Universidad Nacional de Educación a Distancia
Escuela Técnica Superior de Ingeniería Informática

Trabajo Fin de Máster del
Máster en Ingeniería y Ciencia de Datos

Agrupación automática de mensajes de foros

Autor Martín Priego Wood

Directores Álvaro Rodrigo Yuste
Víctor Fresno Fernández

Curso 2023-2024 1.ª Convocatoria

Resumen

Los foros de discusión permiten formular preguntas y obtener respuestas aprovechando la denominada sabiduría de las masas, y se han convertido en herramientas esenciales de cursos en línea, como los de la UNED. Los foros suelen estar divididos en subforos dedicados a temas específicos, pero a menudo los usuarios escriben mensajes en el subforo equivocado, lo que dificulta su visibilidad y puede hacer necesaria una reubicación manual. Para ayudar a prevenir estos errores y aliviar las tareas de mantenimiento, en este trabajo se desarrolla un sistema que agrupa automáticamente foros como los de la UNED y permite medir la similitud semántica entre mensajes. Asimismo, dada una estructura de subforos llena de mensajes y un mensaje nuevo, el sistema es capaz de generar recomendaciones de inserción basadas en similitud.

El trabajo incluye una parte investigativa fundamental en la que se lleva a cabo un análisis exploratorio de 7 foros de la UNED y se experimenta con diversas técnicas de procesamiento de lenguaje natural y de aprendizaje no supervisado. Por ejemplo, se ensaya con representaciones vectoriales de documentos de tipo bolsa de palabras así como con otras más modernas, como los *embeddings* de palabras e incluso de frases. Los mejores resultados se obtienen con versiones ponderadas de la bolsa de palabras y con modelos multilingües de codificación de frases pre-entrenados. En cuanto a la similitud entre mensajes, las métricas coseno y angular producen resultados parecidos, mas la segunda tiene la posible ventaja de ser propiamente una distancia. Por último, se prueban los algoritmos de *clustering* k-medias, aglomerativo y HDBSCAN, que también es jerárquico pero basado en densidad. Los agrupamientos se evalúan usando medidas externas, como la información mutua ajustada, y también internas, como la silueta y el índice de validación basado en densidad. El algoritmo k-medias consigue el mejor alineamiento medio con la estructura de subforos original, pero los otros dos tienen asimismo ventajas, en cuanto a tiempo de ejecución y la información adicional que proporcionan sus jerarquías. El método HDBSCAN destaca por su flexibilidad, robustez y el carácter intuitivo de sus parámetros.

El sistema de agrupación desarrollado es capaz de identificar por sí solo grupos que tienen pleno sentido. En ocasiones, dichos grupos son subconjuntos de un subforo original, e incluso pueden ser parientes cercanos de otros subconjuntos del mismo subforo en un agrupamiento jerárquico. Otras veces, los grupos generados son transversales a la estructura original, debido a la presencia de mensajes semejantes, por ejemplo agradecimientos, a través de los subforos. Aun cuando la estructura original resulte difícil de reproducir automáticamente, el *ranking* de similitud creado por el sistema debería de facilitar la colocación correcta de mensajes nuevos.

Palabras clave: foro, agrupamiento, *clustering*, similitud, bolsa de palabras, *embedding*, procesamiento de lenguaje natural

Abstract

Discussion forums enable asking questions and obtaining answers through the wisdom of the crowd, and have become an essential tool in online courses, such as those run by UNED. Forums are typically divided in subforums devoted to specific topics, but users frequently write messages in the wrong subforum, which can hinder visibility and require manual relocation. To help prevent these errors and ease maintenance tasks, this work presents development of a system that automatically clusters messages from forums like those from UNED and enables measuring semantic similarity between messages. In addition, given a structure of subforums prefilled with messages and a new message, the system can generate recommendations for its assignment based on similarity.

This work includes a fundamental investigative part consisting of an exploratory analysis of 7 UNED forums and experiments with various natural language processing and unsupervised learning techniques. For example, bag-of-words models are tried out along with more modern vector representations, such as word and sentence embeddings. The best results are obtained with weighted versions of the bag of words and with pretrained multilingual sentence encoders. With regard to message similarity, the cosine and angular metrics yield similar results, but the latter has the advantage of being a genuine distance. Lastly, the clustering algorithm trials cover k-means, agglomerative and HDBSCAN, which is also hierarchal but based on density. Clusters are evaluated using external measures, such as adjusted mutual information, and also intrinsic measures, such as the silhouette score and the density-based cluster validation index. The k-means algorithm achieves the best average alignment with the original structure of subforums, but the other two algorithms also have their own advantages, in terms of execution time and the additional information provided by their hierarchies. The HDBSCAN method stands out because of its flexibility, robustness, and the intuitive nature of its parameters.

The developed clustering system is capable of autonomously identifying some meaningful clusters. Sometimes, those clusters are subsets of an original subforum, and may even be close relatives of other subsets of the same subforum in a hierarchal clustering. Other times, the generated clusters are transverse to the original structure, because of the presence of similar messages, for instance thank-yous, across subforums. Even in those cases where the original structure is difficult to reproduce automatically, the similarity raking created by the system should facilitate the correct placement of new messages.

Keywords: forum, clustering, similarity, bag of words, embedding, natural language processing

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

Los foros se han convertido en un medio de comunicación que permite recibir respuestas por parte de otros usuarios aprovechando la denominada sabiduría de las masas. Algunos de los portales de este tipo más conocidos son el generalista Answers.com y el más técnico Stack Overflow. Los foros también son un recurso importante en cursos virtuales, ya sean abiertos, como muchos de Coursera y edX, o restringidos, como los estudios oficiales de la UNED.

Los foros suelen estar divididos en subforos dedicados a temáticas específicas. Sin embargo, a menudo los usuarios escriben mensajes el subforo equivocado, lo que dificulta su visibilidad y en ocasiones hace necesario que los administradores reubiquen estos mensajes manualmente. En virtud de los avances de las últimas décadas, dichas tareas manuales deberían de poder superarse o al menos aliviarse recurriendo a técnicas de procesamiento de lenguaje natural (NLP) y aprendizaje automático.

El principal objetivo de este trabajo es desarrollar un sistema que permita medir la similitud semántica entre los mensajes de un foro de la UNED y agruparlos automáticamente en base a dicha similitud. Asimismo, dada una estructura de subforos con mensajes ya colocados y un mensaje nuevo, el sistema deberá ser capaz de generar recomendaciones de inserción basadas en similitud y ofrecer un *ranking* de los subforos más verosímiles para el mensaje nuevo. El trabajo tiene a la vez un objetivo de investigación, referido a la exploración y caracterización de los foros concretos y a la búsqueda de las técnicas que mejor se adecúen a ellos. El foco se ha puesto en la similitud semántica y el *clustering* porque se considera que los datos disponibles podrían ser insuficientes para el entrenamiento de clasificadores bajo un criterio temático. En parte, se trata de aproximar y comparar la clasificación temática con el agrupamiento basado en similitud, comprendiendo que la correspondencia será normalmente imperfecta debido a la diferencia de criterio. El planteamiento basado en la similitud semántica y la agrupación tiene además la ventaja de ser de aplicación más general, puesto que no requiere datos etiquetados.

Las técnicas de NLP llevan tiempo aplicándose a tareas relacionadas con foros o comunidades de preguntas y respuestas, como pueden ser la identificación de preguntas duplicadas o la búsqueda y clasificación de las respuestas más adecuadas a una cierta pregunta. El estudio de

Patra (2017) repasa las dificultades asociadas a estas dos tareas y los métodos propuestos para resolverlas. Por ejemplo, para la identificación de preguntas duplicadas cita una técnica basada en representación por bolsa de palabras (BOW) y ajuste por frecuencia inversa de documento (IDF), otra basada en modelos probabilísticos de traducción de lenguaje, y dos más basadas en *embeddings* de palabras y combinadas, bien con perceptrones multicapa, o bien con redes neuronales con mecanismo de atención. En cuanto a la búsqueda de las mejores respuestas, el artículo menciona técnicas análogas a las anteriores, basadas en representaciones BOW o en modelos probabilísticos traducción, así como otras que combinan *embeddings* de palabras con redes convolucionales (CNN), redes de memoria larga a corto plazo (LSTM), con o sin atención, o incluso con redes CNN profundas. En algunos casos, los *embeddings* de palabras vienen ya entrenados mediante word2vec (Mikolov *et al.*, 2013a,b), mientras que en otros se entrenan conjuntamente con el modelo. Las pruebas realizadas en el estudio indican que los modelos basados en representación BOW o en modelos de traducción son inferiores a los basados en *embeddings* y redes neuronales especializadas. Esto se debe a que, a diferencia de los dos primeros tipos de modelos, los basados en *embeddings* son capaces de capturar relaciones semánticas. Además, el estudio concluye que los modelos que promedian los *embeddings* por palabra son inferiores a los que los combinan con redes CNN/LSTM, puesto que quizá los últimos eviten la dilución de las palabras más importantes en la media.

En un artículo más reciente, Onan y Toçoğlu (2021) presentan un modelo para la identificación automática de temas de preguntas en foros de discusión de cursos virtuales abiertos masivos (MOOC). Dado el enorme número de potenciales participantes en un MOOC, la identificación automática y priorización de temas de preguntas en los foros puede llegar a resultar esencial. Su estudio está bastante alineado con la temática y objetivos del presente trabajo, pero existen diferencias en cuanto a los foros analizados y algunos de los métodos empleados. Los foros del estudio de Onan y Toçoğlu son todos de cursos edX de programación y ciencia de datos, mientras que los considerados aquí son más variados y no siempre técnicos, como el curso de Acceso para Mayores. Además, los mensajes de Onan y Toçoğlu vienen etiquetados con categorías bastante genéricas, como pregunta de contenido, pregunta técnica, pregunta logística y las correspondientes categorías de respuesta. Aquí, en cambio, muchos subforos se refieren a asuntos más concretos, verdaderos temas, como el análisis léxico o la propiedad intelectual, aunque también los hay genéricos, como el subforo de cuestiones generales y el de estudiantes. En cuanto a los métodos, Onan y Toçoğlu utilizan, para la representación, *embeddings* de palabras ponderados por frecuencia o submuestreo y, para el agrupamiento, algoritmos como el k-medias, los mapas autoorganizados (SOM) y el jerárquico divisivo (DIANA). Los mejores resultados de agrupamiento, según el índice de Rand ajustado (ARI) y las informaciones mutuas normalizada (NMI) y ajustada (AMI), los obtienen mediante el *embedding* doc2vec (Le y Mikolov, 2014) y el *clustering* DIANA. En este trabajo también se ensaya con *embeddings* de palabras ponderados, como word2vec, pero además se prueban la representación tradicional BOW y los modernos *embeddings* de frases. Igualmente, se experimenta con

algoritmos de *clustering* clásicos, como el *k*-medias y el jerárquico aglomerativo, y con otros más recientes, como el jerárquico basado en densidad (HDBSCAN; Campello *et al.*, 2013). Por último, las medidas empleadas aquí para evaluar los agrupamientos no se restringen a las externas citadas anteriormente, sino que incluyen otras internas, como el coeficiente de silueta y el índice de validación basado en densidad (DBCV; Moulavi *et al.*, 2014).

Los *embeddings* de frases, mencionados en el párrafo anterior, pueden ser considerados el estado del arte en la representación de frases, párrafos o textos cortos, como son a menudo los mensajes de foros. A diferencia de los promedios de *emeddings* de palabras, ponderados o no, los *embeddings* de frases pueden interpretar correctamente el sentido de cada palabra en función de su contexto. Es más, puesto que registran el significado, el contexto y las relaciones entre palabras, son capaces de generar representaciones vectoriales semánticamente coherentes de frases completas. Dos de los modelos semánticos de frases más populares en este momento son SBERT (Reimers y Gurevych, 2019) y USE (Cer *et al.*, 2018). Para ambos, existen modelos preentrenados sobre grandes corpus de preguntas y respuestas, foros, noticias y enciclopedias y con aplicaciones en múltiples tareas de NLP, como la recuperación de información basada en similitud semántica y la clasificación y el agrupamiento de textos. De hecho, algunos paquetes de modelización de temas como BERTopic (Grootendorst, 2022) y Top2Vec (Angelov, 2020) integran estos modelos semánticos en *pipelines* de *embedding*, reducción dimensional opcional, agrupamiento y representación de temas. Si bien esta arquitectura se asemeja a la desarrollada aquí para el sistema de agrupación automática, se prescinde de estos paquetes a fin de profundizar en algunos componentes y añadir o suprimir otros.

La estructura del resto del trabajo es la siguiente. En el capítulo 2 se presenta y justifica el diseño del sistema de agrupamiento automático y colocación de mensajes nuevos. También se definen las medidas de similitud de documentos y evaluación de agrupamientos. Los componentes principales del sistema se van explicando y probando en los capítulos siguientes. En el capítulo 3 se detallan los módulos de lectura y limpieza de datos, y se lleva a cabo un análisis exploratorio de los foros de la UNED. El capítulo 4 trata sobre la representación vectorial de documentos. En él se introducen y aplican los 3 tipos de modelo de representación y, además, se prueba a calcular medidas internas sobre la estructura original de los foros, incluyendo una relacionada con la colocación de mensajes nuevos. En el capítulo 5 se explican los 3 algoritmos de agrupamiento considerados y se examinan los resultados de su aplicación en combinación con diferentes representaciones. El análisis incluye la exploración de los parámetros de los algoritmos, así como la inspección de mensajes de subforos y grupos de interés. El capítulo 6 presenta los programas definitivos para la exploración, agrupación automática y colocación de mensajes nuevos en foros con igual formato que los usados en este trabajo. Se describen los principales parámetros de entrada y se exhiben, a modo de ejemplo, capturas de una sesión típica de análisis. El capítulo 7 cierra el trabajo con un resumen de los resultados y objetivos alcanzados y con recomendaciones para futuras líneas de investigación.

Capítulo 2

Fundamentos

El problema de agrupación automática de documentos es bien conocido en la minería de textos y, además de en los artículos de investigación citados en la introducción, viene tratado en profundidad en referencias como Bilbro *et al.* (2018), Han *et al.* (2012) y Aggarwal y Zhai (2012). La agrupación automática, o *clustering*, consiste en la partición de un conjunto de elementos en grupos de modo que los elementos alojados en un mismo grupo se parezcan y los alojados en grupos distintos no se parezcan. El algoritmo de *clustering* es lógicamente la pieza fundamental de todo sistema de agrupación automática, y se tratará más adelante, en un capítulo separado. No obstante, las peculiaridades del texto como forma de representación de información, a diferencia, por ejemplo, de simples datos numéricos, hacen necesario el uso conjunto de otros componentes, como etapas de preproceso y modelos de representación, que también se detallarán en capítulos posteriores. A continuación, se explica cómo encajan unos componentes con otros y cómo, a grandes rasgos, se ha implementado el sistema. Asimismo, hacia el final, se describen las medidas elegidas para evaluar la similitud entre documentos y la calidad de los agrupamientos generados por el sistema.

2.1. Arquitectura

Los sistemas de agrupación automática de textos incluyen, por lo general, los siguientes componentes:

a) Lectura de datos (corpus)
b) Preproceso (opcional)
c) Representación de documentos (vectorización)
d) Reducción dimensional (opcional)
e) Algoritmo de agrupamiento (*clustering*)

Cada componente puede tomar una forma ligeramente diferente dependiendo de la aplicación. Por ejemplo, la lectura de datos puede ser directa, desde un fichero en disco, mediante consulta a una base de datos o incluso por subscripción a una cola de mensajes en tiempo real.

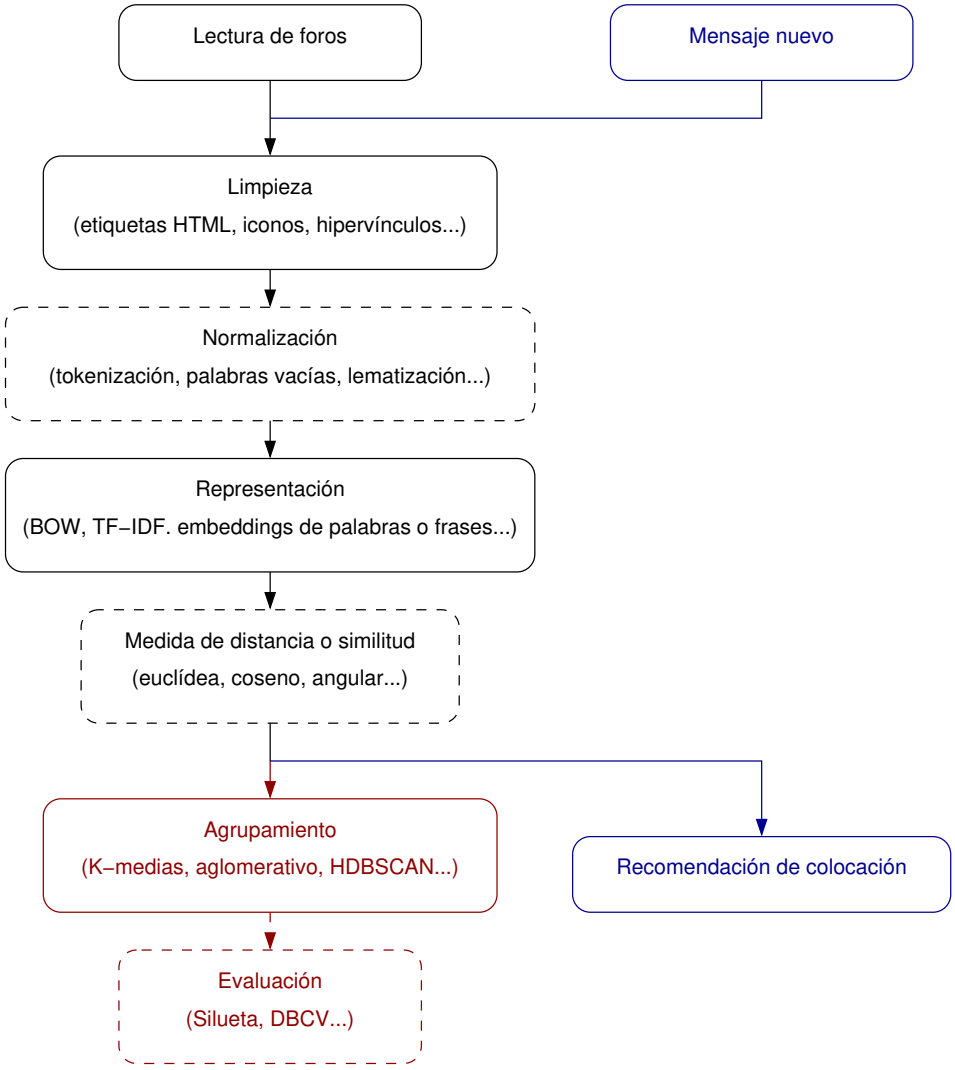


Figura 2.1: Arquitectura del sistema de agrupación automática de foros (izquierda, negro y rojo) y colocación de mensajes nuevos (derecha, negro y azul).

De manera similar, algunos sistemas pueden requerir la limpieza y normalización de los datos una vez leídos, mientras que en otros la robustez del modelo de representación de documentos puede hacer innecesario el preproceso de datos. El modelo de representación es, por supuesto, el encargado de convertir los textos en vectores numéricos, susceptibles de ser comparados y agrupados utilizando técnicas estándar de aprendizaje automático. Este componente también puede tomar formas diversas, desde simple contador para la bolsa de palabras hasta complejas arquitecturas de redes neuronales para los *embeddings* semánticos. Dependiendo del modelo, la dimensión de salida del vectorizador puede llegar a ser enorme, por lo que a veces se incluye una etapa de reducción dimensional previa al *clustering*. Así hacen, por ejemplo, los paquetes de modelización de temas BERTopic y Top2Vec, que recomiendan usar del algoritmo UMAP (McInnes *et al.*, 2020) para mejorar la eficiencia algorítmica y la calidad del *clustering* cuando la dimensión es muy alta. Aquí se ha optado por omitir esta etapa, al menos de primeras, porque añade hiperparámetros que luego hay que ajustar y no ha resultado esencial para obtener

agrupamientos en un tiempo computacional razonable. Por tanto, la salida del vectorizador se ha pasado directamente al algoritmo de *clustering* o, en aquellos casos en los que el algoritmo no soporta nativamente la métrica elegida, a un medidor de distancias interpuesto.

La figura 2.1 muestra la arquitectura del sistema de agrupación automática de foros y colocación de mensajes nuevos desarrollado en este trabajo. Los componentes dibujados con trazo negro son comunes a ambas aplicaciones, mientras que los dibujados en rojo o azul son exclusivos de la agrupación automática o la colocación de mensajes nuevos, respectivamente. Los componentes de normalización de texto y de medida de distancia están dibujados en línea discontinua porque pueden estar incluidos separadamente o no, dependiendo de los modelos de representación y agrupamiento elegidos. Por ejemplo, los *embeddings* de frases no precisan normalización previa y ciertos algoritmos de *clustering* calculan internamente las distancias.

El sistema se ha implementado en Python, que es un lenguaje y ecosistema ideal para este tipo de trabajo, puesto que ofrece multitud de librerías de código abierto para ciencia de datos y aprendizaje automático. Además de los paquetes usados para los diversos componentes, que se detallan en las secciones correspondientes, se destaca aquí el uso de scikit-learn (Pedregosa *et al.*, 2011) para montar el sistema como *pipeline* de transformadores y estimadores, tomando como referencia el capítulo 4 de Bilbro *et al.* (2018). Esta arquitectura modular con interfaces uniformes facilita el desarrollo y la experimentación con diferentes modelos de representación y algoritmos de agrupamiento, tanto en las investigaciones preliminares como con los programas definitivos. Se ha procurado retener la configurabilidad del sistema en todo momento.

2.2. Medidas de similitud

Como se avanzó en la introducción, el foco de este trabajo está puesto en el *clustering* de acuerdo a la similitud semántica. Una vez elegido un modelo de representación y vectorizados los documentos de un corpus, la similitud entre dos documentos se puede calcular en función de su proximidad en el espacio vectorial. Si bien la métrica euclídea es la más habitual en el ámbito general de la agrupación automática, cuando se trata de textos se tiende a preferir la del coseno. Los motivos tienen que ver con la alta dimensión y dispersión de los vectores que aparecen en ese contexto y con que se persiga una medida de similitud semántica independiente de la magnitud o escala de los vectores (véase Han *et al.*, 2012, sec. 2.4.7; Jurafsky y Martin, 2023, sec. 6.7). La definición de la similitud del coseno entre dos vectores $\boldsymbol{u}$ y $\boldsymbol{v}$ no nulos en un espacio euclídeo n-dimensional es sencilla,

$$\cos(\boldsymbol{u}, \boldsymbol{v}) = \frac{\boldsymbol{u} \cdot \boldsymbol{v}}{\|\boldsymbol{u}\|\|\boldsymbol{v}\|} = \frac{\sum_{i=1}^{n} u_i v_i}{\left(\sum_{i=1}^{n} u_i^2 \sum_{i=1}^{n} v_i^2\right)^{1/2}}. \tag{2.1}$$

La similitud toma por tanto valores entre -1 y 1. Estos dos extremos corresponden a vectores alineados con orientación opuesta y coincidente, respectivamente, mientras que para vectores ortogonales la similitud es 0. Para vectores con componentes no negativas, como suele ser el

caso en representación de textos, el rango de la medida reduce a $[0, 1]$. De todo lo anterior se sigue que se puede definir una distancia tomando el complemento a uno de la similitud, es decir,

$$d_{\cos}(\boldsymbol{u}, \boldsymbol{v}) = 1 - \cos(\boldsymbol{u}, \boldsymbol{v}). \tag{2.2}$$

Un problema menor con la similitud y distancia del coseno es que no están definidas para vectores nulos, que pueden aparecer y de hecho surgen en este análisis para mensajes nulos o formados íntegramente por palabras vacías y etiquetas HTML. Este problema se ha resuelto modificando la implementación de la similitud del coseno para que devuelva 1 cuando ambos vectores sean nulos y 0 cuando lo sea sólo uno.

Otro potencial problema, más técnico, es que la distancia del coseno no es estrictamente una métrica, ya que no satisface la desigualdad triangular. Esto no impide que sea utilizada con éxito en multitud de tareas de análisis de texto, pero algunos autores, como Cer *et al.* (2018), han propuesto sustituirla por una distancia angular o de arco derivada de ella,

$$d_{\text{arc}}(\boldsymbol{u}, \boldsymbol{v}) \propto \arccos\big(\cos(\boldsymbol{u}, \boldsymbol{v})\big). \tag{2.3}$$

Además de ser una métrica propiamente dicha, la distancia angular es más regular y distingue con mayor precisión entre vectores casi alineados. Para verificar si una tiene ventajas sobre la otra, en los ensayos de vectorización y *clustering* se considerarán ambas distancias.

2.3. Medidas de evaluación

La evaluación de la calidad de los agrupamientos es parte fundamental del desarrollo del sistema de agrupación automática. Las medidas de evaluación se dividen en dos categorías: las externas, que se basan en comparaciones con el agrupamiento verdadero, y las internas, que se basan en medidas de compacidad y separación y no requieren conocimiento de los grupos verdaderos. Como se verá en el próximo capítulo, los datos para este trabajo vienen anotados, pero en otras aplicaciones los grupos no se conocerían de antemano y las medidas internas serían cruciales para la evaluación del sistema. Además, las medidas internas permitirán evaluar los agrupamientos originales, que debido a su tamaño no han sido revisados manualmente.

2.3.1. Medidas internas

La primera de las tres medidas internas consideradas aquí es el coeficiente de silueta, que mide la calidad en función de las distancias medias intragrupo y con otros grupos (véase, p. ej., Han *et al.*, 2012, sec. 10.6.3). Para cada vector de representación $\boldsymbol{u}$, el coeficiente de silueta $S(\boldsymbol{u})$ viene dado por

$$S(\boldsymbol{u}) = \frac{a(\boldsymbol{u}) - b(\boldsymbol{u})}{\max\{a(\boldsymbol{u}), b(\boldsymbol{u})\}}, \tag{2.4}$$

donde $a(\boldsymbol{u})$ es la distancia media entre $\boldsymbol{u}$ y el resto de vectores en su mismo grupo y $b(\boldsymbol{u})$ es la distancia media entre $\boldsymbol{u}$ y los vectores en el grupo más cercano, excluyendo el propio. El coeficiente de silueta S del agrupamiento es simplemente la media de los de sus elementos.

La silueta toma valores entre -1, que corresponde a un agrupamiento muy malo, en el que hay elementos más cerca de otros grupos que de los suyos propios, y 1, que corresponde a un buen agrupamiento, formado por grupos compactos y bien separados unos de otros. La silueta es una medida interna muy popular y se usa con frecuencia para determinar el número óptimo de grupos. Frente a algunas otras medidas internas similares, tiene la ventaja de que no utiliza conceptos como centroides y medias cuadráticas, que están más orientados a la métrica euclídea. Una desventaja es que es costosa de calcular, pero para los números de documentos manejados aquí su cómputo es perfectamente asumible.

La segunda medida considerada aquí es el índice de validación de agrupamientos basado en densidad (DBCV; Moulavi *et al.*, 2014). Esta medida es más reciente y ha sido diseñada buscando un mejor comportamiento sobre grupos no globulares, esto es, grupos que no sean convexos e isótropos. La mayoría de medidas tradicionales, incluida la silueta, priorizan los agrupamientos globulares. Sin embargo, los algoritmos de agrupamiento basados en densidad, como HDBSCAN, también pueden identificar agrupamientos no globulares con pleno sentido. La medida DBCV, que evalúa la calidad en función de la densidad relativa de las conexiones entre pares y trata de manera natural los elementos considerados como ruido, se adapta mejor al tipo de agrupamiento generado por estos algoritmos. De hecho, el paquete hdbscan que se utiliza más adelante para el *clustering* incluye una implementación de DBCV. Al igual que la silueta, DBCV toma valores entre -1 y 1 y los valores más grandes indican mayor calidad.

Además de las dos medidas anteriores, encontradas en la literatura, se introduce aquí otra medida interna muy sencilla orientada a la tarea de colocación de mensajes nuevos. La idea es que, si de forma bastante natural, los mensajes nuevos se asignasen al grupo más próximo en cuanto a distancia mínima a sus elementos, entonces sería deseable que, de quitar un elemento de un agrupamiento dado, la recomendación de recolocación coincidiese con el grupo original. Denotando el agrupamiento original por $\mathcal{A} = \{A_1, \ldots, A_k\}$ y el total de elementos por N, la proporción de elementos que satisfaría la propiedad anterior sería

$$\varphi = N^{-1} \sum_{i=1}^{k} \sum_{\boldsymbol{u} \in A_i} \theta \left(\min_{\substack{1 \leq j \leq k \\ \boldsymbol{v} \in A_j}} d(\boldsymbol{u}, \boldsymbol{v}) - \min_{\boldsymbol{v} \in A_i \setminus \{\boldsymbol{u}\}} d(\boldsymbol{u}, \boldsymbol{v}) \right), \tag{2.5}$$

donde θ es la función escalón unitaria, o de Heaviside. En la expresión anterior, el segundo mínimo ha de interpretarse como ∞ cuando A_i contenga únicamente a $\boldsymbol{u}$, de modo que tales grupos queden contabilizados como malos. La fracción φ, con rango $[0, 1]$, se denominará aquí consistencia de proximidad, puesto que cuantifica si los elementos seguirían perteneciendo al grupo original si fueran recolocados en base a proximidad. Se considera una medida de consistencia, más que de calidad, porque concede la puntuación máxima a algunos agrupamientos

triviales, como el formado por un único grupo en un conjunto con múltiples elementos.

