

Además, sería beneficioso tener la capacidad de visualizar los datos de manera efectiva mediante gráficos para facilitar la interpretación de los resultados.

3.2. Soluciones

Una vez se haya estudiado todos los requisitos más importantes para el desarrollo del proyecto, debemos hacer una selección de las soluciones que mejor se ajusten a estos que estén a nuestro alcance. A continuación, se indicará las soluciones escogidas junto al motivo de su elección.

3.2.1. Lenguaje de programación

Para abordar este aspecto, se ha optado por seleccionar Python como lenguaje de desarrollo del proyecto, ya que cumple con las características mencionadas anteriormente. Además de eso, es un lenguaje que presenta una sintaxis muy simple que facilita su aprendizaje y cuenta con un montón de librerías, especialmente para las tareas que realizaremos en el presente proyecto, que son el procesamiento del lenguaje humano y el aprendizaje automático.

3.2.2. Formato

Debido a las necesidades presentes en el proyecto con respecto al formato, se ha decidido usar la librería requests para realizar las peticiones a la API de la página del Parlamento de Canarias. En consecuencia se ha optado por el formato JSON usando la librería json de Python para manejar los datos.

3.2.3. Procesamiento del lenguaje natural

En la sección del **PLN** se encuentran varias librerías de las que podríamos disponer, pero en este proyecto se va a hacer uso de la librería NLTK,. Esta herramienta nos permite realizar todas las funciones mencionadas en los requisitos. NLTK es conocida por ser fácil de manejar, potente y compatible con el procesamiento de textos en español, asegurando así la correcta ejecución de las funciones requeridas.

3.2.4. Aprendizaje automático

Para este último punto del aprendizaje automático se usará la librería gensim para el modelo LDA y para el modelo BERTopic la librería bertopic, que nos permite crear los modelos necesarios para el proyecto y que tiene todos los métodos necesarios para trabajar correctamente con los mismos. Como se ha indicado con respecto a las librerias, el tipo de aprendizaje que se utilizará en este proyecto es el no supervisado, por lo tanto, se necesitará un conjunto de datos sin la necesidad de etiquetado previo.

Capítulo 4

Desarrollo de software

En el Capítulo anterior tras analizar los requisitos, se procede a iniciar la explicación del desarrollo del proyecto.

4.1. Estructura y diseño del proyecto

En esta sección se explicará la estructura utilizada durante el desarrollo del proyecto de forma concisa. Un primer vistazo nos muestra que todo el código está estructurado alrededor de los modelos en los que se va a trabajar, los cuales son por un lado, el modelo LDA y por el otro, el modelo BERTopic.

4.1.1. Modelo LDA

El modelo **LDA**[14] (Latent Dirichlet Allocation o en español Asignación Latente de Dirichlet) es una técnica de modelado de tópicos utilizada en procesamiento de lenguaje natural, además de otros campos. Su objetivo es descubrir automáticamente los **temas** (o topics) que están presentes en una colección de documentos y asignar cada palabra en un documento a uno de estos temas.

Este modelo se basa en la concepción de que los temas están ocultos (o **latentes** como dice su nombre) y deben ser inferidos a partir de los datos analizando las frecuencias de las palabras y su coincidencia en los documentos. Una vez encontrados los temas, las palabras se asignan usando asignaciones de Dirichlet para determinar la probabilidad de que una palabra pertenezca a un determinado documento. Este proceso de asignación en la fase inicial hace una asignación aleatoria mientras que en sucesivas iteraciones se mejora esta asignación buscando maximizar la probabilidad conjunta de los datos.

Al momento de hacer la implementación de este modelo, requiere de una serie de pasos, los cuales son:

1. **Preprocesamiento:**
    - Tokenización: Divide el texto en unidades más pequeñas (tokens), como palabras o frases.
    - Eliminación de números: Remueve números que no aportan información relevante para el análisis de temas.
    - Lematización: Reduce las palabras a su forma base o lema.
    - Eliminación de palabras vacías: Filtra palabras comunes (como ɾ", .º", "pero") que no son informativas.

    Estos procesos se detallarán en su sección correspondiente.

2. **Entrenamiento del modelo:**
    - Parámetros de entrada: Se especifican los parámetros necesarios, como el número de temas que se desea identificar.
    - Datos preprocesados: Se utiliza el conjunto de datos que ya ha sido sometido al preprocesamiento.

3. **Evaluación y Visualización de Resultados:**
    - Evaluación cuantitativa: Se mide la calidad de los temas identificados utilizando métricas como la coherencia de los temas.
    - Se crean visualizaciones que facilitan la interpretación de los datos, como gráficos de barras que muestran la distribución de palabras en cada tema.

Por último en este punto llegaría la evaluación de estos resultados obtenidos, cuyos valores usaremos para la evaluación cuantitativa del siguiente capítulo y con la evaluación completada, ya se pueden crear las gráficas con los datos obtenidos una vez evaluados, que permite visualizar estos de una manera más sencilla y más cómoda de lo que sería visualizar los propios datos tal cual se obtienen y que además nos ayudarán en la evaluación cualitativa que se hará en el siguiente capítulo.

Este procedimiento se repetirá para analizar cada grupo parlamentario, separando los extractos según el grupo y entrenando modelos específicos para cada uno.

4.1.2. Modelo BERTopic

BERTopic[15] es otro modelo de aprendizaje no supervisado que combina la potencia de los modelos de lenguaje preentrenados, como **BERT** (Bidirectional Encoder Representations from Transformers o Representación de Codificador Bidireccional de Transformadores), con técnicas avanzadas de *clustering* para identificar y extraer **temas** de un conjunto de documentos. A diferencia de los métodos tradicionales de modelado de temas, BERTopic utiliza *embeddings* de alta calidad para capturar mejor las relaciones semánticas entre las palabras y los documentos.

Este modelado de temas se hace en cuatro pasos, que son:

1. **Generación de Embeddings:** BERT transforma los documentos en *embeddings*, que son representaciones densas y de alta dimensión que capturan el significado contextual de las palabras dentro del documento. Por ejemplo, la palabra 'banco' tendrá una representación distinta si se refiere a una institución financiera o a un asiento, dependiendo del contexto.

2. **Reducción de Dimensionalidad**: Dado que los Dado que los *embeddings* de BERT son de alta dimensión, BERTopic realiza una reducción de dimensionalidad para proyectar estos *embeddings* en un espacio de menor dimensión, utilizando técnicas como UMAP (Uniform Manifold Approximation and Projection). Esto facilita el posterior proceso de clustering.

3. **Clustering:** En el espacio de menor dimensión, se aplican algoritmos de clustering, como HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise), para agrupar documentos con *embeddings* similares. Cada uno de los clusters generados en este paso representa un posible tema.

4. **Extracción de Palabras Clave:** BERTopic extrae las palabras clave más representativas de cada clúster, utilizando técnicas como TF-IDF (Term Frequency-Inverse Document Frequency). Estas palabras clave ayudan a interpretar y etiquetar los tópicos identificados.

La implementación de BERTopic sigue los mismos pasos generales que LDA, con diferencias en el preprocesamiento de los extractos y en la transformación de los textos para que el modelo los procese correctamente. Por ejemplo, el preprocesamiento puede incluir pasos adicionales específicos para el modelo BERT, como la segmentación de oraciones y el manejo de subpalabras.

4.2. Implementación del procesamiento del lenguaje natural

En esta sección se procede a explicar como se ha realizado el procesamiento del lenguaje natural, que constituye el preprocesamiento de los extractos parlamentarios.

4.2.1. Procedimiento

Con el fin de analizar los temas tratados en los extractos parlamentarios, se deben procesar de una manera que los modelos puedan trabajar con ellos. Dado que ambos modelos usan el mismo procedimiento se procederá a explicarlo para ambos casos, obviando la diferencia que se comentó en la sección anterior.

Para que este proceso resulte en una salida que los modelos puedan analizar, es vital realizar una serie de etapas o fases, las cuales son las siguientes:

- **Tokenización:** Es el proceso de dividir un texto en unidades más pequeñas, como palabras o frases, llamados tokens. Es una etapa crucial en el procesamiento del lenguaje natural, facilitando el análisis y manipulación de los datos textuales. Por ejemplo, si tenemos el texto "El parlamento aprobó la ley", la tokenización resultaría en [’El’, ’parlamento’, ’aprobó’, ’la’, ’ley’].

- **Eliminación de números:** En esta fase se procede a identificar y eliminar los caracteres numéricos del texto, ya que se entiende que estos no aportan valor semántico al análisis. Por ejemplo, el texto "La ley 123 fue aprobada en 2020"se transformaría en ’La ley fue aprobada en’.

- **Lematización:** La lematización es un proceso de normalización que consiste en reducir las palabras a su forma base o raíz, conocida como "lema". Este paso es necesario ya que hay que poner todas las palabras en su forma estándar y reducir la variabilidad lingüística. Como ejemplo de este proceso, las palabras ’aprobó’, ’aprobado’, ’aprobar’ se reducen todas a su lema ’aprobar’.

- **Eliminación de palabras vacías:** La eliminación de palabras vacías es un proceso que consiste en eliminar palabras comunes y poco informativas de un texto. Estas palabras, como ’y’, ’o’, ’pero’, no contribuyen significativamente al análisis semántico del texto y, por lo tanto, se eliminan para reducir el ruido y mejorar la eficiencia. En el texto ’El parlamento y la ley’, las palabras ’y’ y ’la’ serían eliminadas, resultando en ’parlamento ley’. En las figuras 4.1 y 4.2 encontramos ejemplos de iniciativas sin PLN y con PLN, respectivamente.

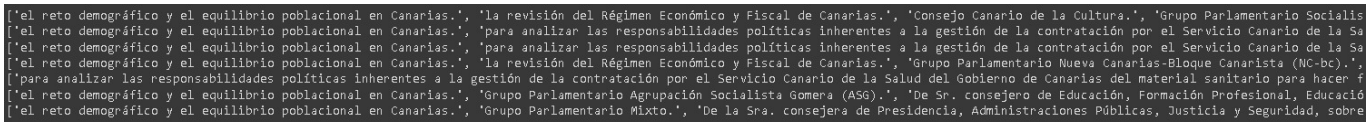


Figura 4.1: Ejemplo de iniciativas sin PLN

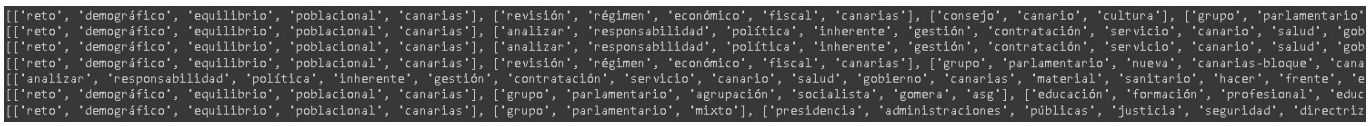


Figura 4.2: Ejemplo de iniciativas con PLN

4.2.2. Lematización o stemming

Cuando se hace el PNL normalmente surge la duda de si utilizar lematización o *stemming*. Esta última recorta las palabras para obtener el lexema (o *stemm* en inglés) de la misma.

Ambos procesos presentan una serie de ventajas y desventajas, por un lado, está el stemming, que es una técnica que reduce las palabras a su base o raíz (stem), cortando sus sufijos. Este proceso es más rápido y simple que la lematización, pero puede ser menos preciso. Esta mayor velocidad del *stemming* permite ser implementado en textos grandes ya que no afectará al tiempo de computo global, y permite averiguar relaciones entre palabras más fácilmente, aunque también puede tener problemas, ya que al quedarse solo con el lexema puede generar relaciones erróneas entre palabras, conduciendo a resultados erróneos.

La lematización por otro lado, la lematización reduce las palabras a su forma base o lema, utilizando un vocabulario y un análisis morfológico de las palabras. Es más precisa que el stemming, pero también más costosa en términos de recursos. Esta operación aunque más costosa, hace que no se pierda tanto contexto como el *stemming*. Debido a que estamos trabajando con textos pequeños, como son las iniciativas parlamentarias, hemos elegido la lematización sobre el stemming para garantizar una mayor precisión y contexto en los resultados. Un ejemplo de como funciona estas dos técnicas es la siguiente figura:

Figura 4.3: Ejemplo de lematización y stemming, extraído de turing.com

4.3. Implementación del aprendizaje automático

Una vez preprocesados los extractos parlamentarios, se pasan a los modelos para su análisis. En esta sección, se describe todo el proceso de entrenamiento de los modelos LDA y BERTopic.

