UNIVERSIDAD NACIONAL DE COLOMBIA

Agrupación de textos cortos para el análisis de temas latentes de investigación en un conjunto de datos de proyectos de investigación

Jorge Mario Carrasco Ortiz

Universidad Nacional de Colombia
Facultad de Ingeniería
Departamento de Ingeniería de Sistemas e Industrial
Bogotá, D. C., Colombia

2017

Agrupación de textos cortos para el análisis de temas latentes de investigación en un conjunto de datos de proyectos de investigación.

Jorge Mario Carrasco Ortiz

Tesis o trabajo de grado presentado como requisito parcial para optar al título de:
Magíster en Ingeniería de Sistemas y Computación

Director:
Ph. D., Fabio Augusto Gonzáles Osorio
Codirectora:
Ph. D., Jenny Marcela Sánchez Torres.

Línea de investigación:
Aprendizaje computacional

Universidad Nacional de Colombia
Facultad de Ingeniería
Departamento de Ingeniería de Sistemas e Industrial
Bogotá, D. C., Colombia
2017

A mis padres,
por haberme brindado apoyo incondicional y por haber soportado todos mis sueños, desesperanzas e ilusiones.

Agradecimientos

Agradezco a mis padres por ser un grandioso ejemplo, a mis hermanos y amigos por su colaboración, por su firme apoyo y por sus consejos en momentos de dificultad.

Doy gracias a mis tutores, el profesor Fabio Augusto González y la profesora Jenny Marcela Sánchez Torres, por orientarme en este largo camino, por su paciencia y dedicación en esta investigación.

Adicionalmente agradezco a los integrantes del Grupo MindLab por la compañía y las discusiones que coadyuvaron en la solución del problema de investigación. También a todas las personas que directa o indirectamente aportaron en el desarrollo de este trabajo.

Finalmente, expreso mi agradecimiento a la Universidad Nacional de Colombia, Sede Bogotá, por darme una vez más la oportunidad para seguirme preparando, como ser humano y como profesional; junto con sus profesores ha hecho la persona que soy.

Resumen

Los documentos de texto son una fuente importante de datos para las técnicas de minería. Normalmente, las bases de datos de texto incluyen documentos suficientemente largos para aplicar técnicas de minería de texto convencionales. Sin embargo, en algunas tareas, como el proceso de identificación de áreas de investigación, se cuenta con bases de datos de textos muy cortos, lo cual representa un desafío para las técnicas convencionales de minería de texto. El problema tiene que ver con el pequeño número de términos que no proporcionan suficiente información estadística para encontrar cualquier tipo de relación entre los documentos de la colección. El objetivo principal de este trabajo es mostrar cómo generar grupos temáticos utilizando solo los títulos de proyectos de investigación de una institución de educación superior.

En esta tesis presentamos un método para agrupar colecciones de textos cortos a partir de representaciones distribucionales de términos. El método utiliza una colección de referencia de textos con mayor extensión, para encontrar una representación distribucional de términos (DTR, por sus siglas en inglés) que codifica relaciones semánticas y sintácticas entre términos. Estas representaciones son utilizadas posteriormente para mejorar los algoritmos de agrupación. Igualmente, exploramos diferentes estrategias para la representación de términos, así como varias estrategias para la agrupación.

El método se evaluó en dos conjuntos de datos. El primero fue construido para este estudio y está compuesto de títulos de artículos científicos, el segundo conjunto de datos corresponde a los títulos de proyectos de investigación de una institución de educación superior. Los resultados fueron evaluados utilizando cuatro medidas extrínsecas (Homogeneity Score, V-measure, Adjusted MI, Pureza) para el primer conjunto de datos, y tres medidas intrínsecas (Davies-Bouldin, QError, Slihouette) para el segundo conjunto de datos. Los resultados muestran que la estrategia de representación distribucional de términos, mejora en gran medida la calidad de las agrupaciones generadas cuando se compara con la producida por las estrategias convencionales de agrupamiento de texto.

Palabras clave: agrupación, textos cortos, representación distribucional de términos, kernel k-medias, NMF, word2Vec, Scopus, ScienceDirect, recuperación de información.

Abstract

Text documents are an important source of data for tech mining techniques. Usually, text databases include documents sufficiently long to apply conventional text mining techniques. However, for some tech mining tasks, such as capabilities identification process, the databases available are comprised of very short texts, which represents a challenge for conventional text mining techniques. The problem in question is that the small number of terms fail to provide enough statistical information to find any kind of relationship among the documents in the collection. The main purpose of this work is to show how to generate thematic clusters

by using only the titles of research projects from a higher education institution.
In this thesis we present a method for clustering very-short-text collections based on distributional text representations. The method uses a reference collection of large texts to find a distributional term representation (DTR) that encodes semantic and syntactic relationships among terms. The DTR is used to represent the very-short texts which are fed to a clustering algorithm. Likewise, we explore different strategies for distributional term representation as well as for clustering.
The method was evaluated in two datasets. The first one was assembled for this study and is composed of scientific paper titles, and the second one corresponds to the titles of a set of research projects from a higher education institution. The results were evaluated by using four extrinsic measures (Homogeneity Score, V-measure, Adjusted MI, Purity) for the first dataset, and three intrinsic measures (Davies-Bouldin, QError, Slihouette) for the second dataset. The results show that the distributional term representation strategy greatly improves the quality of the generated clusterings when compared to the one produced by conventional text clustering strategies.
Keywords: clustering, short texts, distributional term representation, kernel k-means, NMF, word2Vec, Scopus, ScienceDirect, information retrival.

Índice de figuras

Índice de tablas

Contenido

1. Introducción

1.1. Justificación e identificación del problema

Los métodos de agrupamiento son algoritmos que, en general, responden a un problema de aprendizaje no supervisado. En este tipo de problemas se parte de un conjunto de datos para así conformar grupos, con la particularidad de que al interior de estos grupos los objetos que lo conforman tienen la mayor similitud posible (es decir: tienen características similares, responden de manera similar a cierto estímulo, tratan del mismo tema, y tienen similar significado semántico, etc)[35]. En particular, el principal objetivo perseguido por el agrupamiento o *clustering* de textos es encontrar automáticamente, a partir de una colección de documentos, grupos de documentos con contenido semántico similar. Esto se logra, en general, teniendo en cuenta la heterogeneidad de la distribución de cada palabra en cada documentos analizado, para lo cual se generan representaciones de los textos que permiten determinar la relevancia de cada uno de los términos en la colección de documentos [8].

Un corpus corto, es una colección de documentos que se compone, de textos que tienen muy pocos términos. Dicha frecuencia término-documento y la longitud promedio de los textos afecta, directamente las medidas de similitud entre estos y, por ende, los resultados obtenidos por los métodos de agrupamiento [33]. En el estado de arte se encontró, que algunas colecciones usadas con frecuencia para la validación de los métodos de agrupamiento en colecciones de textos denominados como "cortos" son: EasyAbstracts ($\bar{T}_d = 192{,}93$), SEPLN-CICLing ($\bar{T}_d = 65{,}48$), CICLing-2002 ($\bar{T}_d = 70{,}45$), R4-Reuters ($\bar{T}_d = 166{,}4$), R8-Reuters ($\bar{T}_d = 64{,}87$), R52-Reuters ($\bar{T}_d = 64{,}3$), WebKb ($\bar{T}_d = 136{,}26$) y hep-exCERN ($\bar{T}_d = 46{,}53$) [7][33].

El problema con los enfoques tradicionales de agrupación de textos y, en particular, cuando se está analizando la agrupación para conjuntos de datos de longitud corta es la limitación de los métodos para extraer la representación semántica de los documentos [7]. Sin embargo, existen algunos algoritmos que se han venido desarrollando para poder solventar esto, por ejemplo, K-Means, HSCLUST, el método de propagación de afinidad [7].

En esta última década, debido al crecimiento que ha tenido la información textual y gracias también, en gran medida, al crecimiento de internet y las redes sociales, muchos investigadores han centrado esfuerzos en las técnicas de agrupamiento sobre textos cortos [7][52][8][14],

a causa de que este tipo de información es cada vez más usual en internet. Entre las técnicas de agrupamiento no supervisado se destacan las siguientes: Naive Bayes [31], el algoritmo K-nearest neighbor [24], métodos de soporte vectoriales [53][31] y Particle Swarm Optimization [7]. Los métodos de clasificación anteriormente mencionados usan los denominados BoW[1] [14], que son representaciones vectoriales de frecuencias de ocurrencias de cada palabra en cada documento; estos vectores pueden generar representaciones congruentes cuando las cadenas de texto en estudio tienen gran longitud y el vocabulario usado en los objetos de análisis es muy "cercano".

Los métodos usales de agrupación presentan inconvenientes cuando se analizan textos cortos, debido principalmente a que las representaciones de este tipo de textos son más dispersas en comparación con las representaciones de textos de mayor extensión (informes, noticias, libros, artículos, etc.). Los textos cortos tienen representaciones dispersas a causa de la baja frecuencia de las palabras en los textos, lo que supone un vocabulario extenso con pocas ocurrencias de los términos que lo componen [25][6]. Es por esto que los métodos usuales de clasificación no supervisada de textos no son eficientes cuando se trata del análisis de textos cortos.

De esta manera, y como podrá verse más adelante, la representación distribucional es un procedimiento clave en el método de agrupamiento que se propone en esta investigación. Siguiendo esta metodología se busca, por medio de estadísticas de coocurrencia de documentos o términos, tener mayor información para la representación de cada término que se presente en el conjunto de análisis [19][1][36][32]. Con esta representación se puede tener una aproximación mucho más fiable de la semántica de los documentos, lo que genera directamente un mejor resultado en el proceso de agrupación del conjunto de datos que se quiere investigar [19].

Otras investigaciones en el área han llevado a la evaluación de los resultados de la agrupación, en estos estudios se consideran diferentes algoritmos para la tarea de agrupación (K-means, Singular value decomposition (SVD), graph-based approach, Hierarchical Agglomerative Clustering y Spectral Clustering), así como diferentes medidas de representación de los textos (Cosine Similarity, Latent Semantic Analysis, Short Text Vector Space Model, Kullback Leiber Distance) y diferentes medidas de evaluación de los grupos (F-measure, adjusted Rand Index y Indice V): [43][37].