2.3.2. Medidas externas

Para comparar los agrupamientos generados por el sistema con los originales conocidos se han seleccionado dos de las medidas externas usadas en el artículo de Onan y Toçoğlu (2021): el índice de Rand ajustado (ARI) y la información mutua ajustada (AMI). Estas dos medidas no dependen de la forma o geometría de los grupos y tampoco tienen en cuenta las etiquetas absolutas de grupo, es decir, son invariantes con respecto a permutaciones de las etiquetas. Además, ambas medidas dan mayor puntuación cuanto más similares sean los agrupamientos e incluyen ajustes que las hacen tomar valores cercanos a 0 para agrupamientos aleatorios.

Dados dos agrupamientos $\mathcal{A} = \{A_1, \ldots, A_{k_\mathcal{A}}\}$ y $\mathcal{B} = \{B_1, \ldots, B_{k_\mathcal{B}}\}$ en $k_\mathcal{A}$ y $k_\mathcal{B}$ grupos respectivamente, el índice de Rand ajustado se define como

$$\mathrm{ARI}(\mathcal{A}, \mathcal{B}) = \frac{2(ad - bc)}{(a+b)(b+d) + (a+c)(c+d)}, \tag{2.6}$$

donde a es el número de pares de elementos que caen en el mismo grupo tanto en $\mathcal{A}$ como en $\mathcal{B}$, b es el número de pares en el mismo grupo en $\mathcal{A}$ pero no en $\mathcal{B}$, c es el número de pares en el mismo grupo en $\mathcal{B}$ pero no en $\mathcal{A}$ y d es el número de pares en distinto grupo en $\mathcal{A}$ y en $\mathcal{B}$. Cuando los agrupamientos son equivalentes, $b = c = 0$ y el ARI alcanza su valor máximo de 1. Para agrupamientos extremadamente diferentes, el ARI puede alcanzar valores negativos.

Por su parte, la información mutua entre los dos agrupamientos $\mathcal{A}$ y $\mathcal{B}$ se define como

$$\mathrm{MI}(\mathcal{A}, \mathcal{B}) = -\sum_{i=1}^{k_\mathcal{A}} \sum_{j=1}^{k_\mathcal{B}} p_{ij} \log \frac{p_{ij}}{p_i^\mathcal{A} p_j^\mathcal{B}}, \tag{2.7}$$

donde p_{ij} es la probabilidad de que un elemento elegido al azar pertenezca a los grupos A_i y B_j, y $p_i^\mathcal{A}$ y $p_j^\mathcal{B}$ son las probabilidades marginales correspondientes. Para descontar los efectos relacionados con la aleatoriedad y la cantidad de elementos y grupos, se define la información mutua ajustada

$$\mathrm{AMI}(\mathcal{A}, \mathcal{B}) = \frac{\mathrm{MI}(\mathcal{A}, \mathcal{B}) - \mathrm{E}\{\mathrm{MI}(\mathcal{A}, \mathcal{B})\}}{\frac{1}{2}\big(H(\mathcal{A}) + H(\mathcal{B})\big) - \mathrm{E}\{\mathrm{MI}(\mathcal{A}, \mathcal{B})\}} \tag{2.8}$$

donde $H(\mathcal{A})$ y $H(\mathcal{B})$ son las entropías asociadas con cada agrupamiento,

$$H(\mathcal{A}) = -\sum_{i=1}^{k_\mathcal{A}} p_i^\mathcal{A} \log p_i^\mathcal{A} \quad \text{y} \quad H(\mathcal{B}) = -\sum_{i=1}^{k_\mathcal{B}} p_i^\mathcal{B} \log p_i^\mathcal{B} \tag{2.9}$$

En la ecuación (2.8), se ha utilizado la media aritmética de las entropías para la normalización, que es una de las opciones más populares entre las disponibles (Vinh *et al.*, 2010). Como con el ARI, la puntuación máxima de 1 para la AMI implica similitud total entre agrupamientos.

Capítulo 3

Datos

Los datos disponibles para este trabajo son los vertidos de 7 foros de la UNED facilitados por los tutores en forma de ficheros con formato CSV. Por supuesto, casi todo el desarrollo del sistema de agrupación automática en los capítulos posteriores será genérico, aplicable, al menos en teoría, a otros foros cualesquiera. Conviene, no obstante, echar un vistazo temprano a los datos para determinar sus necesidades de preproceso y empezar a caracterizar los foros. La tabla 3.1 muestra los nombres de los ficheros de los foros junto a las asignaturas, estudios y cursos correspondientes. Como se mencionaba en la introducción, la temática de los foros es más variada que en otras investigaciones en la literatura, ya que abarca ciencias de la salud, educación y acceso a la universidad, además de ingeniería informática.

3.1. Lectura

Una inspección rápida de la cabecera del fichero del foro de Procesadores del Lenguaje de 2017–18, mostrada en la figura 3.1, revela que cada línea contiene, además del texto del mensaje, su fecha, título y los nombres del subforo e hilo a los que pertenece. También salta a la vista que los nombres y el texto contienen etiquetas y códigos HTML (p. ej., "
" y """), que habrá que limpiar durante la lectura datos o en la etapa de preproceso.

La lectura de los datos se ha implementado con ayuda de la archiconocida librería pandas,

Tabla 3.1: Nombre de fichero, asignatura, estudios y curso de los foros de la UNED.

Fichero	Asignatura	Estudios	Curso
6201302-_2019_cl.csv	Psicofarmacología	Psicología	2018–19
6390103-_2019_cl.csv	Sociedad del Conocimiento, Tecnología y Educación	Pedagogía	2018–19
ForosPL1_17-18.csv	Procesadores del Lenguaje I	Ingeniería Informática	2017–18
ForosPL1_18-19.csv	Procesadores del Lenguaje I	Ingeniería Informática	2018–19
ForosPL1_19-20.csv	Procesadores del Lenguaje I	Ingeniería Informática	2019–20
ForosPL1_20-21.csv	Procesadores del Lenguaje I	Ingeniería Informática	2020–21
Foros_Acceso_Mayores_cl.csv	Acceso para Mayores	Acceso	2018–19

```
Foro,Hilo,idMensaje,Responde a,Fecha,Titulo mensaje,Texto mensaje
 Foro Análisis léxico (fuera de la práctica)<br/><br/>, Conversion de ER a AFN (Thompson) <br/>,1,,2017-12-30
 ↪  12:15:09,Conversion de ER a AFN (Thompson),"Hola. Tengo una duda con el método de Thompson. Si tenemos
 ↪  la<br/>ER a(a+b), su automata entiendo que sería como la imagen<br/>adjunta.<br/><br/>Sin embargo, en unos
 ↪  apuntes que he encontrado utilizan una<br/>transición epsilon extra para unir las dos ER (a y
 ↪  (a+b)).<br/>¿Es esto necesario?<br/><br/>Muchas gracias.<br/><br/>"
 Foro Análisis léxico (fuera de la práctica)<br/><br/>, Conversion de ER a AFN (Thompson) <br/>,2,1.0,2018-01-03
 ↪  14:01:32,Re: Conversion de ER a AFN (Thompson),"Hola<br/><br/>Me uno a la pregunta sobre la construccion de
 ↪  Thomson en el caso de la<br/>concatenacion de elementos.<br/><br/>En el texto base, se indica
 ↪  &quot;Conectamos el estado de aceptacion de la<br/>máquina de r al estado de inicio de la máquina de<br/>s
 ↪  mediante una transicion &epsilon;&quot;<br/><br/>Sin embargo en toda la documentacion y ejemplos que
 ↪  encuentro en<br/>Internet nunca utiliza la transicion &epsilon; para la concatenacion.<br/><br/>Alguien
 ↪  puede echarnos una mano?<br/><br/>Gracias!<br/><br/>"
```

Figura 3.1: Cabecera del fichero del foro de Procesadores del Lenguaje de 2017–18.

para la ingestión y manejo de tablas, y de la librería Beautiful Soup, para el procesado de HTML. Siguiendo una vez más el tratamiento en Bilbro *et al.* (2018), la lectura se ha implementado como lector de corpus de NLTK (Bird *et al.*, 2009). Esta conocida librería de NLP también se utilizará para el normalizador de texto. En el lector de corpus se han incluido métodos para la eliminación de etiquetas HTML y para el filtrado de mensajes en función el nombre del subforo original. Esta última funcionalidad se usará más adelante para excluir del análisis los subforos de grupos de tutoría.

3.2. Limpieza

Además de las etiquetas y códigos HTML observados en la sección anterior, los mensajes de los foros contienen otros elementos no estrictamente textuales que podrían causar efectos indeseados en ciertos modelos de representación de documentos y es conveniente limpiar. En la parte superior de la figura 3.2 se exponen ejemplos de mensajes con imágenes, ecuaciones, URL, direcciones de correo electrónico y rutas de sistema de archivos. Una vez procesadas las etiquetas HTML, los otros elementos pueden limpiarse por medio de expresiones regulares (Goyvaerts y Levithan, 2012). La parte inferior de la figura muestra el resultado de aplicar las sustituciones especificadas en el limpiador de texto. Los elementos no deseados son reemplazados por indicadores de tipo entre corchetes.

3.3. Exploración

Como se vio en la cabecera del fichero del foro de Procesadores del Lenguaje, en la figura 3.1, los mensajes de los foros de la UNED vienen etiquetados con los nombres de sus subforos e hilos. La tabla 3.2 muestra el número de subforos, hilos y mensajes en cada uno de los foros, con los subforos correspondientes a grupos de tutoría contados por separado. El motivo para separar estos subforos es que son más relativos a la geografía que a la temática, luego quedan fuera del alcance de la agrupación automática y colocación basadas en semántica y podrían añadir ruido en su evaluación. Los grupos de tutoría, que serán en consecuencia descartados,

Originales

```
De nada, está pregunta era de las fáciles [IMAGE:<br/>&apos;smiley&apos;<br/><br/>https://2019.cursosvirtuales.
↪ uned.es/resources/acs-templating/ckeditor/plugins/smiley/images/regular_smile.png<br/>]<br/><br/>
---------------------------------------------------------------------------------------------------------------
Buenos días,<br/><br/>En las transparencias del tema 3, página 25, aparece que la<br/>recursión por la izquierda
↪ \(A::=A\alpha | \beta\) se resuelve<br/>así:<br/><br/>\(\begin{equation} \begin{split} &A::=\alpha A&apos;
↪ \\ &A&apos;::=\beta A&apos; |<br/>\end{split} \end{equation}\)<br/><br/>Pero lo correcto
↪ sería:<br/><br/>\(\begin{equation} \begin{split} &A::=\beta A&apos; \\ &A&apos;::=\alpha A&apos;
↪ |<br/>\end{split} \end{equation}\)<br/><br/>Pues en esa recursión, \(\beta\) representa el caso base de
↪ la<br/>recursión, y en la diapositiva la parte del caso base es<br/>\(\alpha\).<br/><br/>Un saludo.<br/><br/>
---------------------------------------------------------------------------------------------------------------
*SGSI - 01 Conceptos Básicos sobre la Seguridad de la<br/>Información:
↪ https://www.youtube.com/watch?v=zV2sfyvfqik<br/><br/>
---------------------------------------------------------------------------------------------------------------
Estimados estudiantes,<br/><br/>De acuerdo con la guía, los estudiantes que hayan aprobado la<br/>práctica el
↪ curso anterior no tienen que realizarla este curso.<br/>Para saber los estudiantes que se quieren acoger a
↪ esta medida,<br/>necesitamos que cada uno de ellos nos envie un correo<br/>a [redacted]@lsi.uned.es
↪ indicandolo.<br/><br/>Saludos.<br/><br/>
---------------------------------------------------------------------------------------------------------------
Hola [redacted],<br/><br/>Si ejecutas &#39;ant cupTest&#39; desde el directorio donde<br/>está el fichero build.xml,
↪ ¿qué salida produce la<br/>compilación? El directorio es .\doc\config.<br/><br/>Yo estoy ejecutando tanto el
↪ flexTest como el cupTest desde<br/>ahí sin mayor problema.<br/><br/>Saludos,<br/><br/>[redacted]<br/><br/>
```

Limpiados

```
De nada, está pregunta era de las fáciles <{IMAGEN}>
---------------------------------------------------------------------------------------------------------------
Buenos días,

En las transparencias del tema 3, página 25, aparece que la
recursión por la izquierda <{ECUACIÓN}> se resuelve
así:

<{ECUACIÓN}>

Pero lo correcto sería:

<{ECUACIÓN}>

Pues en esa recursión, <{ECUACIÓN}> representa el caso base de la
recursión, y en la diapositiva la parte del caso base es
<{ECUACIÓN}>.

Un saludo.
---------------------------------------------------------------------------------------------------------------
*SGSI - 01 Conceptos Básicos sobre la Seguridad de la
Información: <{URL}>
---------------------------------------------------------------------------------------------------------------
Estimados estudiantes,

De acuerdo con la guía, los estudiantes que hayan aprobado la
práctica el curso anterior no tienen que realizarla este curso.
Para saber los estudiantes que se quieren acoger a esta medida,
necesitamos que cada uno de ellos nos envie un correo
a <{EMAIL}> indicandolo.

Saludos.
---------------------------------------------------------------------------------------------------------------
Hola [redacted],

Si ejecutas 'ant cupTest' desde el directorio donde
está el fichero build.xml, ¿qué salida produce la
compilación? El directorio es <{RUTA}>.

Yo estoy ejecutando tanto el flexTest como el cupTest desde
ahí sin mayor problema.

Saludos,

[redacted]
```

Figura 3.2: Ejemplos de mensajes con imágenes, ecuaciones, URL, correos electrónicos y rutas.

Tabla 3.2: Número de subforos, hilos y mensajes por foro, con grupos de tutoría por separado.

	Excluyendo grupos de tutoría			Grupos de tutoría		
Foro	# subforos	# hilos	# mensajes	# subforos	# hilos	# mensajes
6201302-_2019_cl	9	131	427	33	145	224
6390103-_2019_cl	12	86	588	34	169	312
ForosPL1_17-18	8	88	372	19	52	137
ForosPL1_18-19	8	62	270	18	45	104
ForosPL1_19-20	8	66	176	13	43	96
ForosPL1_20-21	8	82	302	17	49	109
Foros_Acceso_Mayores_cl	7	81	446			

Tabla 3.3: Número de mensajes en cada subforo, excluyendo grupos de tutoría.

	Foro						
Subforo	6201302-_2019_cl	6390103-_2019_cl	ForosPL1_17-18	ForosPL1_18-19	ForosPL1_19-20	ForosPL1_20-21	Foros_Acceso_Mayores_cl
0.1. Foro del Orientador/a de la Comunidad de Acogida Virtual							96
0.2. Foro de estudiantes							261
1. Foro de contenidos y actividades de la Fase 1.							33
2. Foro de contenidos y actividades de la Fase 2.							9
3. Foro de contenidos y actividades de la Fase 3.							10
4. Foro de contenidos y actividades de la Fase 4.							10
Foro de apoyo técnico							27
Foro Análisis léxico (fuera de la práctica)			36	1	7	12	
Foro Análisis sintáctico (fuera de la práctica)			62	7	2	2	
Foro General Práctica			24	46	83	161	
Foro Práctica: análisis léxico			65	55	28	13	
Foro Práctica: análisis sintáctico			97	89	26	59	
Foro de consultas generales			52	28	23	48	
Foro de estudiantes (no moderado por el Equipo Docente)			23	37	2	5	
Cuestiones generales de la asignatura	131						
Foro Tema 1 - Capítulos 4 y 5 del libro de Stahl	114						
Foro Tema 2 - Capítulos 6, 7 y 8 del libro de Stahl	56						
Foro Tema 3 - Capítulo 9 del libro de Stahl	21						
Foro Tema 4 - Capítulo 11 del libro de Stahl	6						
Foro Tema 5 - Capítulos 12 y 13 del libro de Stahl	11						
Foro Tema 6 - Capítulo 14 del libro de Stahl	3						
Foro de estudiantes (Cafetería)	70						
Consultas generales		98					
Prueba de Evaluación Continua (PEC)		82					
Tema 1: Tecnologías y vidas		136					
Tema 2: Base material de Internet		47					
Tema 3: ¿Sociedad de la desinformación? Perspectivas sobre las noticias falsas		62					
Tema 4: Rastros, huellas y filtros digitales		12					
Tema 5: La cutura social en torno a la seguridad de la información		16					
Tema 6: La propiedad intelectual y sus enemigos		12					
Tema 7: Más allá de las pantallas		16					
Tema 8: Bidg Data, educación basada en datos y analítica del aprendizaje		11					
Tema 9: Aprendiendo. Cuando quieras. Donde vayas.		6					
Coordinación tutorial	15	90	13	7	5	2	

Tabla 3.4: Número de mensajes en cada subforo de grupos de tutoría.

Subforo	Foro (véase tabla 3.3)					
Grupo de Tutoria 1			3		16	10
Grupo de Tutoria 2			13	4		9
Grupo de Tutoria 3	2	2	4		11	
Grupo de Tutoria 4	8		4	9	3	3
Grupo de Tutoria 5	2		2	16	16	
Grupo de Tutoria 6		17	2	1		8
Grupo de Tutoria 7	1	6			2	3
Grupo de Tutoria 8	1	7	13		20	
Grupo de Tutoria 9	4			23	3	6
Grupo de Tutoria 10	15	7		1	6	1
Grupo de Tutoria 11	15			2		2
Grupo de Tutoria 12		41	3			2
Grupo de Tutoria 13	5		11			
Grupo de Tutoria 13 - Foro del tutor/a [redacted]			47			
Grupo de Tutoria 14	1	1		14	3	8
Grupo de Tutoria 15	4		2	1	2	6
Grupo de Tutoria 16	2	1	2	2		23
Grupo de Tutoria 17	21			2		
Grupo de Tutoria 18		16	7	3	6	
Grupo de Tutoria 19			1	2		18
Grupo de Tutoria 20	14		4		5	1
Grupo de Tutoria 21	17		1		3	1
Grupo de Tutoria 22	2	5		5		
Grupo de Tutoria 23	16					6
Grupo de Tutoria 24		5		3		
Grupo de Tutoria 25	4	21	8			2
Grupo de Tutoria 26			2			
Grupo de Tutoria 27	2	2		3		
Grupo de Tutoria 28	2		8	2		
Grupo de Tutoria 29				11		
Grupo de Tutoria 30	7					
Grupo de Tutoria 31	7	16				
Grupo de Tutoria 32	1	6				
Grupo de Tutoria 34	9	1				
Grupo de Tutoria 35	5					
Grupo de Tutoria 36		8				
Grupo de Tutoria 37		11				
Grupo de Tutoria 38	1					
Grupo de Tutoria 39	28	7				
Grupo de Tutoria 40	2	2				
Grupo de Tutoria 41	6					
Grupo de Tutoria 42		21				
Grupo de Tutoria 43	1	11				
Grupo de Tutoria 43 - Foro del tutor/a [redacted]	8					
Grupo de Tutoria 44	6	2				
Grupo de Tutoria 45	5					
Grupo de Tutoria 46		22				
Grupo de Tutoria 47		1				
Grupo de Tutoria 48		3				
Grupo de Tutoria 49 - Foro del tutor/a [redacted]		23				
Grupo de Tutoria 50		12				
Grupo de Tutoria 51		1				
Grupo de Tutoria 53		2				
Grupo de Tutoria 54		14				
Grupo de Tutoria 55		8				
Grupo de Tutoria 58 - Foro del tutor/a [redacted]		7				
Grupo de Tutoria 58 - Foro del tutor/a [redacted]		4				

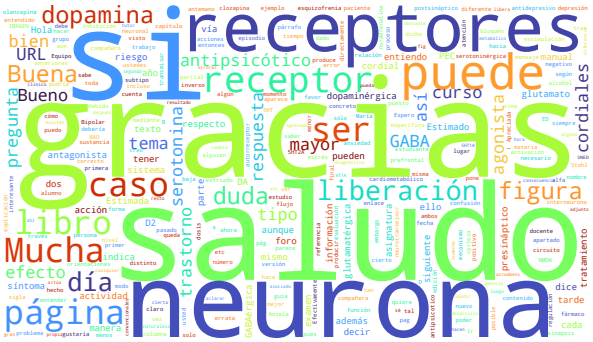
albergan sólo un 27 % de los mensajes, pero suponen una mayoría, el 69 %, de los subgrupos originales. El lado izquierdo de la tabla, que excluye los grupos de tutoría, da por tanto idea del número de subforos que cabe esperar del sistema de agrupación automática, entre 7 y 12, dependiendo del foro. Este número es solamente orientativo, pero conviene saber su orden de magnitud de cara a experimentos con algoritmos de *clustering* que requieren la especificación *a priori* del número de grupos. De acuerdo con ese mismo lado de la tabla, el número de hilos por foro es casi diez veces superior al de subforos, en torno al centenar. Si bien se podría intentar obtener agrupamientos así de finos de forma automática, sería complicado, debido a la cantidad de grupos con poquísimos mensajes.

Las tablas 3.3 y 3.4 muestran la distribución de mensajes por subforo dentro de cada foro con los grupos de tutoría excluidos o aislados, respectivamente. La intención con estas tablas no es guiar el desarrollo del sistema de manera muy concreta, sino realizar alguna observación general y, sobre todo, tenerlas a mano para ayudar a interpretar los resultados de los ensayos. La segunda tabla se incluye más que nada por completitud y para clarificar la implementación mediante expresiones regulares del mecanismo de exclusión de grupos de tutoría en el lector de foros. En cuanto a la primera tabla, los nombres de los subforos parecen confirmar que, a diferencia de los grupos de tutoría, estos agrupamientos son relativos a la temática. Hay, por ejemplo, subforos dedicados a las fases, temas y prácticas de la asignatura, además de otros para cuestiones generales, coordinación tutorial o discusiones informales entre estudiantes. La otra apreciación inmediata es que los subforos pueden ser bastante heterogéneos en cuanto a número de mensajes, incluso dentro de un mismo foro. La mayoría de los subforos tiene entre una decena y una centena de mensajes, pero también hay subforos con 1 o 2 mensajes y, en el otro extremo, uno con 261 mensajes, más de la mitad del total del foro. El tamaño mínimo de los grupos tiene relevancia para la configuración del *clustering* HDBSCAN, de igual manera que el número de grupos es fundamental para algoritmo k-medias.

Como cierre de la exploración inicial de datos, la figura 3.3 muestra las nubes de palabras generadas para cada foro utilizando el paquete wordcloud. Estas representaciones visuales de texto destacan las palabras más importantes o frecuentes escribiéndolas con letra más grande, acentuando su color o situándolas en lugares prominentes (véase, p. ej., Han *et al.*, 2012, sec. 2.3.5). A menudo se normalizan previamente las palabras, pasando plurales al singular, y se eliminan las llamadas palabras vacías, como preposiciones, conjunciones... En este caso no se ha hecho especial esfuerzo con los plurales, dejando el tratamiento por defecto en wordcloud, pero se ha utilizado un listado de palabras vacías específico para el español tomado de NLTK.

Las nubes de palabras de los distintos foros tienen ciertas palabras frecuentes en común, como los saludos y agradecimientos. Al mismo tiempo, en la nube de cada foro se encuentran palabras frecuentes acordes con la materia como, por ejemplo: neurona, receptores y dopamina en Psicofarmacología; información, noticia y falsa en Sociedad del Conocimiento; error, *token* y sentencia en las múltiples instancias de Procesadores del Lenguaje; y curso, acceso y año en Acceso para Mayores. En circunstancias normales, la utilidad de la nube de palabras radica en

que permite, de manera rápida y visual, hacerse una idea del tema de un texto o un corpus del que se sabe poco o nada de antemano.



(a) 6201302-_2019_cl (Psicofarmacología).

(b) 6390103-_2019_cl (Sociedad del Conocimiento).

(c) ForosPL1_17-18 (Procesadores del Lenguaje).

(d) ForosPL1_18-19 (Procesadores del Lenguaje).

(e) ForosPL1_19-20 (Procesadores del Lenguaje).

(f) ForosPL1_20-21 (Procesadores del Lenguaje).

(g) Foros_Acceso_Mayores_cl (Acceso para Mayores).

Figura 3.3: Nubes de palabras para los foros de la UNED.

Capítulo 4

Representación de documentos

La representación de documentos consiste en la transformación de secuencias de palabras o textos de forma libre en vectores numéricos, que resultan más sencillos de analizar aplicando herramientas ordinarias de la ciencia de datos y el aprendizaje automático. Por ejemplo, en la representación por bolsa de palabras, cada componente vectorial corresponde a una palabra en un diccionario reducido y su valor guarda relación con el número de apariciones de la palabra en el documento. A continuación, se explica y ensaya esta técnica vectorización, y más adelante se hace lo propio con otras más modernas basadas en *embeddings* semánticos.

4.1. Bolsa de palabras

La primera técnica de representación de documentos ensayada sobre los foros es de tipo bolsa de palabras (BOW). En este tipo de representación, cada documento se representa como un vector cuya longitud es igual al tamaño del vocabulario del corpus, o en este caso del foro (véase, p. ej., Bilbro *et al.*, 2018, cap. 4). El valor de cada componente vectorial tiene que ver con el número de apariciones del término en el documento, pero como se detalla más adelante, a menudo se ajusta teniendo en cuenta la frecuencia del término en el corpus.

Previamente al conteo de apariciones de términos, las representaciones BOW requieren la segmentación o tokenización de los documentos y la normalización de los términos. Para estas tareas se han empleado los módulos de preproceso de la librería NLTK. Para la segmentación se ha usado un simple tokenizador mediante expresiones regulares, ya que los más sofisticados arrojaban a veces malos resultados a comienzo o fin de frases y párrafos, pese a ser específicos para el español. A continuación de la tokenización, se excluyen las palabras con caracteres no alfabéticos y las vacías de contenido, utilizando el listado para español de NLTK. Por último, las palabras se convierten a minúscula y, opcionalmente, se quitan las tildes y se truncan las terminaciones aplicando el *stemmer* Snowball para español. La finalidad de estas operaciones es reducir el tamaño del vocabulario y, aunque de manera burda, agrupar semánticamente las inflexiones de cada palabra.

Originales

```
Sólo quiere decir que se ha terminado el plazo para<br/>presentarla. A todos nos sale en rojo.<br/><br/>Lo
↪  importante es que en el &quot;Estado de la entrega&quot; te aparece:<br/>&quot;Entregado&quot;.<br/><br/>
--------------------------------------------------------------------------------------------------------------
Pues a mi me aparece así. Y la entregué dos veces, la<br/>segunda con algunos cambios. Y comprobé el
↪  envío<br/>descargando el documento.<br/><br/>
--------------------------------------------------------------------------------------------------------------
Yo entregué mi pec pero ahora no aparece . Y como es fuera de<br/>plazo ya no deja volver a intentarlo , Le he
↪  mandado un correo a mi<br/>tutora pero no me contesta .<br/><br/>Todo esfuerzo para nada?<br/><br/>Es la
↪  primera vez que me sucede algo así .<br/><br/>
```

Normalizados

```
['sol', 'quier', 'dec', 'termin', 'plaz', 'present', 'sal', 'roj', 'import', 'entreg', 'aparec', 'entreg']
--------------------------------------------------------------------------------------------------------------
['pues', 'aparec', 'asi', 'entreg', 'dos', 'vec', 'segund', 'cambi', 'comprob', 'envi', 'descarg', 'document']
--------------------------------------------------------------------------------------------------------------
['entreg', 'pec', 'ahor', 'aparec', 'plaz', 'dej', 'volv', 'intent', 'mand', 'corre', 'tutor', 'contest',
↪  'esfuerz', 'primer', 'vez', 'suced', 'asi']
```

Vectorización

```
[[0.         0.         0.        ]
 [1.58496249 0.         0.        ]
 [0.         0.         0.        ]
 [1.58496249 0.         0.        ]
 [0.58496249 0.         0.58496249]
 [1.58496249 0.         0.        ]
 [1.58496249 0.         0.        ]
 [1.58496249 0.         0.        ]
 [1.58496249 0.         0.        ]
 [1.58496249 0.         0.        ]
 [1.58496249 0.         0.        ]
 [0.         0.58496249 0.58496249]
 [0.         1.58496249 0.        ]
 [0.         1.58496249 0.        ]
 [0.         1.58496249 0.        ]
 [0.         1.58496249 0.        ]
 [0.         1.58496249 0.        ]
 [0.         1.58496249 0.        ]
 [0.         1.58496249 0.        ]
 [0.         1.58496249 0.        ]
 [0.         0.         1.58496249]
 [0.         0.         1.58496249]
 [0.         0.         1.58496249]
 [0.         0.         1.58496249]
 [0.         0.         1.58496249]
 [0.         0.         1.58496249]
 [0.         0.         1.58496249]
 [0.         0.         1.58496249]
 [0.         0.         1.58496249]
 [0.         0.         1.58496249]
 [0.         0.         1.58496249]]
```

Figura 4.1: Ejemplo de normalización y vectorización TF-IDF de tres mensajes de un foro.