4.3.1. Entrenamiento

El proceso de entrenamiento de los modelos de aprendizaje automático implica ajustar los parámetros del modelo para que puedan identificar patrones y temas en los extractos parlamentarios. A continuación, se describen los pasos específicos para entrenar los modelos LDA y BERTopic:

1. **Preparación de los datos:** LDA requiere la creación de una bolsa de palabras (bag-of-words) con los extractos preprocesados. Esto implica transformar los textos en un formato en el que cada documento esté representado como un vector de frecuencias de palabras.

   - Creación de la Bolsa de Palabras: Se cuenta la frecuencia de cada palabra en los documentos.
   - Creación del Diccionario: Se mapean las palabras a identificadores únicos.

2. **Configuración de los modelos:** En este apartado se configurará los modelos para que puedan posteriormente hacer los análisis correctamente.

   Por la parte del modelo LDA, es crucial definir cuántos temas generará el modelo. Para analizar y comparar con BERTopic, este valor será el mismo número de temas que genera este, además de que debemos indicarle los valores de alfa (Representa la dispersión de la distribución de los tópicos sobre los documentos. Un valor alto de alfa implica que los documentos contengan una mezcla de varios temas) y eta (Representa la dispersión de la distribución de las palabras sobre los tópicos. Un valor alto de eta implica que los tópicos están formados por una mezcla de muchas palabras). Estos valores estarán ajustados al tamaño del corpus para que el comportamiento se parezca lo más posible al exhibido por BERTopic.

   Mientras que por la parte de BERTopic, se va a elegir el tipo de tokenizador primero, una de las versiones compatibles con el modelo, que además es una versión de BERT más liviana y en minúsculas. Después cargamos un modelo que ya está pre entrenado, y por último podemos pasarle las configuraciones al modelo que vamos a crear, además de indicarle que los *embeddings* de texto va a ser en español.

3. **Entrenamiento de los modelos:** Ahora que todos los preparativos están completos, se procede con el entrenamiento propiamente dicho en ambos modelos. Esto implica utilizar los datos previamente procesados según la configuración requerida para crear los temas según las características específicas de cada modelo. Se generará el mismo número de temas en ambos casos, conforme a la configuración mencionada. Además, tal como se discutió en el procesamiento del lenguaje natural, este proceso se repetirá en ambos modelos para cada uno de los siete grupos parlamentarios que componen el Parlamento de Canarias.

Con todo esto hecho ya se pasa a hacer la evaluación de los modelos(tanto los entrenados con los datos generales cómo los que se usan para los grupos) usando una serie de métricas que indicarán como de bien están desarrollando los temas ambos modelos y que nos servirá para en el siguiente capítulo poder hacer una evaluación cuantitativa de los resultados. Posterior a esto y con los modelos evaluados se procede a obtener las gráficas, que nos permitirán visualizar todos los temas de una forma fácil y que servirán de cara a hacer el análisis cualitativo de los modelos, como ya se indicó previamente.

Capítulo 5

Evaluación de los resultados

Una vez que los modelos ya devuelven los valores de las evaluaciones y las gráficas, la tarea que queda pendiente por hacer es el análisis de resultados. En este capítulo se hará uso de los resultados en el capítulo anterior para hacer, primero, una evaluación cuantitativa usando los resultados de las evaluación de resultados para comparar los modelos creados en este proyecto, y después realizar una evaluación cualitativa, donde se realizará una evaluación más 'humana' y se intentará también indicar si los temas generados por ambos modelos pueden ser agrupados usando los temas del **Eurovoc**. Por último, en este capítulo se tratará de sacar unas conclusiones con respecto a qué modelo ha hecho mejor la tarea del análisis de extractos y cuál podría ser considerado mejor en esta tarea. Cabe destacar que para hacer este punto y como ya se ha comentado anteriormente, hemos hecho que los parámetros de ambos modelos hayan sido configurados para que sus condiciones iniciales sean las más parecidas posibles y comprobar qué resultados salen con estas características.

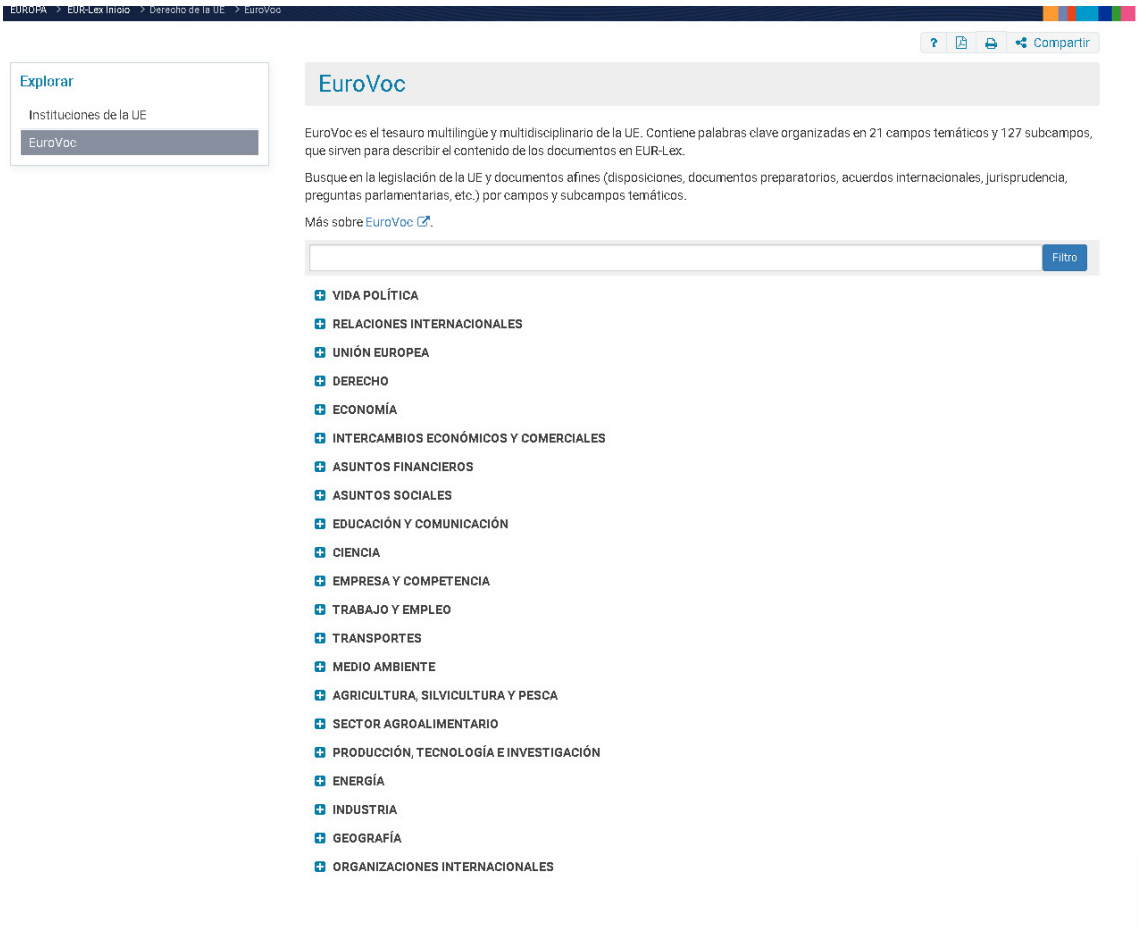


Figura 5.1: Pagina del Eurovoc

5.1. Evaluación cuantitativa

En este apartado utilizaremos, como se ha mencionado anteriormente, los resultados obtenidos en la evaluación de los modelos. Se indicarán para cada modelo las métricas elegidas, que miden y se comparan con las de los otros modelos implementados en el proyecto.

5.1.1. Modelo LDA

Para este modelo se ha elegido como métricas evaluativas la perplejidad, la diversidad y la coherencia, que miden:

- **La perplejidad:** Es una métrica que indica cómo de bien un modelo probabilístico predice un conjunto de datos. En el contexto de nuestro modelo indica cómo de buena es la capacidad de generar una distribución de palabras que se asemeje a la del corpus original. Una perplejidad más baja indica que el modelo LDA genera temas que son más representativos del conjunto de datos. Sin embargo, si este valor es muy bajo puede también indicar que ha sido sobreajustado.

- **La diversidad:** Es una métrica que mide la variedad de palabras que aparecen en los diferentes tópicos generados por el modelo. Se calcula como la proporción de palabras únicas en los tópicos en comparación con el total de palabras en todos los tópicos. Una mayor diversidad indica que los temas generados son más distintos entre sí, lo cual es importante para asegurarse de que el modelo no esté generando temas redundantes o muy similares.

- **La coherencia:** Esta métrica mide la variedad de palabras que aparecen en los diferentes tópicos generados por el modelo. Se calcula como la proporción de palabras únicas en los tópicos en comparación con el total de palabras en todos los tópicos. Una mayor coherencia indica que los temas generados son más interpretables y útiles para los humanos. Esta métrica ofrece varias formas de ser calculada, pero para el proyecto hemos elegido la coherencia de puntos de palabras(C-V).

Una vez explicado qué medidas se usarán para evaluar el rendimiento de los modelos LDA, es el momento de indicar qué los valores obtenidos por cada modelo tras la creación de los temas y compararlos con los de los otros modelos LDA.

- El modelo, cuando extrae los temas de todos los extractos de todos los grupos parlamentarios ha obtenido los siguientes valores en estas tres métricas:

```
Coherencia del modelo LDA: 0.40847280168930306
Perplejidad del modelo LDA: 5856.951499823471
Diversidad de tópicos en LDA: 0.668041237113402
```

- El modelo, empleando los extractos del primer grupo parlamentario obtiene los siguientes valores:

```
Coherencia del modelo: 0.4038033156789261
Perplejidad del modelo LDA: 1699.142418677212
Diversidad de tópicos en LDA: 0.5609756097560976
```

- Los valores del modelo usando los del segundo grupo son:

```
Coherencia del modelo: 0.28454241085379517
Perplejidad del modelo LDA: 711.2404598203952
Diversidad de tópicos en LDA: 0.33043478260869563
```

- Al utilizar los extractos del tercer grupo parlamentario, obtiene los siguientes valores:

```
Coherencia del modelo: 0.2617443856449696
Perplejidad del modelo LDA: 640.550224862316
Diversidad de tópicos en LDA: 0.3130434782608696
```

- Para los siguientes extractos, los del cuarto grupo, los valores obtenidos son:

```
Coherencia del modelo: 0.3434196310698546
Perplejidad del modelo LDA: 289.81792100361884
Diversidad de tópicos en LDA: 0.4846153846153846
```

- El modelo empleando los extractos del quinto grupo parlamentario obtiene los siguientes valores:

```
Coherencia del modelo: 0.34665176547837473
Perplejidad del modelo LDA: 499.185005076285
Diversidad de tópicos en LDA: 0.41875
```

- Con los extractos del sexto grupo parlamentario, el modelo obtiene los siguientes valores:

```
Coherencia del modelo: 0.33221345466119223
Perplejidad del modelo LDA: 386.73175494653515
Diversidad de tópicos en LDA: 0.3625
```

- Utilizado los extractos del último grupo, el séptimo grupo parlamentario obtiene los siguientes valores:

```
Coherencia del modelo: 0.32172299958414796
Perplejidad del modelo LDA: 251.5616056740245
Diversidad de tópicos en LDA: 0.49230769230769234
```

En las figuras 5.2, 5.3 y 5.4 se puede observar mejor los datos presentados anteriormente mediante unas gráficas:

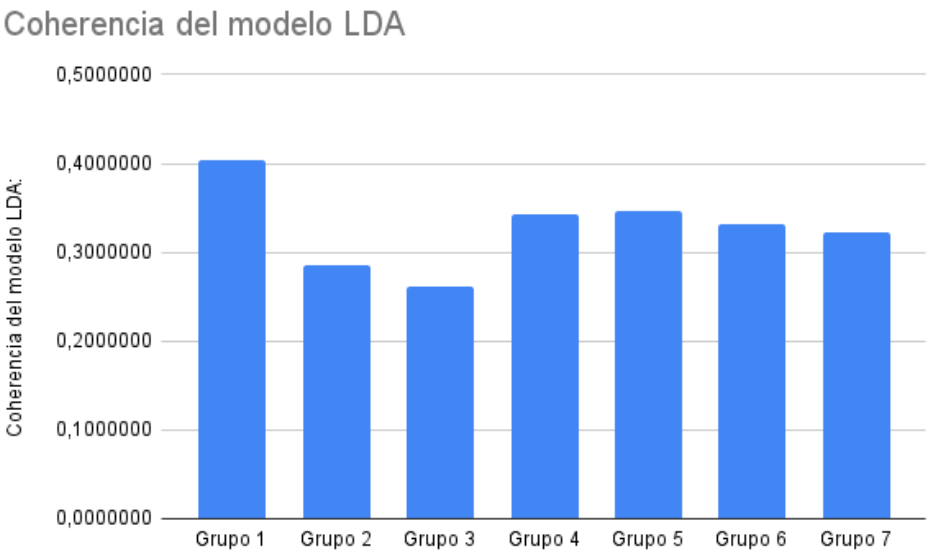


Figura 5.2: Gráfica de coherencia por grupos

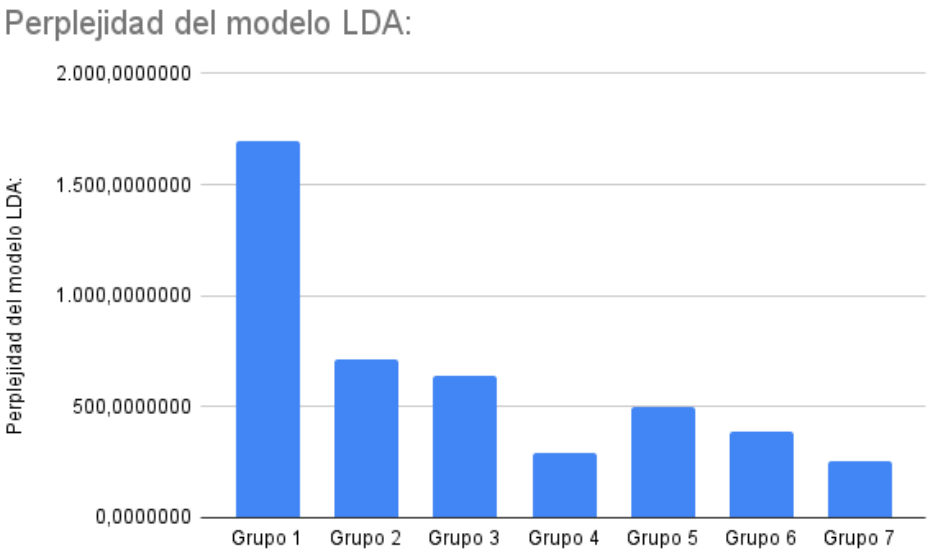


Figura 5.3: Gráfica de perplejidad por grupos

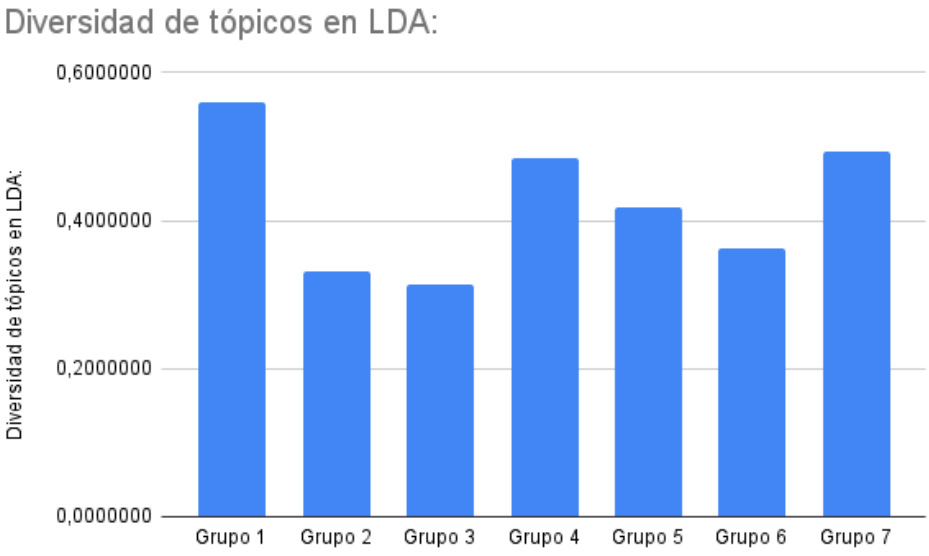


Figura 5.4: Gráfica de diversidad por grupos

Una vez presentado todos los resultados, se procede a analizarlos en conjunto.

En líneas generales, se puede observar que el modelo LDA obtiene valores decentes en las métricas analizadas, aunque muestra un descenso notable en todas las métricas al inicio de la evaluación de los temas del segundo grupo parlamentario. Sin embargo, este descenso se va remontando y el modelo se mantiene estable a partir de la evaluación en el cuarto grupo parlamentario. Aun así, es evidente que las condiciones iniciales con las que crea los temas no favorecen a este modelo, y estas métricas demuestran que necesita ajustes para obtener mejores resultados. Una revisión de cada métrica obtenida en los ocho usos del modelo LDA revela las siguientes conclusiones:

- **Coherencia:** Presenta una variación bastante visible en los diferentes casos evaluados. Esto sugiere que con algunos conjuntos de datos, el modelo produce tópicos más coherentes que con otros. Valores más altos de coherencia indican que los tópicos identificados por el modelo están más estrechamente relacionados entre sí en términos de las palabras que los componen.

- **Diversidad:** También muestra variaciones, bastante más notables que en el caso de la coherencia. En los extractos donde el modelo consigue una diversidad de tópicos más alta sugiere que los tópicos encontrados son más diferentes entre sí, lo cual es en este caso deseable ya que evita la redundancia.

- **Perplejidad:** Es la variable que más varia entre los diferentes grupos parlamentarios. La perplejidad es una medida que analiza cómo de bien el modelo LDA predice un conjunto de datos nuevos. Valores más bajos de perplejidad indican una mejor capacidad predictiva y, por lo tanto, una mejor calidad general del modelo.

  Después de esta evaluación de cada una de las métricas obtenidas en la creación de los temas de los grupos, y revisando todos estos valores en conjunto, se puede concluir que tras crear los temas del séptimo grupo parlamentario, el modelo LDA obtiene los mejores resultados en estas 3 métricas, ya que obtiene la mejor diversidad y la mejor perplejidad, y aunque la coherencia está dentro de un valor normal para los obtenidos, es un valor bueno.

5.1.2. Modelo BERTopic

Analizados los valores obtenidos por el modelo LDA, se procede a analizar los valores del modelo BERTopic, cuyas métricas a evaluar son las siguientes:

- **Coherencia:** Esta métrica se refiere a qué tan coherentes son los tópicos individualmente. Una alta coherencia implica que las palabras dentro de un tópico están estrechamente relacionadas semánticamente.
- **Diversidad:** Es una métrica que mide la variedad de palabras que aparecen en los diferentes tópicos generados por el modelo. La diversidad entre tópicos se mide utilizando la distancia entre las distribuciones de probabilidad de los tópicos.
- **Silhouette score:** El Silhouette Score se utiliza para evaluar la calidad de la agrupación de documentos en tópicos basándose en sus *embeddings* (representaciones vectoriales).Un valor cercano a indica una buena separación entre los temas y que los documentos están bien agrupados dentro de sus respectivos temas, mientras que un valor cercano a 0 sugiere superposición o que los documentos podrían estar en límites entre diferentes tópicos.

Al igual que con el modelo LDA procedemos a mostrar los resultados obtenidos con el modelo BERTopic.

- Empezando por el modelo inicial que recoge todos los extractos parlamentarios, los valores obtenidos son:

```
Coherencia de los temas para el modelo: 0.6609866349373683
diversidad: 0.6396790272416305
Silhouette Score: 0.0048594181425869465
```

- El modelo, empleando los extractos del primer grupo parlamentario obtiene los siguientes valores:

```
Coherencia de los temas para el modelo: 0.670100555971999
diversidad: 0.7226032114915792
Silhouette Score: 0.10889068990945816
```

- Los valores del modelo usando los del segundo grupo son:

```
Coherencia de los temas para el modelo: 0.6551334595226678
diversidad: 0.8224639401929347
Silhouette Score: 0.11357121914625168
```

- Al utilizar los extractos del tercer grupo parlamentario, obtiene los siguientes valores:

```
Coherencia de los temas para el modelo: 0.6398645427207782
diversidad: 0.737764448386827
Silhouette Score: 0.08018354326486588
```

- Para los siguientes extractos, los del cuarto grupo, los valores obtenidos son:

```
Coherencia de los temas para el modelo: 0.5578972631748175
diversidad: 0.7794216284425972
Silhouette Score: 0.09489749372005463
```

- El modelo empleando los extractos del quinto grupo parlamentario obtiene los siguientes valores:

```
Coherencia de los temas para el modelo: 0.6624629364393568
diversidad: 0.7925828443549341
Silhouette Score: 0.09884674102067947
```

- Con los extractos del sexto grupo parlamentario, el modelo obtiene los siguientes valores:

```
Coherencia de los temas para el modelo: 0.6099060808125337
diversidad: 0.8716676405024977
Silhouette Score: 0.13702282309532166
```

- Utilizado los extractos del último grupo, el séptimo grupo parlamentario obtiene los siguientes valores:

```
Coherencia de los temas para el modelo: 0.7195853993510787
diversidad: 0.8865693531035794
Silhouette Score: 0.1224728375673294
```

Al igual que en el modelo BERTopic, las figuras 5.5, 5.6 y 5.7 serán de ayuda a la hora de visualizar de una manera mas sencilla los datos presentados previamente:

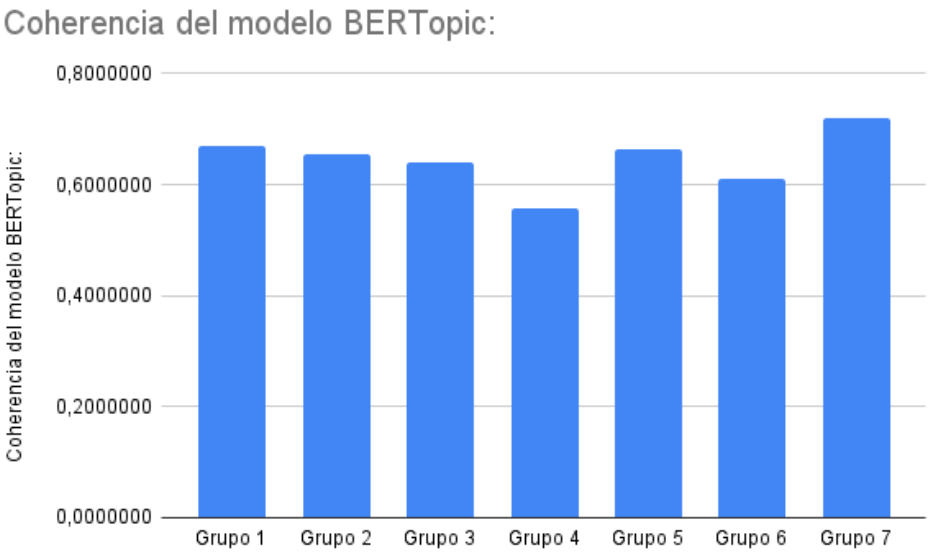


Figura 5.5: Gráfica de coherencia por grupos

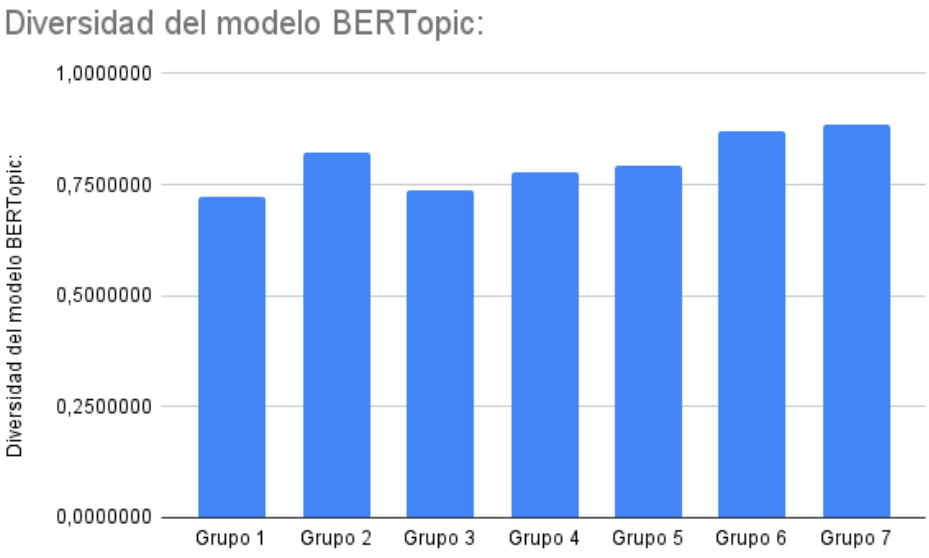


Figura 5.6: Gráfica de perplejidad por grupos

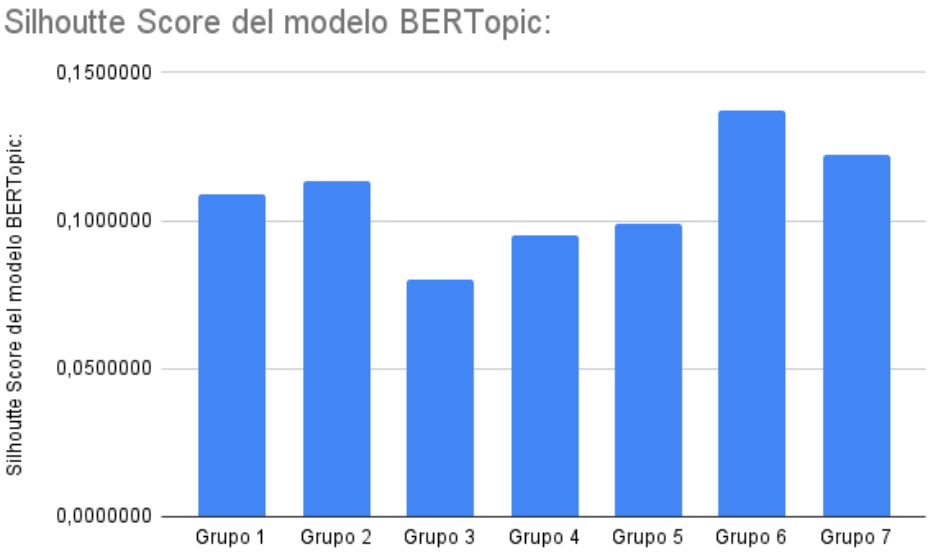


Figura 5.7: Gráfica de diversidad por grupos

Con estos valores podemos observar que en líneas generales estos resultados obtenidos por el modelo BERTopic son bastante positivos, lo que demuestra que el modelo está bastante bien ajustado. En una revisión más pormenorizada de las métricas arroja los siguientes resultados:

- **Coherencia:** Esta métrica varía entre 0.56 y 0.72. Esto indica que, en general, el modelo produce temas bastante coherentes, lo que significa que las palabras dentro de cada tema están bien agrupadas y son relevantes entre sí.
- **Diversidad:** Esta métrica varía entre 0.64 y 0.89, con valores más altos sugiriendo una mayor diversidad en los temas generados. Esto significa que los modelos cubren una amplia gama de tópicos, lo cual es deseable para capturar diferentes aspectos del texto analizado.
- **Silhouette score:** Esta variable varía entre aproximadamente 0.005 y 0.14. Mide qué tan similares son los documentos dentro del mismo tema en comparación con los documentos en otros temas. Los valores en general son bajos, lo que sugiere que los temas podrían no estar claramente diferenciados entre sí.

Además se puede observar gracias a estos datos que el modelo tiene unos resultados buenos a nivel global, siendo los datos de los temas obtenidos por los extractos de último grupo el que tiene el mejor rendimiento global, siendo que este es el modelo más coherente, con mayor diversidad de temas y que consigue que tengan una razonable separación entre ellos.

5.2. Evaluación cualitativa

Una vez hecha la evaluación cuantitativa y evaluados los dos modelos de esta manera, se procede a evaluar ambos modelos usando un análisis realizado por un ser humano para evaluar cómo de coherentes son estos según un observador haciendo uso de los campos temáticos **Eurovoc**. **Eurovoc** es un tesauro multilingüe y multidisciplinario que está disponible en 24 idiomas, ofrece multitud de textos legislativos de la Unión Europea. Este tesauro contiene 21 campos temáticos con los que procederemos a agrupar los temas.

Los campos temáticos disponibles en el Eurovoc son los siguientes:

- VIDA POLÍTICA
- RELACIONES INTERNACIONALES
- UNIÓN EUROPEA
- DERECHO
- ECONOMÍA
- INTERCAMBIOS ECONÓMICOS Y COMERCIALES
- ASUNTOS FINANCIEROS
- ASUNTOS SOCIALES
- EDUCACIÓN Y COMUNICACIÓN
- CIENCIA
- EMPRESA Y COMPETENCIA
- TRABAJO Y EMPLEO
- TRANSPORTES
- MEDIO AMBIENTE
- AGRICULTURA, SILVICULTURA Y PESCA
- SECTOR AGROALIMENTARIO
- PRODUCCIÓN, TECNOLOGÍA E INVESTIGACIÓN
- ENERGÍA
- INDUSTRIA
- GEOGRAFÍA
- ORGANIZACIONES INTERNACIONALES

En esta evaluación se irá revisando las gráficas de los modelos, aunque en este caso se pasará directamente a los temas de cada grupo parlamentario, analizando estas gráficas y luego indicando cuántos temas se agrupan bajo cada campo temático del Eurovoc. Además al final del análisis de cada modelo se presentará una gráfica de los temas por cada grupo parlamentario. Además, al final de la evaluación a cada modelo se podrá visualizar una gráfica con los campos temáticos y los resultados obtenidos.

5.2.1. Gráficas del modelo LDA

Para la evaluación de este modelo, a nivel general y tal y como se decía en la evaluación cuantitativa, se puede observar que aunque los temas generados tienen una coherencia decente en algunos modelos siendo un valor normal o incluso tirando a bajo, tienen problemas al mezclar muchas palabras que en principio no deberían estar en un mismo tema. Además, aunque los temas son bastante diversos, aún hay temas que se solapan de mayor o menor manera. Una vez analizado esto, podemos empezar a analizar los temas de cada grupo parlamentario.

En estos análisis solo se va incluir los campos temáticos que tienen al menos algún tema asociado y se ignorarán aquellos sin ningún tema asociado. Además con cada grupo parlamentario en este apartado previamente se mostrará una gráfica donde se puede visualizar un mapa de distancias entre tópicos y los términos más utilizados con el objetivo de ejemplificar una forma de visualizar los temas.

En el primer grupo parlamentario analizado, que tiene 45 temas, estos se agrupan en los siguientes campos temáticos:

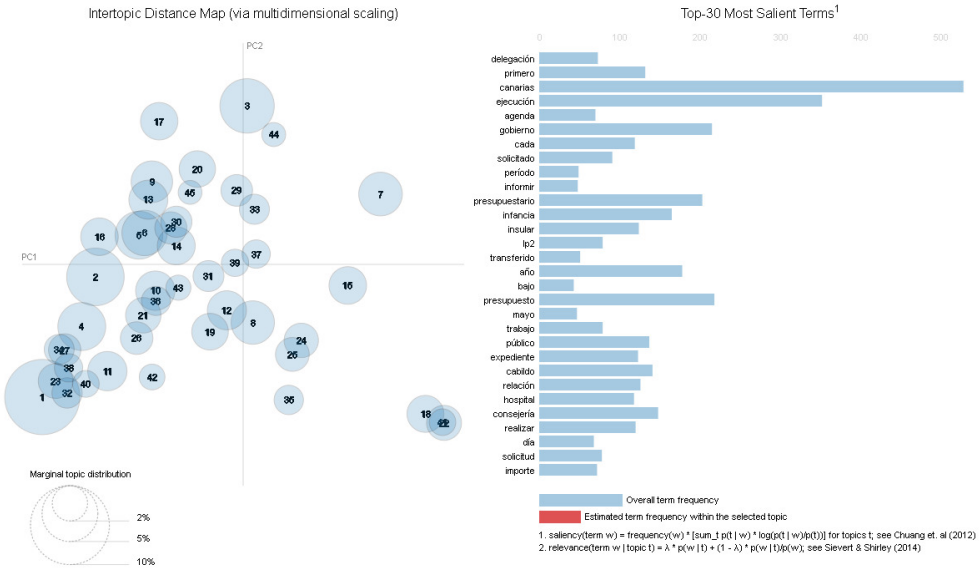


Figura 5.8: Gráfica de temas del primer grupo parlamentario

- **Política:** 3
- **Economía:** 2
- **Finanzas:** 6
- **Asuntos sociales:** 18
- **Educación y comunicación:** 6
- **Trabajo y empleo:** 3
- **Medio Ambiente:** 3
- **Industria:** 4

Este grupo se puede observar que el campo más tratado según LDA y por lo tanto lo que se entiende que es su prioridad principal es el de asuntos sociales, seguido por Finanzas y Educación y comunicación, y ya después los demás temas, lo que nos demuestra que este grupo está más centrado en los problemas sociales y de las personas y en menor medida, la gestión financiera y otros temas económicos. Todo esto nos lleva a ver que el grupo parlamentario aborda tanto las necesidades sociales como las necesidades económicas de la sociedad canaria.

Para el siguiente grupo parlamentario, que tiene 23 temas creados, la agrupación por campos temáticos quedaría de la siguiente forma:

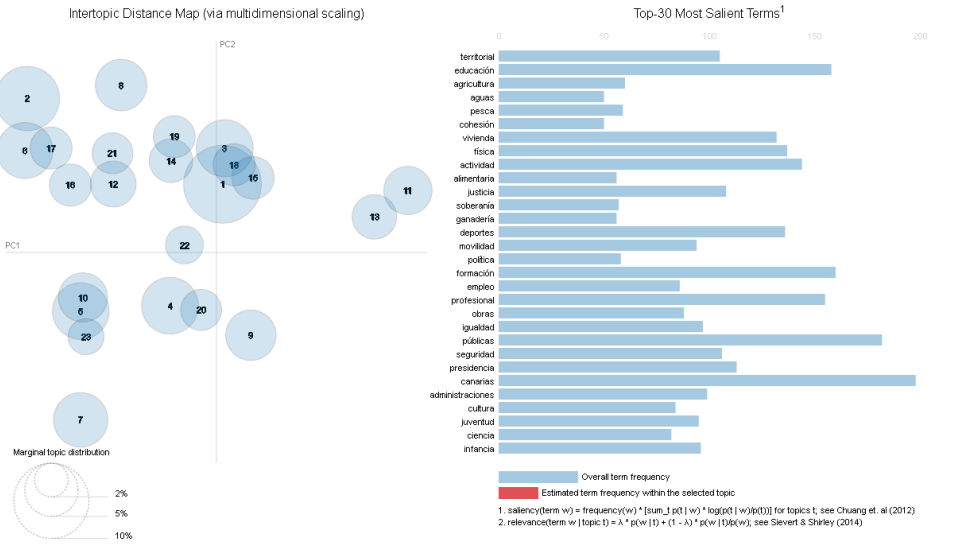


Figura 5.9: Gráfica de temas del segundo grupo parlamentario

- **Política:** 3
- **Derecho:** 2
- **Asuntos sociales:** 9
- **Educación y comunicación:** 5
- **Trabajo y empleo:** 1
- **Medio Ambiente:** 1
- **Agricultura, Silvicultura y pesca:** 2
- **Energía:** 1
- **Industria:** 1

Para este grupo parlamentario se puede observar que la distribución de temas está más diversificada entre varios campos, siendo la excepción a esto el de Asuntos sociales, lo que nos demuestra que si bien este grupo parlamentario prefiere hablar de varios temas importantes para los ciudadanos, como el anterior su prioridad principal son los asuntos sociales de la sociedad canaria para mejorar la calidad de vida de los habitantes de las islas.