Algunos ejemplos claves de textos cortos son: los resúmenes, los títulos y las descripciones. Como ya se mencionó con anterioridad, la agrupación en estos tipos de textos, presentan problemas, debido al bajo volumen de información que contienen. En detalle, los factores que generan inconvenientes en la agrupación de textos cortos son la falta de información o

[1]Siglas en inglés de bag of words, o "paquete de palabras".

representación dispersas [25][6], la superposición y redundancia, significados hiperespecializados de algunos términos, y la presencia de palabras dentro de un dominio muy específico de cada campo de conocimiento [35][7], son inconvenientes que hacen más difícil obtener un buen resultado en la agrupación de los textos.

Ahora bien, dentro del campo del aprendizaje computacional, los métodos de *clustering* o agrupamiento han adquirido mucha relevancia para la investigación científica, por cuanto su finalidad es la descripción de patrones de comportamiento similares entre los individuos de estudio. Este tipo de conclusion es útil en los análisis multivariados de información y puede dar luces sobre comportamientos particulares en las poblaciones de estudio, al igual que aportar en muchas áreas de investigación, como lo son: el aprendizaje automático, la minería de datos, el reconocimiento de patrones, el análisis de imágenes, la cienciometría, entre muchas otras.

El agrupamiento de textos ha sido relevante en procesos que involucran gran cantidad de información muy detallada y, en muchos casos específicos, en forma de texto escrito, sin embargo con el crecimiento de la información textual, los textos de extensión corta son cada vez más usuales. Entre las aplicaciones más destacadas para la agrupación de textos cortos, se encuentran: software como sistemas de extracción de información [7], análisis de sentimientos, marketing , seguridad nacional, la detección de temas latentes [54], indexación de grandes volúmenes de artículos en revistas científicas [7][13] para resolver problemas de atribución de autorías [1], entre muchas otras aplicaciones.

Así las cosas, esta investigación se centrará en el tema específico de la agrupación de textos cuando el volumen de información por unidad de análisis es muy reducido, es decir, cuando la extensión promedio de los textos es muy chica y, por ende, no es adecuado aplicar los enfoques tradicionales de agrupación de textos. Es un campo novedoso de investigación y en este caso particular serviría para agrupar las investigaciones que están en curso en la Universidad Nacional de Colombia usando solamente la descripción corta de las actividades de investigación, la cual tiene una extensión promedio de 20,98 términos por documento. Esto podría facilitar algún tipo de priorización en las políticas de delegación de recurso, al interior de la Universidad o podría servir como marco de referencia para posteriores investigaciones sobre el tema.

En las técnicas de agrupación de textos, como se ha descrito con anterioridad, se busca separar la información por grupos de acuerdo con la similitud de los objetos y, en general, esta característica es medida con una función de similitud [54][28][27], que en las metodologías usuales de agrupación de texto son útiles cuando las cadenas por analizar tienen diferentes niveles de granularidad; por ejemplo, los documentos, que tienen grandes volúmenes de palabras, párrafos y oraciones, lo cual hace algo más “fácil” el problema de clasificación en estos

tipos de texto.

En el caso específico de la agrupación con la información de la Vicerrectoría de Investigación de la Universidad Nacional de Colombia, se cuenta con descripciones muy cortas de cada tema, con textos que tienen una extensión promedio de 21 palabras, es decir, no se tiene suficiente información. Así los métodos de agrupación usuales no han dado los resultados esperados, en la medida en que se generan agrupaciones uno a uno, lo cual puede deberse a que muchas palabras usadas en las descripciones de los temas de investigación están dentro de un dominio muy específico de cada campo de conocimiento [33], y esto dificulta mucho poder hacer asociaciones entre los objetos de análisis (temas de investigación). Una de las aplicaciones del método que se presentará en esta investigación es justamente que partiendo de, los resultados obtenidos en la agrupación automática se puede determinar el tema latente que es común dentro de los elementos de cada grupo [54], así se podría establecer cuáles son las áreas de conocimiento o los temas específicos en los que se están centrando las investigaciones en la Universidad.

1.2. Objetivos

Partiendo de reconocer la particularidad de este conjunto de información, esta investigación busca responder a la siguiente pregunta: ¿Cómo agrupar textos cortos de manera que los grupos conformados reflejen la semántica de los textos analizados?. A partir de esta pregunta surgen otras más específicas como: ¿Cómo realizar correctamente la representación de textos que tienen una extensión de alrededor de 20 palabras? ¿Cuál es el método de agrupación más eficiente, en el problema de agrupación no supervisado de textos cortos? ¿Cómo se pueden detectar temas latentes partiendo de la agrupación de estos tipos de textos?. En el presente trabajo se quiere desarrollar un método alternativo de solución a este tipo de problemas usando los enfoques de representación distribucional. Dichos enfoques, por medio de estadísticas de coocurrencia de términos, permiten tener mayor información para la representación de cada término en el vocabulario o conjunto de análisis, para, a partir de esto, poder hacer más efectivos los métodos de agrupación.

OBJETIVO GENERAL: Desarrollar un método para el agrupamiento de textos cortos que sea aplicable al problema de detección de temas latentes en colecciones de descripciones cortas de proyectos de investigación.

OBJETIVOS ESPECÍFICOS:

Obj 1. Desarrollar una estrategia de representación para textos cortos que capture una mejor representación semántica de los textos y que facilite la tarea de agrupación.

Obj 2. Desarrollar un método de agrupamiento no supervisado que utilice la representación construida para un determinado conjunto de textos de extensión corta.

Obj 3. Evaluar el método desarrollado en un conjunto de datos concreto relacionado con la tarea de detección de temas latentes en una colección de descripciones cortas de proyectos de investigación de la Universidad Nacional de Colombia.

1.3. Metodología

Esta investigación es un estudio de tipo exploratorio, debido a que se prueban distintas estrategias para la representación de los textos de longitud corta, tal como las representaciones distribucionales de los términos y los diferentes métodos de agrupación. Adicionalmente, en el diseño de la investigación se usan datos cuantitativos construidos a partir de las bases de datos: UN Títulos, y a partir de la base Scopus Títulos (ver descripción en la sección 4.1). La investigación es de carácter experimental, dado que recopila evidencias de la validez de los métodos encontrados y de la propuesta metodológica construida para la resolución del problema de investigación.

Así, para alcanzar los objetivos antes mencionados, se definieron tres grandes fases, a continuación, se presenta una pequeña descripción de cada una de las fases del proyecto y una breve descripción de las actividades desarrolladas:

1. **Primera Fase (Incorporación de ampliaciones distribucional de los términos).** En esta fase inicial, se hizo la revisión de la literatura para la identificación de los métodos del estado del arte, que son utilizados en la representación de los textos de longitud corta, con el fin de poder integrar e implementar dichos métodos de base y poder proponer un nuevo esquema para la representación de los textos cortos que son materia de investigación.

   #### Actividades

   - Partiendo de la revisión bibliográfica, se establecieron los esquemas que se utilizaron, basados en fuentes externas, para obtener una representación más adecuada de las unidades de análisis.
   - Se implemento la representación de textos cortos y con base en esta implementación, ejecutar los métodos seleccionados de agrupación de textos, hacer la medición del rendimiento y la validez de los resultados obtenidos.
   - Se hizo la definición de un nuevo esquema y procedimientos que lograran mejorar los resultados de los algoritmos encontrados en la revisión bibliográfica.

2. **Segunda Fase (Comparación de métodos de Cluster).** En esta fase se evaluaron las medidas de validez interna y externa para cada uno de los métodos de agrupamiento, comparando los resultados y evaluando que tanto mejora el desempeño de los métodos cuando se usa el enfoque de ampliación distribucional y las técnicas de selección de términos, después se hicieron evaluaciones sobre el comportamiento de los métodos en términos del tiempo de ejecución.

   Actividades

   - Se hizo una revisión del estado del arte para encontrar los artículos más recientes que tratan sobre la agrupación de textos, en esta revisión bibliográfica se incluyeron aquellos que se especializan en textos de longitud corta.
   - Con base en la revisión bibliográfica previa, se seleccionaron y/o adaptaron los algoritmos más pertinentes para el conjunto de datos de problema, apoyándose en las medidas de validez interna/externa y en costos computacionales de los algoritmos.
   - Realizar la depuración y las consolidaciones de las fuentes de información externas (La base de datos UN Títulos y Scopus Títulos), implementar la experimentación, las medidas de evaluación de validez internas y externas, así como también las medidas de costo de ejecución y recursos computacionales.

3. **Tercera Fase (Evaluación del método en la base de datos UN Títulos)**. Una vez se finalizó la exploración de los métodos para el agrupamiento de textos cortos de la fase previa, trabajaremos con la base de datos depurada de la Universidad, se realizó la implementación del procedimiento seleccionado en un lenguaje de programación que facilitó el procesamiento de texto. Posteriormente se implementaron las validaciones internas de los grupos que arrojo el método desarrollado, así como las comparaciones con los otros métodos en estudio y finalmente algunas representaciones graficas que puedan ayudar a la interpretación de los resultados obtenidos en el agrupamiento.

   Actividades

   - Se procedió a la tarea de pre-procesamiento necesario para la limpieza de términos que puedan afectar los resultados de las técnicas de agrupamiento de la base de datos (Signos de puntuación, verbos, errores de codificación, etc.)
   - Se hizo una medición de validez de los resultados y el desempeño de los métodos que están en estudio.
   - Con base en las fuentes de información, insumo para las dos fases anteriores, se construyen representaciones gráficas para la interpretación de los resultados de la técnica de agrupamiento propuesta.

1.4. Resultados y contribuciones

Los principales resultados y contribuciones de este trabajo pueden ser resumidas como:

- Se desarrollo un metodología para la agrupación de texto con extensión muy corta, basado en la representación distribucional de los términos (DTR), utilizando una fuente de información externa, en nuestro caso, utilizando las bases de datos de Elsevier, logrando una mejora de alrededor del 70 % en los índices de validez externa de los grupos con relación a los métodos básicos de representación de textos, adicionalmente logrando una mejora relativa de al rededor del 10 % con relación a representaciones entrenadas con el popular método word2Vec.