El conteo de términos y los diversos ajustes se han implementado recurriendo a Gensim, que es otra conocida librería para la representación de documentos y la modelización de temas (Řehůřek y Sojka, 2010). La librería incluye múltiples versiones de la representación BOW, expresadas como generalizaciones de la representación por frecuencia de término compensada por frecuencia inversa de documento (TF-IDF). En dicha generalización, el peso de la palabra i en el documento j dentro de un corpus de N documentos viene dado por

$$w_{ij} = w_{\text{local}}(f_{ij})\, w_{\text{global}}(\omega_i, N), \tag{4.1}$$

donde f_{ij} denota el número de apariciones de la palabra i en el documento j y ω_i es el número de documentos en los que aparece la palabra i. En el modelo binario (OH) w_{local} es la función de Heaviside y $w_{\text{global}} = 1$, mientras que en el modelo clásico de frecuencia de término (TF) w_{local} es la función identidad y $w_{\text{global}} = 1$. En el modelo TF-IDF, que atenúa el peso de las palabras compartidas por muchos documentos, w_{local} es también la identidad y $w_{\text{global}}(\omega_i, N) = \log_2(N/\omega_i)$. Una variante habitual de este modelo, a veces denominada LogTF-IDF, también aplica una transformación logarítmica a la frecuencia de término, con $w_{\text{local}}(f_{ij}) = 1 + \log_2 f_{ij}$. Además de permitir elegir estas funciones, Gensim permite fijar umbrales de frecuencia para la exclusión de palabras extremadamente raras o comunes y seleccionar salida en formato denso o disperso. Este último puede resultar más eficiente cuando el vocabulario es muy amplio y, en consecuencia, la dimensión vectorial es muy alta. El vectorizador de texto se ha implementado dejando todas estas opciones expuestas como parámetros.

La figura 4.1 muestra un ejemplo de tokenización, normalización y vectorización TF-IDF de 3 mensajes del foro de Psicofarmacología como si constituyesen un corpus. Los términos normalizados están en minúscula, sin tildes y truncados, y no incluyen palabras vacías, como la conjunción "y". La vectorización se presenta en forma de matriz término–documento, con 34 filas (términos) y 3 columnas (documentos). En la representación matricial, los dos primeros documentos resultan ser ortogonales, si bien ambos tienen una pequeña proyección sobre el tercero. El término "plaz", de "plazo", aparece una vez en los mensajes primero y tercero, y su peso queda algo atenuado por aparecer en $2/3$ de los documentos. Lo mismo pasa con el término "asi" en los mensajes segundo y tercero. Los términos "aparec" y "entreg" son comunes a todos los documentos y, precisamente por ello, sus pesos están totalmente atenuados y las dos filas correspondientes son nulas. Los demás términos aparecen una sola vez en un único documento, por lo que tienen el mismo peso en sus documentos de origen.

La tabla 4.1a muestra las medidas internas medias obtenidas aplicando representaciones tipo BOW a los agrupamientos originales por subforo e hilo de los foros de la UNED. Los coeficientes de silueta obtenidos son todos negativos, aunque pequeños en valor absoluto. Esto indica que los agrupamientos son generalmente malos bajo esta óptica y probablemente exista solapamiento de grupos. Las configuraciones ensayadas incluyen combinaciones de los modelos OH, TF, TF-IDF y LogTF-IDF con frecuencias de término mínimas entre 1 y 3 y distancias coseno y angular. Los resultados menos negativos, aun así malos, son los obtenidos para el agrupamiento por subforos con el modelo TF-IDF, frecuencia mínima de 2 y, por un pequeño margen, medida coseno. Esta configuración parecería razonable incluso *a priori*, dado que el modelo TF-IDF es popular y, por ejemplo, Onan y Toçoğlu (2021) obtienen sus mejores resultados con este pesado. Además, la frecuencia mínima de 2 podría ayudar a reducir el ruido introducido por términos muy raros o erratas de ocurrencia única, como "puedda", que aparece en el primer mensaje de la figura 4.1. En todo caso, el hecho de que el coeficiente de silueta sea malo no significa que el agrupamiento sea con certeza malo. En línea con lo dicho en la sección sección 2.3, podría ser que los agrupamientos fuesen no globulares y la silueta no

se comportase bien con ellos. Más adelante, en el contexto del *clustering* basado en densidad, se probará la medida DBCV, diseñada para grupos de forma arbitraria.

La parte derecha de la tabla muestra las medidas de consistencia de proximidad, que se antojan más aceptables, especialmente para los agrupamientos por subforo. La interpretación es que, pese a la mediocre calidad del agrupamiento a juzgar por la silueta, en un ejercicio de recolocación individual de mensajes basado en máxima proximidad, cerca del 70 % de los mensajes permanecería en el subforo original. Desde este punto de vista, los mejores resultados se obtienen con el modelo LogTF-IDF y frecuencia mínima de 2. Atendiendo a la definición de la consistencia, no es de extrañar que la del agrupamiento por hilos sea menor, puesto que se trata de un refinamiento del agrupamiento por subforos. Además, un 12 % de los mensajes está alojado en hilos con un único mensaje, mientras que para los subforos dicha proporción es inferior al 1 %. Como se advirtió durante la exploración de datos en la sección 3.3, en este trabajo no se intentará la agrupación o recolocación automática de mensajes a nivel de hilo, sino que se agrupará en base a la similitud semántica y si acaso se comparará con los subforos.

Tabla 4.1: Medidas de evaluación internas para los agrupamientos originales por subforo e hilo con representaciones de tipo bolsa de palabras.

(a) Medidas internas promedio para diferentes modelos y frecuencias mínimas.

	Medida	silueta				consistencia	
	Distancia	coseno		angular		coseno/angular	
	Clave agr.	subforo	hilo	subforo	hilo	subforo	hilo
Modelo	Frec. mín.						
OH	1	−0.019	−0.064	−0.015	−0.044	0.670	0.366
	2	−0.019	−0.068	−0.016	−0.047	0.674	0.371
	3	−0.020	−0.074	−0.016	−0.051	0.667	0.360
TF	1	−0.017	−0.069	−0.014	−0.047	0.677	0.391
	2	−0.018	−0.073	−0.014	−0.050	0.680	0.395
	3	−0.019	−0.079	−0.015	−0.054	0.677	0.389
TF-IDF	1	−0.008	−0.031	−0.008	−0.023	0.709	0.436
	2	−0.006	−0.033	−0.007	−0.025	0.712	0.441
	3	−0.007	−0.043	−0.008	−0.031	0.694	0.420
LogTF-IDF	1	−0.008	−0.029	−0.008	−0.022	0.711	0.447
	2	−0.007	−0.031	−0.008	−0.023	0.713	0.450
	3	−0.008	−0.040	−0.008	−0.030	0.695	0.427

(b) Medidas internas por foro para el modelo LogTF-IDF con frecuencia mínima 2.

	Medida	silueta				consistencia	
	Distancia	coseno		angular		coseno/angular	
Foro	Clave agr.	subforo	hilo	subforo	hilo	subforo	hilo
6201302-_2019_cl		−0.003	−0.020	−0.001	−0.014	0.749	0.452
6390103-_2019_cl		−0.073	−0.053	−0.073	−0.048	0.701	0.446
ForosPL1_17-18		0.013	−0.008	0.008	−0.008	0.683	0.484
ForosPL1_18-19		0.020	−0.012	0.014	−0.008	0.696	0.548
ForosPL1_19-20		−0.001	0.019	0.001	0.010	0.693	0.392
ForosPL1_20-21		−0.004	−0.048	−0.002	−0.036	0.758	0.401
Foros_Acceso_Mayores_cl		−0.000	−0.092	−0.000	−0.060	0.711	0.430

Por si existieran diferencias importantes entre foros, la tabla 4.1b muestra las medidas internas de cada uno de ellos para la representación TF-IDF con frecuencia mínima de 2. Las siluetas siguen siendo pequeñas en términos absolutos, pero con este detalle aparecen algunas ligeramente positivas en foros de Procesadores del Lenguaje. Las diferencias no son suficientes para decir que los agrupamientos originales sean buenos para unos foros y malos para otros, sino que siguen siendo generalmente mediocres de acuerdo con la silueta. Tampoco se observan diferencias notables entre las consistencias de los distintos foros.

4.2. *Embedding* de palabras

Los *embeddings* semánticos de palabras constituyen el siguiente nivel de sofisticación en representación de documentos después de la bolsa de palabras. En este tipo de representación, una red neuronal se encarga de asignar a cada palabra un punto en un espacio vectorial denso de dimensión relativamente baja, típicamente de unos pocos cientos (véase, p. ej., Jurafsky y Martin, 2023, cap. 6). Estas redes se entrenan sobre enormes corpus de manera tal que las palabras con significado parecido, entendido como que suelen aparecer en contextos similares, quedan colocadas próximas como vectores en el espacio. Dos de los algoritmos de aprendizaje y cómputo de *embeddings* de palabras más conocidos son word2vec (Mikolov *et al.*, 2013a,b) y fastText (Bojanowski *et al.*, 2017). Ambos tienen dos versiones, una basada en la bolsa de palabras continua (CBOW) y otra en el modelo *skip-gram*, que vienen incluidas en la librería Gensim. Una diferencia importante entre word2vec y fastText es que el segundo maneja *n*-gramas además de palabras, lo que le procura un mejor comportamiento frente a palabras o inflexiones desconocidas, que son sencillamente descartadas por el primero.

El corpus de foros de la UNED recibido para el desarrollo de este trabajo, con unas 260 905 palabras en total, sería del todo insuficiente para el entrenamiento de *embeddings* semánticos. Por fortuna, en estos casos se puede recurrir a modelos preentrenados sobre corpus gigantes, aunque no siempre del mismo ámbito, disponibles libremente en la red. Para los ensayos de esta sección se han seleccionado el modelo word2vec para español de Almeida y Bilbao (2018), entrenado sobre un corpus de más de 3 000 millones de palabras y con dimensión de salida 400, y uno de los modelos oficiales de fastText para español (Grave *et al.*, 2018), entrenado sobre un gigantesco corpus de unos 73 000 millones de palabras y con dimensión de salida 300. Estos modelos pueden cargarse en Gensim como correspondencias entre palabras y vectores y utilizarse conjuntamente con el contador de palabras generalizado de la sección anterior para obtener representaciones semánticas de los mensajes. Sin embargo, aquí tiene menos sentido truncar las terminaciones o aplicar un umbral mínimo de frecuencia, puesto que, para cada palabra, el *embedding* conocerá su significado y le asignará un vector, o si no la descartará sin problema. Por tanto, algunos parámetros del normalizador de texto y del contador de palabras se ajustarán de forma diferente en los *pipelines* que utilicen esta representación.

La tabla 4.2a muestra las medidas internas medias obtenidas aplicando los *embeddings* de

palabras word2vec y fastText con ponderación OH, TF, TF-IDF y LogTF-IDF a los agrupamientos originales por subforo. Los coeficientes de silueta son de nuevo negativos e indican mala calidad de agrupamiento o presencia de grupos no globulares. En términos relativos, las siluetas son más negativas que las logradas con la representación BOW, especialmente para la distancia coseno. No se aprecia gran diferencia a nivel de silueta entre los modelos word2vec y fastText. La consistencia de proximidad es asimismo peor que la alcanzada con BOW, con resultados en torno al 60 %, frente al 70 % visto en la sección anterior. Los mejores valores de consistencia y casi de silueta son los obtenidos con ponderación TF, que se detallan por foro en la tabla 4.2b. En esta ocasión se observa algo más de variación entre foros, al menos en la consistencia, que va del 48 % al 69 % y es un poco mejor con fastText que con word2vec.

Es una lástima que los resultados obtenidos usando *embeddings* de palabras no mejoren los obtenidos con representaciones BOW. Como se dice en Onan (2019), si bien estos *embeddings* son capaces de capturar y representar relaciones semánticas y sintácticas entre palabras, para que den buenos resultados es importante que su entrenamiento esté alineado con la tarea o el problema que se pretende resolver. En este caso, los modelos utilizados vienen preentrenados sobre corpus enormes, pero generalistas, y quizá no sean del todo adecuados para la temática de los foros. En el mismo artículo se comenta que otros investigadores han tratado de mejorar el rendimiento de los *embeddings* de palabras preentrenados incorporando un contexto local o combinándolos con otras técnicas de NLP. Aquí no se profundizará más en las limitaciones de los *embeddings* de palabras y se pasará directamente a los modelos de última generación

Tabla 4.2: Medidas de evaluación internas para los agrupamientos originales por subforo con representaciones por *embedding* de palabras ponderado por frecuencia.

(a) Medidas internas promedio para diferentes *embeddings* y ponderaciones.

	Medida	silueta				consistencia	
	Distancia	coseno		angular		coseno/angular	
	Modelo	word2vec	fastText	word2vec	fastText	word2vec	fastText
Ponderación	Frec. mín.						
OH	1	−0.091	−0.101	−0.055	−0.056	0.561	0.598
TF	1	−0.084	−0.084	−0.051	−0.047	0.601	0.621
TF-IDF	1	−0.087	−0.082	−0.052	−0.046	0.587	0.609
LogTF-IDF	1	−0.090	−0.088	−0.054	−0.050	0.579	0.609

(b) Medidas internas por foro con ponderación TF y frecuencia mínima 1.

	Medida	silueta				consistencia	
	Distancia	coseno		angular		coseno/angular	
Foro	Modelo	word2vec	fastText	word2vec	fastText	word2vec	fastText
6201302-_2019_cl		−0.155	−0.119	−0.089	−0.066	0.564	0.564
6390103-_2019_cl		−0.125	−0.119	−0.101	−0.096	0.566	0.636
ForosPL1_17-18		−0.016	−0.018	−0.009	−0.009	0.489	0.527
ForosPL1_18-19		−0.027	−0.046	−0.011	−0.020	0.648	0.648
ForosPL1_19-20		−0.055	−0.076	−0.028	−0.035	0.574	0.591
ForosPL1_20-21		−0.180	−0.203	−0.101	−0.104	0.669	0.679
Foros_Acceso_Mayores_cl		−0.032	−0.006	−0.015	0.002	0.693	0.700

para la representación de textos cortos.

4.3. *Embedding* de frases

Los *embeddings* de frases generan representaciones vectoriales densas y de tamaño fijo de frases, párrafos o textos cortos utilizando todo el contexto. De esta manera, son capaces de interpretar correctamente el sentido de cada palabra, las relaciones entre palabras y el significado del texto al completo. Las representaciones que generan son semánticamente coherentes, de modo que frases con significado parecido acaban próximas en el espacio vectorial. Los modelos de este tipo se implementan usando sofisticadas arquitecturas de redes neuronales, que a menudo incorporan mecanismos de autoatención en forma de transformadores (Vaswani *et al.*, 2017). Los *embeddings* se entrenan sobre corpus tan grandes que su uso habitual es a través de modelos preentrenados y si acaso refinados para la tarea concreta, sea ésta de clasificación, agrupación, búsqueda semántica u otra de las muchas aplicaciones de estos *embeddings*.

Dos de los modelos semánticos de frases más avanzados y utilizados en la actualidad son *Sentence*-BERT (SBERT, Reimers y Gurevych, 2019) y *Universal Sentence Encoder* (USE, Cer *et al.*, 2018). Ambos disponen de modelos multilingües susceptibles de ser usados para la representación de los mensajes de los foros. Por ejemplo, el paquete SentenceTransformers da acceso al modelo SBERT paraphrase-multilingual-mpnet-base-v2 (Reimers y Gurevych, 2020), que reconoce más de 50 lenguas, incluido el español, y tiene dimensión de salida 768. Una limitación importante de este modelo y todos los multilingües incluidos en el paquete es que, por defecto, la longitud máxima de entrada es de 128 palabras. Aunque dicho límite puede subirse hasta 512, el mensaje más largo de los foros contiene 546 palabras, por lo que ese mensaje y quizá algún otro quedarán en efecto truncados en su representación. Por otra parte, el popular paquete TensorFlow (Abadi *et al.*, 2016) facilita el acceso al modelo USE multilingüe (Yang *et al.*, 2019), que admite 16 lenguas, incluido del español, y tiene dimensión de salida 512. Este modelo no tiene límite estricto sobre la longitud del texto de entrada, pero

Tabla 4.3: Medidas de evaluación internas para los agrupamientos originales por subforo con representaciones por *embedding* de frases.

Medida	silueta				consistencia	
Distancia	coseno		angular		coseno/angular	
Foro / Modelo	SBERT	USE	SBERT	USE	SBERT	USE
6201302-_2019_cl	−0.032	−0.042	−0.015	−0.023	0.689	0.717
6390103-_2019_cl	−0.106	−0.076	−0.089	−0.073	0.697	0.713
ForosPL1_17-18	−0.003	0.013	0.000	0.010	0.645	0.685
ForosPL1_18-19	0.016	0.037	0.013	0.024	0.704	0.763
ForosPL1_19-20	−0.058	−0.010	−0.031	−0.003	0.642	0.688
ForosPL1_20-21	−0.126	−0.026	−0.072	−0.014	0.775	0.725
Foros_Acceso_Mayores_cl	−0.029	0.014	−0.015	0.009	0.693	0.722
promedio	−0.048	−0.013	−0.030	−0.010	0.692	0.716

cuanto más largo sea, más diluido quedará su significado. A diferencia de los de las secciones anteriores, este vectorizador toma como entrada texto sin segmentar, luego los *pipelines* que lo usan no llevan normalizador de texto.

La tabla 4.3 muestra las medidas internas obtenidas aplicando los *embeddings* SBERT y USE a los agrupamientos originales por subforo. Los coeficientes de silueta medios son una vez más negativos, pero en general mejores que los conseguidos con los *embeddings* de palabras. A su vez, las siluetas obtenidas con USE son casi todas mejores que las logradas con SBERT, y no quedan demasiado lejos de las menos malas alcanzadas con representaciones BOW. Los resultados de consistencia de proximidad son también mejores con USE que con SBERT. De hecho, el promedio de 71.6 % conseguido con USE es el mejor resultado de consistencia hasta el momento, superando mínimamente el 71.3 % logrado con el modelo LogTF-IDF y frecuencia mínima 2. La variación de la consistencia con el foro es ligeramente mayor que la vista para LogTF-IDF, pero aun así significativamente mejor que la de los *embeddings* de palabras.

Una última diferencia entre los dos *embeddings* de frases ensayados, importante desde el punto de vista computacional, es que el modelo multilingüe preentrenado USE corre entre 20 y 30 veces más rápido que el correspondiente modelo SBERT. Por ejemplo, en un portátil con 8 núcleos, USE vectoriza todos los mensajes de los foros, excluyendo grupos de tutoría, en menos de 8 segundos, mientras que SBERT se demora más de 4 minutos en la misma tarea. Unido a los resultados de medidas internas, esto colocaría de momento al modelo USE como opción preferente de representación por *embedding* de frases para el sistema de agrupación automática y colocación de mensajes nuevos.

Capítulo 5

Algoritmos de agrupamiento

Una vez discutidas la representación de documentos y las medidas de similitud, se puede pasar al desarrollo del último componente principal del sistema, que es el agrupamiento. Los algoritmos probados a continuación van desde los más clásicos y sencillos hasta otros bastante recientes y sofisticados, basados en la densidad. Estos últimos podrían resultar convenientes, dado que los resultados de silueta vistos hasta ahora sugieren que los agrupamientos podrían ser no globulares. También se incluyen algoritmos jerárquicos, que quizá ayuden a comprender la estructura de los foros y faciliten la determinación del número óptimo de grupos.

5.1. K-medias

El primero de los métodos de agrupamiento ensayados es el tradicional k-medias, que es un algoritmo de partición iterativo basado en centroides (véase, p. ej., Han *et al.*, 2012, sec. 10.2.1). El algoritmo define los centroides como las medias de los vectores en cada grupo y trata de minimizar la inercia, esto es, la media cuadrática de las distancias de los centroides a los vectores de cada grupo. Al inicio se asignan k grupos al azar, tomando k como parámetro de entrada conocido. En cada iteración, primero se recalculan los k centroides y acto seguido se actualizan los grupos, asignando cada vector al centroide más cercano. Este proceso continúa hasta que los grupos se estabilizan o se llega a un número máximo de iteraciones. Para mitigar la dependencia con respecto a la configuración inicial, a menudo se corren varias repeticiones del proceso agrupamiento y se escoge el resultado con menor inercia. El agrupador k-medias se ha implementado como envoltorio del correspondiente agrupador de la librería NLTK, que es de los pocos que permite seleccionar una función de distancia distinta a la euclídea.

La tabla 5.1 muestra las medidas de evaluación promedio obtenidas aplicando el algoritmo k-medias con semilla aleatoria fija y 10 repeticiones a las representaciones de los foros más prometedoras de las secciones anteriores. Para poder atisbar al menos la influencia del parámetro fundamental del algoritmo, se examinan valores de k comprendidos entre el número de grupos original menos 2 y más 2, parametrizados por la diferencia $\tilde{k} = k - k_{\text{orig.}}$. Comenzando por

las medidas internas, los coeficientes de silueta promedio son relativamente pequeños, pero a diferencia de los vistos para los agrupamientos originales, positivos. Esto tiene sentido, puesto que, al tratar de minimizar la inercia, el algoritmo k-medias tiende a incrementar la silueta. Como se dijo anteriormente, el que las siluetas de los agrupamientos originales sean predominantemente negativas podría deberse a que dichos agrupamientos fueran no globulares o tuvieran solapamientos. Los mejores resultados de silueta se consiguen con los *embeddings* de frases y la distancia coseno. Igualmente, los mejores resultados de consistencia de proximidad corresponden a los *embeddings* de frases, que alcanzan valores cercanos al 80 % con el número de grupos original. Naturalmente, la consistencia aumenta cuando se reduce dicho número.

En cuanto a las medidas externas, el ARI y la AMI toman valores entre 0.095 y 0.225 y entre 0.164 y 0.289, respectivamente. Dado que ambas medidas incorporan ajustes de aleatoriedad, estos valores expresan un alineamiento no despreciable, mejor que azaroso, entre los agrupamientos generados automáticamente y los originales. Los mejores resultados de ARI y AMI se alcanzan con el *embedding* de frases USE, el número de grupos original y, por poco, la distancia angular. Los valores de la AMI son en todos los casos mayores que los del ARI.

Tabla 5.1: Medidas de evaluación promedio de los agrupamientos obtenidos con el algoritmo k-medias para diferentes representaciones y números de grupos, parametrizados por la diferencia $\tilde{k} = k - k_{\text{orig.}}$ con respecto al número de subforos original.

		Medida	silueta		consistencia		ARI		AMI	
		Distancia	coseno	angular	coseno	angular	coseno	angular	coseno	angular
Modelo	Frec. mín.	$\tilde{k}$								
LogTF-IDF	2	-2	0.037	0.027	0.675	0.658	0.157	0.152	0.211	0.208
	2	-1	0.039	0.027	0.657	0.654	0.130	0.129	0.196	0.187
	2	0	0.039	0.029	0.640	0.659	0.128	0.133	0.194	0.198
	2	$+1$	0.041	0.031	0.614	0.622	0.121	0.128	0.193	0.200
	2	$+2$	0.043	0.031	0.615	0.615	0.122	0.122	0.205	0.205
word2vec-TF	1	-2	0.096	0.070	0.744	0.762	0.113	0.130	0.164	0.173
	1	-1	0.083	0.057	0.731	0.739	0.111	0.122	0.174	0.184
	1	0	0.083	0.056	0.741	0.734	0.121	0.122	0.182	0.180
	1	$+1$	0.084	0.058	0.718	0.712	0.102	0.103	0.175	0.166
	1	$+2$	0.075	0.051	0.709	0.709	0.095	0.095	0.168	0.168
fastText-TF	1	-2	0.097	0.064	0.782	0.780	0.133	0.129	0.190	0.196
	1	-1	0.072	0.048	0.753	0.733	0.116	0.117	0.192	0.192
	1	0	0.072	0.057	0.711	0.731	0.114	0.128	0.184	0.190
	1	$+1$	0.071	0.056	0.724	0.724	0.117	0.118	0.190	0.191
	1	$+2$	0.071	0.057	0.713	0.727	0.125	0.123	0.199	0.203
SBERT		-2	0.134	0.087	0.815	0.816	0.170	0.175	0.238	0.238
		-1	0.125	0.081	0.803	0.803	0.159	0.159	0.248	0.248
		0	0.130	0.084	0.799	0.799	0.163	0.163	0.245	0.245
		$+1$	0.123	0.081	0.779	0.780	0.140	0.144	0.240	0.241
		$+2$	0.124	0.080	0.773	0.764	0.151	0.143	0.255	0.244
USE		-2	0.136	0.087	0.843	0.841	0.225	0.221	0.288	0.274
		-1	0.125	0.082	0.802	0.802	0.191	0.191	0.268	0.268
		0	0.122	0.083	0.791	0.800	0.203	0.220	0.283	0.289
		$+1$	0.119	0.078	0.773	0.773	0.180	0.180	0.259	0.259
		$+2$	0.111	0.074	0.765	0.777	0.161	0.169	0.267	0.274

Por supuesto, son medidas diferentes, una basada en el conteo de pares y la otra en la teoría de la información. No obstante, la tendencia podría deberse a que el ARI prime los agrupamientos formados por grupos grandes de tamaño similar y la AMI se comporte mejor frente agrupamientos de referencia no balanceados y con grupos pequeños (Romano *et al.*, 2016). Como se vio en la tabla 3.3, los agrupamientos originales son heterogéneos e incluyen subforos minúsculos, por lo que la AMI se antoja más idónea como medida de validación externa.

La tabla 5.2 presenta los resultados de validación por foro para los *clustering* k-medias con el número de grupos original. Si bien las medidas no son homogéneas a través de los 7 foros, la mayoría de las observaciones anteriores sigue en pie. Así, en casi todos los casos, los mejores resultados de silueta y consistencia de proximidad se obtienen con uno de los *embeddings* de

Tabla 5.2: Medidas de evaluación por foro de los agrupamientos obtenidos con el algoritmo k-medias y el número de grupos original para diferentes representaciones.

		Medida	silueta		consistencia		ARI		AMI	
Modelo	Foro	Distancia	coseno	angular	coseno	angular	coseno	angular	coseno	angular
LogTF-IDF (2)	6201302-_2019_cl		0.021	0.013	0.632	0.632	0.132	0.132	0.189	0.189
	6390103-_2019_cl		0.084	0.077	0.600	0.600	0.157	0.157	0.197	0.197
	ForosPL1_17-18		0.028	0.019	0.769	0.769	0.208	0.208	0.259	0.259
	ForosPL1_18-19		0.036	0.024	0.641	0.641	0.138	0.138	0.231	0.231
	ForosPL1_19-20		0.044	0.023	0.602	0.653	0.098	0.080	0.192	0.147
	ForosPL1_20-21		0.021	0.021	0.599	0.685	0.050	0.101	0.117	0.189
	Foros_Acceso_Mayores_cl		0.039	0.025	0.635	0.635	0.111	0.111	0.176	0.176
word2vec-TF (1)	6201302-_2019_cl		0.086	0.053	0.726	0.726	0.164	0.164	0.202	0.202
	6390103-_2019_cl		0.123	0.107	0.723	0.723	0.104	0.104	0.207	0.207
	ForosPL1_17-18		0.086	0.055	0.769	0.769	0.101	0.101	0.148	0.148
	ForosPL1_18-19		0.058	0.047	0.700	0.663	0.150	0.170	0.214	0.179
	ForosPL1_19-20		0.099	0.052	0.716	0.705	0.088	0.076	0.111	0.128
	ForosPL1_20-21		0.077	0.048	0.801	0.801	0.108	0.108	0.201	0.201
	Foros_Acceso_Mayores_cl		0.050	0.032	0.753	0.753	0.135	0.135	0.195	0.195
fastText-TF (1)	6201302-_2019_cl		0.065	0.032	0.653	0.698	0.126	0.162	0.189	0.210
	6390103-_2019_cl		0.112	0.103	0.692	0.692	0.124	0.124	0.206	0.206
	ForosPL1_17-18		0.066	0.063	0.739	0.809	0.090	0.124	0.161	0.165
	ForosPL1_18-19		0.054	0.034	0.715	0.715	0.185	0.185	0.220	0.220
	ForosPL1_19-20		0.083	0.055	0.727	0.727	0.081	0.081	0.158	0.158
	ForosPL1_20-21		0.086	0.052	0.728	0.728	0.064	0.064	0.159	0.159
	Foros_Acceso_Mayores_cl		0.039	0.063	0.724	0.744	0.130	0.154	0.193	0.211
SBERT	6201302-_2019_cl		0.139	0.086	0.848	0.848	0.210	0.210	0.330	0.330
	6390103-_2019_cl		0.186	0.146	0.815	0.815	0.197	0.197	0.306	0.306
	ForosPL1_17-18		0.118	0.072	0.753	0.753	0.093	0.093	0.166	0.166
	ForosPL1_18-19		0.150	0.092	0.852	0.852	0.278	0.278	0.371	0.371
	ForosPL1_19-20		0.098	0.061	0.761	0.761	0.107	0.107	0.169	0.169
	ForosPL1_20-21		0.122	0.074	0.768	0.768	0.068	0.068	0.175	0.175
	Foros_Acceso_Mayores_cl		0.095	0.057	0.794	0.794	0.189	0.189	0.198	0.198
USE	6201302-_2019_cl		0.096	0.060	0.728	0.728	0.204	0.204	0.276	0.276
	6390103-_2019_cl		0.186	0.146	0.808	0.808	0.192	0.192	0.283	0.283
	ForosPL1_17-18		0.095	0.059	0.831	0.831	0.221	0.221	0.307	0.307
	ForosPL1_18-19		0.122	0.075	0.811	0.811	0.283	0.283	0.411	0.411
	ForosPL1_19-20		0.110	0.070	0.784	0.784	0.178	0.178	0.201	0.201
	ForosPL1_20-21		0.141	0.089	0.805	0.805	0.167	0.167	0.277	0.277
	Foros_Acceso_Mayores_cl		0.106	0.081	0.769	0.832	0.179	0.295	0.223	0.271

frases. En las columnas de ARI y AMI, y también en la de consistencia, salta a la vista que, en multitud de casos, la elección de la distancia coseno o angular no altera el resultado, casi con seguridad porque el agrupamiento final es idéntico. Aun así, de promedio, la distancia angular arroja resultados ligeramente mejores para estas medidas de evaluación. Cabe recordar que la distancia angular es una función creciente de la distancia coseno, con la posible ventaja de ser una métrica en sentido estricto y distinguir con mayor precisión entre vectores casi alineados.