El tercer grupo parlamentario, que tiene 23 temas al igual que el anterior, tiene los temas agrupados de la siguiente manera:

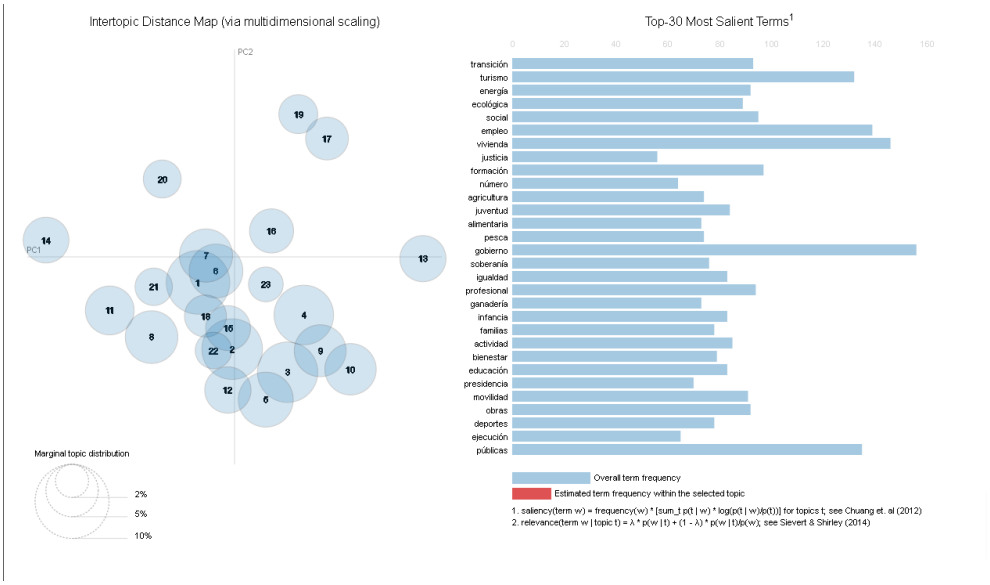


Figura 5.10: Gráfica de temas del tercer grupo parlamentario

- **Política:** 2
- **Unión Europea:** 1
- **Derecho:** 1
- **Finanzas:** 1
- **Asuntos sociales:** 6
- **Educación y comunicación:** 2
- **Trabajo y empleo:** 5
- **Medio Ambiente:** 1
- **Agricultura, Silvicultura y pesca:** 1
- **Energía:** 3

Este grupo presenta una fuerte inclinación por los temas sociales y laborales, siendo las dos prioridades principales de este grupo parlamentario y siendo también dos temas que tienen preocupados actualmente a toda la sociedad en general, y especialmente a la canaria, distribuyéndose el resto de temas de manera más uniforme. Estos temas que trata demuestra que el grupo tiene una especial preocupación por la sociedad y el medio ambiente de las islas. Cabe reseñar que en este grupo aparece el campo de la Unión europea cuando en los anteriores grupos parecía que no se trataba este tema.

Para el cuarto grupo parlamentario, que tiene 13 temas, las agrupaciones de temas quedan de la siguiente forma:

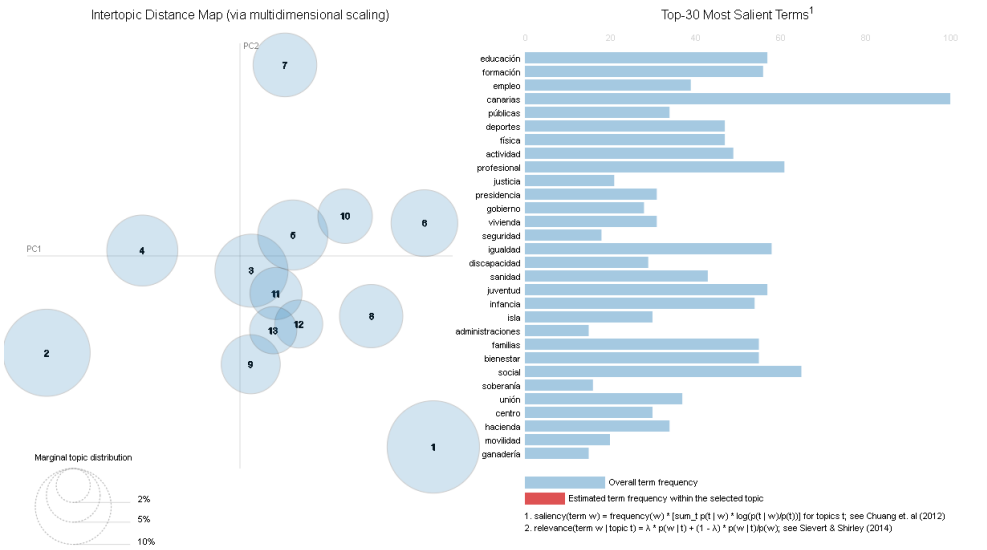


Figura 5.11: Gráfica de temas del cuarto grupo parlamentario

- **Política:** 1
- **Unión Europea:** 2
- **Asuntos sociales:** 4
- **Educación y comunicación:** 1
- **Trabajo y empleo:** 2
- **Energía:** 1

Este grupo parlamentario tiene menos temas, pero aún así vemos que se mantiene la línea general que la prioridad principal del grupo son los asuntos sociales, y después en menor medida tenemos otros temas relevantes como el trabajo y empleo y la educación dos temas que siempre están considerados importantes.

En el quinto grupo, con 16 temas, las agrupaciones de temas serían de la siguiente forma:

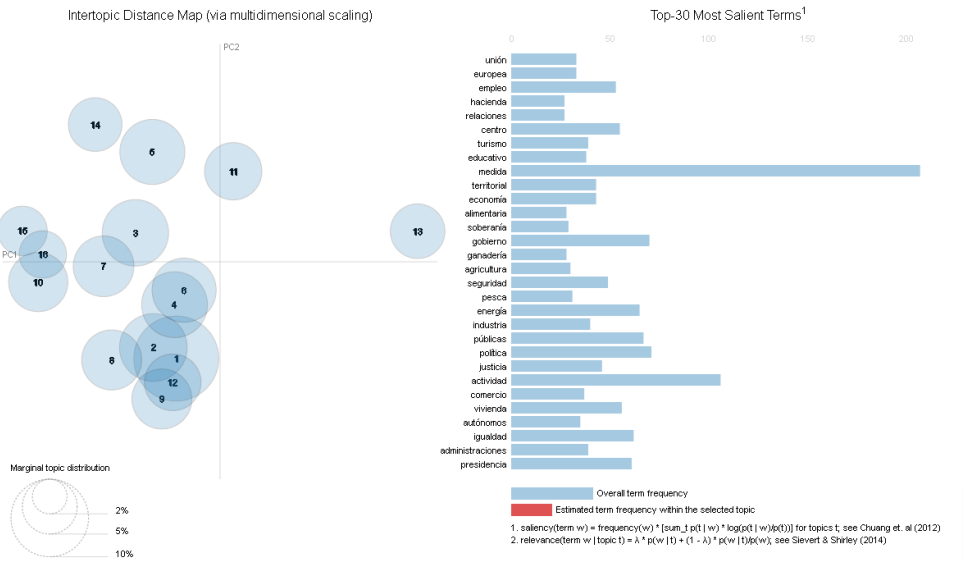


Figura 5.12: Gráfica de temas del quinto grupo parlamentario

- **Política:** 3
- **Unión Europea:** 1
- **Asuntos sociales:** 2
- **Educación y comunicación:** 5
- **Trabajo y empleo:** 2
- **Medio Ambiente:** 1
- **Agricultura, Silvicultura y pesca:** 1

Y aquí podemos ver como se rompe la tendencia de ser los asuntos sociales la prioridad principal en favor de la Educación y comunicación, que para este grupo se convierte en la prioridad principal, demostrando así que este grupo está más interesado en la formación de las personas de las islas, por encima de cuestiones sociales, que también aparecen pero en menor medida.

Para el sexto grupo parlamentario, que tiene 13 temas, las agrupaciones de temas quedan de la siguiente forma:

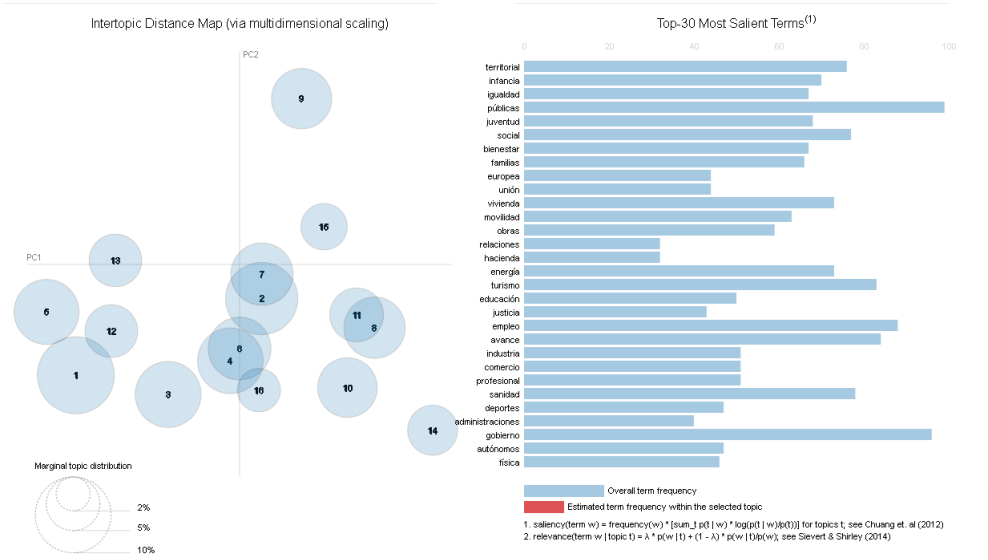


Figura 5.13: Gráfica de temas del sexto grupo parlamentario

- **Política:** 2
- **Unión Europea:** 2
- **Derecho:** 1
- **Asuntos sociales:** 2
- **Educación y comunicación:** 1
- **Trabajo y empleo:** 2
- **Energía:** 2
- **Industria:** 1

En este caso podemos comprobar como según el modelo este grupo parlamentario no presenta ningún tema como prioridad principal, siendo que todos los temas que tratan en este grupo parlamentario obtienen la misma prioridad, lo que indica que el grupo intenta abarcar múltiples temas que son importantes para la sociedad, como la educación o el empleo.

En el caso del último grupo, el séptimo, tiene 13 temas y las agrupaciones de temas serían:

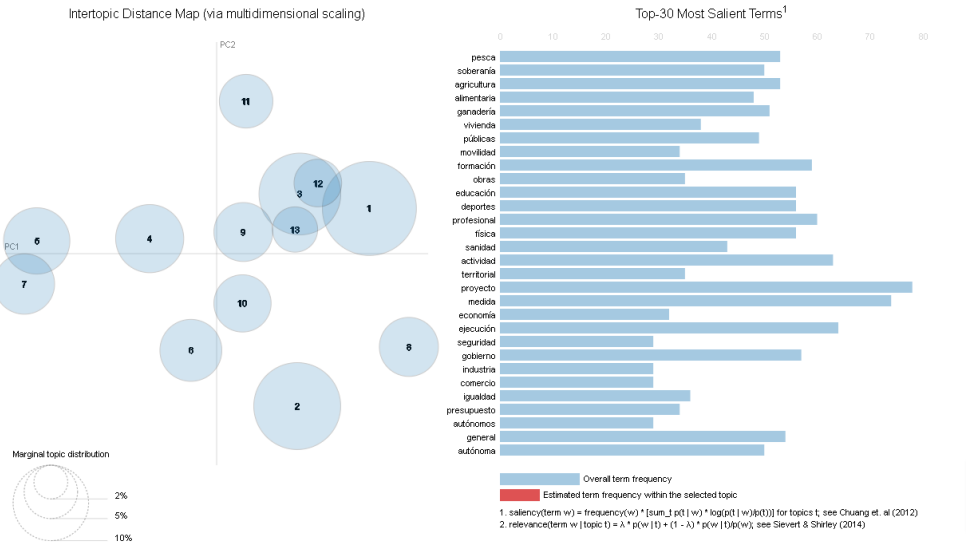


Figura 5.14: Gráfica de temas del séptimo grupo parlamentario

- **Política:** 2
- **Economía:** 1
- **Finanzas:** 1
- **Asuntos sociales:** 5
- **Educación y comunicación:** 1
- **Trabajo y empleo:** 1
- **Agricultura, Silvicultura y pesca:** 2

Este grupo vuelve seguir la tendencia de tener como prioridad el campo de los asuntos sociales, por lo que buscan el bienestar social de los ciudadanos de las islas, y además también buscan una buena economía y gestión monetaria, tanto para el sector público como el privado, entendiéndose por lo tanto que también están preocupados por la situación laboral y económica en las islas.