  La figura **3-1** presenta la visión general del método desarrollado en esta investigación, presenta sus diferentes etapas y procesos. El método se basa en la idea detrás de la representación distribucional de los términos, como ya se ha mencionado, la idea es representar una palabra por medio de estadísticas de coocurrencia de documentos o términos, intuitivamente el método se base en buscar la mejor representación semantica partiendo de la búsqueda de relaciones entre los termino o documentos que existen en una colección externa de documentos, esta representación enriquecerá la semantica de los términos[6]. El método consta de cuatro etapas principales:

  1. El pre-procesamiento de los textos originales, lo cual consiste en la depuración y normalización de los términos de los textos.
  2. Partiendo de los textos depurados se construyen consultas, que se utilizan para la expansión de los términos, esto es, para encontrar elementos textuales de mayor extensión, por ejemplo, resúmenes de artículos, títulos, citas, y/u otros elementos que pueden estar asociados a la temática de los textos originales.
  3. Se extrae la representación distribucional de los textos a partir de los elementos recuperados en la etapa anterior, dicha representación se usa para ampliar el sentido semántico de los términos y poder encontrar relaciones entre los textos que componen la colección.
  4. Se aplican los diferentes métodos de agrupamiento sobre la colección de textos cortos después de aplicar la expansión de los términos.

- Carrasco, J. M., Sánchez-Torres, J. M. and Fabio A. González. Oral Presentation - "A very-short-text clustering method based on distributed representation to identify research capabilities of a Higher Education Institution" En la *6-th Global TechMining Conference* 2016. (Ver anexo C).

  En este trabajo, se propuso la utilización de varios métodos de representación distribucional para el problema de investigación, se realizo la comparación con los métodos

estándar de agrupación de textos. Las contribuciones en este trabajo incluyeron la participación en el desarrollo de los programas/códigos, la ejecución de los experimentos, la realización de presentación final del proyecto y una presentación oral del trabajo en el evento de la conferencia que se menciono con anterioridad.

- La mayor parte del proceso desarrollado está disponible en el siguiente repositorio web: Los códigos de los métodos presentados en el articulo "*A very-short-text clustering method based on distributed representation to identify research capabilities of a Higher Education Institution*" están disponibles en https://github.com/JoraJora/shortTextCluster.

1.5. Estructura del documento

La presente tesis está estructurada de la siguiente forma. El capítulo 2 detalla de manera general la teoría y los conceptos básicos involucrados en los métodos de agrupamiento de textos. Primero se describe el trabajo relacionado sobre la representación semántica de los textos; luego, en esta misma sección se explican de manera general los componentes y la definición del método de agrupación de textos, y se describe cuáles son la problemáticas identificadas en los diferentes métodos seleccionados para la experimentación; finalmente, se exponen las metodologías de evaluación del desempeño de los algoritmos. La discusión sobre estos temas y las descripciones de los métodos contextualizarán la aproximación metodológica propuesta, esto permitirá entender de mejor forma el contenido de este documento.

En el capítulo 3 se explica el esquema metodológico propuesto. Dicho esquema se divide en las siguientes secciones: la primera consiste en describir al detalle el procesamiento de los textos antes de encontrar la representación semántica propuesta para los términos que componen los textos; en la segunda sección se detalla el esquema de consulta en la fuente externa, así como la posterior representación y agrupación de los textos; finalmente, en la tercera se explicara la implementación de la metodología propuesta, describiendo los pasos y el diseño de las funciones creadas para este tarea.

Posteriormente en el capítulo 4 se relacionan los diferentes grupos de datos que se utilizarán en la evaluación del método propuesto, y se da una breve descripción de la configuración experimental haciendo referencia a los métodos descritos en el capítulo 2. Al final de esta sección del documento se presentan los resultados experimentales encontrados y se complementa con un análisis de estos. La finalidad de esta sección del documento es describir el desempeño de los métodos identificados en los conjuntos de datos seleccionados.

Finalmente en el capítulo 5 se presentan las conclusiones del presente trabajo y el planteamiento de algunas ideas de investigación para trabajos futuros.

2. Marco teórico

Esta sección tiene como objetivo introducir las definiciones y las nociones básicas de los métodos utilizados en el nuestro proyecto de investigación. Las definiciones y los temas presentados, se espera, contribuyan a en una mejor comprensión del lector de este documento. En la primera parte, sección 2.1, se presenta los trabajos relacionados con la agrupación de textos. En la sección 2.2 se presentan las diferentes formas de representar a los textos vía ampliación distribucional de los términos. Posteriormente, en la sección 2.3, se describen los conceptos y los métodos de la agrupación de textos, como son la extracción de características, métodos de aprendizaje y por último, en la sección 2.4, se presenta la forma en que se evalúa la tarea de agrupación, presentando, las diferentes medidas de validez internas y externas que serán utilizadas en la evaluación experimental.

2.1. Trabajos relacionados

La importancia de trabajar con textos cortos ha sido reconocida en estudios recientes, en particular, la mayoría de estos trabajos se apoya en nuevos métodos para resolver problemas de clasificación. La principal preocupación con estos enfoques es que requieren un número suficientemente grande de ejemplos de entrenamiento para lograr una alta precisión en textos no clasificados [42, 1]. La alternativa es utilizar una técnica no supervisada para textos cortos que probablemente mostrará resultados más eficientes, como un método de agrupación. En esta sección, por un lado, se enuncian los tres aspectos que se deben someter a revisión a la hora de evaluar un proceso de agrupación y, por otro, se ofrece una visión general de los enfoques encontrados en la literatura que proponen una solución para resolver el problema de la agrupación de textos de extensión corta.

Si planeamos usar un método de agrupación, se deben considerar tres aspectos relevantes: primero, se escoge la representación del documento o la manera de ponderar los términos dentro de los textos originales con suficiente información [52]. Ante el inconveniente de los textos de corta extensión, una posible solución es la expansión de los términos del documento, lo cual modifica el peso de estos, con el fin de incorporar la representación de los términos que no se presentan en el documento, como de poder contribuir a la representación semántica de los términos que realmente aparecen en el texto original [1, 20, 32, 36]. En la actualidad, una estrategia relevante en la expansión de documentos es utilizar las representaciones distribucionales de los términos que los componen. Esta estrategia trata de resolver

el problema de dispersión de las representaciones tradicionales (como *df* y *tf-idf*) y la baja frecuencia de aparición en las colecciones con textos de extensión corta. Este enfoque utiliza la frecuencias de los términos en el documento y la distribución de co-ocurrencia de cada término para generar su respectiva representación, para lo cual también se usa información contextual tomada de una fuente externa [26, 6, 19].

Otro método representativo que se encuentra en el estado del arte es el denominado Word2Vec, que fue introducido por Mikolov et al. [29, 30]. Esta estrategia propone una representación basada en redes neuronales en la que una palabra está representada por un vector (comúnmente conocido en inglés como *word embeding* o CBOW[1]) que capta las relaciones sintácticas y semánticas de palabras que componen los textos de entrenamiento. El modelo está formulado para predecir el contexto que rodea una determinada palabra. Los vectores y las representaciones construidos con esta arquitectura, o similares, son utilizados para resolver problemas de procesamiento de lenguaje natural: segmentación del usuario, representación del conocimiento, respuesta de preguntas, extracción de temas, entre otras tareas; recientemente, una extensión de este método es presentada por Bojanowski et al [5].

El segundo aspecto relevante en el proceso de agrupación es determinar la matriz de similitud o disimilitud que el algoritmo utilizará para encontrar las relaciones entre los documentos analizados. En la literatura encontramos muchas estrategias para crear la matriz: similaridad de coseno (CS) usando representación *tf-idf*, análisis semántico Latente (LSA), maquina de vectores de soporte (SVSM) y divergencia de Kullback-Leibler (KLD) [33, 43].

Finalmente, el tercer y último aspecto que se debe tener en cuenta es la selección del método más apropiado y su aplicación en el proceso de agrupación de textos cortos, de acuerdo con la representación seleccionada y la matriz de similitud que se definieron en las etapas anteriores [43]. Desde la última década, debido al crecimiento de la información de texto y el creciente uso de Internet y las redes sociales, muchos investigadores han centrado sus esfuerzos en técnicas de agrupamiento en textos cortos (Cagnina2014, Yuan2011, Carretero-Campos2013 y Faguo2010). Debido a que este tipo de información es cada vez más habitual en Internet, se mencionan las técnicas de agrupación supervisadas más destacadas: Naive Bayes [31], el algoritmo del vecino más cercano K [24], Support Vector Machines [53, 31], Optimización de Enjambre de Partículas [7], entre otros. Estos métodos de agrupación descritos anteriormente hacen uso de lo que se denomina BoW [14], que son representaciones de vector de frecuencia de ocurrencias de cada palabra en cada documento. Estos vectores pueden generar representaciones congruentes cuando las cadenas de texto en estudio tienen una amplia gama de longitudes y el vocabulario utilizado en los objetos de análisis es muy çercano".

El objetivo final de esta investigación es, así, agrupar textos cortos de tal manera que los

[1] Siglas en inglés de *continuos bag of words*

textos relacionados con cierta categoría puedan encontrarse en un grupo mismo grupo, lo que significa que la similitud dentro de los grupos de texto será mucho más fuerte. Para lograr este objetivo, y para la evaluación de las metodologías propuestas de agrupación (como la medición de la similitud dentro de los grupos), realizamos búsquedas en diferentes artículos para identificar posibles conjuntos de datos de referencia. En el estado de arte se encontró que algunas colecciones usadas con frecuencia para la validación de los métodos de agrupamiento en colecciones de textos denominados como "cortos" son: EasyAbstracts ($\bar{\mathcal{T}}_d = 192{,}93$), SEPLN-CICLing ($\bar{\mathcal{T}}_d = 65{,}48$), CICLing-2002 ($\bar{\mathcal{T}}_d = 70{,}45$), R4-Reuters ($\bar{\mathcal{T}}_d = 166{,}4$), R8-Reuters ($\bar{\mathcal{T}}_d = 64{,}87$), R52-Reuters ($\bar{\mathcal{T}}_d = 64{,}3$), WebKb ($\bar{\mathcal{T}}_d = 136{,}26$) y hep-exCERN ($\bar{\mathcal{T}}_d = 46{,}53$) [7, 33]. Por desgracia, el conjunto de datos que contienen los títulos de los proyectos de investigación no tiene las mismas características del conjunto de datos de referencia. En nuestro problema, el número promedio de términos en las colecciones es de $\bar{\mathcal{T}}_d = 20{,}91$, el cual resulta ser mucho menor que cualquier otro conjunto de datos de referencia.