El mejor resultado de AMI es el 0.411 obtenido con la representación USE para el foro de Procesadores del Lenguaje de 2018–19, que parece decente para lo encontrado hasta ahora. Lo curioso es que la AMI se reduce a menos de la mitad para el mismo foro de 2019–20, que debería de tener una temática y estructura similar. En la tabla 5.3, que muestra la matriz de confusión para el agrupamiento con representación USE del foro de Procesadores del Lenguaje de 2018–19, se ve cómo el sistema identifica casi a la perfección el subforo de estudiantes como el grupo 4. La figura 5.1 muestra los 37 mensajes del foro de estudiantes divididos entre los que acaban en los grupos 4 y 5. Los 35 del grupo 4 tienen que ver con peticiones de adhesión a un grupo de Telegram y suelen incluir alias y agradecimientos. Los otros 2 que acaban en el grupo 5 junto a mensajes omitidos de otros subforos, en un caso se refieren a un asunto distinto, y en otro quizá sean suficientemente largos para que su significado sea diluido por el *embedding*. En el agrupamiento del foro de Procesadores del Lenguaje de 2019–20 no existe ese grupo grande y puro referido a Telegram dentro del subforo de estudiantes. De hecho, según la tabla 3.3, el subforo de estudiantes de ese año sólo contiene 2 mensajes.

Por dar un ejemplo de distribución a la inversa, la figura 5.2 muestra divididos por subforo original los mensajes del grupo 2 del mismo agrupamiento con matriz de confusión 5.3. En este caso, los mensajes contienen mayormente agradecimientos por alguna respuesta anterior. El segundo del foro de consultas generales contiene una petición concreta adicional, pero por lo demás y firmas aparte, los mensajes parecen bastante genéricos y sin especial relación con el tema o al menos título del subforo. De hecho, hay mensajes prácticamente idénticos alojados en subforos diferentes. Por tanto, el algoritmo parecería haber encontrado un grupo válido en sentido semántico, pero transversal al agrupamiento original, por lo que no contribuiría muy

Tabla 5.3: Matriz de confusión del agrupamiento del foro de Procesadores del Lenguaje de 2018–19 con algoritmo *k*-medias, representación USE y número de grupos original.

Subforo \ Grupo	0	1	2	3	4	5	6	7	Total
Foro Análisis léxico (fuera de la práctica)	0	0	0	0	0	1	0	0	1
Foro Análisis sintáctico (fuera de la práctica)	0	4	0	2	0	1	0	0	7
Foro General Práctica	2	5	7	1	0	8	22	1	46
Foro Práctica: análisis léxico	1	10	8	12	0	1	4	19	55
Foro Práctica: análisis sintáctico	25	36	6	12	0	0	3	7	89
Foro de consultas generales	0	0	8	0	0	17	3	0	28
Foro de estudiantes (no moderado por el Equipo Docente)	0	0	0	0	35	2	0	0	37
Coordinación tutorial	0	0	0	0	0	7	0	0	7
Total	28	55	29	27	35	37	32	27	270

Grupo 4

```
Buenas,¶¶Te dejo mi alias: [redacted]¶¶Un saludo¶¶
Buenas,¶¶Corrección: @[redacted]¶¶¶¶Gracias¶¶
Hola,¶¶Mi alias es @[redacted].¶¶Muchas gracias!!!¶¶
Hola!¶¶Yo también estaría interesado en entrar, mi alias es¶@[redacted]¶¶Muchas gracias¶¶
Buenas. Mi alias es @[redacted]¶¶
Hola.¶¶¿Puedes agregarme al grupo de telegram?¶¶@[redacted]¶¶Gracias¶¶
Buenos días a tod@s,¶¶Mi usuario es @[redacted]¶¶Gracias por agregarme.¶¶Saludos¶¶
Hola!¶¶Yo también estaría interesado en entrar, mi alias es¶@[redacted]¶¶Muchas gracias¶¶
Buenos días,¶¶Mi usuario es [redacted]¶¶Un saludo y gracias¶¶
Hola,¶¶mi alias: @[redacted]¶¶Gracias¶¶
Mi alias es @[redacted]¶¶Gracias :D¶¶
Hola¶¶Mi usuario es @[redacted]¶¶Muchas gracias¶¶
Dejo mi alias para que me incluyais, gracias!¶¶[redacted]¶¶
El mio es @[redacted].¶¶¶¶Gracias¶¶
Saludos¶¶Mi alias es @[redacted]¶¶Muchas gracias¶¶
El mio es @[redacted], gracias¶¶
Otro más por aqui:¶¶@[redacted]¶¶¶¶Gracias!¶¶
¿Podeis agregarme por favor? @[redacted]¶¶Muchas gracias!¶¶
Muy buenas, por favor incluidme en el grupo, mi alias es [redacted]¶¶Saludos¶¶
Mi telegran es: <{URL}>¶¶
mi alias es @[redacted]¶¶
Mi alias es [redacted]¶¶Gracias!¶¶
Hola!¶¶Mi alias @[redacted]¶¶¶¶Gracias!¶¶
Buenas [redacted],¶¶Mi usuario es @[redacted], ¿me podrías añadir?¶Gracias! Un saludo¶¶
Hola¶¶Mi alias, @[redacted]¶¶Gracias,¶Un saludo¶¶
Hola, mi alias es @[redacted], ¿me podríais agregar?¶Gracias.¶¶
Buenos días:¶¶Mi alias es @[redacted]¶¶Gracias¶¶
Buenas noches, si podeis añadirme..¶¶Alias: [redacted]¶¶Gracias, un saludo.¶¶
Buenas tardes,¶¶Yo también estaria interesado en apuntarme mi alias es¶@[redacted]¶¶Muchas gracias.¶¶
A ti no te encuentro. ¿Es correcto tu alias?¶¶
Creo que estais todos agregados. Avisadme en caso contrario.¶¶
Podrías añadirme por favor? @[redacted]¶¶
Hola, me gustaría que me añadierais si es posible. Mi¶alias es @[redacted]. Muchas gracias!¶¶
Hola mi usuario de Telegram es @[redacted] por si me podeis añadir.¶¶Gracias.¶¶
Hola, sé que llego un poco tarde pero me gutaria entrar en el¶grupo, mi alias es @[redacted]¶¶Muchas gracias.¶¶
```

Grupo 5

```
Buenas compañeros,¶¶Escribo para informaros de que tenemos un grupo en Telegram para¶discutir temas de esta
↪  asignatura y resolvernos dudas entre nosotros.¶Y también claro para conocernos un poco, que la¶educación a
↪  distancia es muy solitaria. Si a alguien le¶interesa unirse sólo tiene que dejar aquí su alias y se¶le
↪  agregará.¶¶Todo el mundo es bienvenido <{IMAGEN}>¶¶
Con vistas a preparar el examen de la segunda semana,¶¿podría alguien que haya hecho ya el examen compartir el¶de
↪  la primera?¶¶
```

Figura 5.1: Distribución de los mensajes del subforo de estudiantes en el agrupamiento del foro de Procesadores del Lenguaje de 2018–19 con algoritmo *k*-medias, representación USE y número de grupos original.

positivamente a una medida de evaluación externa. Este ejemplo es importante de cara a las expectativas puestas sobre el sistema de agrupación automática, que atendiendo únicamente al texto de los mensajes, difícilmente podría asignar mensajes casi idénticos a grupos distintos alineados con los originales.

5.2. Aglomerativo

El segundo algoritmo de *clustering* ensayado pertenece a la familia de métodos jerárquicos. Estos métodos agrupan objetos a diferentes niveles de granularidad en una estructura de tipo árbol, que puede resultar útil para el resumen, la visualización y la búsqueda de datos (véase, p. ej., Aggarwal y Zhai, 2012; Han *et al.*, 2012, sec. 10.3). Específicamente, los algoritmos

Foro General Práctica

```
Gracias y un saludo¶¶
Hola,¶¶Sí, eso era. Muchas gracias por la aclaración.¶¶Saludos.¶¶
reservadas.¶Gracias por responder.¶¶Un saludo.¶¶
Gracias a ambos por la respuesta.¶¶
Gracias por las contestaciones.¶¶Saludos.¶¶
Gracias.¶¶
Muchas gracias [redacted]. ¡Saludos!¶¶
```

Foro Práctica: análisis léxico

```
Hola¶¶Muchas gracias por la aclaración.¶¶Saludos.¶¶
Entendido, no me habia fijado que lo especificaba en el enunciado,¶ahora lo revisaré y retocaré la
↪  practica.¶¶Muchas gracias [redacted]¶¶
Aclarado, muchas gracias.¶¶Un saludo.¶¶[redacted]¶¶
Muchas gracias alvaro por la aclaración¶¶Un saludo¶¶
Gracias.¶¶
expanding {{1}}¶Gracias por la información [redacted].¶¶
expanding {{1}}¶Lo he comprobado y estan los mismos, gracias¶¶
expanding {{1}}¶Creo que he conseguido solucionarlo ya. Muchas gracias¶¶Un saludo¶¶[redacted]¶¶
```

Foro Práctica: análisis sintáctico

```
Vale, perfecto. Andaba buscando acerca de esta cuestión y he¶visto este hilo próximo a ella.¶¶Muchas gracias¶¶
Hola [redacted].¶¶Entendido.¶¶Gracias por la aclaración.¶¶Saludos¶¶[redacted]¶¶
Muchas gracias por la respuesta. Me ha quedado claro¶¶Un saludo¶¶
Le había dado un enfoque más complejo al paso de¶parámetros de lo que lo era en realidad.¶¶Me ha quedado claro,
↪  ahora, lo que se pide.¶¶Gracias y un saludo.¶¶
Aclarado.¶¶Muchas gracias.¶¶Un saludo.¶¶
Aclarado, gracias!!¶¶
```

Foro de consultas generales

```
Gracias por el comentario.¶¶
Muchas gracias por el vídeo de teoría, pero¶¿sería posible que nos adjuntaráis también¶las
↪  diapositivas?¶¶¶¶Muchas gracias¶¶Un saludo¶¶
Muchas gracias [redacted]¶¶
Muchas gracias [redacted]. Son de mucha ayuda.¶¶
Creo que puede servirte lo que adjunto¶¶
Gracias [redacted],¶¶Muy amable.¶¶
Gracias [redacted],¶¶Está muy detallado.¶¶Lo miro.¶¶
De acuerdo, duda resuelta, gracias¶¶
```

Figura 5.2: Distribución de los mensajes del grupo 2 en el agrupamiento del foro de Procesadores del Lenguaje de 2018–19 con k-medias, representación USE y número de grupos original.

aglomerativos comienzan asignando un grupo por objeto y después van uniendo grupos en función de su similitud hasta llegar a un solo grupo o cumplirse una condición de terminación, como un número predefinido de grupos. A diferencia del algoritmo k-medias, la configuración inicial es determinista y el proceso concluye en un máximo de iteraciones igual al número de objetos. Al mismo tiempo, una vez unidos dos grupos, éstos ya no se pueden separar, por lo que resulta crucial elegir una función de similitud o enlace adecuada para decidir qué grupos unir en cada paso. Aquí se probarán las habituales basadas en las distancias intergrupales

$$\text{mínima} \qquad d_{\text{mín}}(A,B) = \min_{\boldsymbol{u}\in A,\boldsymbol{v}\in B} d(\boldsymbol{u},\boldsymbol{v}), \tag{5.1}$$

$$\text{promedio} \quad d_{\text{prom}}(A,B) = \frac{\sum_{\boldsymbol{u}\in A,\boldsymbol{v}\in B} d(\boldsymbol{u},\boldsymbol{v})}{|A||B|} \tag{5.2}$$

$$\text{y máxima} \quad d_{\text{máx}}(A,B) = \max_{\boldsymbol{u}\in A,\boldsymbol{v}\in B} d(\boldsymbol{u},\boldsymbol{v}), \tag{5.3}$$

también conocidas como enlace simple, enlace promedio y enlace completo. Todas éstas están entre las opciones disponibles en el módulo de *clustering* jerárquico de la librería SciPy (Virtanen *et al.*, 2020), que además incluye funciones para visualizar el árbol de agrupamiento o dendrograma y cortarlo a diferentes niveles. Esto facilita el análisis de la estructura jerárquica y del impacto del número de grupos sobre las medidas de evaluación.

La tabla 5.4 muestra las medidas de evaluación promedio obtenidas aplicando el algoritmo aglomerativo con los tres tipos de enlace a los mismos modelos de representación ensayados con el algoritmo k-medias en la sección anterior. En todos los casos, el árbol de agrupamiento se ha cortado a la altura correspondiente al número de subforos original. Los coeficientes de silueta medios alcanzan valores significativamente mayores a los encontrados con el *clustering* k-medias, sobre todo para los *embeddings* de palabras y el enlace promedio. Sin embargo, con enlace simple aparecen siluetas muy pequeñas e incluso negativas para la representación BOW y los *embeddings* de frases. Esto no debería sorprender demasiado, ya que el enlace simple no promueve en absoluto el carácter globular del agrupamiento. Más bien, por su diseño, puede formar grupos de forma arbitraria siempre que en cada unión el objeto o grupo adicional sea cercano según la distancia mínima. Por este mismo motivo, corre el riesgo de ir encandenando objetos y grupos hasta formar grupos en los que haya pares de elementos muy poco similares.

Los resultados de consistencia de proximidad superan también los del *clustering* k-medias, sobrepasando el 90 % en la mayoría de los casos y llegando hasta el 98 % para el modelo LogTF-IDF con enlace simple. De nuevo, era de esperar que la consistencia fuera mayor para el enlace simple, porque su uso de la distancia mínima está alineado con aquel en la definición (2.5) de la consistencia. Conviene recordar que la consistencia no es propiamente una medida de calidad y que podría conseguirse una consistencia perfecta con un agrupamiento trivial consistente en

Tabla 5.4: Medidas de evaluación promedio de los agrupamientos obtenidos con el algoritmo aglomerativo y el número de grupos original para diferentes representaciones y tipos de enlace.

		Medida	silueta		consistencia		ARI		AMI	
		Distancia	coseno	angular	coseno	angular	coseno	angular	coseno	angular
Modelo	Frec. mín.	Enlace								
LogTF-IDF	2	simple	−0.035	−0.021	0.981	0.981	−0.003	−0.003	0.002	0.002
	2	promedio	0.041	0.030	0.930	0.928	0.186	0.179	0.221	0.217
	2	completo	0.006	0.006	0.906	0.906	0.065	0.065	0.109	0.109
word2vec-TF	1	simple	0.121	0.079	0.979	0.979	−0.008	−0.008	−0.009	−0.009
	1	promedio	0.212	0.142	0.957	0.952	0.011	0.017	0.050	0.054
	1	completo	0.107	0.070	0.810	0.810	0.072	0.072	0.126	0.126
fastText-TF	1	simple	0.245	0.155	0.979	0.979	−0.006	−0.006	−0.002	−0.002
	1	promedio	0.332	0.210	0.968	0.970	0.012	0.013	0.050	0.050
	1	completo	0.206	0.127	0.876	0.876	0.049	0.049	0.101	0.101
SBERT		simple	−0.015	−0.005	0.980	0.980	−0.004	−0.004	0.001	0.001
		promedio	0.125	0.077	0.951	0.954	0.082	0.084	0.168	0.164
		completo	0.095	0.061	0.854	0.854	0.118	0.118	0.206	0.206
USE		simple	0.025	0.018	0.979	0.979	−0.006	−0.006	−0.005	−0.005
		promedio	0.120	0.080	0.957	0.956	0.084	0.082	0.124	0.122
		completo	0.107	0.071	0.863	0.863	0.180	0.180	0.223	0.223

un solo grupo. Como se explicó en su momento, su introducción y uso en este trabajo se debe a su relación con el mecanismo propuesto para la colocación de mensajes nuevos.

Las medidas externas toman valores entre -0.008 y 0.180 para el ARI y -0.009 y 0.223 para la AMI. Los mejores resultados, aunque algo inferiores a los alcanzados con k-medias, son los obtenidos con el *embedding* USE y enlace completo y con la bolsa de palabras LogTF-IDF y enlace promedio. Una vez más, los valores de la AMI son generalmente mayores que los del ARI, lo que podría atribuirse a la heterogeneidad de los grupos de referencia. Saltan a la vista los promedios negativos o minúsculos de estas medidas externas para las configuraciones con enlace simple. Esto quiere decir que, aunque la consistencia sea buena y alguna silueta no sea del todo mala, esos agrupamientos generados automáticamente guardan poca relación con los

Tabla 5.5: Medidas de evaluación por foro de los agrupamientos obtenidos con el algoritmo aglomerativo, enlace promedio, número de grupos original y diferentes representaciones.

		Medida	silueta		consistencia		ARI		AMI	
Modelo	Foro	Distancia	coseno	angular	coseno	angular	coseno	angular	coseno	angular
LogTF-IDF (2)	6201302-_2019_cl		0.024	0.016	0.944	0.944	0.243	0.243	0.275	0.275
	6390103-_2019_cl		0.084	0.074	0.903	0.900	0.151	0.135	0.187	0.179
	ForosPL1_17-18		0.022	0.014	0.941	0.941	0.091	0.091	0.128	0.128
	ForosPL1_18-19		0.042	0.028	0.941	0.941	0.303	0.303	0.292	0.292
	ForosPL1_19-20		0.056	0.038	0.903	0.903	0.222	0.222	0.307	0.307
	ForosPL1_20-21		0.032	0.022	0.954	0.954	0.059	0.059	0.194	0.194
	Foros_Acceso_Mayores_cl		0.026	0.018	0.924	0.917	0.232	0.197	0.162	0.145
word2vec-TF (1)	6201302-_2019_cl		0.175	0.108	0.977	0.970	0.000	0.002	0.010	0.019
	6390103-_2019_cl		0.253	0.193	0.940	0.939	0.039	0.039	0.081	0.082
	ForosPL1_17-18		0.265	0.166	0.978	0.978	0.030	0.030	0.064	0.064
	ForosPL1_18-19		0.201	0.133	0.944	0.937	0.108	0.109	0.124	0.127
	ForosPL1_19-20		0.181	0.136	0.932	0.909	−0.017	0.018	0.037	0.055
	ForosPL1_20-21		0.174	0.112	0.957	0.957	−0.045	−0.045	0.035	0.035
	Foros_Acceso_Mayores_cl		0.238	0.148	0.973	0.973	−0.036	−0.036	−0.004	−0.004
fastText-TF (1)	6201302-_2019_cl		0.235	0.144	0.953	0.958	0.011	0.013	0.012	0.024
	6390103-_2019_cl		0.388	0.269	0.976	0.983	0.033	0.032	0.079	0.067
	ForosPL1_17-18		0.325	0.202	0.970	0.970	0.001	0.001	0.013	0.013
	ForosPL1_18-19		0.366	0.225	0.970	0.970	0.113	0.113	0.158	0.158
	ForosPL1_19-20		0.313	0.199	0.966	0.966	−0.004	−0.004	0.058	0.058
	ForosPL1_20-21		0.299	0.182	0.957	0.960	−0.059	−0.056	0.035	0.033
	Foros_Acceso_Mayores_cl		0.398	0.249	0.984	0.984	−0.009	−0.009	−0.006	−0.006
SBERT	6201302-_2019_cl		0.177	0.095	0.937	0.948	0.238	0.227	0.299	0.266
	6390103-_2019_cl		0.065	0.053	0.959	0.966	0.019	0.022	0.085	0.086
	ForosPL1_17-18		0.133	0.078	0.965	0.965	0.054	0.052	0.120	0.113
	ForosPL1_18-19		0.102	0.065	0.963	0.967	0.173	0.192	0.306	0.313
	ForosPL1_19-20		0.117	0.075	0.915	0.915	0.084	0.084	0.184	0.184
	ForosPL1_20-21		0.139	0.086	0.950	0.950	0.055	0.055	0.182	0.182
	Foros_Acceso_Mayores_cl		0.137	0.084	0.971	0.969	−0.052	−0.042	0.000	0.006
USE	6201302-_2019_cl		0.122	0.078	0.951	0.951	0.227	0.227	0.245	0.245
	6390103-_2019_cl		0.123	0.110	0.969	0.963	0.030	0.029	0.070	0.073
	ForosPL1_17-18		0.122	0.077	0.973	0.976	0.030	0.029	0.058	0.053
	ForosPL1_18-19		0.136	0.086	0.941	0.941	0.125	0.125	0.172	0.172
	ForosPL1_19-20		0.093	0.058	0.920	0.920	0.147	0.130	0.184	0.166
	ForosPL1_20-21		0.117	0.074	0.964	0.964	0.048	0.048	0.145	0.145
	Foros_Acceso_Mayores_cl		0.130	0.081	0.978	0.980	−0.017	−0.015	−0.003	−0.003

originales. La razón podría ser el mencionado fenómeno de encadenamiento (Aggarwal y Zhai, 2012, sec. 3.1), que no afecta a los otros dos enlaces, o de manera más general, la sensibilidad de los enlaces simple y completo frente al ruido y los valores extremos (Han *et al.*, 2012, sec. 10.3.2). Estos fenómenos hacen que suela recomendarse el enlace promedio como compromiso relativamente robusto, siempre que resulte factible desde el punto de vista computacional.

Las tablas 5.5 y 5.6 muestran las medidas de validación por foro para los agrupamientos obtenidos con el número de grupos original y los enlaces promedio y completo. Los resultados son otra vez inhomogéneos a través de los foros. Por ejemplo, para el foro de Procesadores del Lenguaje de 2018–19, ese que contenía un subforo de estudiantes referido casi exclusivamente a un grupo de Telegram, las medidas externas ARI y AMI son siempre positivas, y

Tabla 5.6: Medidas de evaluación por foro de los agrupamientos obtenidos con el algoritmo aglomerativo, enlace completo, número de grupos original y diferentes representaciones.

		Medida	silueta		consistencia		ARI		AMI	
Modelo	Foro	Distancia	coseno	angular	coseno	angular	coseno	angular	coseno	angular
LogTF-IDF (2)	6201302-_2019_cl		−0.011	−0.007	0.941	0.941	0.012	0.012	0.083	0.083
	6390103-_2019_cl		0.003	0.005	0.912	0.912	0.002	0.002	0.067	0.067
	ForosPL1_17-18		−0.002	−0.001	0.914	0.914	0.037	0.037	0.092	0.092
	ForosPL1_18-19		0.034	0.024	0.926	0.926	0.145	0.145	0.213	0.213
	ForosPL1_19-20		0.013	0.012	0.807	0.807	0.128	0.128	0.129	0.129
	ForosPL1_20-21		0.012	0.011	0.924	0.924	0.080	0.080	0.143	0.143
	Foros_Acceso_Mayores_cl		−0.005	−0.003	0.917	0.917	0.050	0.050	0.034	0.034
word2vec-TF (1)	6201302-_2019_cl		0.154	0.096	0.836	0.836	0.156	0.156	0.201	0.201
	6390103-_2019_cl		0.132	0.116	0.747	0.747	0.065	0.065	0.135	0.135
	ForosPL1_17-18		0.088	0.053	0.833	0.833	0.113	0.113	0.118	0.118
	ForosPL1_18-19		0.061	0.037	0.744	0.744	0.117	0.117	0.195	0.195
	ForosPL1_19-20		0.122	0.074	0.795	0.795	−0.003	−0.003	0.101	0.101
	ForosPL1_20-21		0.115	0.070	0.834	0.834	−0.021	−0.021	0.077	0.077
	Foros_Acceso_Mayores_cl		0.078	0.046	0.877	0.877	0.078	0.078	0.053	0.053
fastText-TF (1)	6201302-_2019_cl		0.170	0.099	0.852	0.852	0.184	0.184	0.200	0.200
	6390103-_2019_cl		0.266	0.191	0.847	0.847	0.044	0.044	0.087	0.087
	ForosPL1_17-18		0.292	0.174	0.944	0.944	0.023	0.023	0.062	0.062
	ForosPL1_18-19		0.101	0.058	0.830	0.830	0.143	0.143	0.213	0.213
	ForosPL1_19-20		0.232	0.142	0.903	0.903	0.001	0.001	0.080	0.080
	ForosPL1_20-21		0.110	0.066	0.828	0.828	−0.021	−0.021	0.066	0.066
	Foros_Acceso_Mayores_cl		0.274	0.161	0.926	0.926	−0.031	−0.031	−0.004	−0.004
SBERT	6201302-_2019_cl		0.128	0.080	0.892	0.892	0.243	0.243	0.308	0.308
	6390103-_2019_cl		0.103	0.077	0.837	0.837	0.123	0.123	0.219	0.219
	ForosPL1_17-18		0.068	0.042	0.852	0.852	0.106	0.106	0.176	0.176
	ForosPL1_18-19		0.140	0.086	0.904	0.904	0.295	0.295	0.394	0.394
	ForosPL1_19-20		0.083	0.052	0.795	0.795	0.030	0.030	0.147	0.147
	ForosPL1_20-21		0.117	0.072	0.904	0.904	−0.002	−0.002	0.123	0.123
	Foros_Acceso_Mayores_cl		0.026	0.017	0.794	0.794	0.034	0.034	0.074	0.074
USE	6201302-_2019_cl		0.113	0.070	0.852	0.852	0.214	0.214	0.261	0.261
	6390103-_2019_cl		0.131	0.110	0.832	0.832	0.160	0.160	0.210	0.210
	ForosPL1_17-18		0.079	0.049	0.892	0.892	0.176	0.176	0.221	0.221
	ForosPL1_18-19		0.127	0.079	0.889	0.889	0.357	0.357	0.338	0.338
	ForosPL1_19-20		0.109	0.070	0.824	0.824	0.172	0.172	0.229	0.229
	ForosPL1_20-21		0.134	0.084	0.917	0.917	0.073	0.073	0.178	0.178
	Foros_Acceso_Mayores_cl		0.058	0.036	0.836	0.836	0.107	0.107	0.127	0.127

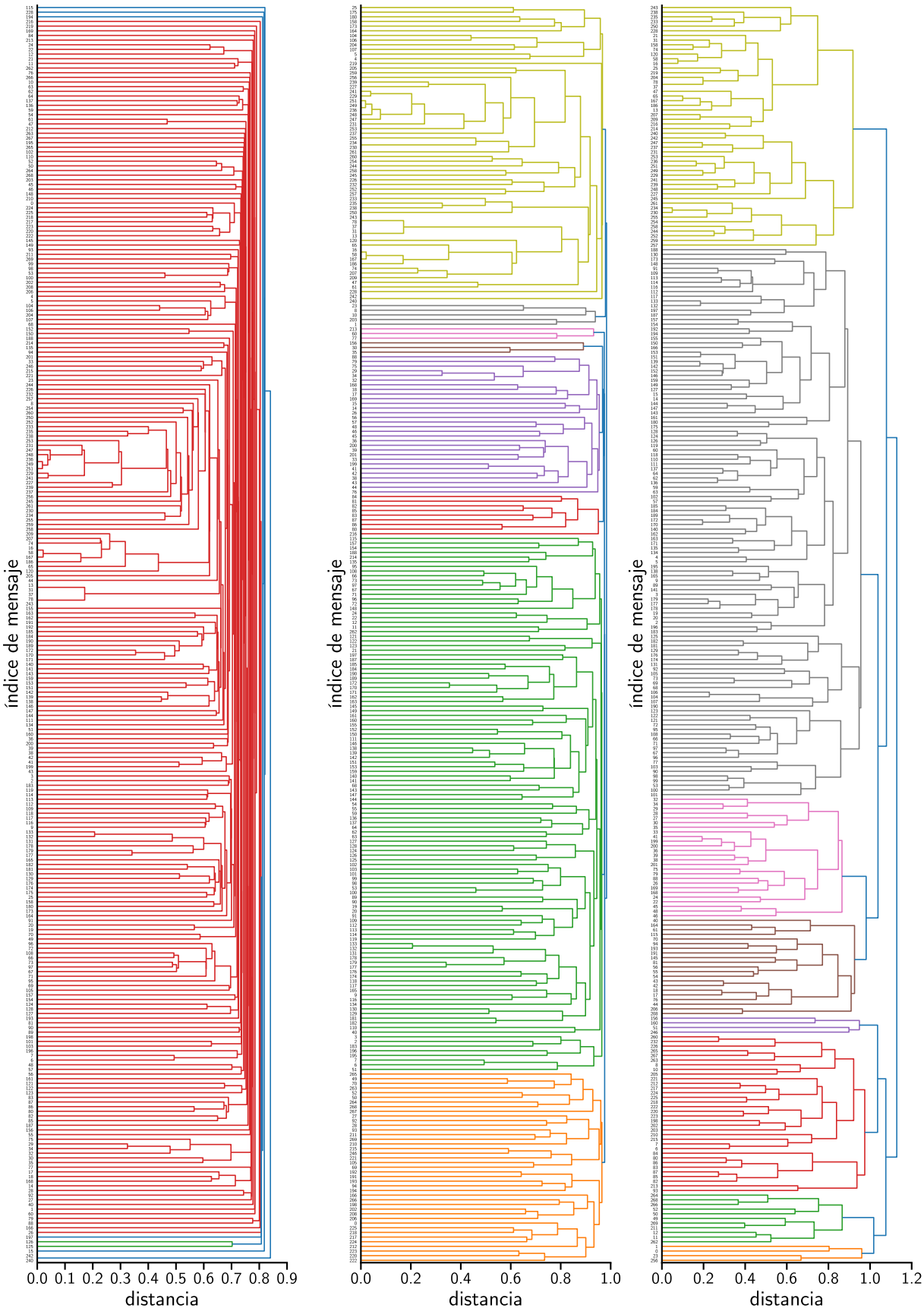
van de 0.108 a 0.394. En cambio, para el foro de Acceso para Mayores, dichas medidas son negativas o diminutas en todos los agrupamientos con enlace promedio salvo los logrados con representación LogTF-IDF. Como era de prever tras el análisis de los resultados promedio, buena parte de los mejores resultados individuales de ARI y AMI aparece en configuraciones con representación LogTF-IDF, para el enlace promedio, y con *embedding* USE, para el enlace completo. Estas dos configuraciones tienen la virtud adicional de producir medidas externas positivas para todos los foros.