En líneas generales se podría decir que el análisis de los temas creados por el modelo LDA ha mostrado como todos los grupos parlamentarios están especialmente preocupados por los asuntos sociales del archipiélago, en mayor o menor medida debido a la diferencias en el número de propuestas presentadas. Siendo este campo seguido de cerca por los temas de trabajo y educación, unos temas que actualmente tienen muy preocupados a la sociedad canaria, además del tema medioambiental que en los últimos años está siendo cada vez mas seguido por la ciudadanía, consiguiendo este interés que se trate con mayor frecuencia en nuestro parlamento.

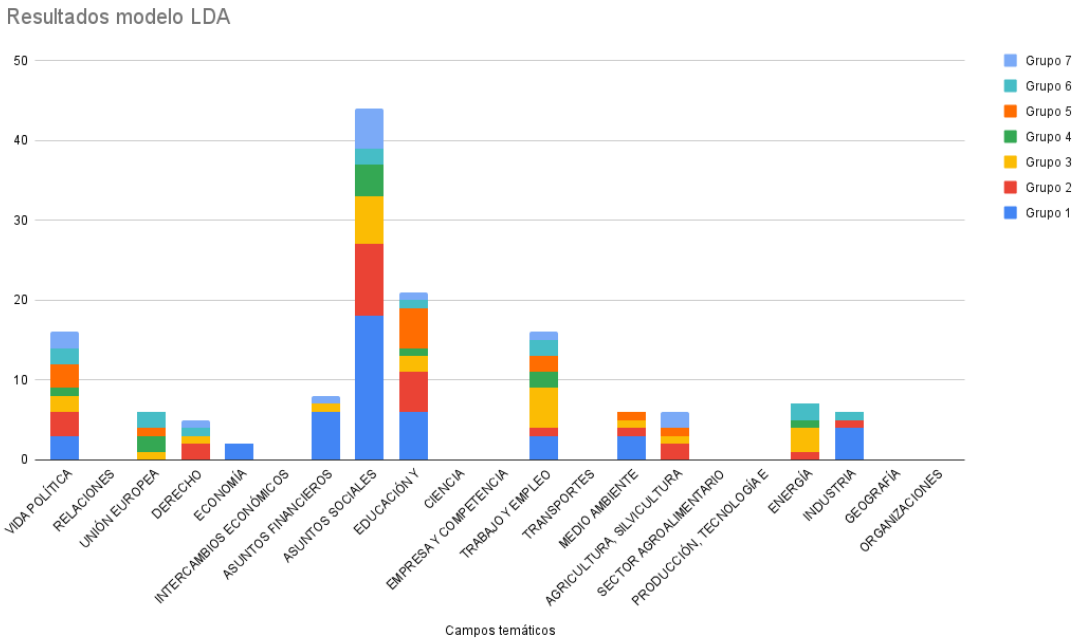


Figura 5.15: Distribución de temas por campo temático y grupo parlamentario

5.2.2. Gráficas del modelo BERTopic

Una vez evaluado los temas del modelo LDA, se hará la evaluación de los temas obtenidos a través del modelo BERTopic, un modelo que se observa que obtiene unos modelos en principio más coherentes que en el caso del modelo anterior, tal y como se indicó en la evaluación cuantitativa. De la misma manera que se hizo anteriormente, se procederá a analizar los temas obtenidos por cada grupo parlamentario. Teniendo en cuenta que se ha buscado que los dos modelos trabajen en las mismas condiciones y generando el mismo número de temas, en este modelo los grupos tienen el mismo número de estos que en el modelo LDA. Como en la sección anterior, por cada grupo parlamentario, se ejemplificará la visualización de temas con un extracto de la gráfica de temas con las palabras más relevantes.

En el primer grupo parlamentario analizado los temas se agrupan en los siguientes campos temáticos:

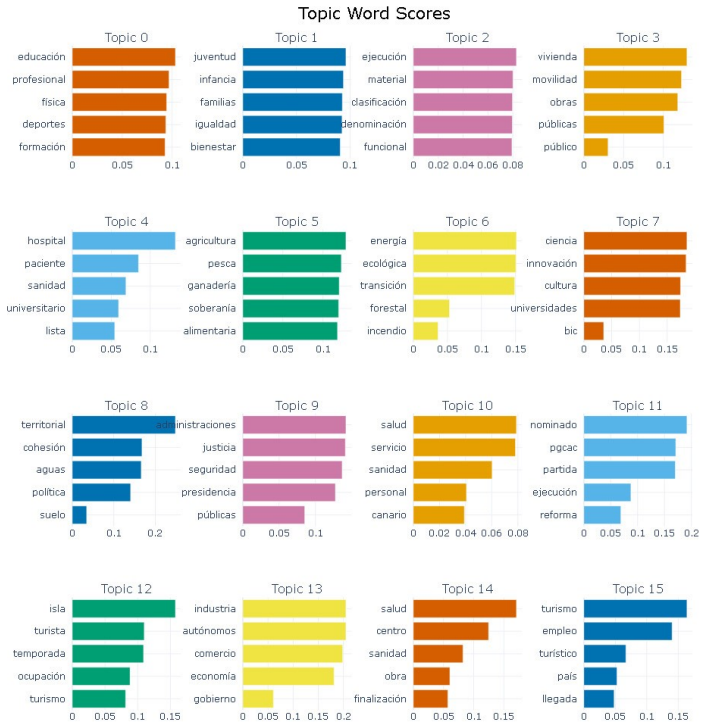


Figura 5.16: Ejemplo de temas del primer grupo parlamentario

- **Política:** 4
- **Unión Europea:** 1
- **Economía:** 1
- **Finanzas:** 2
- **Asuntos sociales:** 15
- **Educación y comunicación:** 2
- **Ciencia:** 1
- **Empresa y competencia:** 2
- **Trabajo y empleo:** 5
- **Transportes:** 3
- **Medio Ambiente:** 2
- **Agricultura, Silvicultura y pesca:** 1
- **Producción,tecnología e investigación:** 2
- **Energía:** 2

Se puede ver que el campo más prioritario para este grupo es el de asuntos sociales, siendo el que más temas contiene, pero aun así no es el único que se trata ya que además se presentan una gran variedad de campos, indicando que el grupo presenta un amplio espectro de intereses y que trata de llevar al parlamento todos los aspectos que mas están importando a la sociedad en la actualidad.

Para el segundo grupo parlamentario, se obtiene la siguiente clasificación temática:

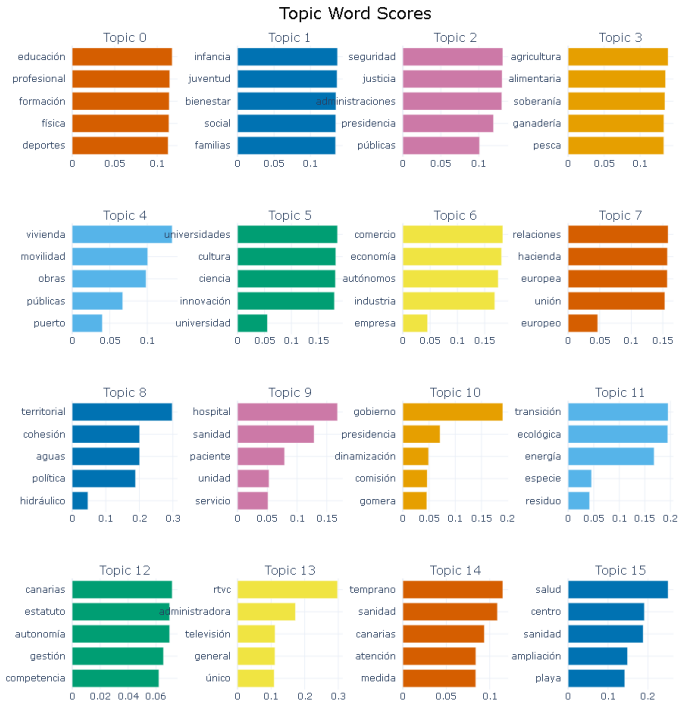


Figura 5.17: Ejemplo de temas del segundo grupo parlamentario

- **Política:** 3
- **Unión Europea:** 1
- **Derecho:** 2
- **Comercio:** 1
- **Asuntos sociales:** 8
- **Educación y comunicación:** 3
- **Trabajo y empleo:** 1
- **Transportes:** 1
- **Medio Ambiente:** 1
- **Agricultura, Silvicultura y pesca:** 1
- **Energía:** 1

Este grupo también prioriza los asuntos sociales, y trata varios temas, no tan variados como el primero, aunque eso pueda deberse al menor número de temas generados para este grupo, pero aún así todos los temas en mayor o menor medida resultan interesantes y son temas importantes para la sociedad canaria.

Para el siguiente grupo, los temas se ordenan de la siguiente forma:

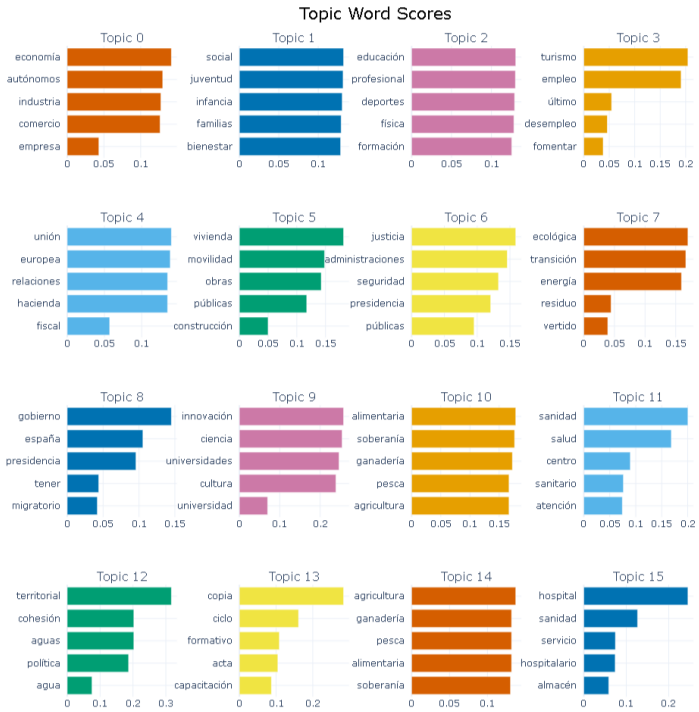


Figura 5.18: Ejemplo de temas del tercer grupo parlamentario

- **Política:** 1
- **Unión Europea:** 1
- **Derecho:** 2
- **Finanzas:** 1
- **Asuntos sociales:** 7
- **Educación y comunicación:** 3
- **Ciencia:** 1
- **Trabajo y empleo:** 3
- **Transportes:** 1
- **Medio Ambiente:** 1
- **Agricultura, Silvicultura y pesca:** 2
- **Sector agroalimentario:** 1
- **Energía:** 1

En este grupo a parte de su prioridad principal, podemos ver como destacan ligeramente 2 temas, que son educación y trabajo, dos temas bastante importantes actualmente y que rápidamente están cogiendo fuerza dentro de la sociedad.