2.2. Representaciones distribucionales de términos (DTR)

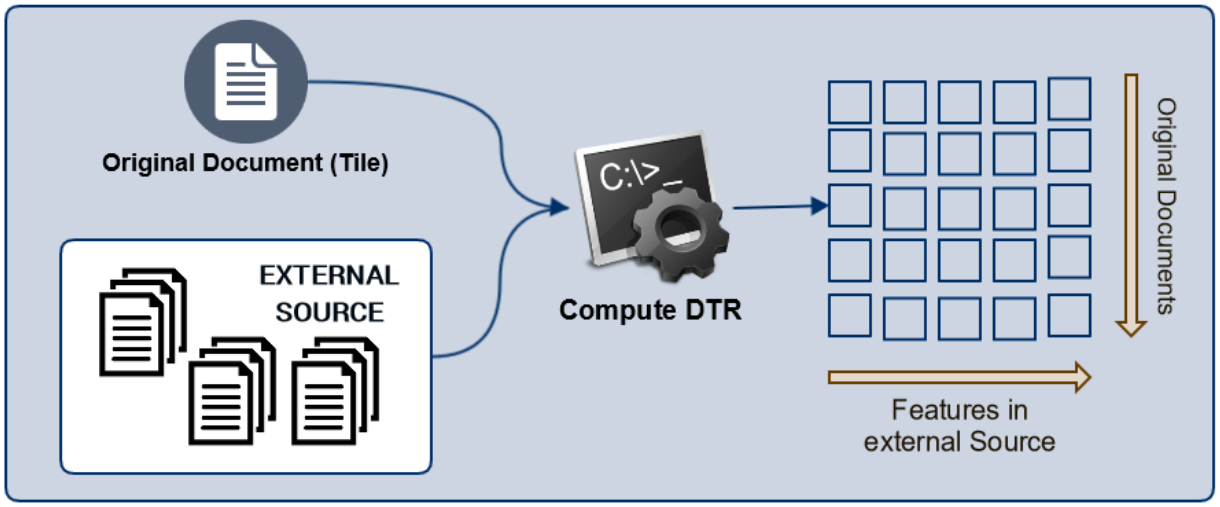


Figura 2-1.: Estrategia de representación distribucional de los términos.

Como podemos ver en la figura **2-1**, la idea principal de usar representaciones distribucionales de términos (DTR, por sus siglas en inglés) es construir un nuevo vector de representación numérica que pueda capturar el significado semántico de los documentos usando una fuente externa de información. Con base en esta fuente se emplea una función para la representación de los términos que incluye el documento; dicha función usualmente utiliza "bolsas de documentos" y/o "bolsas de palabras", las cuales aparecen simultáneamente en la fuente y en

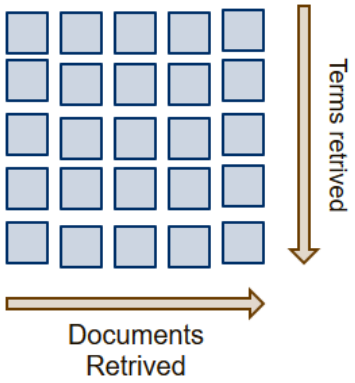
la representación original del documento que se quiere clasificar o en este caso agrupar [25]. Esta interacción entre las diferentes palabras y la fuente externa,al igual que las respectivas representaciones, es llamada usualmente en la literatura *semantic word embeddings*. Como explicaremos más adelante, DTR es una manera de expandir la representación semántica de los documentos, que consiste en encontrar principalmente w_{jk}, que se calcula con el texto de la fuente relacionada. Más formalmente, dados los pesos de los términos w_{jk}, la representación final de un documento d_i basado en una representación de una fuente externa estará dado por [6]:

$$d_i^{dtr} = \sum_{t_j \in d_i} \alpha_j w_{ji} \tag{2-1}$$

Donde α_j es un número real que mide la contribución del término $t_j \in d_i$ dentro de la representación del documento. Existen varias posibilidades para definir α_j. De esta manera, a lo largo de esta sección, se describirán en detalle los diferentes enfoques para la expansión distribucional de términos que fueron considerados en nuestro trabajo de investigación, también se describirán en detalle los esquemas probados de pesado para los términos (es decir encontrar $w_{t_j i}$).

2.2.1. Representación por ocurrencias de documentos

La representación de ocurrencias de documentos (comúnmente llamada DOR, por sus siglas en inglés) puede definirse como la representación dual de la tradicional medida $tf-idf$, la cual es reconocida en el contexto de minería de texto, como uno de los modelos de representación mas básicos y mas utilizados [6, 25]. La idea central de este enfoque es asumir que la semántica de un documento y, más específicamente, la semántica de los términos que componen este documento pueden ser representadas por la distribución estadística de las co-ocurrencias de documentos en el corpus que contienen dicho término, es decir, un documento estará representado por otros en los cuales los términos co-ocurren. Esto resulta tener sentido, a consecuencia de que los documentos en donde aparece frecuentemente un término pueden caracterizar la semántica de dicho término, así como también, si un documento tiene muchos términos diferentes con respecto a los de los otros documentos, o poco distintivos, menor será la contribución para la representación semántica del documento.
El objetivo de esta aproximación es encontrar el vector $\mathbf{w}$ de pesos asociados a cada uno de los términos t_j, formalmente $\mathbf{w}_j = (w_{j1}, \ldots, w_{Nj})$, en donde N es el número de documentos en la nueva colección, es decir, los documentos que fueron recuperados de una fuente externa de información (en nuestro caso, todos los resúmenes de artículos que fueron recuperados en el proceso de recuperación de información). Sea $w_{jk} \in [0, 1]$ la representación de la contribución del documento k-ésimo a la representación semántica del texto j-ésimo, en la figura **2-2** se encuentra la definición de esta representación [25]:



$$w_{k,j} = \text{dfitf}\,(d_k, t_j)$$
$$= df(d_k, t_k)\,.\,log\left(\frac{|T|}{\pi(d_k)}\right)$$

Figura 2-2.: Matriz de representación de los términos recuperados y definición inicial de representación DOR.

Los pesos que se obtienen en la representación DOR suelen ser normalizados, bajo la normalización de coseno, entonces, la ecuación para calcular los pesos de un documento propiamente se define como sigue [6]:

$$w_{k,j} = \frac{\text{dfitf}\,(d_k, t_j)}{\sqrt{\sum_{s=1}^{|D|} \text{dfitf}\,(d_s, t_j)}}$$

Se puede construir una representación del peso DOR de manera matricial, esto se logra definiendo una matriz diagonal de pesos D a partir de las frecuencias de ocurrencia de los términos, de la siguiente manera:

$$D = \text{diag}\left(\frac{|T|}{\pi(d_1)}, \ldots, \frac{|T|}{\pi(d_N)}\right)$$
$$DOR = \underbrace{(1 + \log(A^T))}_{A'}\, D \tag{2-2}$$

En esta ecuación, la matriz A está conformada por $A_{ij} = df(d_j, t_i)$, es decir, la frecuencia del término t_i en el documento d_j (df, por sus siglas en inglés). Como se muestra en la ecuación 2-3, esta medida será positiva si al menos aparece una vez el término dentro del documento.

$$df(d_j, t_i) = \begin{cases} 1 + log(\pi(d_j, ti)) & \text{si } \pi(d_j, ti) > 0 \\ 0 & \text{e.o.c} \end{cases} \tag{2-3}$$

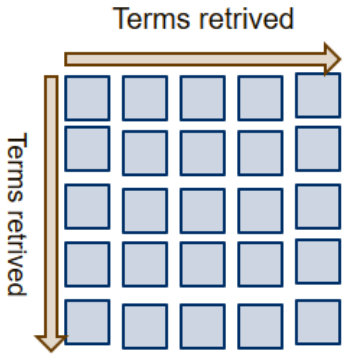
Adicionalmente, se denota $\pi(d_k)$ el número de diferentes términos que conforman el diccionario[2] T y que aparecen en al menos una vez en el documento d_k, más formalmente podemos

[2] Este contendrá todas las palabras en los conjuntos de pruebas y en los textos recuperados.

definirlo como el cardinal del siguiente conjunto: $\pi(d_k) = |\{t_i \mid t_i \in d_k \wedge t_i \in T\}|$.

2.2.2. Representación por co-ocurrecias de términos

La idea detrás de la representación por co-ocurrencias de términos (comúnmente llamada TCOR, por sus siglas en inglés), es que un término puede caracterizarse por la distribución de frecuencia de las palabras que co-ocurren con él en la colección de documentos. En nuestro caso, por ejemplo, una palabra puede ser caracterizada mediante los términos que aparecen frecuentemente con esta palabra en los diferentes resúmenes que se recuperaron. En un lenguaje más formal, la semántica de una palabra t_j puede verse como un función de la bolsa de términos que co-ocurren con t_j en la colección de documentos. Se puede encontrar que el vector de representación $\vec{w_j} =< w_{j1}, \ldots, w_{j|T|} >\in R^{|T|}$, tal que $t_j \in T$ es el conjunto de diferentes términos que ocurren en el documento y $w_{kj} \in [0, 1]$ corresponde a la contribución del k-ésimo término a la representación semántica de la palabra j-ésima del texto [25, 6]. En la figura **2-3** se presenta la definición de la representación TCOR.



$$w_{kj} = \text{tf}(t_k, t_j)\ \log \frac{|T|}{\#\tau\ (t_k)}$$

Figura 2-3.: Matriz de representación de los términos recuperados y definición inicial de representación TCOR.