Con respecto a las medidas internas, los mejores resultados individuales de silueta y consistencia de proximidad aparecen con el *embedding* fastText con ponderación TF, tanto para el enlace promedio como para el completo. Por ejemplo, para el foro de acceso para mayores, la configuración con representación fastText-TF, distancia coseno y enlace promedio consigue valores de 0.398 para la silueta y 0.984 para la consistencia. Sin embargo, las medidas externas correspondientes a esta configuración son ligeramente negativas, y la matriz de confusión de la tabla 5.7 destapa que el agrupamiento concentra un 97.5 % de los mensajes en el grupo 0. Esta concentración extrema explica el alto valor de la consistencia así como las malas medidas externas, puesto que en ese gran grupo mezcla mensajes de todos los subforos originales.

Como se mencionó al inicio de la sección, el proceso aglomerativo de agrupación, en el que objetos y grupos van uniéndose paso a paso en función de su similitud, genera una estructura de tipo árbol que puede representarse gráficamente como dendrograma. La figura 5.3 muestra tres ejemplos para el foro de Procesadores del Lenguaje de 2018–19: dos con representación LogTF-IDF y enlace simple o promedio, y otro con representación USE y enlace completo. En los tres casos, la distancia utilizada para comparar documentos es la del coseno. El proceso de agrupamiento tiene lugar de izquierda a derecha, empezando con un grupo por objeto y terminando con un único grupo. La abscisa indica la distancia intergrupal en cada unión, de acuerdo con el tipo de enlace, y el color señala el grupo asignado cuando el dendrograma se trunca en el nivel correspondiente al número de grupos original. Así pues, el dendrograma 5.3a para el agrupamiento con enlace simple evidencia la formación de un grupo enormemente dominante, al que acompañan varios microgrupos de uno o dos elementos. Además, puede entreverse que las últimas adhesiones al grupo dominante son asimismo uniones de objetos

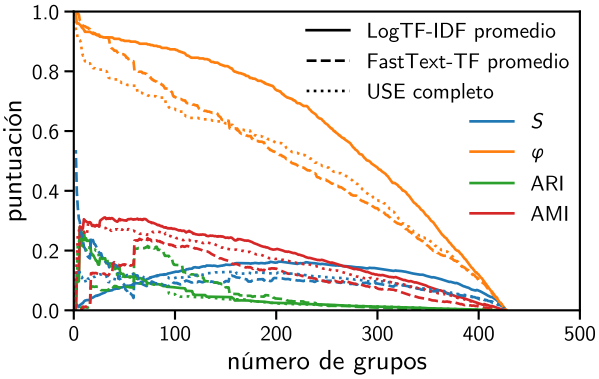
Tabla 5.7: Matriz de confusión del agrupamiento del foro de Acceso para Mayores con algoritmo aglomerativo, enlace promedio, representación fastText-TF y número de grupos original.

Subforo \ Grupo	0	1	2	3	4	5	6	Total
0.1. Foro del Orientador/a de la Comunidad de Acogida Virtual	95	1	0	0	0	0	0	96
0.2. Foro de estudiantes	253	0	3	1	1	2	1	261
1. Foro de contenidos y actividades de la Fase 1.	32	0	0	0	1	0	0	33
2. Foro de contenidos y actividades de la Fase 2.	9	0	0	0	0	0	0	9
3. Foro de contenidos y actividades de la Fase 3.	10	0	0	0	0	0	0	10
4. Foro de contenidos y actividades de la Fase 4.	10	0	0	0	0	0	0	10
Foro de apoyo técnico	26	0	1	0	0	0	0	27
Total	435	1	4	1	2	2	1	446

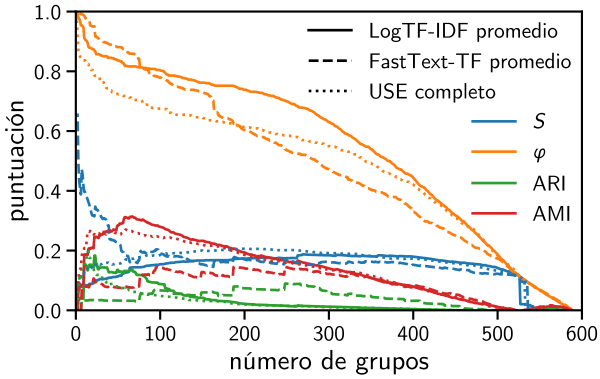


(a) LogTF-IDF con enlace simple. **(b)** LogTF-IDF con enlace promedio. **(c)** USE con enlace completo.

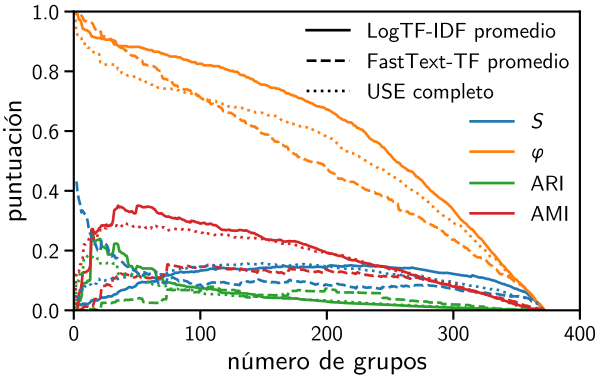
Figura 5.3: Dendrogramas para tres agrupamientos aglomerativos del foro de Procesadores del Lenguaje de 2018–19 con distancia coseno y diferentes representaciones y tipos de enlace.



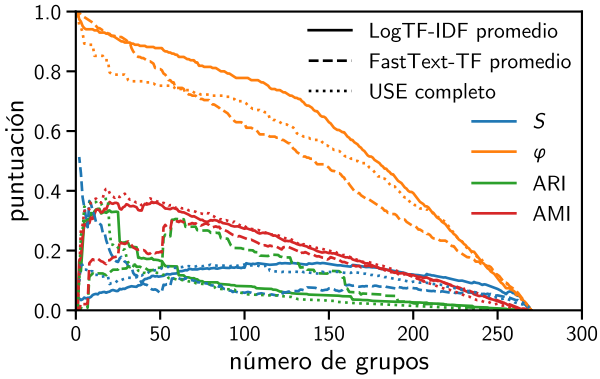
(a) 6201302-_2019_cl (Psicofarmacología).



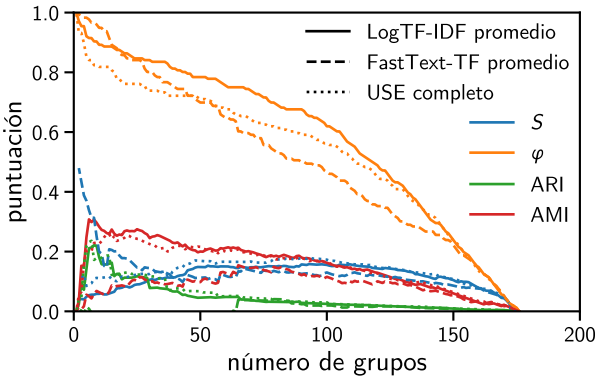
(b) 6390103-_2019_cl (Sociedad del Conocimiento).



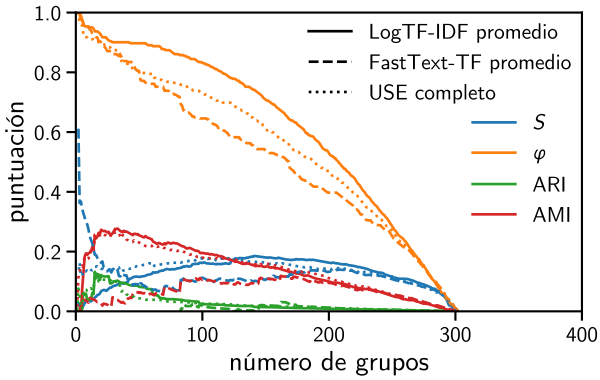
(c) ForosPL1_17-18 (Procesadores del Lenguaje).



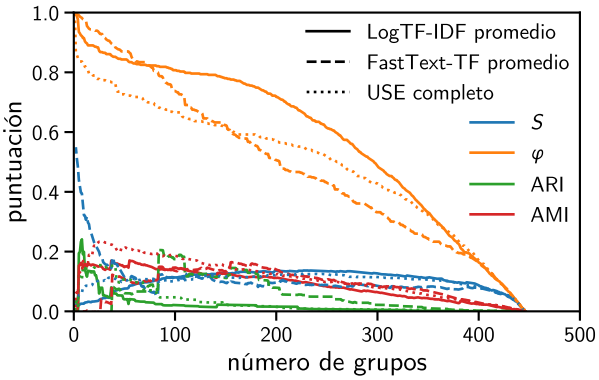
(d) ForosPL1_18-19 (Procesadores del Lenguaje).



(e) ForosPL1_19-20 (Procesadores del Lenguaje).



(f) ForosPL1_20-21 (Procesadores del Lenguaje).



(g) Foros_Acceso_Mayores_cl (Acceso para Mayores).

Figura 5.4: Medidas de evaluación en función del número de grupos para los agrupamientos obtenidos con el algoritmo aglomerativo, la distancia coseno y las representaciones LogTF-IDF (enlace promedio; continua), fastText-TF (promedio; discontinua) y USE (completo; punteada).

individuales. Como se dijo en el análisis de los resultados medios, los agrupamientos generados con enlace simple tienen alta consistencia de proximidad, pero poca relación con los originales o deseables, y no se ensayarán más. Los otros dos dendrogramas, obtenidos con enlaces promedio y completo, presentan una estructura genuinamente jerárquica, con múltiples niveles de grupos y subgrupos no triviales. En estos dos ejemplos existe también un grupo dominante, pero éste no aloja ni la mitad de los mensajes, y tampoco hay grupos de un solo mensaje entre los coloreados para los cortes de 8 grupos.

La otra ventaja del algoritmo aglomerativo es que, una vez corrido el proceso y generado el árbol de agrupamiento, es fácil y barato cortarlo a diferentes niveles para evaluar la calidad del agrupamiento a través de la jerarquía. La figura 5.4 muestra las medidas de validación por foro en función del número de grupos para los agrupamientos obtenidos con la distancia coseno y las representaciones LogTF-IDF y fastText-TF, con enlace promedio, y la representación USE, con enlace completo. La primera y la última de estas configuraciones se han identificado como las más prometedoras en cuanto a medidas externas. El *embedding* de palabras se incluye por completar los tipos de representación y por haber arrojado los mejores resultados de silueta y consistencia con enlace distinto al simple. Como se podía esperar, la consistencia va cayendo desde 1 hasta 0 a medida que aumenta el número de grupos, mientras que, por lo general, las medidas externas empiezan y acaban muy pequeñas, alcanzando su máximo entre medias. Si bien la posición del máximo no coincide exactamente con el número de grupos original, sobre todo para la representación fastText-TF, los máximos absolutos del ARI y la ARI no superan por mucho los encontrados hasta ahora, llegando a 0.373 y 0.408 para el foro de Procesadores del Lenguaje de 2018–19 con representación USE. Finalmente, la silueta tiene un comportamiento algo más variado. Para las representaciones LogTF-IDF y USE, alcanza su máximo en un número de grupos intermedio. En cambio, para la representación fastText-TF, continúa creciendo a medida que el número de grupos se reduce hasta 2, y alcanza ahí su máximo global de 0.659 para el foro de Sociedad del Conocimiento. Sin embargo, como en casos anteriores, esta silueta más que decente no se traduce en buenos resultados de medidas externas, por lo que el agrupamiento tiene poco que ver con el original.

5.3. HDBSCAN

El último algoritmo de agrupamiento ensayado es uno basado en densidad. Los métodos de este tipo identifican los grupos como regiones densas del espacio de datos separadas por otras dispersas y, de esta manera, son capaces de encontrar grupos de forma arbitraria, a diferencia de los métodos partitivos y jerárquicos (Han *et al.*, 2012, sec. 10.4). Esta característica se antoja conveniente para el *clustering* de los foros en vista de la multitud de siluetas negativas encontradas hasta ahora, incluidas las de los agrupamientos originales, que sugieren que los grupos podrían ser no globulares o estar solapados. El algoritmo probado aquí, HDBSCAN (Campello *et al.*, 2013), es una evolución del DBSCAN clásico capaz de identificar grupos

de distinta densidad e integrarlos en una estructura similar a la de los algoritmos jerárquicos. De hecho, el algoritmo construye primero un árbol recubridor mínimo, con pesos sobre las aristas derivados de la densidad y la distancia entre puntos, y después lo corta efectivamente a varios niveles de densidad maximizando la estabilidad del agrupamiento. Al igual que otras técnicas basadas en densidad, HDBSCAN tiene en cuenta la posible existencia de ruido en los datos y deja al margen del agrupamiento los puntos reconocidos como ruido, típicamente en regiones dispersas. Una ventaja adicional de HDBSCAN sobre otros métodos basados en densidad es que sus parámetros son algo más intuitivos y robustos. Los principales son la muestra mínima y el tamaño mínimo de grupo. El primero guarda relación con la estimación de la densidad, que se lleva a cabo utilizando la distancia al enésimo vecino más cercano. La muestra mínima es el número de vecinos usado en dicha estimación y afecta a su suavidad y sensibilidad. También influye en la cantidad de puntos que acaban siendo considerados ruido y, en la práctica, se ajusta para que el agrupamiento resulte más o menos conservador. El valor mínimo de 1 corresponde al *clustering* más agresivo posible. Por su parte, el tamaño mínimo de grupo afecta a la consolidación del árbol recubridor mínimo en una jerarquía de grupos más manejable. En dicho proceso, después de cada división, se descarta cualquier descendiente de cardinalidad inferior al parámetro, aumentando así la persistencia del progenitor, reduciendo la complejidad de la jerarquía e impidiendo la aparición de grupos muy pequeños en el *clustering* final. El segundo parámetro tiene por tanto un significado incluso más intuitivo que el primero. La librería hdbscan (McInnes *et al.*, 2017) contiene una implementación eficiente del algoritmo en forma de agrupador de scikit-learn, que se puede incorporar directamente en los *pipelines* del sistema. Además, incluye diversas herramientas de análisis y visualización, entre ellas una implementación de la medida interna DBCV, que se presentó en la sección 2.3.1 y debería de adaptarse bien a los agrupamientos basados en densidad.

La tabla 5.8 muestra las medidas de evaluación promedio de los agrupamientos obtenidos aplicando HDBSCAN con diversas combinaciones de parámetros a las mismas representaciones de los foros examinadas en las secciones precedentes. La muestra mínima se ha tomado entre 1, el mínimo admisible, y 5, que resultará suficientemente conservador como para considerar como ruido el 40 % de los mensajes o más, según el caso. Conviene recordar que, de acuerdo con la tabla 3.2, cada foro contiene unos pocos centenares de mensajes y que las dimensiones de las representaciones también van desde los pocos cientos, para los *embeddings* de palabras, hasta los pocos miles, para algunas representaciones BOW. Por tanto, los mensajes podrían estar bastante desperdigados por el espacio de datos. Para cada muestra mínima, el tamaño mínimo de grupo se ha tomado entre ese mismo valor y 7. Tal y como se vio en la tabla 3.3, la mayoría de los subforos originales tiene entre 10 y 100 mensajes, pero también hay 4 con 2 mensajes y 1 subforo con 1 único mensaje. En todo caso, el rango de parámetros considerado debería de ser suficiente para vislumbrar tendencias.

Empezando por el lado izquierdo de la tabla, las dos primeras columnas de resultados contienen lo que podría llamarse la cobertura del agrupamiento, es decir, la fracción de mensajes

Tabla 5.8: Medidas de evaluación promedio para los agrupamientos obtenidos con HDBSCAN.

Modelo	Muestra mínima	Tamaño mínimo	cobertura		# grupos		DBCV		silueta		consistencia		ARI		AMI	
			cos	ang.	cos	ang.	cos	ang.	cos	ang.	cos	ang.	cos	ang.	cos	ang.
LogTF-IDF (2)	1	3	0.602	0.610	40.4	40.1	0.107	0.068	0.121	0.087	0.602	0.610	0.051	0.051	0.184	0.184
		4	0.533	0.538	26.1	25.9	0.089	0.056	0.100	0.073	0.533	0.538	0.066	0.072	0.178	0.181
		5	0.505	0.505	16.6	16.6	0.068	0.043	0.078	0.058	0.505	0.505	0.103	0.103	0.186	0.186
		6	0.482	0.490	12.0	11.9	0.053	0.032	0.067	0.048	0.482	0.490	0.107	0.107	0.176	0.177
		7	0.463	0.463	8.4	8.4	0.041	0.025	0.058	0.042	0.463	0.463	0.109	0.109	0.168	0.168
	2	3	0.405	0.405	24.0	24.0	0.094	0.061	0.101	0.074	0.399	0.399	0.061	0.061	0.144	0.144
		4	0.389	0.389	15.1	15.1	0.077	0.050	0.085	0.063	0.386	0.386	0.086	0.086	0.150	0.150
		5	0.365	0.371	11.0	10.9	0.066	0.041	0.074	0.056	0.363	0.369	0.098	0.102	0.150	0.151
		6	0.362	0.362	8.0	8.0	0.054	0.034	0.065	0.048	0.360	0.360	0.101	0.101	0.143	0.143
		7	0.368	0.368	6.7	6.7	0.040	0.025	0.058	0.042	0.366	0.366	0.106	0.106	0.142	0.142
	3	3	0.307	0.307	12.9	12.9	0.074	0.049	0.075	0.057	0.299	0.299	0.086	0.086	0.126	0.126
		4	0.290	0.293	10.1	10.0	0.072	0.048	0.072	0.055	0.285	0.289	0.087	0.088	0.124	0.127
		5	0.294	0.297	7.6	7.4	0.061	0.039	0.066	0.050	0.289	0.293	0.097	0.099	0.127	0.129
		6	0.292	0.292	6.4	6.4	0.054	0.035	0.061	0.046	0.287	0.287	0.091	0.091	0.121	0.121
		7	0.293	0.293	5.9	5.9	0.048	0.030	0.057	0.043	0.288	0.288	0.088	0.088	0.117	0.117
	4	4	0.272	0.272	7.1	7.1	0.057	0.037	0.065	0.050	0.267	0.267	0.096	0.096	0.113	0.113
		5	0.265	0.265	6.0	6.0	0.046	0.029	0.061	0.047	0.263	0.263	0.095	0.095	0.110	0.110
		6	0.259	0.259	4.9	4.9	0.042	0.027	0.057	0.043	0.258	0.258	0.093	0.093	0.109	0.109
		7	0.257	0.257	4.4	4.4	0.039	0.024	0.055	0.041	0.255	0.255	0.091	0.091	0.104	0.104
	5	5	0.232	0.232	4.6	4.6	0.046	0.029	0.056	0.043	0.230	0.230	0.086	0.086	0.097	0.097
		6	0.220	0.220	3.9	3.9	0.043	0.028	0.055	0.042	0.217	0.217	0.082	0.082	0.092	0.092
		7	0.242	0.242	3.1	3.1	0.043	0.027	0.048	0.036	0.239	0.239	0.075	0.075	0.087	0.087
word2vec-TF (1)	1	3	0.543	0.617	11.7	10.6	0.075	0.052	0.140	0.105	0.543	0.617	0.056	0.049	0.104	0.097
		4	0.690	0.753	5.0	4.4	0.115	0.074	0.235	0.164	0.690	0.753	0.042	0.024	0.071	0.057
		5	0.763	0.763	3.6	3.6	0.134	0.075	0.268	0.175	0.763	0.763	0.031	0.031	0.062	0.062
		6	0.757	0.766	3.1	2.9	0.143	0.075	0.271	0.181	0.757	0.766	0.030	0.034	0.059	0.066
		7	0.686	0.695	3.0	2.7	0.121	0.062	0.261	0.174	0.686	0.695	0.039	0.044	0.068	0.076
	2	3	0.568	0.684	7.6	4.9	0.192	0.116	0.200	0.150	0.566	0.683	0.035	0.027	0.074	0.064
		4	0.706	0.706	3.3	3.3	0.187	0.108	0.265	0.175	0.705	0.705	0.024	0.024	0.052	0.052
		5	0.707	0.707	2.7	2.7	0.187	0.107	0.274	0.180	0.706	0.706	0.023	0.023	0.053	0.053
		6	0.714	0.714	2.3	2.3	0.206	0.118	0.292	0.193	0.713	0.713	0.027	0.027	0.054	0.054
		7	0.640	0.640	2.3	2.3	0.166	0.097	0.272	0.179	0.639	0.639	0.028	0.028	0.047	0.047
	3	3	0.681	0.681	2.9	2.9	0.206	0.123	0.266	0.176	0.680	0.680	0.022	0.022	0.046	0.046
		4	0.686	0.686	2.6	2.6	0.194	0.115	0.283	0.188	0.686	0.686	0.024	0.024	0.050	0.050
		5	0.686	0.686	2.3	2.3	0.195	0.116	0.289	0.191	0.686	0.686	0.025	0.025	0.051	0.051
		6	0.686	0.686	2.3	2.3	0.195	0.116	0.289	0.191	0.686	0.686	0.025	0.025	0.051	0.051
		7	0.543	0.559	2.4	2.3	0.154	0.090	0.240	0.161	0.543	0.559	0.040	0.034	0.057	0.057
	4	4	0.677	0.677	2.1	2.1	0.195	0.115	0.283	0.187	0.676	0.676	0.023	0.023	0.048	0.048
		5	0.677	0.677	2.1	2.1	0.195	0.115	0.283	0.187	0.676	0.676	0.023	0.023	0.048	0.048
		6	0.677	0.677	2.1	2.1	0.195	0.115	0.283	0.187	0.676	0.676	0.023	0.023	0.048	0.048
		7	0.547	0.547	2.1	2.1	0.154	0.090	0.237	0.157	0.546	0.546	0.032	0.032	0.054	0.054
	5	5	0.600	0.600	2.1	2.1	0.213	0.123	0.258	0.172	0.599	0.599	0.031	0.031	0.059	0.059
		6	0.587	0.600	2.3	2.1	0.212	0.123	0.254	0.172	0.586	0.599	0.036	0.031	0.059	0.059
		7	0.521	0.534	2.3	2.1	0.179	0.103	0.234	0.158	0.520	0.534	0.036	0.031	0.051	0.051
fastText-TF (1)	1	3	0.719	0.719	7.3	7.3	0.140	0.080	0.216	0.141	0.719	0.719	0.034	0.034	0.085	0.085
		4	0.688	0.754	4.7	4.1	0.170	0.100	0.262	0.189	0.688	0.754	0.038	0.037	0.086	0.087
		5	0.746	0.746	3.1	3.1	0.194	0.110	0.339	0.218	0.746	0.746	0.036	0.036	0.076	0.076
		6	0.713	0.713	3.1	3.1	0.190	0.108	0.327	0.210	0.713	0.713	0.032	0.032	0.069	0.069
		7	0.668	0.668	2.9	2.9	0.146	0.080	0.303	0.195	0.668	0.668	0.039	0.039	0.072	0.072
	2	3	0.640	0.675	5.9	5.6	0.184	0.129	0.226	0.163	0.639	0.674	0.036	0.034	0.080	0.079
		4	0.696	0.696	3.6	3.6	0.218	0.127	0.310	0.200	0.696	0.696	0.035	0.035	0.071	0.071
		5	0.691	0.691	2.9	2.9	0.254	0.149	0.340	0.221	0.689	0.689	0.032	0.032	0.062	0.062
		6	0.693	0.693	2.6	2.6	0.253	0.148	0.341	0.221	0.691	0.691	0.033	0.033	0.058	0.058
		7	0.616	0.616	2.6	2.6	0.203	0.117	0.301	0.195	0.614	0.614	0.038	0.038	0.061	0.061
	3	3	0.695	0.695	3.3	3.3	0.192	0.111	0.291	0.187	0.694	0.694	0.033	0.033	0.066	0.066
		4	0.721	0.721	3.0	3.0	0.200	0.114	0.332	0.214	0.720	0.720	0.035	0.035	0.066	0.066
		5	0.682	0.682	2.4	2.4	0.225	0.130	0.341	0.222	0.682	0.682	0.033	0.033	0.060	0.060
		6	0.682	0.682	2.4	2.4	0.225	0.130	0.341	0.222	0.682	0.682	0.033	0.033	0.060	0.060
		7	0.648	0.648	2.1	2.1	0.165	0.091	0.324	0.210	0.648	0.648	0.035	0.035	0.063	0.063

Tabla 5.8 (cont.): Medidas de evaluación promedio para los agrupamientos obtenidos con HDBSCAN.