El cuarto grupo parlamentario obtiene la siguiente clasificación a partir de los temas obtenidos:

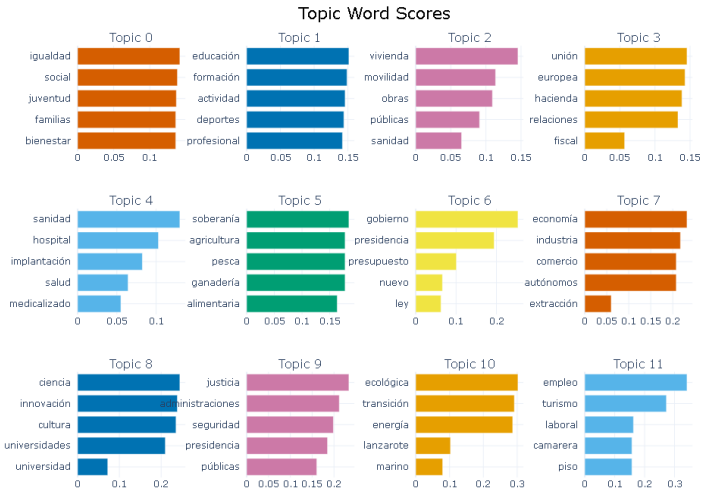


Figura 5.19: Ejemplo de temas del cuarto grupo parlamentario

- **Política:** 1
- **Unión Europea:** 1
- **Derecho:** 1
- **Finanzas:** 1
- **Asuntos sociales:** 3
- **Educación y comunicación:** 2
- **Trabajo y empleo:** 1
- **Agricultura, Silvicultura y pesca:** 1
- **Energía:** 1

En este grupo, se encuentra que no hay una clara prioridad por un campo en concreto, sino que tratan de llevar al parlamento varios temas que pueden ser bastante importantes para todo el archipiélago, como el trabajo o el control financiero de las cuentas.

Para el quinto grupo, se consigue la siguiente clasificación:

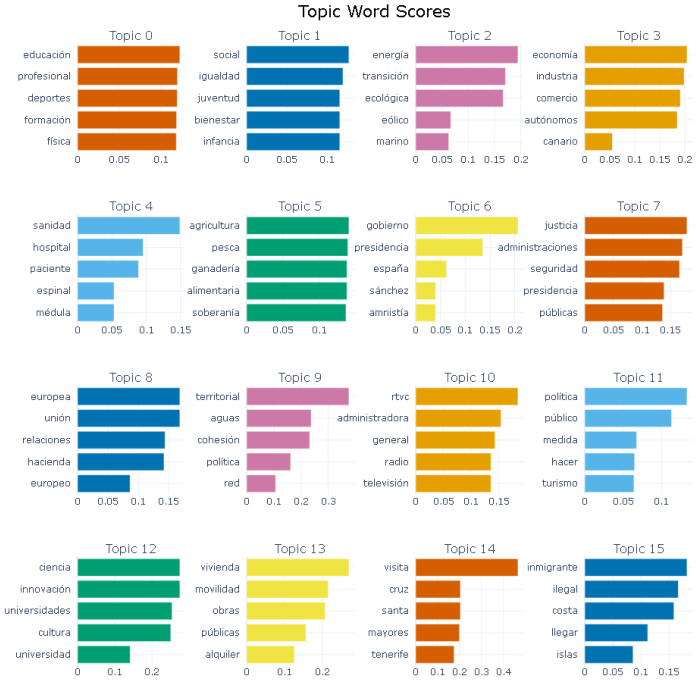


Figura 5.20: Ejemplo de temas del quinto grupo parlamentario

- **Política:** 2
- **Unión Europea:** 1
- **Derecho:** 2
- **Economía:** 1
- **Asuntos sociales:** 6
- **Educación y comunicación:** 2
- **Ciencia:** 1
- **Agricultura, Silvicultura y pesca:** 1
- **Energía:** 1

En este grupo volvemos a ver como los asuntos sociales vuelve a ser prioritario para este grupo parlamentario, aun así vemos que este grupo también vemos que toca otros temas, siendo los temas igualmente importantes y de gran relevancia.

El sexto grupo parlamentario obtiene la siguiente clasificación:

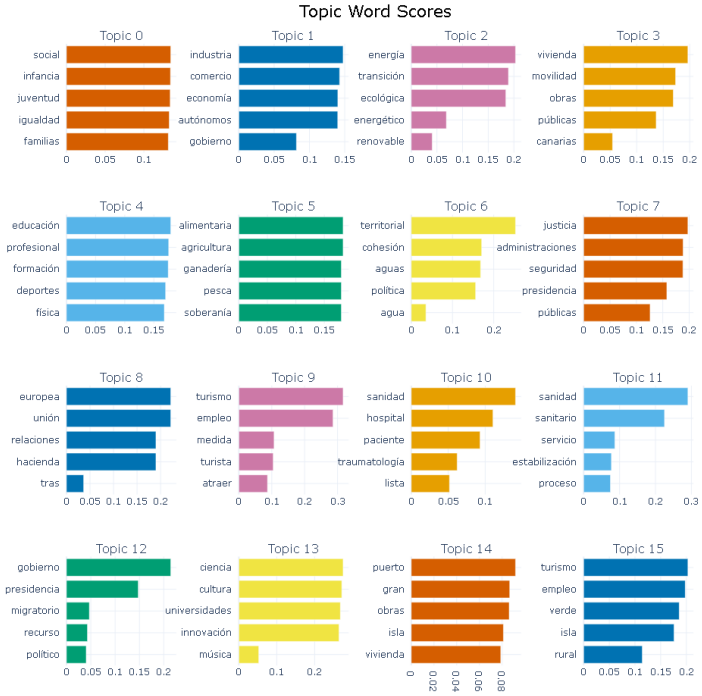


Figura 5.21: Ejemplo de temas del sexto grupo parlamentario

- **Política:** 2
- **Unión Europea:** 1
- **Derecho:** 1
- **Asuntos sociales:** 5
- **Educación y comunicación:** 2
- **Trabajo y empleo:** 1
- **Transporte:** 1
- **Agricultura, Silvicultura y pesca:** 1
- **Energía:** 1
- **Industria:** 1

Este grupo muestra una gran variedad de temas, más allá de la prioridad general que se presenta en casi todos los grupos, además presenta otros temas igual de interesantes e importantes, como el de la agricultura, un tema que está siendo de actualidad al verse afectado varios productos que se producen en este sector, aumentando su precio o en el lado más extremo, escaseando.

Por último, se procede a mostrar la clasificación de temas del séptimo grupo parlamentario:

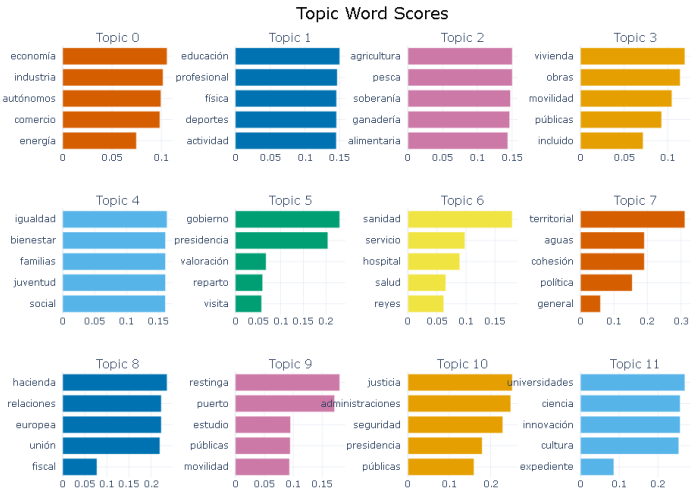


Figura 5.22: Ejemplo de temas del séptimo grupo parlamentario

- **Política:** 1
- **Unión Europea:** 1
- **Derecho:** 2
- **Economía:** 1
- **Asuntos sociales:** 3
- **Educación y comunicación:** 2
- **Transporte:** 1
- **Agricultura, Silvicultura y pesca:** 1

En este grupo se ve una equidad en el trato de los diferentes campos, a pesar de que este grupo tiene pocos, despuntando ligeramente los asuntos sociales. Estos valores demuestran que este grupo intenta tratar varios temas relevantes para la sociedad.

Los grupos parlamentarios tratan, según los temas obtenidos por el modelo BERTopic, una amplia variedad de temas. Si bien casi todos los grupos despuntan en el campo de los asuntos sociales(que es un tema que tiene una gran importancia para cualquier sociedad) también podemos abordar otros temas que son muy importantes para la sociedad canaria, como el sector primario, el trabajo o la economía y los presupuestos públicos.

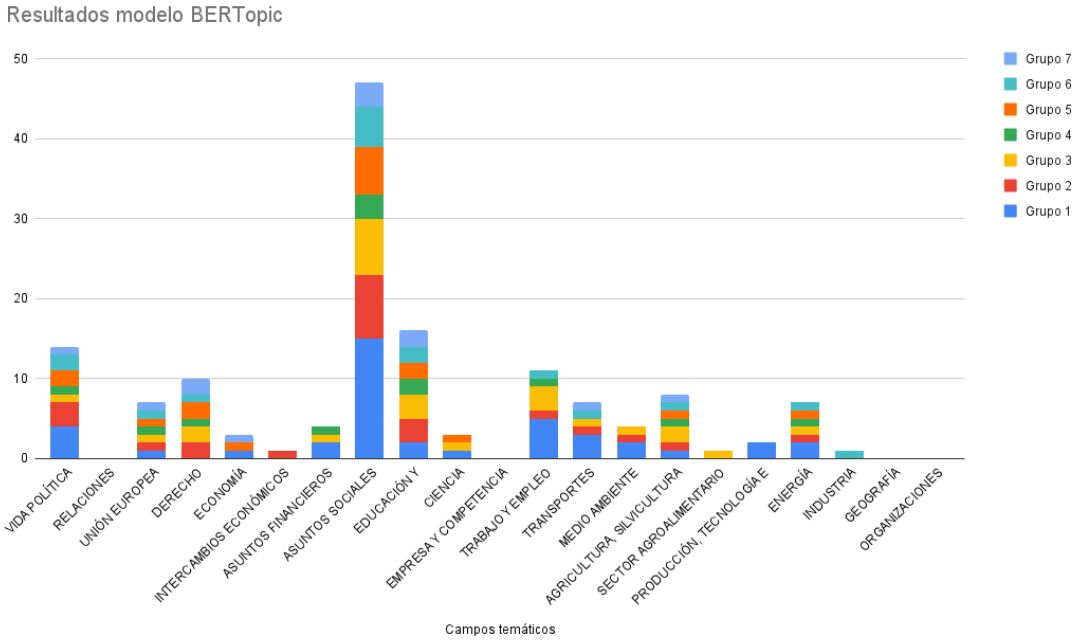


Figura 5.23: Distribución de temas por campo temático y grupo parlamentario

Como se indicó al inicio de esta sección, se ha creado una gráfica para mejorar la visualización de estos resultados:

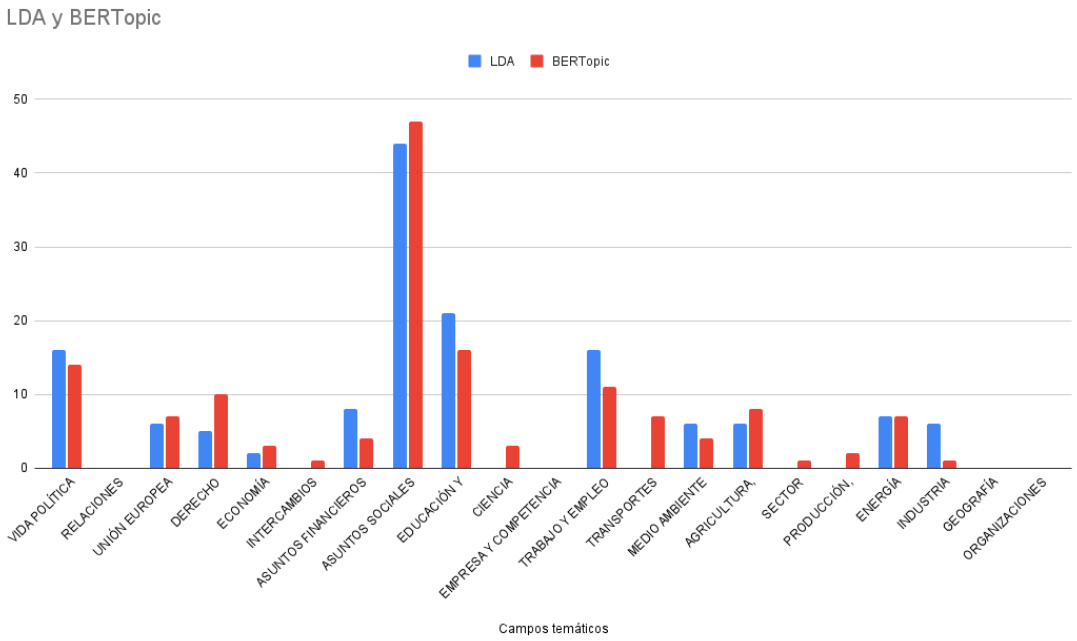


Figura 5.24: Distribución de temas por campo temático y modelo

Capítulo 6

Conclusiones y líneas futuras

6.1. Conclusiones

Habiendo analizado en el capítulo anterior los resultados de las métricas obtenidas por ambos modelos, se puede proceder a analizar cual de los dos modelos ha tenido mejores resultados en la evaluación cuantitativa. Para esta evaluación final solo se utilizará las dos métricas que comparten ambos modelos, que son coherencia y diversidad, que además ambas pueden ser iguales en los dos modelos, implicando que ambos modelos pueden lograr resultados similares en términos de cohesión y variedad de tópicos . La perplejidad no es una buena métrica para evaluar en BERTopic, ya que este modelo no es probabilístico, siendo la perplejidad un evaluador de la calidad de este tipo de modelos y el *silhoutte score* no es una métrica apropiada para evaluar el modelo LDA, ya que esta evaluación se basa en calcular la cohesión interna de un *cluster* y la separación entre estos, y teniendo en cuenta que LDA es un modelo probabilístico, no hace una asignación única y discreta de documentos a *cluster*, como sí hace BERTopic.