Donde T_k es el número de diferentes términos en el diccionario que aparecen junto con el término t_j, al menos una vez en un documento, $tf\ (.)$ queda definida como:

$$tf(t_k, t_j) = [1 + \log\left(\#(t_k, t_j)\right)]\ \mathbb{1}_{\#(t_k,t_j)>0} = \begin{cases} 1 + \log\left(\#(t_k, t_j)\right) & \text{si } \#(t_k, t_j) > 0 \\ 0 & \text{e.o.c} \end{cases}$$

En la anterior definición, $\#(t_k, t_j)$ corresponde al número de veces que co-ocurre el término t_j con t_k. Los pesos que se obtienen en la representación TCOR al igual que en la representación DOR suelen ser normalizados, bajo la normalización de coseno:

$$w_{kj}^{'} = \frac{\mathrm{tf}(t_k, t_j)\ \log \frac{|T|}{\#\tau(t_k)}}{\sqrt{\sum_{s=1}^{|T|} w_{sj}}}$$

Se pude obtener una expresión análoga para la representación semántica TCOR haciendo uso de operaciones matriciales y construyendo una matriz diagonal D de la siguiente forma:

$$D = \mathrm{diag}\left(\frac{|T|}{\gamma(t_1)}, \ldots, \frac{|T|}{\gamma(t_{|T|})}\right)$$
$$TCOR = DB^{'} = D(1 + \log(B^t * B)) \tag{2-4}$$

En 2-5, $B_{ij} = \mathbb{1}_{t_j \in d_i}$ es una función que indica si t_j pertenece al documento d_i, $\gamma(t_j)$ representa el número de términos diferentes en el diccionario T que co-ocurren con t_j en al menos un documento. Como se mencionó en la anterior sección, se puede normalizar la representación de co-ocurrencia de términos haciendo uso de la normalización de coseno $\|.\|$.

2.2.3. Representación Word2Vec

Como ya hemos mencionado con anterioridad, el principal propósito de usar el enfoque de representación semántica Word2Vec es encontrar una representación que pueda capturar el valor semántico y sintáctico de las palabras que componen un determinado texto. Esta representación mapea cada palabra a un vector continuo $\mathbf{w}_t$ es la representación de los textos y a un vector $\mathbf{w}_c$ es la representación del contexto, dada las palabras que lo componen, usando el modelo de bolsa continua de palabras (considerablemente citado en la literatura como *skip-gram model*) [29]. Dado un corpus de entrenamiento extenso ($\mathcal{T} = \{w_1, \ldots, w_T\}$) se puede definir la función objetivo como la log-verosimilitud:

$$\sum_{t=1}^{T} \sum_{c \in C_t} \log\, p\left(w_c | w_t\right) = \sum_{t=1}^{T} \sum_{c \in C_t} \frac{e^{s(w_t, w_c)}}{\sum_{j=1}^{W} e^{s(w_t, w_j)}} \tag{2-5}$$

Aquí, el contexto C_t es el conjunto de índices de las palabras que rodean a una palabra w_t y se define s como un producto escalar entre la palabras y la representación del contexto $s\left(w_t, w_c\right) = \mathbf{u}_{w_t}^{'} \mathbf{v}_{w_c}$.
Se usa el algoritmo de gradiente descendiente para la optimización de la función objetivo 2-5, esta metodología usa la función softmax para parametrizar la probabilidad de observar el contexto de una palabra w_c dada una palabra w_t [30]. El resultado final es una matriz $W2V = \left[\mathbf{w}_1, \ldots, \mathbf{w}_{|\mathcal{T} \cap \mathcal{O}|}\right] \in R^{|T|}$ donde cada fila contiene la representación vectorial asociada a cada uno de los términos que fueron usados en el proceso de entrenamiento del modelo Word2Vec, pero al mismo tiempo pertenece a la base de datos original $\mathcal{O}$.

2.3. Algoritmos de agrupación

Los desarrollos sobre métodos de agrupamiento y la literatura que se puede encontrar sobre estos es increíblemente inmensa [21]. Hacer una descripción detallada de los diferentes métodos de agrupamiento se considera que está fuera de los objetivos de esta investigación. Sin embargo, teniendo en cuenta que los algoritmos de agrupación son una parte fundamental en el presente trabajo, aunque no sean propiamente el punto central de la investigación, en esta sección se describirá, por una parte, la función principal de estos algoritmos, su definición formal en términos de un problema de clasificación no supervisada [18] y una breve caracterización de los métodos. Por otra parte, se explicarán varios de los algoritmos estándar en la tarea de agrupación, más específicamente, en el contexto de la agrupación de textos.

2.3.1. El problema de agrupación

Un algoritmo de agrupamiento es un método que genera particiones dentro un conjunto de datos para crear subconjuntos o *clusters*, de tal forma que los objetos que conforman un grupo compartan rasgos comunes, a menudo la proximidad, de acuerdo con métricas específicas y definidas [45][15]. Este tipo de métodos, también llamados segmentación de datos, tiene varios propósitos. Como ya se menciono uno de ellos es asignar los objetos de análisis a los diferentes grupos, pero más importante aún, esta segmentación permite describir un objeto (en nuestro caso un texto) a partir de los objetos con los cuales está relacionado [18].
Otro de los propósitos que se encuentran en la literatura es determinar, por medio de las medidas de disimilaridad entre los objetos de cada grupo, si los datos consisten o no en un conjunto de subgrupos distintos [18]. Es decir, se podría concluir si el formato compacto que arroja los métodos de agrupación sigue siendo una versión informativa de la totalidad del conjunto de datos. Por lo tanto, estas técnicas también tienen como finalidad generar un resumen de la categorización de la información del conjunto original de datos [15].
Lograr la segmentación de datos resulta ser una tarea relevante en la actualidad debido a que nos enfrentamos a un enorme volumen de información almacenada y presentada como datos. Unos de los elementos esenciales para el análisis y la gestión de los datos consiste en clasificarlos o agruparlos, para así entender nuevos objetos, conceptos, fenómenos o características y poder compararlos con otros elementos ya conocidos [15][49].
En el contexto de análisis de información textual, como los títulos de proyectos de investigación de la Universidad, son variados los enfoques que se encuentran en la literatura. Los métodos de agrupamiento, en este tipo de información, pueden utilizarse en la organización de documentos sin supervisión, extracción automática de temas y tareas de recuperación de información [9].
Existen diferentes taxonomías para clasificar los métodos de *cluster* [21, 15, 49, 4]. La siguientes son las categorías generales que describen los enfoques más utilizados en el contexto de análisis de texto, algunos de los cuales serán explicados más ampliamente en la sec-

ción 2.3.1.1.

- **Métodos particionales:** los métodos de agrupamiento por particiones buscan dividir una colección de documentos en un conjunto de grupos que no se superponen, para maximizar la medida de diferencia de los grupos en el proceso de agrupación [9, 21].
- **Métodos jerárquicos:** este grupo de métodos tienen su origen en el estudio de taxonomías. Es decir, se basan principalmente en la noción de aaglomeración y división de grupos.
- **Métodos según su función objetivo:** estas técnicas de agrupación operan de acuerdo con la definición de un criterio de amortización o con base en un criterio de búsqueda de la mejor partición posible.

Aunque los algoritmos de agrupamiento no abarcan todos los objetivos de esta tesis, son importantes para la determinación de los grupos temáticos de los textos objeto de esta investigación. Para poder evaluar los posibles modelos de representación semántica propuestos para el tratamiento de textos cortos y planteados al inicio de este capítulo, en este trabajo se utilizan tres algoritmos de agrupamiento, los cuales podrían proveer ventajas en el tratamiento de información textual: el algoritmo k-medias, con el uso de transformaciones kernel; la factorización de matrices no-negativas; y la agrupación espectral. El resto de esta sección los revisa.

2.3.1.1. Definición formal

Para tener cierta claridad en la términología y posteriores definiciones de los métodos, en esta corta sección se describirá formalmente cuál es la finalidad de una técnica de agrupación. Considere un conjunto de datos $\mathbf{X}$ (según lo descrito por Berkhin [4], pueden representar objetos, instancias, casos, patrones, etc.) $\mathbf{x}_i = (x_{i1}, ..., x_{id})$, $i \in \{1 \dots N\}$, en el cual N es el número de observaciones en el conjunto de datos. Cada componente $x_{il} \in A_l$, $l \in \{1 \dots d\}$, se trata de las observación numéricas o categóricas de cada uno de los d atributos en un espacio de componentes que denotaremos como A; dichos atributos también son llamados componentes, variables, características, dimensiones, etc. La finalidad de las técnicas de agrupamiento es asignar los conjuntos de datos a un sistema finito de k conjuntos. Los algoritmos seleccionados en esta investigación cumplen las siguientes restricciones [49]:

1. Los subconjuntos de datos originales no pueden ser vacíos $C_i \neq \varnothing$.
2. La unión de los conjuntos encontrados es igual al conjunto de datos $\bigcup_{i=1}^{k} C_i = \mathbf{X}$. Con la posible excepción de valores atípicos.
3. Los subconjuntos encontrados en los algoritmos son disyuntos es decir $\forall i, j \in 1, \dots, K$ y $i \neq j$ tenemos que $C_i \cap C_j = \varnothing$.

2.3.2. K-medias usando tranformaciones Kernel

De acuerdo con la descripción dada en Dhillon[11], este algoritmo resulta ser una extensión del tradicional método de agrupación k-medias. La idea principal es mejorar el rendimiento del algoritmo haciendo uso de transformaciones Kernel, lo cual se puede definir como una función no lineal para transformar los datos desde el espacio de representación original a un espacio dimensional superior, en el cual los objetos de estudio sean linealmente separables [11]. Como se indica en Dhillon[11], la función objetivo, usando una función de kernel ϕ, está definida como:

$$\mathcal{D}\left(\mathcal{C}_1, \ldots, \mathcal{C}_k\right) = \sum_{j=1}^{k} \sum_{\mathbf{a} \in \mathcal{C}_j} w(\mathbf{a}) ||\phi\left(\mathbf{a}\right) - \mathbf{m_j}||^2$$

$$\mathbf{m_j} = \frac{\sum_{\mathbf{b} \in \mathcal{C}_j} w(\mathbf{b}) \phi(\mathbf{b})}{\sum_{\mathbf{b} \in \mathcal{C}_j} w(\mathbf{b})}$$

De acuerdo con la intuición general detrás del algoritmo de k-medias, la tarea de agrupación parte del problema de encontrar los "mejores" grupos. Estos se pueden encontrar, minimizando, mediante la siguiente expresión:

$$m_j = \text{argmin}_{\mathbf{z}} \sum_{\mathbf{a} \in \mathcal{C}_j} w(\mathbf{a}) ||\phi\left(\mathbf{a}\right) - \mathbf{z}||^2$$

El proceso general está descrito en el algoritmo que se presenta a continuación (algoritmo 2.1[3]). El paso principal consiste en calcular la matriz Kernel. Esta matriz contiene en sus registros los productos $\phi(a)\phi(b) = K(a, b)$, que, por su parte, se usan para estimar la distancia euclidiana $||\phi\left(\mathbf{a}\right) - \mathbf{m_j}||$, y así minimizar la función objetivo.