Modelo	Muestra mínima	Tamaño mínimo	cobertura		# grupos		DBCV		silueta		consistencia		ARI		AMI	
			cos	ang.	cos	ang.	cos	ang.	cos	ang.	cos	ang.	cos	ang.	cos	ang.
fastText-TF (1)	4	4	0.667	0.667	2.6	2.6	0.261	0.150	0.341	0.222	0.666	0.666	0.031	0.031	0.057	0.057
		5	0.666	0.666	2.4	2.4	0.260	0.150	0.341	0.222	0.665	0.665	0.031	0.031	0.057	0.057
		6	0.666	0.666	2.4	2.4	0.260	0.150	0.341	0.222	0.665	0.665	0.031	0.031	0.057	0.057
		7	0.593	0.593	2.3	2.3	0.209	0.117	0.305	0.199	0.591	0.591	0.033	0.033	0.055	0.055
	5	5	0.681	0.681	2.1	2.1	0.253	0.147	0.357	0.233	0.680	0.680	0.026	0.026	0.049	0.049
		6	0.681	0.681	2.1	2.1	0.253	0.147	0.357	0.233	0.680	0.680	0.026	0.026	0.049	0.049
		7	0.587	0.587	1.9	1.9	0.196	0.111	0.364	0.238	0.586	0.586	0.021	0.021	0.039	0.039
SBERT	1	3	0.487	0.571	27.1	23.4	0.108	0.059	0.133	0.093	0.487	0.571	0.098	0.090	0.170	0.158
		4	0.546	0.615	14.6	12.1	0.100	0.072	0.145	0.106	0.546	0.615	0.124	0.112	0.169	0.147
		5	0.481	0.547	10.9	8.9	0.068	0.038	0.119	0.075	0.481	0.547	0.137	0.110	0.176	0.154
		6	0.515	0.528	6.9	6.6	0.056	0.033	0.122	0.082	0.515	0.528	0.141	0.145	0.187	0.191
		7	0.505	0.538	5.6	5.1	0.064	0.034	0.122	0.080	0.505	0.538	0.133	0.129	0.175	0.174
	2	3	0.561	0.642	12.1	9.4	0.124	0.070	0.151	0.099	0.558	0.639	0.079	0.055	0.115	0.090
		4	0.463	0.487	9.6	8.9	0.084	0.050	0.129	0.089	0.460	0.484	0.104	0.097	0.136	0.140
		5	0.378	0.475	7.3	5.4	0.068	0.040	0.113	0.081	0.376	0.474	0.122	0.111	0.153	0.147
		6	0.405	0.422	5.4	5.3	0.070	0.039	0.116	0.082	0.404	0.421	0.122	0.124	0.158	0.164
		7	0.436	0.436	4.0	4.0	0.086	0.050	0.118	0.081	0.435	0.435	0.109	0.109	0.147	0.147
	3	3	0.440	0.544	7.4	6.1	0.105	0.059	0.138	0.090	0.437	0.540	0.086	0.068	0.125	0.096
		4	0.481	0.595	5.6	4.3	0.116	0.077	0.138	0.096	0.477	0.593	0.089	0.047	0.126	0.080
		5	0.414	0.482	4.3	3.7	0.088	0.056	0.129	0.089	0.411	0.480	0.107	0.075	0.142	0.113
		6	0.408	0.476	3.9	3.3	0.087	0.055	0.130	0.090	0.406	0.474	0.107	0.075	0.139	0.110
		7	0.402	0.474	3.6	3.1	0.085	0.055	0.128	0.089	0.400	0.471	0.107	0.076	0.137	0.110
	4	4	0.383	0.429	4.6	4.1	0.076	0.045	0.122	0.080	0.382	0.428	0.108	0.086	0.132	0.109
		5	0.450	0.496	3.6	3.1	0.081	0.047	0.127	0.084	0.449	0.494	0.079	0.056	0.105	0.083
		6	0.447	0.492	3.3	2.9	0.079	0.046	0.125	0.083	0.446	0.491	0.079	0.057	0.104	0.081
		7	0.452	0.492	3.1	2.9	0.089	0.046	0.129	0.083	0.451	0.491	0.076	0.057	0.104	0.081
	5	5	0.419	0.467	3.4	3.0	0.087	0.064	0.125	0.084	0.418	0.465	0.076	0.054	0.100	0.078
		6	0.420	0.465	3.1	2.9	0.097	0.064	0.127	0.083	0.419	0.463	0.074	0.055	0.100	0.078
		7	0.418	0.418	3.0	3.0	0.095	0.058	0.127	0.086	0.417	0.417	0.073	0.073	0.099	0.099
USE	1	3	0.479	0.482	26.4	26.3	0.093	0.054	0.121	0.082	0.479	0.482	0.103	0.106	0.186	0.188
		4	0.486	0.529	15.4	13.4	0.081	0.043	0.118	0.078	0.486	0.529	0.139	0.123	0.204	0.192
		5	0.576	0.636	7.7	6.4	0.077	0.036	0.138	0.090	0.576	0.636	0.122	0.108	0.186	0.178
		6	0.503	0.699	7.7	4.9	0.067	0.032	0.126	0.086	0.503	0.699	0.134	0.082	0.188	0.153
		7	0.560	0.641	6.3	5.0	0.053	0.025	0.128	0.083	0.560	0.641	0.119	0.101	0.174	0.165
	2	3	0.443	0.566	14.7	10.9	0.087	0.048	0.104	0.073	0.441	0.564	0.105	0.096	0.158	0.157
		4	0.588	0.588	7.0	7.0	0.076	0.044	0.126	0.084	0.586	0.586	0.092	0.092	0.155	0.155
		5	0.550	0.550	5.1	5.1	0.067	0.037	0.131	0.087	0.549	0.549	0.097	0.097	0.151	0.151
		6	0.567	0.656	4.7	3.6	0.066	0.034	0.136	0.086	0.566	0.654	0.103	0.083	0.155	0.135
		7	0.581	0.652	4.1	3.3	0.087	0.043	0.144	0.090	0.579	0.650	0.107	0.080	0.146	0.130
	3	3	0.539	0.539	6.0	6.0	0.080	0.045	0.120	0.081	0.537	0.537	0.094	0.094	0.141	0.141
		4	0.551	0.551	4.9	4.9	0.079	0.045	0.127	0.086	0.549	0.549	0.098	0.098	0.148	0.148
		5	0.516	0.516	4.9	4.9	0.086	0.049	0.137	0.092	0.512	0.512	0.126	0.126	0.166	0.166
		6	0.477	0.539	4.6	3.7	0.096	0.050	0.134	0.091	0.474	0.537	0.136	0.096	0.167	0.138
		7	0.462	0.492	3.9	3.6	0.099	0.055	0.135	0.085	0.460	0.491	0.121	0.098	0.159	0.142
	4	4	0.452	0.532	4.7	4.0	0.089	0.059	0.114	0.085	0.449	0.530	0.112	0.107	0.142	0.141
		5	0.503	0.534	4.0	3.7	0.100	0.060	0.124	0.087	0.502	0.532	0.101	0.106	0.137	0.140
		6	0.488	0.532	3.9	3.6	0.101	0.059	0.126	0.087	0.486	0.530	0.105	0.105	0.139	0.140
		7	0.509	0.554	3.3	3.0	0.099	0.055	0.134	0.095	0.508	0.552	0.104	0.096	0.140	0.136
	5	5	0.508	0.508	3.6	3.6	0.121	0.072	0.145	0.098	0.506	0.506	0.115	0.115	0.148	0.148
		6	0.512	0.512	3.4	3.4	0.121	0.072	0.148	0.100	0.511	0.511	0.115	0.115	0.149	0.149
		7	0.507	0.507	3.1	3.1	0.120	0.071	0.151	0.102	0.506	0.506	0.112	0.112	0.144	0.144

que no son etiquetados como ruido por el algoritmo. Al igual que en las tablas de las secciones anteriores, los promedios se han tomado por separado para los agrupamientos obtenidos con distancia coseno y angular. Como se anticipaba en el párrafo anterior, por lo general, a tamaño mínimo de grupo fijo, la cobertura desciende a medida que se incrementa la muestra mínima. Las coberturas logradas en el rango de parámetros examinado van de 22 % a 77 %. El segundo par de columnas se refiere al número de grupos. Toca recordar ahora que los agrupamientos originales tienen entre 7 y 12 subforos, con un promedio de 8.6, excluyendo grupos de tutoría. El número de grupos disminuye cuando se incrementa cualquiera de los dos parámetros, pero al igual que con la cobertura, el rango de variación depende de la representación. Por ejemplo, con la LogTF-IDF se alcanza un promedio superior a los 40 grupos con muestra mínima 1 y tamaño mínimo 3, mientras que con el *embedding* fastText el promedio queda por debajo de 8 en todos los casos. La ausencia de un número prefijado de grupos es una virtud del algoritmo, en cuanto a que deja que sean los datos los que lo determinen, aunque con cierta influencia de los parámetros. Sólo en casos relativamente poco frecuentes como éste, donde se conoce de antemano un agrupamiento de referencia, podría parecer que dificultase su replicación.

A continuación, se examinan los promedios de las medidas internas, abriendo con el índice de validación basado en densidad. Sus promedios son todos positivos, y alcanzan máximos de 0.261 y 0.150 para las distancias coseno y angular en la configuración con representación fastText-TF, muestra mínima 4 y tamaño mínimo 4. Al igual que los máximos, los promedios del DBCV son en general mayores con la distancia coseno que con la angular. Puesto que el rango teórico de la medida es $[-1, 1]$, los promedios obtenidos no son terribles, pero tampoco puede decirse que sean especialmente buenos. Si acaso, el hecho de que sean todos positivos respalda que el algoritmo HDBSCAN trabaja en una dirección alineada con el índice DBCV. Además, el índice trata de igual manera que el algoritmo los objetos considerados como ruido, a diferencia de las medidas venideras, que requerirán de ajuste. De hecho, el mismo artículo de Moulavi *et al.* (2014) dedicado al DBCV menciona varias formas de adaptar las medidas tradicionales al ruido, y termina recomendando excluir los puntos de ruido del cómputo de la medida y penalizar después multiplicando por la cobertura. Éste es el método adoptado aquí para calcular el resto de las medidas de la tabla, tanto internas como externas.

Los coeficientes de silueta promedio son todos positivos también, lo que no se daba por hecho, ya que los agrupamientos basados en densidad pueden tener forma arbitraria. Es más, los promedios máximos obtenidos para cada representación superan los encontrados con los algoritmos k-medias y aglomerativo. Sin embargo, hay que recordar que en los agrupamientos obtenidos con HDBSCAN el número de grupos no es fijo y puede haber ruido, por lo que los agrupamientos pueden tener muy poco que ver con los examinados en las secciones anteriores. Por ejemplo, la silueta media máxima de 0.364 se alcanza en una configuración con *embedding* fastText, para la que la cobertura es del 68 % y el número de grupos promedio es tan sólo 1.9. Un análisis detallado de ese promedio revela que los agrupamientos de los foros tienen 2 o 3 grupos cada uno, salvo el del foro de Procesadores del Lenguaje de 2019–20, que está

Tabla 5.9: Medidas de evaluación por foro para los agrupamientos obtenidos con HDBSCAN con muestra mínima 1 y tamaño mínimo 5.

Modelo	Muestra mínima	Tamaño mínimo	Foro \ Medida / Distancia	cobertura cos	cobertura ang.	# grupos orig.	# grupos cos	# grupos ang.	DBCV cos	DBCV ang.	silueta cos	silueta ang.	consistencia cos	consistencia ang.	ARI cos	ARI ang.	AMI cos	AMI ang.
LogTF-IDF (2)	1	5	6201302-_2019_cl	0.557	0.557	9	17	17	0.039	0.023	0.050	0.033	0.557	0.557	0.139	0.139	0.196	0.196
			6390103-_2019_cl	0.401	0.401	12	19	19	0.097	0.065	0.130	0.117	0.401	0.401	0.050	0.050	0.125	0.125
			ForosPL1_17-18	0.530	0.530	8	22	22	0.071	0.043	0.077	0.050	0.530	0.530	0.090	0.090	0.215	0.215
			ForosPL1_18-19	0.563	0.563	8	14	14	0.065	0.042	0.083	0.059	0.563	0.563	0.256	0.256	0.299	0.299
			ForosPL1_19-20	0.494	0.494	8	9	9	0.074	0.045	0.079	0.054	0.494	0.494	0.068	0.068	0.150	0.150
			ForosPL1_20-21	0.623	0.623	8	17	17	0.069	0.049	0.079	0.059	0.623	0.623	0.087	0.087	0.211	0.211
			Foros_Acceso_Mayores_cl	0.368	0.368	7	18	18	0.059	0.036	0.051	0.034	0.368	0.368	0.032	0.032	0.105	0.105
word2vec-TF (1)	1	5	6201302-_2019_cl	0.742	0.742	9	6	6	0.130	0.068	0.223	0.143	0.742	0.742	0.040	0.040	0.084	0.084
			6390103-_2019_cl	0.541	0.541	12	5	5	0.129	0.081	0.251	0.189	0.541	0.541	0.031	0.031	0.065	0.065
			ForosPL1_17-18	0.761	0.761	8	4	4	0.139	0.078	0.259	0.161	0.761	0.761	0.014	0.014	0.056	0.056
			ForosPL1_18-19	0.804	0.804	8	2	2	0.176	0.096	0.333	0.214	0.804	0.804	0.110	0.110	0.147	0.147
			ForosPL1_19-20	0.841	0.841	8	2	2	0.206	0.118	0.240	0.158	0.841	0.841	0.027	0.027	0.062	0.062
			ForosPL1_20-21	0.841	0.841	8	4	4	0.146	0.079	0.264	0.169	0.841	0.841	−0.001	−0.001	0.016	0.016
			Foros_Acceso_Mayores_cl	0.809	0.809	7	2	2	0.012	0.006	0.305	0.193	0.809	0.809	−0.004	−0.004	0.004	0.004
fastText-TF (1)	1	5	6201302-_2019_cl	0.660	0.660	9	5	5	0.133	0.068	0.221	0.135	0.660	0.660	0.053	0.053	0.124	0.124
			6390103-_2019_cl	0.944	0.944	12	3	3	0.248	0.142	0.514	0.362	0.944	0.944	0.035	0.035	0.078	0.078
			ForosPL1_17-18	0.626	0.626	8	3	3	0.090	0.052	0.284	0.173	0.626	0.626	−0.009	−0.009	0.000	0.000
			ForosPL1_18-19	0.693	0.693	8	3	3	0.283	0.163	0.362	0.231	0.693	0.693	0.038	0.038	0.079	0.079
			ForosPL1_19-20	0.812	0.812	8	2	2	0.390	0.234	0.347	0.222	0.812	0.812	0.032	0.032	0.068	0.068
			ForosPL1_20-21	0.758	0.758	8	2	2	0.128	0.069	0.415	0.267	0.758	0.758	−0.021	−0.021	0.017	0.017
			Foros_Acceso_Mayores_cl	0.731	0.731	7	4	4	0.083	0.044	0.231	0.139	0.731	0.731	0.123	0.123	0.169	0.169
SBERT	1	5	6201302-_2019_cl	0.403	0.864	9	17	3	0.080	0.038	0.122	0.037	0.403	0.864	0.181	−0.011	0.198	0.047
			6390103-_2019_cl	0.410	0.410	12	14	14	0.089	0.053	0.180	0.146	0.410	0.410	0.100	0.100	0.167	0.167
			ForosPL1_17-18	0.433	0.433	8	13	13	0.063	0.035	0.106	0.066	0.433	0.433	0.075	0.075	0.142	0.142
			ForosPL1_18-19	0.578	0.578	8	10	10	0.105	0.060	0.155	0.097	0.578	0.578	0.299	0.299	0.319	0.319
			ForosPL1_19-20	0.511	0.511	8	8	8	0.049	0.028	0.120	0.079	0.511	0.511	0.173	0.173	0.190	0.190
			ForosPL1_20-21	0.480	0.480	8	9	9	0.058	0.032	0.102	0.064	0.480	0.480	0.068	0.068	0.115	0.115
			Foros_Acceso_Mayores_cl	0.552	0.552	7	5	5	0.033	0.019	0.050	0.033	0.552	0.552	0.066	0.066	0.100	0.100
USE	1	5	6201302-_2019_cl	0.384	0.445	9	11	10	0.051	0.037	0.105	0.077	0.384	0.445	0.127	0.127	0.169	0.191
			6390103-_2019_cl	0.415	0.415	12	14	14	0.059	0.035	0.156	0.127	0.415	0.415	0.066	0.066	0.148	0.148
			ForosPL1_17-18	0.815	0.815	8	4	4	0.083	0.034	0.137	0.086	0.815	0.815	0.033	0.033	0.120	0.120
			ForosPL1_18-19	0.500	0.863	8	10	2	0.096	0.006	0.119	0.051	0.500	0.863	0.168	0.071	0.264	0.184
			ForosPL1_19-20	0.506	0.506	8	6	6	0.074	0.041	0.124	0.081	0.506	0.506	0.184	0.184	0.199	0.199
			ForosPL1_20-21	0.523	0.523	8	7	7	0.087	0.052	0.146	0.094	0.523	0.523	0.159	0.159	0.235	0.235
			Foros_Acceso_Mayores_cl	0.888	0.888	7	2	2	0.089	0.050	0.181	0.114	0.888	0.888	0.119	0.119	0.167	0.167

íntegramente formado por ruido.

Los mejores promedios de consistencia de proximidad se obtienen con los *embeddings* de palabras, llegando al 76 % para la representación word2vec-TF con muestra mínima 1 y tamaño mínimo 5 o 6. Es notable que la consistencia coincida con la cobertura en esas configuraciones y muchas otras, lo que significa que esos agrupamientos son perfectamente consistentes si se descarta sin más el ruido. La ponderación de la consistencia por la cobertura explica también su tendencia descendente, por lo general, con los dos parámetros del algoritmo, así como el amplio rango de variación de la consistencia observado para la representación LogTF-IDF. Ya se vio que esa representación presentaba mayor variabilidad de cobertura y número de grupos.

En cuanto a las medidas externas, los promedios del ARI y la AMI son positivos y varían entre 0.021 y 0.145 y entre 0.039 y 0.204, respectivamente. Los mejores resultados, obtenidos con los *embeddings* de frases y muestra mínima 1, son inferiores a los conseguidos usando los algoritmos *k*-medias y aglomerativo. Los resultados obtenidos con la bolsa de palabras son un poco peores, y a la cola se sitúan los producidos con *embeddings* de palabras, que se reducen aproximadamente a la mitad. Parcialmente por efecto de la cobertura, los mejores resultados de validación externa para todas y cada una de las representaciones se obtienen con muestra mínima 1. No es el caso para los valores del DBCV y, de hecho, y por desgracia, no se aprecia relación directa entre esta medida interna específica y las medidas de evaluación externas. Al menos los parámetros con los que se logran los mejores alineamientos con los agrupamientos originales parecen razonables: muestra mínima 1, que se justificaría por la alta dispersión de los datos en un vasto espacio, y tamaño mínimo entre 3 y 6, que sería suficientemente grande para definir grupos estables y pequeño para capturar la mayoría de los subforos originales.

La tabla 5.9 presenta los resultados de validación por foro para los agrupamientos obtenidos con muestra mínima 1 y tamaño mínimo 5. El valor del tamaño mínimo se ha seleccionado como ejemplo de compromiso entre los maximizan los promedios de las medidas externas para

Tabla 5.10: Matriz de confusión del agrupamiento del foro de Sociedad del Conocimiento con algoritmo HDBSCAN, representación fastText-TF, muestra mínima 1 y tamaño mínimo 5.

Subforo \ Grupo	−1	0	1	2	Total
Consultas generales	10	0	88	0	98
Prueba de Evaluación Continua (PEC)	4	1	76	1	82
Tema 1: Tecnologías y vidas	5	3	117	11	136
Tema 2: Base material de Internet	3	1	37	6	47
Tema 3: ¿Sociedad de la desinformación? Perspectivas sobre las noticias falsas	2	1	48	11	62
Tema 4: Rastros, huellas y filtros digitales	0	0	6	6	12
Tema 5: La cutura social en torno a la seguridad de la información	0	0	8	8	16
Tema 6: La propiedad intelectual y sus enemigos	0	0	7	5	12
Tema 7: Más allá de las pantallas	2	0	10	4	16
Tema 8: Bidg Data, educación basada en datos y analítica del aprendizaje	0	0	5	6	11
Tema 9: Aprendiendo. Cuando quieras. Donde vayas.	0	0	0	6	6
Coordinación tutorial	7	3	80	0	90
Total	33	9	482	64	588

Grupo 0

```
<{IMAGEN}>¶¶
<{IMAGEN}>¶¶
<{IMAGEN}>¶¶
<{IMAGEN}>¶¶
<{IMAGEN}>¶¶
<{IMAGEN}><{IMAGEN}>¶¶
<{IMAGEN}><{IMAGEN}>¶¶[redacted]¶¶
<{IMAGEN}>¶¶
<{IMAGEN}>¶¶
```

Grupo 2

```
Planteamiento de dudas, preguntas y cuestiones.¶¶
Preguntas, dudas y cuestiones.¶¶
Preguntas, dudas y cuestiones.¶¶
Preguntas, dudas y cuestiones.¶¶
Preguntas, dudas y cuestiones.¶¶
Preguntas, dudas y cuestiones.¶¶
coevolución.¶Preguntas, dudas y cuestiones.¶¶
Preguntas, dudas y cuestiones.¶¶
ciberespacio¶Preguntas, dudas y cuestiones.¶¶
Preguntas, dudas y cuestiones.¶¶
Preguntas, dudas y cuestiones.¶¶
Una progresiva aceleración.¶Preguntas, dudas y cuestiones.¶¶
Preguntas, dudas y cuestiones.¶¶
...
```

Figura 5.5: Mensajes de los grupos 0 y 2 en el agrupamiento del foro de Sociedad del Conocimiento con HDBSCAN, representación fastText-TF, muestra mínima 1 y tamaño mínimo 5.

las distintas representaciones. Las medidas internas siguen siendo todas positivas, pero para el ARI y la AMI aparecen valores mínimamente negativos para algunas representaciones y foros, quedando a salvo las representaciones LogTF-IDF y USE. Llama la atención la silueta de 0.514 alcanzada con la representación fastText-TF y distancia coseno sobre el foro de Sociedad del Conocimiento, que viene acompañada de valores de cobertura y consistencia en torno al 94 %. De acuerdo con la matriz de confusión 5.10, se trata de un agrupamiento sencillo. Un 6 % de mensajes se considera ruido y viene con etiqueta de grupo -1. Existe un grupo 1, mayoritario y seguramente variopinto, que alberga el 82 % de los mensajes. El 12 % restante se divide en dos grupos, 0 y 2, transversales a la estructura de subforos. Todo esto explica que el ARI y la AMI no sean particularmente buenos, con valores respectivos de 0.035 y 0.078. No obstante, la figura 5.5 demuestra que los dos grupos pequeños tienen perfecto sentido. Los mensajes del grupo 0 son todos imágenes (emoticonos y GIFs), mientras que los del grupo 2 son textos cortos, casi idénticos y con extrañas réplicas. Éste es otro ejemplo en el que, mirando sólo el texto de los mensajes, no habría manera de reproducir su distribución original por subforos.

Merece la pena inspeccionar uno más de estos agrupamientos, por ejemplo el obtenido con representación USE y distancia angular para el foro de Psicofarmacología. Éste es uno de los pocos casos en la tabla 5.9 en los que la elección de una distancia u otra altera las etiquetas de grupo y, por ende, las medidas externas. Aquí, la distancia angular logra mejor AMI, 0.191 frente a 0.169 para la distancia coseno, mientras que el ARI es aproximadamente 0.127 para ambas distancias. La matriz de confusión del agrupamiento con distancia angular se muestra en la tabla 5.11. Un 56 % de los mensajes se reconoce como ruido y el otro 44 % se distribuye en 10 grupos con tamaños que van desde mínimo prefijado de 5 hasta 72. Varios grupos son

puros, en cuanto se componen de mensajes provenientes de un solo subforo. Es el caso de los mostrados, algunos, por su extensión, truncados, en las figuras 5.6 y 5.7. Un examen rápido de los mensajes confirma que los grupos tienen sentido. El grupo 3 trata del libro de la asignatura y de sus temas y ediciones, mientras que el 4 se refiere a algunos problemas con una descarga. Desde luego parecen temas distintos, pese a que los mensajes de ambos vienen del subforo de cuestiones generales. De igual manera, los mensajes de los grupos 5 y 9 provienen del subforo del primer tema de la asignatura, pero el grupo 5 se centra en la esquizofrenia y sus síntomas, y el grupo 9 lo hace en los antipsicóticos, como la clozapina. De hecho, los capítulos 4 y 5 del libro de Stahl *et al.* (2014) a los que alude el título del subforo se titulan justamente "Psicosis y esquizofrenia" y "Agentes antipsicóticos". Por último, el grupo 7, formado por mensajes del subforo de estudiantes, contiene peticiones de adhesión por número de teléfono a un grupo de WhatsApp, similares a las referidas a Telegram en los foros de Procesadores del Lenguaje.

Al inicio de la sección se explicó que detrás del agrupamiento producido por HDBSCAN yace una estructura jerárquica parecida a la de los algoritmos aglomerativos y divisivos. Cabe preguntarse si dicha estructura podría reflejar o dilucidar relaciones como la encontrada entre los grupos 5 y 9 del foro de Psicofarmacología, cuyos mensajes provienen de un mismo subforo, pero se centran en aspectos distintos. La figura 5.8 muestra el árbol condensado del *clustering* discutido en el párrafo anterior y plasmado en la matriz de confusión 5.11. El árbol condensado es una representación resumida de la jerarquía construida durante el proceso de agrupamiento en la que se dejan caer los grupos con tamaño inferior al mínimo que se sueltan a medida que sube el umbral de densidad y se eliminan aristas del árbol recubridor mínimo (véase McInnes y Healy, 2017). El árbol refleja este tipo de pérdida como una disminución gradual del tamaño de grupo, que se indica por el color y ancho de línea. Por supuesto, también captura la división de grupos en subgrupos de tamaño admisible, que se distribuyen espaciadamente a lo largo de la horizontal. Los grupos seleccionados por su estabilidad o persistencia para el agrupamiento final se suelen rodear y en ocasiones etiquetar. Volviendo a la cuestión original, en el árbol de la figura 5.8, los grupos 5 y 9 no son hijos de un mismo progenitor, pero el hermano del 5 es

Tabla 5.11: Matriz de confusión del agrupamiento del foro de Psicofarmacología con algoritmo HDBSCAN, representación USE, distancia angular, muestra mínima 1 y tamaño mínimo 5.

Subforo \ Grupo	−1	0	1	2	3	4	5	6	7	8	9	Total
Cuestiones generales de la asignatura	90	1	3	16	6	5	0	8	0	2	0	131
Foro Tema 1 - Capítulos 4 y 5 del libro de Stahl	57	0	0	0	0	0	7	5	0	33	12	114
Foro Tema 2 - Capítulos 6, 7 y 8 del libro de Stahl	32	0	0	0	0	0	0	4	0	20	0	56
Foro Tema 3 - Capítulo 9 del libro de Stahl	13	0	0	0	0	0	0	1	0	7	0	21
Foro Tema 4 - Capítulo 11 del libro de Stahl	4	0	0	0	0	0	0	0	0	2	0	6
Foro Tema 5 - Capítulos 12 y 13 del libro de Stahl	6	0	1	0	0	0	0	1	0	3	0	11
Foro Tema 6 - Capítulo 14 del libro de Stahl	0	0	0	0	0	0	0	0	0	3	0	3
Foro de estudiantes (Cafetería)	28	3	0	0	0	0	0	6	33	0	0	70
Coordinación tutorial	7	1	1	4	0	0	0	0	0	2	0	15
Total	237	5	5	20	6	5	7	25	33	72	12	427

Grupo 3

Estimado ED:¶¶Me gustaría saber si la 5ª **edición** del **libro** *_¶Psicofarmacología Esencial de Stahl. Guía del¶Prescriptor_* sirve o sería mejor la 6ª (actual).¶¿En qué cambian?¶¶Gracias de antemano.¶¶ [redacted], particularmente no dispongo de la 6ª**¶edición** (mayo 2018) de la guía del prescriptor, pero con¶la 5ª (octubre 2015) les es más que suficiente para¶preparar la asignatura (como material complementario, claro).¶¶No creo que haya cambiado mucho en tres años, aunque ya le digo¶que hablo sin conocimiento de causa; pero es probable que hayan¶actualizado algún psicofármaco nuevo de reciente¶creación y/o aceptación de algún otro agente¶psicofarmacológico -en estudio preclínico previo- con¶posterioridad a 2015 (5ª **edición**); o que el autor haya¶considerado eliminar alguno/s otro/s que en EEUU se hayan retirado del¶mercado por diversas causas esgrimidas según los criterios¶establecidos por la oficina de la FDA (Food and Drug Association). O,¶incluso, que se hayan hecho añadidos respecto a las dosis¶recomendadas y efectos secundarios descubiertos como nuevos. A veces,¶las nuevas **ediciones**, responden a intereses comerciales de las¶editoriales exclusivamente.¶¶Siempre, si no tiene la quinta **edición**, es conveniente adquirir¶la última **edición** disponible, por lo que le acabo de¶comentar; pero no es imprescindible, bajo mi punto de vista.¶¶

Buenas tardes, voy a usar el nuevo **libro** para estudiar pero¶también me voy a apoyar en los apuntes que una compañera¶ha compartido conmigo, estos son del curso 2016/2017, me¶gustaría saber si se han hecho cambios desde entonces,para¶poder estudiarlos segura. Gracias.¶¶

Buenas tardes:¶¶En la guía de la asignatura de este curso, pone que la¶**edición** del **libro** que debemos tener es la cuarta, la del¶año 2014, sin embargo tengo entendido que hay una¶versión más reciente. Alguien me podría aclarar¶qué **edición** debo comprar?¶¶Muchas gracias¶¶

Buenos días.¶¶Mi pregunta es porqué solo hay dos foros que engloban solo¶algunos temas del **libro**. ¿Qué pasa con los otros temas ?¶¿Estamos siguiendo la ultima versión del **libro** de Stahl¶verdad?¶¶Gracias.¶¶

Me he liado un poco, ¿son los mismos temas y **libro** que el¶año pasado ? porque solo veo dos foros para todos los temas?¶¶

Grupo 4

Buenas tardes! Yo tampoco puedo **descargarme** el documento en pdf de las¶adendas. Cada vez que lo intento me salta un aviso¶comunicándome que la página da error o no se encuentra.¶¶

Buenas tardes,¶¶Yo tampoco puedo **descargar** el archivo.¶¶Un saludo,¶¶[redacted].¶¶

Buenos días,¶¶Desconozco cual es el problema dado que yo sí puedo¶**descargarla**, pero para que puedan acceder al fichero se lo adjunto en¶este correo.¶¶Saludos cordiales,¶¶

Gracias por la sugerencia, [redacted]; pero yo estoy entrando en el curso¶actual y no he tenido problemas para la **descarga** tampoco.¶¶Aún así, lo comunicaremos, por si hubiera algún¶error en los enlaces.¶¶Si siguen teniendo problemas para **descargar** algún documento¶háganoslo saber para darle solución.¶¶¶¶Saludos.¶¶

Buenos días:¶¶En relación a este asunto, efectivamente, el enlace de **descarga¶**"apuntaba" al curso del año pasado. Es por ello que daba el¶error que se indica, no pudiendo accder a los archivos.¶¶Ya está solucionado.¶¶Sentimos las molestias que esta indicencia les haya podido ocasionar.¶En nombre de todo el Equipo Docente, les pido disculpas.¶¶Un saludo cordial,¶¶[redacted]¶¶

Grupo 5

esquizofrenia¶Buenas tardes,¶¶Me surge una duda sobre los circuitos cerebrales y su respectiva¶correspondencia sintomatica.En la página 85 del manual se¶explica que el cortex prefrontal ventromedial y mesocortical se¶relacionan con síntomas negativos y afectivos. Sin embargo, en¶el dibujo de la página 84 los síntomas negativos (no los¶sintomas afectivos)se deben a una disfunción de circuitos¶mesocorticales y los sintomas afectivos se relacionan con el cortex¶prefrontal ventromedial. Mi pregunta sería si el cortex¶prefrontal ventromedial y el mesocortical si ambos se relacionan¶tanto con síntomas negativos como con síntomas¶afectivos¶¶Gracias.¶¶

esquizofrenia¶Estimada [redacted],¶¶En la **esquizofrenia**, se considera que los síntomas negativos¶son debidos a una disfunción de circuitos mesocorticales,¶mientras que los síntomas afectivos serían causados por¶una disfunción de circuitos de la corteza prefrontal¶ventromedial.¶¶Saludos cordiales,¶¶

Apreciado [redacted]:¶¶Le sugiero que consulte la siguiente FAQ [1]("No entiendo bien la¶hipótesis de la hipofunción de los receptores NMDA en la¶**esquizofrenia**").¶¶Si sigue teniendo dudas, por favor, háganoslo saber.¶¶Un saludo,¶¶[redacted]¶¶¶¶[1] ¶<{URL}>¶¶

Estimado equipo docente y compañeros,¶¶Dada la hipótesis de la hipofuncion del receptor NMDA en la¶**esquizofrenia**, y en concreto en cuanto a la desconectividad de las¶neuronas glutamatergicas anteriores en el hipocampo, quiere esto decir¶que una activación de estás vías, por ejemplo por¶recuerdos traumáticos, estaría influyendo en los¶síntomas positivos de la **esquizofrenia**? Esto se relaciona con¶el hecho de que determinados sucesos vitales estresantes sean caldo de¶cultivo para posibles **esquizofrenias** al unirse con otros factores¶desencadenantes?¶¶Gracias de antemano,¶¶Saludos cordiales,¶¶[redacted]¶¶

Estimado equipo docente,¶¶Mi pregunta anterior va más encaminada al desencadenamiento de¶crisis psicóticas, más que al desarrollo de la¶**esquizofrenia**, ya que esto último se trata en el¶epígrafe de neurodesarrollo y genética de la¶**esquizofrenia**.¶¶Saludos cordiales,¶¶[redacted]¶¶

Estimado equipo docente,¶¶Los síntomas afectivos estarían también¶explicados por la hipótesis de hipofuncionalidad de NMDA con la¶hipótesis dopaminergica de la **esquizofrenia**, verdad? Es que no¶se hace referencia a ello ni en el texto ni en la figura 4.32...¶¶Gracias de antemano,¶¶Saludos cordiales,¶¶[redacted]¶¶

Estimada [redacted],¶¶Efectivamente se piensa que la hipofunción de los ecpetores¶NMDA podría explicar también los síntomas¶afectivos. A este respecto, en la página 107, se indica¶que la acción del antagonista del receptor NMDA fenciclidina¶(PCP) "mimetiza los síntomas cognitivos, negativos y afectivos¶de la **esquizofrenia** como aislamiento social y disfunción¶ejecutiva".¶¶Saludos cordiales,¶¶

Figura 5.6: Mensajes de los grupos 3, 4 y 5 en el agrupamiento del foro de Psicofarmacología con HDBSCAN, representación USE, distancia angular, muestra mínima 1 y tamaño mínimo 5.