Por un lado, tenemos la coherencia de los temas, que en el modelo LDA varía entre los valores 0.26 y 0.41 mientras que BERTopic presenta una variación en esta métrica de entre 0.56 y 0.72, con lo que podemos ver que BERTopic obtiene mejores resultados que LDA en este aspecto, sino que también podemos decir que el modelo es más coherente.

En el otro lado, tenemos la diversidad de los temas, que en LDA varia entre 0.31 y 0.67, mientras que en BERTopic varía entre 0.64 y 0.89, lo que aunque LDA pueda llegar a alcanzar los valores de BERTopic, en líneas generales este último crea temas más diferentes en comparación con el primero, lo que también es beneficioso para análisis de los temas ya que permite al lector tener una visión más completa de los datos.

Es por todo esto que se ha explicado que podemos concluir en que BERTopic en este caso obtiene una mejor evaluación cuantitativa que LDA y por lo tanto, también se puede decir que en este caso en concreto que manejamos que es mejor que LDA.

Es evidente, dado el volumen mencionado que sigue en aumento, que el *machine learning*, combinado con un buen procesamiento del lenguaje, capacita para abordar la tarea presentada en este proyecto de una manera precisa y en un tiempo considerablemente menor al requerido por un ser humano para hacer esta misma tarea con la misma exactitud.

En otro orden, siendo esta una herramienta que analiza propuestas parlamentarias para identificar los temas más abordados por los diferentes grupos políticos, es prudente considerar que, junto con los beneficios que puede aportar, como la capacidad de entender mejor a cada partido alejándose de sesgos preconcebidos sobre el mismo, también es necesario contemplar posibles usos inapropiados. Por ejemplo, podría utilizarse para analizar temas previamente abordados por un partido con el fin de influir en negociaciones para obtener ventajas injustas. Incluso se puede conseguir que la herramienta realice predicciones sobre las tendencias políticas que tomaran todos los grupos parlamentarios, abriendo un abanico de posibilidades que, si no se manejan adecuadamente, podrían generar más problemas de los que resuelven.

Además, tras la evaluación cualitativa y habiendo revisado los temas de cada grupo generado por ambos modelos, además de todas las gráficas que se han creado para ambos modelos, como los mapas de distancias entre tópicos o los árboles de *clustering* jerárquico, se puede concluir que los resultados obtenidos por la evaluación cuantitativa son bastante acertados, pues aunque los temas obtenidos por el modelo LDA aunque son medianamente coherentes, y por lo tantos una persona podría comprenderlos, genera los temas combinando muchos términos que en varias ocasiones no están relacionados, generando temas que pueden llegar a ser complicados de entender.

Todos los problemas que se vieron reflejados en la evaluación anterior se pueden ver también en esta, obteniendo temas mejor estructurados en BERTopic y que generan mejores resultados en el análisis, más cercanos a la realidad y también mostrando más diversidad de campos temáticos, cosas que no se podían conseguir con el modelo LDA, consiguiendo que en todos los gráficos comprobados, el modelo BERTopic consiga mejores temas.

Con todo los resultados obtenidos en el capítulo anterior, se puede concluir que el modelo BERTopic constituye una mejor opción en este contexto que el modelo LDA, obteniendo en ambas evaluaciones unos resultados mejores siendo en algunos casos diferencias bastante significativas.

En este proyecto hemos comprobado que con un conjunto de datos extenso y en constante crecimiento puede generar resultados sumamente interesantes, permitiendo conocer mejor las tendencias políticas de los parlamentos. En el caso concreto que se maneja en este proyecto se ha podido averiguar que en el Parlamento de Canarias se tratan temas que no solo son de actualidad, sino también temas que se mantienen en el tiempo como una preocupación por la sociedad, siendo esta una tendencia que comparten todos los grupos políticos, observándose que los políticos están, desde sus respectivas posiciones ideológicas, también preocupados por los mismos problemas que el resto de ciudadanos del archipiélago.

Todo lo comentado anteriormente genera un debate que podría convertirse en un tema crucial en los próximos años, especialmente considerando que ya se están explorando implementaciones en diversos parlamentos. Es probable que el análisis e implementación del *machine learning* lleguen a formar parte de los propios procesos parlamentarios, y este será un tema que se debatirá y estudiará detenidamente en los próximos años.

6.2. Líneas y futuras

Este proyecto presenta múltiples posibilidades de ser mejorado en la parte del software.

Para empezar, la forma más evidente de mejora sería permitir la visualización de los datos a cualquier ciudadano en tiempo real, usando como forma más rápida y accesible una aplicación web o una aplicación móvil, desarrollada para que vuelque los datos que se obtienen como resultado en este trabajo permitiendo a cualquier persona ver los resultados sin tener mucho conocimiento del mundo de la informática y la programación.

Otra mejora es el desarrollo de una mayor visualización con información adicional que mejore el entendimiento de los datos mostrados, lo que podría combinarse con lo expuesto anteriormente para mejorar la experiencia de usuario y que los gráficos sean fácilmente comprensible por los usuarios que los vean.

También sería una mejora hacer un análisis previo a realizar las dos mejoras anteriores sería añadir nuevos modelos para comprobar si hay mejores modelos que puedan ser utilizados en sustitución de los ya implementados.

Finalmente, se podría intentar crear un chatbot que con los datos analizados por los modelos explique todos los resultados obtenidos y al que le puedas preguntar cualquier duda con respecto a los mismos o sobre cualquier aspecto relacionado con su obtención.

Capítulo 7

Summary and Conclusions

7.1. Summary

This project starts with the idea to analize parliamentary initiatives of the Parliament of the Canary Islands using one of the most requested branch of the artificial intelligence in this days, which is machine learning, a type of artificial intelligence that enables computers to learn from data without being programmed.

The software that we are going to develop must be able to connect to the Parliament's API and extract the initiatives and then extract the topics to make all the graphics that allow us to analize this information.

Before beginig the development we need to specify all the requirements necesary to success in the development of this project, thorough analysis of the available tools.

After the first analysis, we can set out a group of various tasks, first of all obtaining the data through the Parliament of the Canary Islands' API,and when we start the development, we must refine this iniciatives using Natural Languaje Processing obtaining all the words in a format that the models can use, and after that we can start the development of the models.

Finally we need to evaluate the results that we obtained from the models to establish which one obteined better results.

7.2. Conclusions

In this project we have verified that with an extensive and constantly growing data set it can generate quite interesting results, allowing us to better understand the political tendencies of parliaments.

It is evident seeing the size that, as already mentioned, continues to grow, that machine learning, combined with good language processing, allows us to carry out the task presented in this project in a precise way and in a much shorter time than it would cost a human being to do this same task with the same precision.

In another order, this being a tool that analyzes parliamentary proposals to find out the issues that the different parliamentary groups deal with the most, it is logical to think that along with all the good things that can be obtained from it, such as the ability to better understand each party by moving away of preconceived biases about it, it is also easy to see that in the wrong hands it can result in things such as analyzing topics used until now by a party to approach negotiations with it, taking advantage of it to come out favored, and you can even get the tool to carry out predictions about the political trends that all parliamentary groups will take, opening a range of possibilities that if poorly controlled could generate more problems than it solves.

Everything mentioned above generates a debate that in the coming years could become an important topic since we are already witnessing how it is being studied and how to implement it in different parliaments and it is even likely that the analysis and implementation of machine learning will reach the parliaments themselves. parliaments to be debated in the coming years.

Capítulo 8

Presupuesto

Análisis y diseño			
Tarea	**Nº de horas**	**Precio por hora**	**Precio Total**
Análisis de requisitos	10	9€	90€
Estudio de conjunto de datos	15	10€	150€
Precio total = 240€			

Tabla 8.1: Presupuesto Análisis y diseño

Implementación y desarrollo			
Tarea	**Nº de horas**	**Precio por hora**	**Precio Total**
Obtención de datos	10	11€	110€
Desarrollo de software	64	13€	832€
Resolución de errores	50	13€	650€
Documentación	6	10€	60€
Pruebas	40	13€	520€
Precio total = 2172€			

Tabla 8.2: Presupuesto Implementación y desarrollo

Precio total del proyecto = 240€ + 2172€ = 2312€

Bibliografía

[1] *Inteligencia artificial: ejemplos de dilemas éticos*. 24 de abr. de 2023. url: https://www.unesco.org/es/artificial-intelligence/recommendation-ethics/cases.

[2] Francisco Gonzalez. *El Rol de la Inteligencia Artificial en la Actualidad*. 24 de mar. de 2023. url: https://www.linkedin.com/pulse/el-rol-de-la-inteligencia-artificial-en-actualidad-francisco-gonz%C3%A1lez/.

[3] Muthanna Saari. "IR 4.0 in Parliament: Conceptualising the application of artificial intelligence and machine learning in the Parliament of Malaysia's parliamentary questions". En: *International Journal of Law Government and Communication* 5.20 (2020), págs. 124-137.

[4] Jörn von Lucke y Fotios Fitsilis. "Using Artificial Intelligence in Parliament-The Hellenic Case". En: *International Conference on Electronic Government*. Springer. 2023, págs. 174-191.

[5] Iberdrola Corporativa. *Descubre los principales beneficios del Machine Learning*. url: https://www.iberdrola.com/innovacion/machine-learning-aprendizaje-automatico.

[6] Vladimir Nasteski. "An overview of the supervised machine learning methods". En: *Horizons. b* 4.51-62 (2017), pág. 56.

[7] Samreen Naeem et al. "An unsupervised machine learning algorithms: Comprehensive review". En: *International Journal of Computing and Digital Systems* (2023).

[8] YCAP Reddy, P Viswanath y B Eswara Reddy. "Semi-supervised learning: A brief review". En: *Int. J. Eng. Technol* 7.1.8 (2018), pág. 81.

[9] Leslie Pack Kaelbling, Michael L Littman y Andrew W Moore. "Reinforcement learning: A survey". En: *Journal of artificial intelligence research* 4 (1996), págs. 237-285.

[10] Outside Insight. *How AI helps Spotify win in the music streaming world - Outside Insight*. 26 de nov. de 2018. url: https://outsideinsight.com/insights/how-ai-helps-spotify-win-in-the-music-streaming-world/.

[11] Allen Yu. "How Netflix Uses AI, Data Science, and Machine Learning — From A Product Perspective". En: (8 de dic. de 2021). url: https://becominghuman.ai/how-netflix-uses-ai-and-machine-learning-a087614630fe.

[12] Ahmad Abdulkader, Aparna Lakshmiratan y Joy Zhang. *Introducing DeepText: Facebook's text understanding engine*. 26 de jun. de 2018. url: https://engineering.fb.com/2016/06/01/core-infra/introducing-deeptext-facebook-s-text-understanding-engine/.

[13] *¿Qué es el procesamiento del lenguaje natural (PLN)?* url: https://www.ibm.com/es-es/topics/natural-language-processing.

[14] GeeksforGeeks. *Latent Dirichlet Allocation*. 6 de jun. de 2021. url: https://www.geeksforgeeks.org/latent-dirichlet-allocation/.

[15] Maarten P. Grootendorst. *The Algorithm - BERTopic*. url: https://maartengr.github.io/BERTopic/algorithm/algorithm.html.