2.3.3. Factorización de matrices no-negativas

Otro de los algoritmos seleccionados para la experimentación con los textos en estudio es la factorización de matrices no-negativas (NMF, por sus siglas en inglés). Varios trabajos recientes demuestran la importancia de este algoritmo para la detección simultánea de grupos de documentos y de agrupaciones de palabras. Este enfoque ha sido ampliamente utilizado en el agrupamiento de documentos, sin embargo, al utilizarlo con textos cortos no ha mostrado resultados, como consecuencia de la baja frecuencia de los términos y el problema de la dispersión de las medidas en las representaciones de los textos cortos [50]. Se espera que este tipo de problemas se pueda resolver con lo descrito en la sección de representaciones distribucionales (ver sección 2.2).

La factorización de matrices no-negativas se basa en la descomposición de matrices descrita inicialmente por Ding, He y Simon [12]. Tal como se expone en Kim y Park [22], se parte de

[3] Tomado de Dhillon[11]

Algoritmo 2.1: K-medias usando transformación Kernel

Datos: K: la matriz kernel, k: número de grupos, w: pesos de cada punto.
Resultado: $C_1, \ldots, C_k$, correspondiente a los k grupos de la partición de los datos.

Inicializar $C_1^{(0)}, \ldots, C_k^{(k)}$
Definir $t = 0$
repetir
  para $\mathbf{a} \in X$ **hacer**
    Encontrar j*, el nuevo grupo de* $\mathbf{a}$*, con:*

$$j^*(\mathbf{a}) = argmin_j \left|\phi(\mathbf{a}) - \mathbf{m}_j\right|^2$$

  Calcular nuevos grupos como:

$$C_j^{(t+1)} = \{\mathbf{a} \ : \ j^*(\mathbf{a}) = j\}$$

  Definir $t = t + 1$
hasta que *Alcanzar criterio de convergencia*

la definición de X como una matriz no-negativa de tamaño $n \times p$, (v. g. con $x_{ij} \geq 0$, denota $X \geq 0$), y $r > 0$. La factorización de matrices no-negativas (NMF) consiste en encontrar una aproximación dada por:

$$X \approx W_{n\times r}\, H_{r\times p}$$

En donde W y H son al igual matrices no-negativas, que resumen la información contenida en la matriz original X dentro de r factores. Así, para el caso de estudio de esta investigación se resumirá la representación de los textos en r diferentes factores [46].

2.3.4. Agrupación espectral

El agrupamiento espectral ha sido ampliamente utilizado en una serie de problemas de agrupación de textos [7, 11, 23, 48]. Este algoritmo parte de la definición de un grafo G, totalmente conectado, como representación del conjunto de datos original. En este caso, el grafo $G = (V, E, \mathbf{W})$ está dado por: V, un conjunto de vértices; E, el conjunto de aristas; y $\mathbf{W} = w_{ii'}$, que representa la matriz de afinidad construida a partir de una medida de similaridad, usualmente denotada como matriz de adyacencia [18].
Partiendo de un grafo G, y de definir $D = d_{ii}$, en donde $d_{ii} = \sum_j w_{ij}$ se define como el *grado*, la suma de los pesos de las conexiones hacia el vértice v_i, partiendo de la definición de esta matriz diagonal, se define el grafo *laplaciano* como:

$$\mathbf{L} = \mathbf{D} - \mathbf{W}$$

La versión normalizada de esta matriz es:

$$\mathbf{L}' = \mathbf{D}^{-\frac{1}{2}}\mathbf{L}\mathbf{D}^{-\frac{1}{2}} = I_n - \mathbf{D}^{-\frac{1}{2}}\mathbf{W}\mathbf{D}^{-\frac{1}{2}}$$

La función objetivo en este caso es maximizar $\epsilon(X) = \frac{1}{k}\sum_{j=1}^{k}\left(X_j^t\mathbf{W}X_j\right) / \left(X_j^tDX_j\right)$. Una solución bien conocida de este problema se obtiene calculando los k autovectores propios más grandes de la matriz $\mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2}$. Como paso final, se usan estos vectores propios para calcular una partición discreta del conjunto de datos, el algoritmo de k-medias es usualmente utilizado para esta tarea [11, 18]. Más abajo podrá encontrarse el algoritmo 2.2, que presenta el proceso general que se lleva a cabo en este método.
Así las cosas, la definición de $\mathbb{W}$, en los experimentos se usó la similaridad de coseno entre los documentos para encontrar el grado de afinidad entre los textos. Esta medida se basa en el angulo entre dos vectores numéricos, y se puede comprobar que es independiente del largo de los documentos que se están comparando [42].

Algoritmo 2.2: Agrupación epectral.

Datos: X matriz de datos original, k: número de grupos.
Resultado: $C_1, \ldots, C_k$, correspondiente a los k grupos de la partición de los datos.

1 *Inicializar* $\mathbf{W}$*, usando la matriz de datos original.*
2 *Calcular el grado de la matriz de adyacencia, usando:* $\mathbf{D} = Diag\left(\mathbf{W}\mathbf{1}_n\right)$.
3 *Encontrar la solución del problema, calculando:*

$$\mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2}V_{[k]} = V_{[k]}\,diag\left(\{s_1, \ldots, s_k\}\right), s_k \text{ es el valor propio } k.$$
$$\mathbf{Z} = D^{-\frac{1}{2}}V_{[k]}, \text{ con } [k] = \{1, \ldots, k\}$$
$$\mathbf{Z}^* = diag\left(ZZ_{1,1}^t, \ldots, ZZ_{n,n}^t\right)^{-\frac{1}{2}}\mathbf{Z} \text{ (normalización)}$$

4 *Encontrar los grupos, usando* $\mathbf{Z}^*$ *y el algoritmo* k*-medias.*

2.4. Medición del desempeño y validez de los grupos

El último aspecto referente al proceso de agrupación de textos cortos es la evaluación interna y externa (ver figura **3-1** y capitulo 3). Este procedimiento permite la valoración de los

grupos encontrados, para determinar dos resultados claves: el primero es la definición del mejor esquema de representación semántica para el caso de textos cortos (ecuaciones (2-2), (2-4) o (2-5), enunciadas en la sección 2.2); el segundo, que es el más relevante, es evaluar la calidad de los grupos encontrados, hallando el número óptimo de grupos y el método de agrupación con mejor desempeño, para lo cual se evalúan los resultados en términos de la similaridad intragrupo (coeficiente Davies-Bouldin, índice QError, coeficiente Silueta) y en términos de medidas de homogeneidad de los grupos, como los son: índice Rand ajustado, índice de Homogeneidad, índice V, información mutua ajustada (Adjusted MI) y la pureza [46]. Las metodologías usadas en el proceso de medición del desempeño se describirán en esta sección. Para las definiciones de los distintos métodos de validez se tendrán las siguientes convenciones [4]:

- K_h: representa el conjunto de la categoría h, de g categorías que han sido previamente etiquetadas ($h \in \{1, \ldots, g\}$).
- C_l: el grupo l-ésimo encontrado por el algoritmo de agrupación, con $l \in \{1, \ldots, k\}$.
- n_j: el número total de objetos que están en el grupo j-ésimo.
- $n^{(h)}$: el número total de objetos que están clasificados en la clase h-ésima.
- $n_j^{(h)} = |\{x_i | x_i \in C_l \wedge x_i \in K_h\}|$: el número de objetos clasificados en el *cluster* j que pertenecen a la clase h.
- λ ($X \to \lambda$): el vector de etiquetas de los *cluster*, la indicadora de pertenencia del grupo de los objetos encontradas por el algoritmo.
- κ: el vector de etiquetas de la clase de grupos "verdaderos"..

2.4.1. Medidas de validez externas

Las medidas de validez externa requieren, como su nombre lo indica, de información externa para la evaluación del desempeño de los algoritmos. Este tipo de medidas se basan en la idea de comparar numéricamente los grupos obtenidos con respecto a las categorías que se suponen "verdaderas", es decir, por ejemplo, una clasificación dada por un experto [34]. Sin embargo, a diferencia de los problema de clasificación supervisada, dicha verdad no está disponible para el algoritmo de agrupamiento. Esta clase de medidas de evaluación puede utilizarse para comparar el rendimiento de principio a fin de cualquier agrupación independientemente de los modelos o las similitudes utilizados.
Dado que los algoritmos de agrupación responden a un problema de carácter no supervisado, el rendimiento de este tipo de métodos no puede ser juzgado de la misma manera que un problema de clasificación supervisado, debido a que la clasificación para la validación podría no ser la mas óptima [4]. En los siguientes apartados se presentan, las cuatro medidas de

validez externas que fueron utilizadas para la evaluación de los diferentes métodos propuestos en el trabajo de investigación realizado en esta tesis de maestría.

2.4.1.1. Índice de pureza

Como mencionan Ghosh and Strehl [17] el índice de pureza corresponde a la precisión con que nuestros modelos de agrupación agrupan los objetos de análisis, bajo el supuesto de que todos los objetos de un determinado grupo identificado se clasifican como miembros de una clase; por ejemplo, para el caso del problema no supervisado los grupos encontrados deben corresponder a las clases dadas por los expertos. Formalmente, dado un grupo C_l, la pureza es la razón entre el número de objetos de la clase dominante y el total de elementos que compone el grupo, es decir:

$$\phi_{pureza}(C_l, \kappa) = \frac{1}{n_l} \max_h(n_l^{(h)})$$

Para evaluar esta medida a través de los diferentes grupos se calcula el promedio de las purezas agrupadas ponderadas por el tamaño del grupo. Entre más cercano a 1, indicará que el número de elementos correctamente clasificados en cada una de las clases se aproxima a n_l, lo que indicaría que la mayoría de los objetos están correctamente asignados a los grupos, formalmente este índice está definido por:

$$\phi_{pureza}(\lambda, \kappa) = \frac{1}{n} \sum_{l=1}^{k} \max_h(n_l^{(h)}) \tag{2-6}$$

2.4.1.2. Índice Rand ajustado

Con el fin de comprobar los resultados de los algoritmos seleccionados, se puede hacer uso de una medida de concordancia entre dos particiones, en este caso, entre la partición encontrada por el método de agrupación y la partición inducida por las clases que provienen de un criterio externo [51]. Teniendo en cuenta los grupos encontrados y las clases, se definen a y b como:

- $a = \|\{(x_i, x_j) | x_i \in C_l \wedge x_j \in C_l \wedge x_i \in K_h \wedge x_j \in K_h\}\|$, corresponden al número de pares de objetos que están en el mismo grupo y también están en la misma clase.
- $b = \|\{(x_i, x_j) | x_i \in C_l \wedge x_j \notin C_l \wedge x_i \in K_h \wedge x_j \notin K_h\}\|$, esto es, el número de pares de elementos que están en conjuntos diferentes en C y que también están en conjuntos diferentes en K.