Grupo 7

hola, me gustaria que me pudieran añadir, gracias.¶¶[redacted]¶¶
Buenos dias, me gustaria ser añadida en el grupo [redacted]¶¶Gracias¶¶
Buenos días.¶¶También me gustaría ser añadida: [redacted].¶¶Muchas gracias.¶¶
B días!¶¶Este es mi número para añadirme al grupo [redacted]¶¶Gracias¶¶
Me gustaría mucho ser añadida a vuestro grupo [redacted] Gracias¶y un saludo a todos¶¶
Hola, me gustaría estar en el grupo.¶¶[redacted]¶¶Gracias!!¶¶
Hola !!!¶¶Me gustaría añadirme al grupo de Wassap.¶¶Mi número [redacted].¶¶Gracias¶¶
Si me podéis añadir [redacted]¶¶Muchas gracias¶¶
a mi también me interesa [redacted]¶¶gracias¶¶
Yo quiero entrar, [redacted].¶¶Gracias¶¶
El grupo me interesa. Puede añadirme?¶¶Mi telefono : [redacted]¶¶gracias¶¶
Hola, por favor si me pueden agregar al grupo de whatsaap [redacted]¶¶Gracias¶¶
Hola, por favor si alguien me pudiera agregar al grupo, mi¶número es [redacted]¶¶
Podríais admitirme en el grupo de WhatsApp? Gracias [redacted]¶¶
...

Grupo 9

Buenos dias¶¶En la pag. 182 del libro de texto nos habla de la Clozapina:¶"..posiblemente tenga el mayor riesgo
↪ cardiometabólico de entre¶todos los **antipsicóticos**".¶¶Luego, en el siguiente apartado,nos habla de la
↪ Olanzapina: "Es el¶**antipsicótico** con mayor riesgo cardiometabólico, dado¶que.."¶¶Busco la respuesta a la
↪ pregunta: Cuál de los dos¶**antipsicóticos** mencionados tiene mayor riesgo¶cardiometabólico?¶¶Gracias¶¶
En referencia a la Zotepina (pág 190) me asaltan varias dudas a¶consecuencia de leer " la dosis de Zotepina
↪ prolonga el ciclo QTc¶proporcionalmente, y generalmente se administra tres veces al¶día".¶¶He leído lo
↪ relacionado al ciclo QT en el apartado de¶preguntas, porque no sé si había pasado algo por alto,¶pero este
↪ término no me sonaba de nada.¶¶Tras leer en qué consiste, claro está con una idea muyyy¶general, me pregunto
↪ si aquellos **antipsicoticos** que generen un mayor¶riesgo metabólico, también tienden a prolongar el ciclo¶...
Estimada [redacted]¶¶En persona sanas sin medicación, existen algunos estudios que¶relacionan la alteración del
↪ intervalo QTC con la edad, el¶índice de masa corporal, la presión arterial, y los¶niveles plasmáticos de
↪ glucosa y triglicéridos, todos¶ellos relacionados con el síndrome metabólico. Sin¶embargo, en relación a su
↪ pregunta le comento que los¶**antipsicóticos** pueden tener provocar distintos riesgos para¶padecer síndrome
↪ metabólico y alteraciones del intervalo¶QTc. Así, la clozapina y la olanzapina son los¶**antipsicóticos** que...
Buenas, en la página 182, se indica en la primera columna, al¶final del primer párrafo que *la clozapina tiene el
↪ mayor¶riesgo de cardiometabólico de entre todos los¶**antipsicóticos***. Y en la misma página, segunda
↪ columna,¶primer párrafo se dice que *la olanzapina es el¶**antipsicótico** con mayor riesgo
↪ cardiometabólico*.¶¶Entonces, ¿cual de los dos es el de mayor riesgo?¶¶
Estimada [redacted],¶¶Tanto la clozapina como la olanzapina son **antipsicóticos** con un¶alto riesgo cardiometabólico
↪ porque aumentan los¶triglicéridos, y producen intolerancia a la glucosa, y ganancia¶de peso corporal. En el
↪ libro se indica que la clozapina "_*¶está entre los **antipsicóticos** más destacados a la¶hora de incrementar el
↪ riesgo cardiometabólico, incluyendo el¶incremento en plasma de los triglicéridos en ayunas así¶como de la
↪ resistencia a la insulina*_" mientras que de la olanzapina¶se afirma que "*_es el **antipsicótico** con mayor...
Estimado profesor [redacted]:¶¶Debo decirle que la citada página del manual no lo expresa¶exactamente como
↪ usted ha expuesto.¶¶Tal como indica la compañera se puede constatar que en la¶página 182 del manual viene
↪ explicado textualmente sobre la¶clozapina que "posiblemente tenga el mayor riesgo¶cardiometabólico de entre
↪ todos los **antipsicóticos**". Y¶en la misma página sobre la olanzapina también nos¶indica que efectivamente "es
↪ el **antipsicótico** con mayor riesgo¶cardiometabólico".¶¶Después de su aclaración y de puntualizar que es la¶...
Buenos días.¶¶Al igual que el compañero [redacted], opino que el ED¶debería aclarar tal extremo, toda vez que en el
↪ libro se cita¶textualmente que ambas "pinas" tienen el mayor riesgo¶cardiometabólico de entre los
↪ **antipsicóticos**.¶¶Saludos.¶¶
Buenos días!¶¶Al igual que el compañero [redacted], opino que el ED¶debería aclarar tal extremo, toda vez que en el
↪ libro cita¶textualmente, cada una en su apartado correspondiente, que ambas¶"pinas" son las que tienen el
↪ mayor riesgo cardiometabólico de¶entre los **antipsicóticos** (página 182, línea 12 de¶la primera columna y 5 de
↪ la segunda).¶¶Saludos.¶¶
Buenos días,¶¶Tal y como he indicado, ambos **antipsicóticos** clozapina y¶olanzapina muestran el mayor grado de
↪ riesgo cardiometabólico,¶claramente por encima de otros **antipsicóticos** y, ambos¶fármacos tienen un riego
↪ similar. Asi pues, en el examen no¶preguntaremos si la clozapina tiene mayor riesgo¶cardiometabólico que la
↪ olazapina o viceversa porque no se¶podría contestar de manera correcta e inequívoca.¶¶Saludos cordiales,¶¶
Buenas noches, mi pregunta es la siguiente:¶¶¿Cuando en los **antipsicóticos** atípicos (pinas,¶donas, etc) se nos
↪ indica que provocan resistencia a la insulina y¶dislipidemia, se nos está indicando que suponen un
↪ riesgo¶cardiometabólico (entendido como mayor probabilidad de¶desarrollar diabetes o una enfermedad
↪ cardiovascular)?¶¶Un saludo.¶¶
Buenos días.¶¶En la página 182, párrafos 2 y 3 se dice en ambos que el¶**antipsicótico** con mayor riesgo
↪ cardiometabólico es¶clozapina y olanzapina respectivamente, así como resistencia a¶la insulina. ¿En cuál de
↪ los dos es correcto lo¶mencionado? ¿Es quizás una errata?¶¶Gracias.¶¶
Parece ser que, en términos generales de todos los¶**antipsicóticos** de segunda generación (clozapine,¶olanzapina,
↪ quetiapina), la clozapina es la que mayor riesgo¶metabólico suele presentar en la mayoría de los¶pacientes
↪ y, sobre todo, la que parece ser menos segura -menos¶saludable, para entendernos; no nos olvidemos que puede
↪ producir¶agranulocitosis- de todos ellos aunque sí un¶psicofármaco más eficaz como **antipsicótico**.
↪ No¶obstante, hay que decir que, al igual que el autor del libro suele¶equiparar en riesgo metabólico a la...

Figura 5.7: Mensajes de los grupos 7 y 9 en el agrupamiento del foro de Psicofarmacología con HDBSCAN, representación USE, distancia angular, muestra mínima 1 y tamaño mínimo 5.

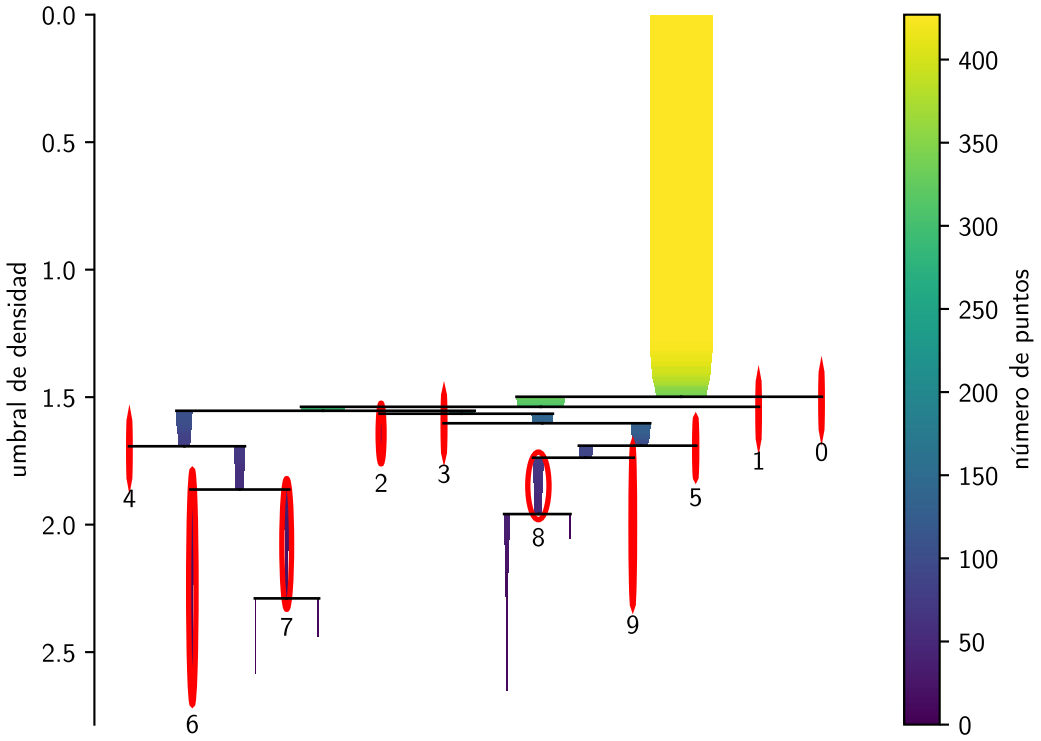


Figura 5.8: Árbol condensado del agrupamiento del foro de Psicofarmacología con algoritmo HDBSCAN, representación USE, distancia angular, muestra mínima 1 y tamaño mínimo 5.

padre del 8 y el 9, luego tampoco quedan lejos en la jerarquía. Según la matriz de confusión, el grupo 8 es un grupo grande y transversal, con 72 mensajes de los cuales 33 provienen del primer tema, al igual que los de los grupos 5 y 9. Podría parecer extraño que dicho grupo, hermano del 9 y sobrino del 5, albergue mensajes de subforos de dudosa relación con el primer tema, como el de coordinación tutorial. Sin embargo, y a modo de ejemplo, la figura 5.9 revela que los dos mensajes del grupo 8 provenientes del subforo de coordinación tutorial mencionan la esquizofrenia y los antipsicóticos. Es más, contienen referencias explícitas al primer tema de la asignatura y al capítulo 4 del libro de Stahl, por lo que cabría preguntarse si estos mensajes están colocados en el subforo más idóneo en la estructura original. Por casualidad, el segundo de los mensajes de la figura 5.9 incluye un fragmento en inglés, pero eso no debería de estorbar al *embedding* multilingüe preentrenado usado aquí para su representación.

Por ofrecer un ejemplo más de la utilidad del árbol condensado para el análisis de agrupamientos, la figura 5.8 indica que grupo 6 es hermano del 7, aquel proveniente del subforo de estudiantes y compuesto por solicitudes de adhesión a un grupo de WhatsApp. El grupo 6 es un grupo transversal, con 25 mensajes de los cuales 6 vienen del subforo de estudiantes. Tal como muestra el extracto de mensajes en la figura 5.10, el grupo 6 contiene agradecimientos diversos, sin demasiado contexto, y varias peticiones de adhesión al grupo de WhatsApp. No queda claro por qué estas últimas terminan en un grupo distinto al 7, pero al menos quedan en un grupo hermano en la estructura jerárquica, y su mezcla con mensajes de agradecimiento resulta comprensible por cuanto las solicitudes vienen acompañadas de cortesías.

```
Estimados compañeros y compañeras,¶¶Escribo para hacer constar una apreciación sobre el **primer tema**¶de la
↪ asignatura. Primero que nada decir que no dispongo de la¶versión actualizada del manual Stahl, sino la
↪ versión de¶2014, y que éste es mi primer año como PT, así¶que desconozco si esto lo han discutido en años
↪ anteriores.¶¶En la página 86 del *__capítulo 4__ *del libro:¶Psicósis y **esquizofrenia**, habla de
↪ neurotransmisores y¶circuitos en la **esquizofrenia** y explica en concreto el sistema¶dopaminérgico. Cito
↪ textualmente: "Quizá el receptor de¶dopamina más extensamente investigado sea el receptor de¶dopamina 2
↪ (D2), ya que es estimulado por los agonistas¶dopaminérgicos para el tratamiento de la EP, y bloqueado por¶los
↪ **antipsicóticos** antagonistas de dopaminérgicos en el¶tratamiento de la **esquizofrenia**". Luego en la Figura 4-7
↪ dice: "Hay¶además, un receptor de dopamina del subtipo 2 en la presinapsis¶que funciona como autorreceptor,
↪ regulando la liberación de¶dopamina de la neurona presináptica. También hay varios¶tipos de receptores de
↪ dopamina de los subtipos 1, 2, 3, 4 y 5".¶Más adelante solo cita D2 cuando se refiere a
↪ los¶autorreceptores.¶¶Aunque al principio se dudaba de si D3R funcionaba como autorreceptor,¶actualmente se
↪ conoce que D3R es también *autorreceptor¶dopaminérgico* y regula la actividad de¶DAT (Castro-Hernández,
↪ Neurobiol dis. 2015). Los receptores¶dopaminérgicos se agrupan en D2-like (D2, D3 y D4) y D1-like¶(D1 y D5).
↪ Es posible que en el libro haya habido una mala¶traducción y cuando habla de autorreceptores D2 se refiera
↪ a¶D2-like, aunque el papel de D4 es menos conocido.¶¶Además, el receptor D3 es diana de uno de los
↪ fármacos¶empleados actualmente en tratamiento de la EP, el pramipexol.¶¶Saludos a todos y todas desde
↪ UNED-Tenerife y buen comienzo de curso!¶¶[redacted].¶¶
Apreciado [redacted]:¶¶En primer lugar quiero pedirte disculpas, al igual que al resto de¶compañeras y compañeros,
↪ por mi ausencia en estos foros¶hasta ahora, pero acabo de incorporarme tras una licencia fuera
↪ de¶España.¶¶Aunque ya lo ha hecho [redacted], en nombre de todos nosotros te¶agradezco tu mensaje, la acertada
↪ consideración que haces en¶él y el habernos proporcionado la cita de ese interesante¶trabajo -que
↪ desconocía- sobre la relación entre el¶autoreceptor D3 y el DAT en la regulación de la¶liberación de dopamina
↪ en el estriado, que tienes publicado con¶otros autores en _Neurobiology of Disease_; trabajo por el que,
↪ al¶igual que [redacted], te felicito muy sinceramente.¶¶A raíz de tu mensaje he consultado la versión inglesa
↪ de¶la Cuarta Edición del libro de Stahl. En ella se indica lo¶siguiente: \*_Receptors for dopamine also
↪ regulate dopaminergic¶neurotransmission (Figure 4-7). The DA transporter DAT and the¶vesicular transporter
↪ VMAT2 are both types of receptors. A plethora of¶additional receptors exist, including at least five
↪ pharmacological¶subtypes and several more molecular isoforms. Perhaps the most¶extensively investigated
↪ dopamine receptor is the dopamine 2 (D2)¶receptor, as it is stimulated by dopamine agonists for the
↪ treatment¶of Parkinson's disease, and blocked by dopamine antagonist¶**antipsychotics** for the treatment of
↪ **schizophrenia**_*."¶¶En concusión no parece que haya habido un fallo en la¶traducción, pues prácticamente es
↪ literal a lo que en la¶versión española se indica: \*_Los receptores de¶dopamina además regulan la
↪ neurotransmisión¶dopaminérgica (Figura 4-7). El transportador de DA (TDA) y el¶transportador vesicular TVMA2
↪ son ambos tipos de receptores. Existe¶toda una batería de receptores dopaminérgicos¶adicionales, incluidos
↪ al menos cinco subtipos farmacológicos y¶varias isoformas moleculares más. Quizá el receptor de¶dopamina más
↪ extensamente investigado sea el receptor de¶dopamina 2 (D2), ya que es estimulado por los
↪ agonistas¶dopaminérgicos para el tratamiento de la enfermedad de¶Parkinson, y bloqueado por los
↪ **antipsicóticos** antagonistas¶dopaminérgicos en el tratamiento de la **esquizofrenia**._*"¶¶En el manual se ...
```

Figura 5.9: Mensajes provenientes del subforo de coordinación tutorial en el grupo 8 del agrupamiento del foro de Psicofarmacología con HDBSCAN, representación USE, distancia angular, muestra mínima 1 y tamaño mínimo 5.

```
Muchísimas gracias, [redacted], por todas las aclaraciones.¶¶¶¶Saludos cordiales.¶¶
Mil gracias.¶¶¶¶Un saludo.¶¶
Muchas gracias, [redacted].¶¶Un saludo cordial¶¶
Muchas gracias, [redacted].¶¶¶¶Un saludo.¶¶
Muchas gracias.¶¶¶¶Ya le he enviado el correo.¶¶¶¶Un saludo.¶¶
Hola,¶¶Lo tendré en cuenta.¶¶Muchísimas gracias.¶¶Un Saludo.¶¶
Buenas tardes!¶¶Arreglado, muchas gracias!¶¶
Perfecto, muchas gracias.¶¶
¡Muchas gracias!¶¶
Es cierto.¶¶Muchas gracias.¶¶
...
Entendido ¡Gracias!¶¶
Gracias, es justo lo que querría saber.¶¶
Gracias.¶¶
Gracias.¶¶Un saludo.¶¶
Por favor, ¿podéis incluirme? [redacted]¶¶¶¶Gracias¶¶
Hola [redacted], parece que no funciona el enlace. Me podriais anadir por¶favor?¶¶[redacted]¶¶Muchas gracias de
↪ antemano.¶¶
Muchas gracias! las buscaré!!!¶¶
Hola!! Podrias anadirme al [redacted] por favor?¶¶Muchas gracias.¶¶
Hola. Podriais anadirme al [redacted].¶¶Muchas gracias!¶¶
Y a mí, gracias¶¶[redacted]¶¶
```

Figura 5.10: Mensajes del grupo 6 en el agrupamiento del foro de Psicofarmacología con HDBSCAN, representación USE, distancia angular, muestra mínima 1 y tamaño mínimo 5.

Capítulo 6

Programas

A lo largo de los capítulos anteriores, se han desarrollado, probado y evaluado los componentes del sistema de agrupación automática y colocación de mensajes nuevos. Dando por concluida la parte investigativa del trabajo, en este capítulo se presentan los programas desarrollados como puntos de entrada de dicho sistema. Estos programas, que integran los diversos componentes en *pipelines* de scikit-learn, podrían ser ejecutados con facilidad por usuarios no expertos sobre otros ficheros de foros con el mismo formato o ser adaptados para correr como servicios en un sistema en línea.

6.1. Exploración de foros

El primero de los programas está orientado a tareas de exploración de foros, como la llevada a cabo en la sección 3.3. El programa consta de dos comandos, uno para generar una tabla resumen con el número de subforos, hilos y mensajes de uno o varios foros, y otro para inspeccionar en más detalle la estructura de subforos. El programa admite opciones que permiten especificar el directorio de datos y los foros a explorar. El comando de exploración de subforos admite además una opción de modo interactivo que permite consultar los mensajes de ficheros o subforos concretos.

La figura 6.1 muestra un ejemplo de sesión de exploración, en el que primero se obtiene el resumen de los foros de Procesadores del Lenguaje y después se inspeccionan interactivamente los subforos de los cursos 2019–20 y 2020–21. Tanto la tabla de resumen como la de subforos tienen apartados separados para los subforos normales y los de grupos de tutoría. La expresión regular utilizada para determinar el tipo de subforo está expuesta como opción del programa. En el ejemplo mostrado, se consultan los mensajes de los subforos de los grupos de tutoría 7 y 12 del curso 2020–21, que contienen 3 y 2 mensajes, respectivamente. Ambos subforos incluyen mensajes referidos a la obligatoriedad de las sesiones de control y a su planificación. Se recuerda que las divisiones entre grupos de tutoría tienen que ver más con la geografía que con la temática, por lo cual esos subforos han sido excluidos de la mayor parte del análisis.

```
$ python unedforums/scripts/explora.py -d data -f ForosPL1.*\\.csv resumen
Tutoría                 No                      Sí
Entidad            Subforos Hilos Mensajes Subforos Hilos Mensajes
Fichero
ForosPL1_17-18.csv        8    88      372       19    52      137
ForosPL1_18-19.csv        8    62      270       18    45      104
ForosPL1_19-20.csv        8    66      176       13    43       96
ForosPL1_20-21.csv        8    82      302       17    49      109

$ python unedforums/scripts/explora.py -d data -f ForosPL1_..-2.\\.csv subforos -i
idFichero                                                                  0                 1
Fichero                                                  ForosPL1_19-20.csv ForosPL1_20-21.csv
Tutoría idSubforo Subforo
No      A         Foro Análisis léxico (fuera de la práctica)              7                12
        B         Foro Análisis sintáctico (fuera de la práctica)          2                 2
        C         Foro General Práctica                                   83               161
        D         Foro Práctica: análisis léxico                          28                13
        E         Foro Práctica: análisis sintáctico                      26                59
        F         Foro de consultas generales                             23                48
        G         Foro de estudiantes (no moderado por el Equipo ...       2                 5
        H         Coordinación tutorial                                    5                 2
Sí      I         Grupo de Tutoria 1                                      16                10
        J         Grupo de Tutoria 10                                      6                 1
        K         Grupo de Tutoria 11                                      0                 2
        L         Grupo de Tutoria 12                                      0                 2
        M         Grupo de Tutoria 14                                      3                 8
        N         Grupo de Tutoria 15                                      2                 6
        O         Grupo de Tutoria 16                                      0                23
        P         Grupo de Tutoria 18                                      6                 0
        Q         Grupo de Tutoria 19                                      0                18
        R         Grupo de Tutoria 2                                       0                 9
        S         Grupo de Tutoria 20                                      5                 1
        T         Grupo de Tutoria 21                                      3                 1
        U         Grupo de Tutoria 23                                      0                 6
        V         Grupo de Tutoria 25                                      0                 2
        W         Grupo de Tutoria 3                                      11                 0
        X         Grupo de Tutoria 4                                       3                 3
        Y         Grupo de Tutoria 5                                      16                 0
        Z         Grupo de Tutoria 6                                       0                 8
        BA        Grupo de Tutoria 7                                       2                 3
        BB        Grupo de Tutoria 8                                      20                 0
        BC        Grupo de Tutoria 9                                       3                 6

Selecciona un subforo y/o fichero (p. ej., "A", "0", "A0", ...): BA1
Buenos días, me gustaría saber si fuera posible entregar<br/>la práctica de cara a la convocatoria de Septiembre,
↪  y si esto<br/>fuera posible cómo se realizaría la sesión de<br/>control.<br/><br/>Muchas
↪  gracias,<br/><br/>[redacted].<br/><br/>

Buenos días.<br/><br/>Esta asignatura requiere que se supere una sesión de control<br/>obligatoria y, por
↪  supuesto, entregar la práctica obligatoria,<br/>pero me han surgido varias dudas. La primera de ellas es la
↪  fecha de<br/>la sesión de control y si, en estos momentos, se puede saber si<br/>se realizará de manera
↪  presencial u online. La siguiente duda<br/>es la fecha de entrega de la propia práctica.<br/><br/>Gracias de
↪  antemano.<br/><br/>Un saludo.<br/><br/>

Buenas, estaba interesado en saber si han resuelto su duda.<br/>Necesitaria saber cuando y como va ser la la
↪  sesión de control<br/>obligatoria. Muchas gracias<br/><br/>

Selecciona un subforo y/o fichero (p. ej., "A", "0", "A0", ...): L1
Hola, buenos días.<br/><br/>Como sabéis, en esta asignatura debéis presentaros a una<br/>sesión de control de la
↪  práctica.<br/><br/>Los que no la habéis hecho ya, podéis hacerla en hora<br/>de tutoría, cuando queráis en
↪  los dos lunes que quedan:<br/>11 y 18 de enero. Si no es así, podemos intentar encontrar otro<br/>momento
↪  para hacerla. Escribidme por favor a<br/>[redacted] para concretar fecha y hora para
↪  la<br/>sesión.<br/><br/>Un saludo,<br/><br/>Inés<br/><br/>

Hola,<br/><br/>Hoy teníamos la primera sesión con el tutor y no ha<br/>aparecido nadie, no se si ha pasado algo o
↪  es que se mueve la<br/>tutoría a otro día y no ha dado tiempo a avisar.<br/><br/>En cualquier caso, para los
↪  que no hayan podido asistir, no ha habido<br/>tutoría.<br/><br/>

Selecciona un subforo y/o fichero (p. ej., "A", "0", "A0", ...):
```

Figura 6.1: Salida del programa de exploración de foros.

```
$ python unedforums/scripts/agrupa.py -i -r use -c hdbscan --min_samples 1 --min_cluster_size 5 --angular
↪  data/6201302-_2019_cl.csv
grupo                                                    -1  0  1   2  3  4  5   6   7   8   9  Total
idSubforo Subforo
A         Cuestiones generales de la asignatura          90  1  3  16  6  5  0   8   0   2   0    131
B         Foro Tema 1 - Capítulos 4 y 5 del libro de Stahl  57  0  0   0  0  0  7   5   0  33  12    114
C         Foro Tema 2 - Capítulos 6, 7 y 8 del libro de Stahl  32  0  0   0  0  0  0   4   0  20   0     56
D         Foro Tema 3 - Capítulo 9 del libro de Stahl     13  0  0   0  0  0  0   1   0   7   0     21
E         Foro Tema 4 - Capítulo 11 del libro de Stahl     4  0  0   0  0  0  0   0   0   2   0      6
F         Foro Tema 5 - Capítulos 12 y 13 del libro de Stahl  6  0  1   0  0  0  0   1   0   3   0     11
G         Foro Tema 6 - Capítulo 14 del libro de Stahl     0  0  0   0  0  0  0   0   0   3   0      3
H         Foro de estudiantes (Cafetería)                 28  3  0   0  0  0  0   6  33   0   0     70
I         Coordinación tutorial                            7  1  1   4  0  0  0   0   0   2   0     15
Total                                                    237  5  5  20  6  5  7  25  33  72  12    427

                      Valor
Medida
Cobertura          0.444965
Núm. grupos       10.000000
Núm. grupos orig.  9.000000
Silueta            0.076536
Consistencia       0.444965
DBCV               0.037120
ARI                0.126805
AMI                0.191366

Selecciona un subforo y/o grupo (p. ej., "A", "0", "A0", ...): A1
Estimado ED:<br/><br/>Me gustaría saber quién es mi tutor para la PEC y por<br/>dónde puedo ponerme en contacto
↪  con él para consultar<br/>algunas dudas.<br/><br/>Pertenezco al CA de Cáceres -
↪  Plasencia.<br/><br/><br/><br/>Gracias de antemano.<br/><br/>Un saludo.<br/><br/>

Estimado [redacted]:<br/><br/><br/><br/>El día 30 de noviembre me puse en contacto con [redacted],
↪  tutor de mi CA (Cáceres- Plasencia), para<br/>consultar algunas dudas sobre la PEC y aún no he
↪  obtenido<br/>respuesta.<br/><br/>No sé si puedo o debo dirigirme a otro tutor/a para que me<br/>orienten con
↪  el informe de la Pec.<br/><br/><br/><br/>Gracias por su atención.<br/><br/><br/><br/>Saludos.<br/><br/>

Apreciado [redacted]:<br/><br/>Según me han indicado en el Centro Asociado de Plasencia, el<br/>horario de atención
↪  de su Profesor Tutor son los lunes de 18:00<br/>a 19.00 h. No obstante, estoy intentado hablar con él para
↪  que<br/>me indique cuál es la mejor manera para que usted pueda ponerse<br/>en contacto con él a efectos de
↪  hacerle esas consultas que<br/>indica, ya que el foro de su tutoría veo que está<br/>vacío.<br/><br/>Un
↪  saludo cordial,<br/><br/>

Selecciona un subforo y/o grupo (p. ej., "A", "0", "A0", ...): F1
Disculpe, pero tiene que contactar con su Profesor tutor del centro<br/>asociado al que le corresponde. O bien,
↪  escribir en el foro de<br/>alumnos.<br/><br/><br/><br/>Saludos.<br/><br/>

Selecciona un subforo y/o grupo (p. ej., "A", "0", "A0", ...): I1
Buenos días<br/><br/>Soy nueva como tutora en el CA de Talavera de la Reina, Quería<br/>saber si para corregir la
↪  PEC nos envían una plantilla con los<br/>resultados del caso y las respuestas o es a criterio del
↪  tutor.<br/><br/>Gracias, un cordial saludo<br/><br/>[redacted]<br/><br/>
```

Figura 6.2: Salida del programa de agrupación automática sobre el foro de Psicofarmacología.