El índice de Rand no ajustado está dado por la razón del número de concordancias del algoritmo y la clasificación externa (numerador), entre el número de posibles parejas (denominador), formalmente:

$$\phi_{RI} = (a+b)\binom{n}{k}^{-1}$$

Se puede definir una formulación alternativa del índice Rand que se ajusta por el posible agrupamiento casual de las parejas. El número de parejas que concuerdan puede estar sesgado por efecto de la aleatoriedad, para esto se asume una distribución hipergeométrica para la distribución de ocurrencias de los grupos y las clases. Con base en la distribución asumida se puede encontrar $E(\phi_{RI})$, y así puede definirse el índice ajustado de Rand como:

$$\phi_{ARI} = \frac{\phi_{RI} - E(\phi_{RI})}{\text{máx}(\phi_{RI}) - E(\phi_{RI})} \tag{2-7}$$

El índice ϕ_{ARI} toma un valor entre 0 y 1, un valor de 0 indicaría que la partición inducida por las clases no concuerda, en ningún par de objetos, con la agrupación encontrada por el algoritmo, por otro parte si el índice toma el valor de 1 indicaría que la partición encontrada por el algoritmo es exactamente igual a la partición real de los datos.

2.4.1.3. Información mutua

El índice de información mutua es una de las medidas de validación externa que más ha sido utilizada, debido principalmente a su fundamento teórico [17]. Esta basada en medidas de entropía[4], las cuales, para el caso de este índice, son los siguientes [47]:

$$\begin{aligned} H(C) &= \sum_{l=1}^{k} P(i)\log(P(i)) = \sum_{i=1}^{l} \frac{|C_l|}{N}\log\left(\frac{|C_l|}{N}\right) \\ H(K) &= \sum_{h=1}^{g} P'(j)\log(P'(j)) = \sum_{j=1}^{g} \frac{|K_h|}{N}\log\left(\frac{|C_l|}{N}\right) \end{aligned} \tag{2-8}$$

En 2-8, C y K representan dos variables aleatorias asociadas a los grupos y a las clases, respectivamente. En la formulación de la entropía, $P(i)$ y $P'(j)$ representan la probabilidad de que un objeto de análisis extraído aleatoriamente de C o de K sea clasificado en el grupo l-ésimo o en la clase H-ésima, respectivamente. Estas probabilidades, tal como se puede ver en la anterior formulación, pueden ser estimadas por las frecuencias de ocurrencias de los grupos y las clases. Si $H(.)$ es cercana, indica que la entropía es máxima, es decir que las categorías no están concentradas en ninguna clase o grupo.
El índice de validez externa, denominada "información mutua", se define como:

[4] Medida de la incertidumbre existente ante un conjunto de datos, en este caso, el conjunto de etiquetas (grupos o clases).

$$\phi_{MI}(\lambda,\kappa) = \sum_{l=1}^{k}\sum_{h=1}^{g} P\left(l,h\right)\log\left(\frac{P\left(l,h\right)}{P(l)P'(h)}\right)$$
$$= \sum_{l=1}^{k}\sum_{h=1}^{g}\frac{n_l^{(h)}}{N}\log\left(\frac{n_l^{(h)}}{n^{(h)}.n_l}\right) \tag{2-9}$$

2.4.1.4. Información mutua ajustada

La información mutua, al igual que su variante utilizando normalización, no se ajusta al azar. Según se menciona en la literatura, a medida que aumenta el número de diferentes grupos también aumenta el índice, independientemente de la cantidad real de "información mutua" entre las grupos encontrados y las clases reales [47]. Por esta razón se hace el ajuste de la información utilizando $E[\mathrm{MI}(U,V)]$, la siguiente esperanza matemática [46]:

$$E[\mathrm{MI}(C,K)] = \sum_{i=1}^{k}\sum_{j=1}^{g}\sum_{n_{ij}=(a_i+b_j-N)}^{\min(a_i,b_j)}\frac{n_{ij}}{N}\log\left(\frac{N.n_{ij}}{a_i b_j}\right)\frac{a_i!b_j!(N-a_i)!(N-b_j)!}{N!n_{ij}!(a_i-n_{ij})!(b_j-n_{ij})!(N-a_i-b_j+n_{ij})!}$$

Utilizando el valor esperado, la información mutua ajustada puede entonces calcularse utilizando una forma similar a la del índice Rand ajustado visto en la sección 2.4.1.2:

$$\phi_{AMI}(\lambda,\kappa) = \frac{\mathrm{MI} - E[\mathrm{MI}]}{\max(H(C),H(K)) - E[\mathrm{MI}]} \tag{2-10}$$

Los índices ϕ_{MI} y ϕ_{AMI} son simétricos en términos de U y V. Estos índices se encuentran entre 0 y 1, donde 0 indicará que no existe una buena agrupación de los datos, es decir no existe información común o mutua entre la partición real de los objetos y la partición encontrada por el método. Un valor de 1 indicará que la agrupación encontrada por el método tiene una "correlación perfecta" con las clases reales de los datos.

2.4.1.5. Índice V

Este índice se base en dos aproximaciones de entropía: la primera es un índice de homogeneidad (h) y la segunda es de la completitud (c). Formalmente estos índices están dados por:

$$h = 1 - \frac{H(C|K)}{H(C)}$$
$$= 1 - \frac{1}{H(C)}\left[\sum_{l=1}^{k}\sum_{h=1}^{g}\frac{n_l^{(h)}}{N}\cdot\log\left(\frac{n_l^{(h)}}{n_l}\right)\right]$$

Aquí, h representa el índice de homogeneidad de los grupos encontrados por el algoritmo. Estos grupos satisfacen el criterio de homogeneidad si al interior de los grupos los objetos pertenecen a una misma clase [40]. Esto se cumple cuando el desorden al interior de los grupos es mínimo con relación a las clases, es decir, si la distribución de las clases al interior de un grupo determinado está concentrada en una sola etiqueta. El desorden de esta distribución puede ser medida por $H(U|V)$ (la entropía condicional de las clases dadas las asignaciones de los grupos). Si esta entropía es cercana a 0 y en consecuencia h es cercana a 1, indicaría que solo existe una clase prevalente en cada grupo encontrado. $H(C)$ es la entropía de los grupos y está dada por la expresión 2-8, que fue introducida con anterioridad.

$$\begin{aligned} c &= 1 - \frac{H(K|C)}{H(K)} \\ &= 1 - \frac{1}{H(K)} \left[\sum_{l=1}^{k} \sum_{h=1}^{g} \frac{n_l^{(h)}}{N} \cdot \log \left(\frac{n_l^{(h)}}{n^{(h)}} \right) \right] \end{aligned}$$

La medida c representa el índice de completitud de los grupos encontrados por el algoritmo de agrupación. Contrario a la formulación del índice de homogeneidad, la completitud se refiere a la característica ideal de que todos los objetos clasificados en una clase h sean agrupados como elementos de un mismo grupo o *cluster* [40]. Para evaluar la completitud, examinamos la distribución de asignaciones de *cluster* dentro de cada clase, esto se traduce en evaluar la entropía condicional de los *cluster* dada la asignación a las clases ($H(K|C)$). c es máxima cuando $H(K|C)$ es mínima, esto ocurre, por ejemplo, cuando se asignan todos los objetos a un único grupo y, como consecuencia, todos los elementos de una clase estarán en la clase y la entropía será cero.
Tal como lo describen Rosenberg y Hirschberg [40], el índice V (V-measure, como es comúnmente conocido en inglés) se define como la media armónica de las medidas presentadas anteriormente, es decir:

$$\phi_V = 2 \cdot \frac{h \cdot c}{h + c} \tag{2-11}$$

Análogo a los otros índices de validez externas expuestos con anterioridad, el índice ϕ_V, está acotado entre 0,0 y 1,0, en donde 1 significa que las etiquetas o grupos, encontrados por el método, concuerdan perfectamente con las clases reales de los objetos.

2.4.2. Medidas de validez internas

En oposición a las medidas de validez externa, los criterios de validación internas formulan la calidad como una función de los datos o de similaridad de los puntoso o grupos. Por ejemplo, cuando usamos el criterio de error cuadrático medio, el algoritmo puede evaluar su propio rendimiento y afinar sus resultados en consecuencia [17].

Cuando se utilizan criterios internos, el agrupamiento se convierte en un problema de optimización. Por ejemplo, al usar el criterio de mínimos cuadraros se busca encontrar los grupos que minimizan el error bajo un criterio de distancia, y no bajo una estructura preespecificada de los datos, que es impuesta normalmente por un experto, como las clases verdaderas [34]. En los siguientes apartados se presentan las tres medidas de validez interna que fueron utilizadas para la evaluación de los diferentes métodos propuestos en el trabajo de investigación.