6.2. Agrupación automática

El segundo programa se encarga de la agrupación automática de los mensajes de un foro. Además de llevar a cabo el agrupamiento, el programa calcula las medidas de evaluación vistas anteriormente y las presenta junto a la matriz de confusión con respecto al agrupamiento de referencia. Al igual que el programa de exploración, dispone de un modo interactivo que permite inspeccionar los mensajes correspondientes a las diferentes entradas de la matriz de confusión, esto es, a diferentes subforos originales y grupos. El programa admite multitud de opciones para ajustar la representación de documentos, el algoritmo de agrupamiento y la medida de similitud. Las claves de las opciones pueden consultarse corriendo el programa con la opción

de ayuda habitual, que aparece también al correr el programa sin argumentos.

La figura 6.2 muestra un ejemplo de sesión de agrupación automática con representación USE, algoritmo HDBSCAN y distancia angular sobre el foro de Psicofarmacología. Se trata de la misma configuración analizada hacia el final de la sección 5.3, empezando en la página 58. Aquí, se usa el modo interactivo para inspeccionar el grupo 1, que es un grupo de 5 mensajes que provienen de los subforos de cuestiones generales, del tema 5 y de coordinación tutorial. Todos los mensajes mencionan la figura del tutor. La mayoría tiene que ver con peticiones de contacto, pero hay otros temas comunes, como la práctica y el centro asociado. El grupo tiene por tanto cierto sentido, a pesar de incluir mensajes de diferentes subforos. Como curiosidad, en el mensaje que viene del subforo del tema 5, parece que el profesor, o quizá el moderador, aconseja al estudiante que escriba su pregunta en el subforo de estudiantes y no en el subforo del tema 5. Justamente, la prevención de sucesos de este tipo forma parte de la motivación del desarrollo del sistema de agrupación automática y colocación de mensajes nuevos.

6.3. Colocación de mensajes nuevos

El tercer programa es el encargado de generar recomendaciones de colocación basadas en similitud. El programa toma como argumentos el fichero del foro y otro fichero con el texto del mensaje nuevo. Esta interfaz es bastante genérica, porque venga de donde venga el mensaje nuevo, siempre podrá volcarse en un archivo temporal en el momento de la ejecución y borrarse después. El programa cuenta con opciones para ajustar la representación de texto y la función de similitud. La salida del programa es un *ranking* de subforos ordenado por distancia mínima y, por si ocurriesen empates, por distancia media en segundo lugar. Además, la salida incluye el texto del mensaje nuevo y el del mensaje más próximo en todo el foro. El programa también dispone de un modo interactivo que permite ver el mensaje más próximo en cada subforo.

La figura 6.3 muestra un par de ejemplos de ejecución. En el primero, se pretende colocar un mensaje referido a problemas con una descarga en el foro de Psicofarmacología usando una vez más la representación USE y la distancia angular. Salvo en casos de empate, la elección de la distancia coseno o angular no altera la clasificación, puesto que una es función creciente de la otra. Como se pudo ver en la tabla 5.11 y la figura 5.6, el subforo de cuestiones generales de la asignatura tiene al menos 5 mensajes referidos a problemas con una descarga. No sorprende entonces que dicho subforo ocupe, con cierta holgura, el primer lugar en el *ranking* de subforos más similares semánticamente, y constituya pues la recomendación de colocación. El mensaje más próximo encontrado en ese subforo alude, por supuesto, a problemas con la descarga. En cambio, el más cercano en el segundo subforo del *ranking*, que se consulta mediante el modo interactivo, es una solicitud de adhesión al grupo de WhatsApp, o sea, guarda poca relación.

En el segundo ejemplo, el mensaje pregunta por la sesión de control en el foro de Procesadores del Lenguaje de 2020–21. Como se vio en la demostración del programa de exploración, en la figura 6.1, este tipo de mensaje es típico de los subforos de grupos de tutoría, que están

```
$ TMPFILE=$(mktemp); echo "Tengo problemas para descargarme el archivo." > $TMPFILE; python
↪  unedforums/scripts/coloca.py -r use --angular -i data/6201302-_2019_cl.csv $TMPFILE; rm $TMPFILE
Distancia                                                      mín  promedio
idSubforo Subforo
A         Cuestiones generales de la asignatura               0.636335  0.932717
B         Foro de estudiantes (Cafetería)                     0.849156  0.918977
C         Foro Tema 3 - Capítulo 9 del libro de Stahl         0.874644  0.993748
D         Foro Tema 1 - Capítulos 4 y 5 del libro de Stahl    0.895361  0.999274
E         Foro Tema 2 - Capítulos 6, 7 y 8 del libro de Stahl 0.908716  0.990805
F         Foro Tema 4 - Capítulo 11 del libro de Stahl        0.915112  0.968861
G         Foro Tema 5 - Capítulos 12 y 13 del libro de Stahl  0.932322  0.986703
H         Coordinación tutorial                               0.941270  0.995378
I         Foro Tema 6 - Capítulo 14 del libro de Stahl        0.996223  1.014161

Mensaje nuevo:
Tengo problemas para descargarme el archivo.

Mensaje más próximo en el subforo Cuestiones generales de la asignatura:
Buenas tardes,<br/><br/>Yo tampoco puedo descargar el archivo.<br/><br/>Un saludo,<br/><br/>[redacted].<br/><br/>

Selecciona un subforo (p. ej., "A", "B", ...): B
Mensaje más próximo en el subforo Foro de estudiantes (Cafetería):
Si me podeis añadir, que lo acabo de ver: [redacted]<br/><br/>Gracias.<br/><br/>

$ TMPFILE=$(mktemp); echo "¿Cuándo tenemos la sesión de control?" > $TMPFILE; python
↪  unedforums/scripts/coloca.py -t "" -r use data/ForosPL1_20-21.csv -i $TMPFILE; rm $TMPFILE
Distancia                                                          mín  promedio
idSubforo Subforo
A         Grupo de Tutoria 7                                   0.547584  0.639590
B         Grupo de Tutoria 19                                  0.578697  0.828659
C         Grupo de Tutoria 11                                  0.622777  0.632370
D         Grupo de Tutoria 25                                  0.645616  0.684800
E         Grupo de Tutoria 16                                  0.655022  0.882816
F         Grupo de Tutoria 23                                  0.657689  0.836879
G         Grupo de Tutoria 6                                   0.657700  0.790161
H         Grupo de Tutoria 1                                   0.662916  0.744861
I         Foro de consultas generales                          0.692389  0.887738
J         Foro General Práctica                                0.699065  0.951780
K         Grupo de Tutoria 2                                   0.704249  0.811940
L         Grupo de Tutoria 12                                  0.720374  0.791605
M         Grupo de Tutoria 9                                   0.724068  0.836336
N         Grupo de Tutoria 4                                   0.742720  0.803295
O         Foro de estudiantes (no moderado por el Equipo Docente)  0.807164  0.825413
P         Grupo de Tutoria 21                                  0.818500  0.818500
Q         Foro Análisis léxico (fuera de la práctica)          0.826558  0.945025
R         Foro Práctica: análisis sintáctico                   0.831208  0.948186
S         Grupo de Tutoria 15                                  0.839028  0.915572
T         Foro Práctica: análisis léxico                       0.854577  0.947340
U         Grupo de Tutoria 14                                  0.855246  0.925850
V         Grupo de Tutoria 20                                  0.862791  0.862791
W         Coordinación tutorial                                0.866843  0.907519
X         Grupo de Tutoria 10                                  0.894868  0.894868
Y         Foro Análisis sintáctico (fuera de la práctica)      0.956152  0.988994

Mensaje nuevo:
¿Cuándo tenemos la sesión de control?

Mensaje más próximo en el subforo Grupo de Tutoria 7:
Buenas, estaba interesado en saber si han resuelto su duda.<br/>Necesitaria saber cuando y como va ser la la
↪  sesión de control<br/>obligatoria. Muchas gracias<br/><br/>

Selecciona un subforo (p. ej., "A", "B", ...): B
Mensaje más próximo en el subforo Grupo de Tutoria 19:
09/12/2020<br/>Hola.<br/><br/>Me apunto a esta sesión de control.<br/><br/>Un saludo.<br/><br/>

Selecciona un subforo (p. ej., "A", "B", ...): H
Mensaje más próximo en el subforo Grupo de Tutoria 1:
Podéis realizar la sesión de control en el Centro que<br/>mejor os venga, siempre y cuando el Centro os
↪  acepte.<br/><br/>Saludos<br/><br/>
```

Figura 6.3: Salida del programa de colocación de mensajes para mensajes nuevos en los foros de Psicofarmacología y de Procesadores del Lenguaje de 2020–21.

excluidos por defecto en los programas de agrupamiento y colocación. No obstante, se puede ajustar aquí la opción relativa a la identificación de los grupos de tutoría, haciéndola nula para desactivar el descarte. Así, las 8 primeras posiciones en el *ranking* de proximidad corresponden a grupos de tutoría. Las distancias mínimas a los 2 subforos más cercanos son parecidas, y los mensajes más próximos en estos dos subforos preguntan por o mencionan la fecha de la sesión de control. Bajando hasta el octavo puesto en la clasificación, el mensaje del grupo de tutoría 1 también trata sobre la sesión de control, pero en relación con el centro donde se tiene que realizar, no con la fecha. El número de subforos de grupos de tutoría similarmente próximos a este mensaje de prueba ilustra el problema con los grupos de tutoría mencionado varias veces. La división en grupos de tutoría no suele tener que ver con la temática, sino con la geografía, que no siempre se explicita en los mensajes. Por tanto, sería difícil reproducir esas divisiones automáticamente o recomendar con certeza un grupo de tutoría sobre otro para la colocación de un mensaje nuevo.

Capítulo 7

Conclusiones

En este trabajo, se ha desarrollado un sistema de agrupación automática y colocación de mensajes nuevos para los foros de la UNED u otros foros de discusión con el mismo formato. La motivación ha sido facilitar, y en la medida de lo posible evitar, las tareas de mantenimiento derivadas de la colocación de mensajes en subforos equivocados por parte de los usuarios. El problema se ha abordado con un enfoque basado en similitud semántica y *clustering*, en parte porque los ejemplos disponibles podrían quedar cortos para entrenar clasificadores, y también para retener generalidad de cara a un eventual uso con datos no etiquetados. No obstante, la estructura temática original de los 7 foros analizados ha servido para comparar agrupamientos cualitativamente y calcular medidas externas con fines informativos.

El trabajo ha tenido un marcado carácter prospectivo, y buena parte del tiempo se ha dedicado a ensayar diferentes técnicas de representación de documentos y agrupación automática, para tratar de encontrar las que mejor se adecúen a las tareas y los foros concretos. Se ha experimentado con tres de las técnicas de representación de documentos más populares en la actualidad: la bolsa de palabras, con diversos tipos de compensación por frecuencia, los *embeddings* de palabras, similarmente compensados, y los *embeddings* de frases. Para los dos últimos, que tienen la ventaja de capturar al menos parcialmente la semántica, se ha recurrido a modelos preentrenados, ya que el tamaño del corpus es mucho menor que el requerido para entrenar esos modelos. Las distintas representaciones se han probado separadamente, antes de proceder al *clustering*, calculando medidas de validación interna sobre la estructura original de subforos. En ese punto se han usado el coeficiente de silueta y la consistencia de proximidad, definida pensando en la colocación de mensajes nuevos basada en máxima similitud.

Las siluetas de los agrupamientos originales han resultado ser pobres y en muchos casos negativas, independientemente de que se use la métrica coseno o la angular. Tampoco existía garantía de que los subforos originales fuesen compactos, estuviesen bien separados y, además, tuviesen forma globular, como haría falta para obtener buenas siluetas medias. Sin embargo, los resultados de consistencia de proximidad son bastante decentes, y alcanzan promedios del 71 % con la representación LogTF-IDF y el *embedding* de frases USE. Esto significa que en un

ejercicio de recolocación individual de mensajes de según la máxima similitud, esa proporción de mensajes recaería en el subforo original. Los resultados correspondientes a los *embeddings* de palabras son algo inferiores, probablemente por la falta de contextualización y la agregación más simple en comparación con los *embeddings* de frases. Dentro de los *embeddings* de palabras, se han conseguido puntuaciones ligeramente mejores con fastText que con word2vec, quizá debido al mejor comportamiento del primero frente a palabras desconocidas por su recurso a *n*-gramas. Entre los *embeddings* de frases, el modelo multilingüe preentrenado USE ha arrojado mejores resultados que el modelo SBERT análogo, y tiene la ventaja de no tener límite estricto sobre el tamaño del mensaje y de correr mucho más rápido.

La exploración de técnicas de agrupamiento también ha incluido tres tipos de algoritmo: partitivo (*k*-medias), jerárquico aglomerativo y jerárquico basado en densidad (HDBSCAN). Los agrupamientos se han evaluado utilizando las medidas internas ya mencionadas, así como el índice de validación basado en densidad, sólo para HDBSCAN, y, como medidas externas, el índice de Rand ajustado y la información mutua ajustada. La mejor correspondencia con la estructura original de subforos se ha obtenido con el algoritmo *k*-medias, combinado con el *embedding* USE y con número de grupos igual o ligeramente menor que el original, según la medida externa contemplada. En la mayoría de casos, la AMI ha resultado mayor que el ARI, seguramente por el tamaño heterogéneo de los subforos de referencia, que hace que la AMI se considere más idónea.

Volviendo a las medidas internas, los mayores coeficientes de silueta con número de grupos similar al original se han conseguido con el algoritmo aglomerativo con enlace promedio y representación fastText-TF. Esos agrupamientos tienen además valores de consistencia muy altos, pero su inspección ha revelado que esto se debe a una concentración extrema de mensajes en un grupo principal. Más razonables y mejor alineados con los subforos originales han resultado los agrupamientos obtenidos con enlace promedio y representación LogTF-IDF y con enlace completo y *embedding* USE. Una virtud del algoritmo aglomerativo es que construye un árbol que permite estudiar de manera eficiente la calidad del agrupamiento en función del número de grupos. Así se ha comprobado que la silueta y las medidas externas evolucionan de forma bien distinta y sus máximos no coinciden ni siquiera aproximadamente. Conviene recordar que los agrupamientos originales son temáticos, tienen siluetas pequeñas o negativas, y que, aunque su consistencia de proximidad haya resultado ser decente, no debería esperarse una correspondencia rigurosa con agrupamientos basados en similitud semántica.

Si bien los algoritmos *k*-medias y aglomerativo han demostrado tener virtudes, HDBSCAN se ha destacado para propósito general, con número de grupos desconocido, por su flexibilidad, robustez y el carácter intuitivo de sus parámetros. El algoritmo es capaz de identificar grupos con densidad y forma arbitraria e integrarlos en una jerarquía, descartando por el camino los puntos reconocidos como ruido. Los promedios de medidas externas obtenidos con HDBSCAN han quedado por debajo de los alcanzados con *k*-means y el algoritmo aglomerativo, pero hay que tener en cuenta que los primeros han sido atenuados por la cobertura, que es la fracción

complementaria al ruido, y los segundos han sido generados con número de grupos conocido. Al igual que con los otros dos algoritmos, los mejores promedios de AMI se han obtenido en combinación con el *embedding* USE, y tampoco se ha notado mucha diferencia al sustituir la métrica coseno por la angular, a pesar de las ventajas teóricas de la angular. No obstante, los máximos de silueta y DBCV se han conseguido con la representación fastText-TF y, aunque ciertos grupos de esos agrupamientos tengan pleno sentido, están globalmente peor alineados con la estructura original de subforos.

A lo largo del desarrollo del componente de agrupación automática, además de analizarse las medidas de validación, se han inspeccionado en detalle varios agrupamientos, examinando sus matrices de confusión e incluso los textos de mensajes seleccionados. De esta manera, se ha comprobado que los grupos formados pueden ser perfectamente válidos desde el punto de vista semántico, aun cuando sean transversales a la estructura temática original. Por ejemplo, se han visto grupos formados íntegramente por mensajes de agradecimiento, o por imágenes, procedentes de subforos muy diversos. Los mensajes de este tipo dejan clara, una vez más, la imposibilidad de alcanzar alineamientos precisos entre los agrupamientos basados en similitud semántica y los subforos temáticos originales. A la vez, se han encontrado múltiples instancias de grupos con mensajes provenientes de un único subforo, como las peticiones de adhesión a grupos de Telegram o WhatsApp, en subforos de estudiantes, o los problemas de descarga, en un subforo de cuestiones generales. En ciertos casos el grupo abarca la práctica totalidad del subforo, mientras que en otros se trata de subconjunto pequeño centrado en algún particular. En un caso verdaderamente interesante, el algoritmo HDBSCAN, operando con el *embedding* USE, ha detectado dos grupos puros formados por mensajes del subforo del primer tema de Psicofarmacología, uno centrado en la esquizofrenia y el otro en los antipsicóticos. Estos dos grupos han resultado ser parientes cercanos en el árbol de agrupamiento, con otros familiares en común que mencionan ambos asuntos, que son precisamente el tema de ese subforo. Este ejemplo ha puesto de manifiesto el funcionamiento adecuado del algoritmo y la utilidad de la estructura jerárquica para la interpretación y el análisis del agrupamiento.

Para facilitar la investigación llevada a cabo aquí y cualquiera futura sobre foros similares, así como para servir como puntos de entrada del sistema y aplicaciones de referencia, se han preparado tres programas: uno para la exploración de foros, otro para la agrupación automática y un último para asistir en la colocación de mensajes nuevos mediante un *ranking* de similitud. Los programas integran los diversos componentes del sistema en *pipelines* configurables desde la línea de comandos y podrían adaptarse sin mucha dificultad para prestar esa funcionalidad dentro de un sistema en línea. Se han proporcionado varios ejemplos de ejecución, incluyendo casos en los que se preservan los subforos de grupos de tutoría. Dichos subforos se han dejado al margen de análisis principal porque sus divisiones están más relacionadas con la geografía que con la temática, como se ha podido verificar en esos ejemplos de ejecución.

La variedad de técnicas de representación de documentos y *clustering* ensayadas aquí y los detallados análisis llevados a cabo sobre sus resultados no impiden poder recomendar algunas

líneas de investigación interesantes para futuros trabajos. Una obvia sería extender el estudio a un conjunto mayor de foros de la UNED, que incluyese múltiples instancias anuales de cada asignatura, como sucede aquí con la de Procesadores del Lenguaje. Esto permitiría observar con mayor nitidez diferencias sistemáticas en el comportamiento del sistema frente a distintos tipos de foro. Asimismo, con un corpus más grande podría empezar a plantearse, no entrenar los *embeddings* sobre el corpus, pero quizá aplicar un ajuste fino cuando el modelo lo permita. Otra investigación interesante sería probar a incluir una etapa de reducción dimensional previa al *clustering*, por ejemplo UMAP, que se utiliza en varios paquetes de modelización de temas. Aquí no se ha hecho porque añade hiperparámetros y no ha resultado esencial desde un punto de vista computacional, pero quizá ayudaría a mejorar la calidad de los agrupamientos y, desde luego, facilitaría su examen visual. Una última sugerencia sería profundizar en la relación entre la agrupación automática y la colocación de mensajes nuevos. Aquí, las recomendaciones de colocación se han basado directamente en la máxima similitud semántica con los mensajes de cada grupo. Los resultados de consistencia de proximidad para los agrupamientos generados de forma automática han confirmado que ésta es una opción razonable. Una alternativa sería basarlas en medidas internas, aunque las probadas aquí han arrojado resultados cuestionables sobre los agrupamientos y, además, tendría mayor coste computacional. Otra posibilidad sería usar procedimientos de colocación específicamente alineados con los algoritmos de *clustering*. Por ejemplo, en agrupamientos creados con k-medias se podría asignar el grupo con centroide más cercano, mientras que en los creados con HDBSCAN se podría usar la predicción basada en el árbol condensado, tal como viene implementada en la librería utilizada en este trabajo. Sería interesante probar algunas de estas alternativas y comparar sus resultados con los de la asignación basada en máxima similitud semántica.

Bibliografía

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., y otros (2016). TensorFlow: Large-scale machine learning on heterogeneous distributed systems. *arXiv:1603.04467*.

Aggarwal, C. C. y Zhai, C. (2012). *Mining Text Data*, capítulo A Survey of Text Clustering Algorithms, pp. 77–128. Springer US, Boston, MA.

Almeida, A. y Bilbao, A. (2018). Spanish 3B words word2vec embeddings (versión 1.0). Conjunto de datos doi:10.5281/zenodo.1155474.

Angelov, D. (2020). Top2Vec: Distributed representations of topics. *arXiv:2008.09470*.

Bilbro, R., Ojeda, T., y Bengfort, B. (2018). *Applied Text Analysis with Python*. O'Reilly Media.

Bird, S., Klein, E., y Loper, E. (2009). *Natural Language Processing with Python*. O'Reilly Media, Sebastopol, CA, USA.

Bojanowski, P., Grave, E., Joulin, A., y Mikolov, T. (2017). Enriching word vectors with subword information. *arXiv:1705.04009*.

Campello, R. J. G. B., Moulavi, D., y Sander, J. (2013). Density-based clustering based on hierarchical density estimates. En Pei, J., Tseng, V. S., Cao, L., Motoda, H., y Xu, G., editores, *Advances in Knowledge Discovery and Data Mining*, pp. 160–172, Berlin, Heidelberg. Springer Berlin Heidelberg.

Cer, D., Yang, Y., yi Kong, S., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., Sung, Y.-H., Strope, B., y Kurzweil, R. (2018). Universal Sentence Encoder. *arXiv:1803.11175*.

Goyvaerts, J. y Levithan, S. (2012). *Regular Expressions Cookbook*. O'Reilly Media, 2.ª edición.

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., y Mikolov, T. (2018). Learning word vectors for 157 languages. En *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. ELRA.

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

Han, J., Kamber, M., y Pei, J. (2012). *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers, Waltham, MA, USA, 3.ª edición.

Jurafsky, D. y Martin, J. H. (2023). *Speech and Language Processing*. Borrador 3.ª edición.

Le, Q. V. y Mikolov, T. (2014). Distributed representations of sentences and documents. *arXiv:1405.4053*.

McInnes, L. y Healy, J. (2017). Accelerated hierarchical density based clustering. En *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, pp. 33–42.

McInnes, L., Healy, J., y Astels, S. (2017). hdbscan: Hierarchical density based clustering. *Journal of Open Source Software*, 2(11):205.

McInnes, L., Healy, J., y Melville, J. (2020). UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv:1802.03426*.

Mikolov, T., Chen, K., Corrado, G., y Dean, J. (2013a). Efficient estimation of word representations in vector space. *arXiv:1301.3781*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., y Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. *arXiv:1310.4546*.

Moulavi, D., Jaskowiak, P. A., Campello, R. J. G. B., Zimek, A., y Sander, J. (2014). Density-based clustering validation. En *Proceedings of the 2014 SIAM International Conference on Data Mining (SDM)*, pp. 839–847.

Onan, A. (2019). Two-stage topic extraction model for bibliometric data analysis based on word embeddings and clustering. *IEEE Access*, 7:145614–145633.

Onan, A. y Toçoğlu, M. A. (2021). Weighted word embeddings and clustering-based identification of question topics in MOOC discussion forum posts. *Computer Applications in Engineering Education*, 29(4):675–689.

Patra, B. (2017). A survey of community question answering. *arXiv:1705.04009*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., y Édouard Duchesnay (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(85):2825–2830.

Řehůřek, R. y Sojka, P. (2010). Software framework for topic modelling with large corpora. En *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pp. 45–50, Valletta, Malta. ELRA.

Reimers, N. y Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using siamese BERT-networks. *arXiv:1908.10084*.

Reimers, N. y Gurevych, I. (2020). Making monolingual sentence embeddings multilingual using knowledge distillation. *arXiv:2004.09813*.

Romano, S., Vinh, N. X., Bailey, J., y Verspoor, K. (2016). Adjusting for chance clustering comparison measures. *Journal of Machine Learning Research*, 17(134):1–32.

Stahl, S. M., Muntner, N., y García de León, A. (2014). *Psicofarmacología Esencial de Stahl: Bases Neurocientíficas y Aplicaciones Prácticas*. UNED, Madrid, 4.ª edición.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., y Polosukhin, I. (2017). Attention is all you need. En Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., y Garnett, R., editores, *Advances in Neural Information Processing Systems*, volumen 30. Curran Associates, Inc.

Vinh, N. X., Epps, J., y Bailey, J. (2010). Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research*, 11(95):2837–2854.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., y otros (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272.

Yang, Y., Cer, D., Ahmad, A., Guo, M., Law, J., Constant, N., Abrego, G. H., Yuan, S., Tar, C., Sung, Y.-H., Strope, B., y Kurzweil, R. (2019). Multilingual universal sentence encoder for semantic retrieval. *arXiv:1907.04307*.

Nomenclatura

AMI	Información mutua ajustada (*Adjusted Mutual Information*), página 14
ARI	Índice de Rand ajustado (*Adjusted Rand Index*), página 14
BOW	Bolsa de palabras (*Bag Of Words*), página 14
CBOW	Bolsa de palabras continua (*Continuous Bag Of Words*), página 35
CNN	Red neuronal convolucional (*Convolutional Neural Network*), página 14
CSV	Valores separados por comas (*Comma Separated Values*), página 23
DBCV	Índice de validación de agrupamientos basado en densidad (*Density-Based Cluster Validation*), página 15
DIANA	Agrupamiento de análisis divisivo (*DIvisive ANAlysis clustering*), página 14
GIF	Formato de Intercambio de Gráficos (*Graphics Interchange Format*), página 58
HDBSCAN	Agrupamiento espacial jerárquico basado en densidad de aplicaciones con ruido (*Hierarchal Density-Based Spatial Clustering of Applications with Noise*), página 15
HTML	Lenguaje de marcado de hipertexto (*HyperText Markup Language*), página 23
IDF	Frecuencia inversa de documento (*Inverse Document Frequency*), página 14
LogTF-IDF	Frecuencia de término logarítmica compensada por frecuencia inversa de documento (*Log Term Frequency-Inverse Document Frequency*), página 33
LSTM	Memoria larga a corto plazo (*Long Short-Term Memory*), página 14
MOOC	Curso virtual abierto masivo *Massive Open Online Course*, página 14
NLP	Procesamiento de lenguaje natural (*Natural Language Processing*), página 13
NLTK	Caja de herramientas de lenguaje natural (*Natural Language Toolkit*), página 24
NMI	Información mutua normalizada (*Normalized Mutual Information*), página 14

OH Codificación binaria (*One-Hot*), página 33

SBERT *Embedding* de frases utiliza redes BERT siamesas (*Sentence-Bidirectional Encoder Representations from Transformers*), página 15

SOM Mapa autoorganizado (*Self-Organizing Map*), página 14

TF Frecuencia de término (*Term Frequency*), página 33

TF-IDF Frecuencia de término compensada por frecuencia inversa de documento (*Term Frequency-Inverse Document Frequency*), página 32

UMAP Aproximación uniforme y proyección a variedad para reducción dimensional (*Uniform Manifold Approximation and Projection*), página 18

URL Localizador de recursos uniforme (*Uniform Resource Locator*), página 24

USE Codificador de Frases Universal (*Universal Sentence Encoder*), página 15