2.4.2.1. Índice Davies-Bouldin

De acuerdo con la descrito por Davies y Bouldin [10], existen ciertas propiedades que son importantes para la construcción de una medida de similaridad entre grupos. Para ello, se construye una función de la dispersión al interior de los *cluster*, denotada S_i, que depende directamente de los objetos clasificados en C_i, es decir, $S_i = S(x_1, \ldots, x_{n_i})$, en donde $C_i = \{x_1, \ldots x_{n_i}\}$. El objetivo descrito por Davies y Bouldin [10] es encontrar $R_{S_i,S_j,M_{i,j}}$, una medida de similaridad entre dos grupos, tal que:

- $R_{S_i,S_j,M_{i,j}} \geqslant 0$. La función de similaridad es no-negativa.
- $R_{S_i,S_j,M_{i,j}} = R_{C_j,C_i}$. Esta función cumple con la propiedad de simetría.
- $R_{S_i,S_j,M_{i,j}} = 0$, si y solo si $S_i = S_j = 0$. La medida de similaridad es 0 si la dispersión de ambos grupos es 0.
- Cuando $S_j = S_k$ y $M_{i,j} \leqslant M_{i,k}$, entonces $R_{S_i,S_j} > R_{S_i,S_k}$. Esta propiedad indica que cuando la dispersión se mantiene constante entre los grupos, a medida que la distancia $M(.)$ aumenta la similaridad, como es esperado, disminuirá.
- Cuando $S_j \geqslant S_k$ y $M_{i,k} = M_{i,k}$, entonces $R_{S_i,S_j,M_{i,j}} > R_{S_i,S_k}$. Esta propiedad enuncia que cuando la distancia entre los grupos se mantiene constante, pero la dispersión aumenta, la medida de similaridad también aumentará.

Esta medida de similaridad, por definición, tiene que tener en cuenta $M_{i,j}$, esto es, la distancia entre el i-ésimo y el j-ésimo grupo. En el caso de una agrupación perfecta la distancia entre grupos debe ser lo más grande posible. En nuestro caso, está dada por la expresión 2-12, donde A_i representa el centroide de C_i, y n_i es el tamaño del i-ésimo grupo, cuando $p = 2$ representa la distancia euclidiana de los puntos. La funciones de dispersión y similaridad de un grupo describirse mediante las siguientes expresiones [10]:

$$S_i = \left(\frac{1}{n_i} \sum_{j=1}^{T_i} |X_j - A_i|^p \right)^{1/p}$$

$$M_{i,j} = \left(\sum_{k=1}^{n} \|a_{k,i} - a_{k,j}\|^p \right)^{\frac{1}{p}}$$

$$R_{S_i,S_j,M_{i,j}} = \frac{S_i + S_j}{M_{i,j}} \tag{2-12}$$

Es posible utilizar otras métricas en $M_{i,j}$, especialmente en problemas con datos de dimensionalidad alta. Para casos en los que la distancia euclidiana no permite capturar la noción de distancia, lo recomendable es usar la misma métrica que usa el algoritmo de agrupación que se pretende evaluar. Finalmente, la medida presentada en Davies y Bouldin [10] está definida como:

$$D_i \equiv \max_{j \neq i} R_{i,j}$$

$$\phi_{DB} = \frac{1}{k} \sum_{i=1}^{k} D_i \tag{2-13}$$

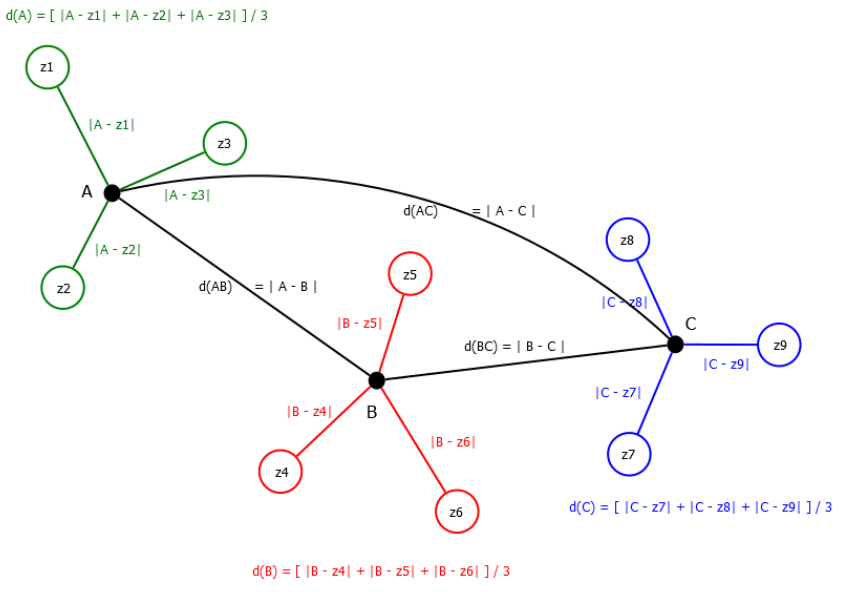


Figura 2-4.: Descripción del calculo para el índice Davies-Bouldin.

Un pequeño ejemplo del cálculo del índice ϕ_{DB} se presenta en la figura **2-4**. En este se encuentra la distancia promedio de cada punto a su centroide correspondiente, así se calcula

S_i; luego, se calculan las distancias entre los centroides de los diferentes grupos, para luego calcular $R_{S_i,S_j,M_{i,j}}$ de acuerdo con las ecuaciones presentadas con anterioridad; a continuación se calcula D_i, que correspondería a la medida de similaridad con el *cluster* más "cercano" bajo la distancia definida. La mejor selección de los grupos será entonces aquella que logre minimizar ϕ_{DB}, es decir minimizar el promedio de las distancias D_i [10].

2.4.2.2. Índice de cuantificación de error

Este índice, más conocido en inglés como cuantificación del error (QE, por sus siglas en inglés), no es más que una modificación del conocido error cuadrático medio, en el cual se mide la dispersión de los puntos en cada uno de los grupos encontrados usando el centroide de cada grupo y la distancia de este a cada uno de los puntos, que conforman el grupo [17]. Formalmente, de acuerdo con el número de objetos en el *cluster* C_l según λ, el centroide del grupo y la suma cuadrada de los errores (SSE, por sus siglas en inglés), se puede definir como:

$$\mathbf{c}_l = \frac{1}{n_l}\sum_{\lambda_j=l}\mathbf{x}_j \Rightarrow SSE(\mathbf{X},\lambda) = \sum_{l=1}^{k}\sum_{\mathbf{x}\in C_l}\|\mathbf{x}-\mathbf{c}_l\|^2$$

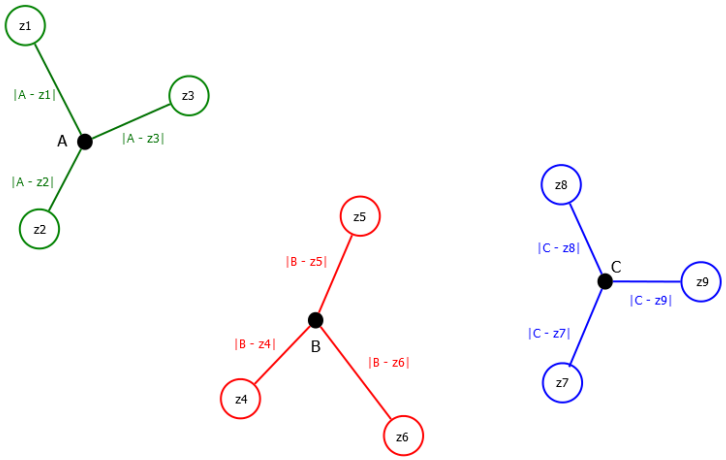


Figura 2-5.: Descripción del calculo para el índice Quantization-Error (QE).

Como se pretende construir una medida de calidad de los grupos que esté dentro del intervalo 0 a 1, donde 1 indicaría una asociación perfecta, esto ocurriría si SSE tiende a 0. En este caso se define el índice como:

$$\phi_{QE} = \exp^{-SSE(\mathbf{X},\lambda)} \tag{2-14}$$

Es conocido que minimizar $SSE(.)$ es uno de los objetivos del algoritmo heurístico de K-medias, o también puede ser considerado análogo a la minimización de la función de verosimilitud de los datos, cuando se asume que los datos y, por ende, la agrupación se generan

por mixturas de funciones de distribución normales multivariadas [10]; como se mencionó con anterioridad, ϕ_{QE} es máxima cuando SSE es mínimo. Un ejemplo de dicho cálculo se presenta en la figura **2-5**, en donde se calculan las diferencias $|\mathbf{x}-\mathbf{c}_l|$, cuando los grupos son homogéneos y estas diferencias son mínimas.

2.4.2.3. Coeficiente Silueta

Una de las estrategias cuando no se cuenta con las clases reales sobre un conjunto, es decir, cuando no se cuenta con la información de una fuente externa, se conoce como coeficiente Silueta. Este coeficiente puede calcularse para cada una de las muestras y se compone de dos elementos esenciales [41, 44]:

- a_i: la distancia promedio entre un objeto y todos los demás puntos del i-ésimo grupo.
- b_i: la distancia promedio entre un objeto y todos los demás objetos que conforman el grupo más cercano al grupo i-ésimo.

El coeficiente Silueta para una muestra simple se define entonces como en la expresión 2-15, gráficamente se puede ver en la figura **2-6** un ejemplo con tres grupos encontrados por un algoritmo, primero se encuentra el coeficiente silueta para cada uno de los puntos y luego se calcula el promedio de dicha medidas.

$$\begin{aligned} s_i &= \frac{b_i - a_i}{max(a_i, b_i)} \\ \phi_{Silueta} &= \frac{1}{N}\sum_{i=1}^{N} s_i \end{aligned} \tag{2-15}$$

Por la definición 2-15 se puede deducir que el índice cumple: $-1 \leq \phi_{Silueta} \leq 1$. Así, una puntuación alta cercana a uno, para $\phi_{Silueta}$, se relaciona con un partición en donde los grupos están mejor definidos, es decir mayor diferencia promedio entre grupos y menores diferencias promedios intragrupos (homogeneidad dentro del grupo), esto ocurre cuando $a_i \ll b_i$. Si $\phi_{Silueta}$ es cercano a cero, indicaría que algunos objetos están en la frontera de dos grupos, implicaría mayores diferencias promedio intragrupos y menores diferencias promedios entre grupos. [41].

Breve síntesis del capítulo En esta sección de la tesis, se presentaron los algoritmos involucrados en la construcción de una nueva metodología para agrupación de textos cortos. Las metodologías y las estrategias presentadas en este capítulo se concentraron principalmente en apoyar el objetivo número 1 de desarrollar la representación para textos cortos que logre un mejor semántica de los términos que componen los textos y que facilite la tarea

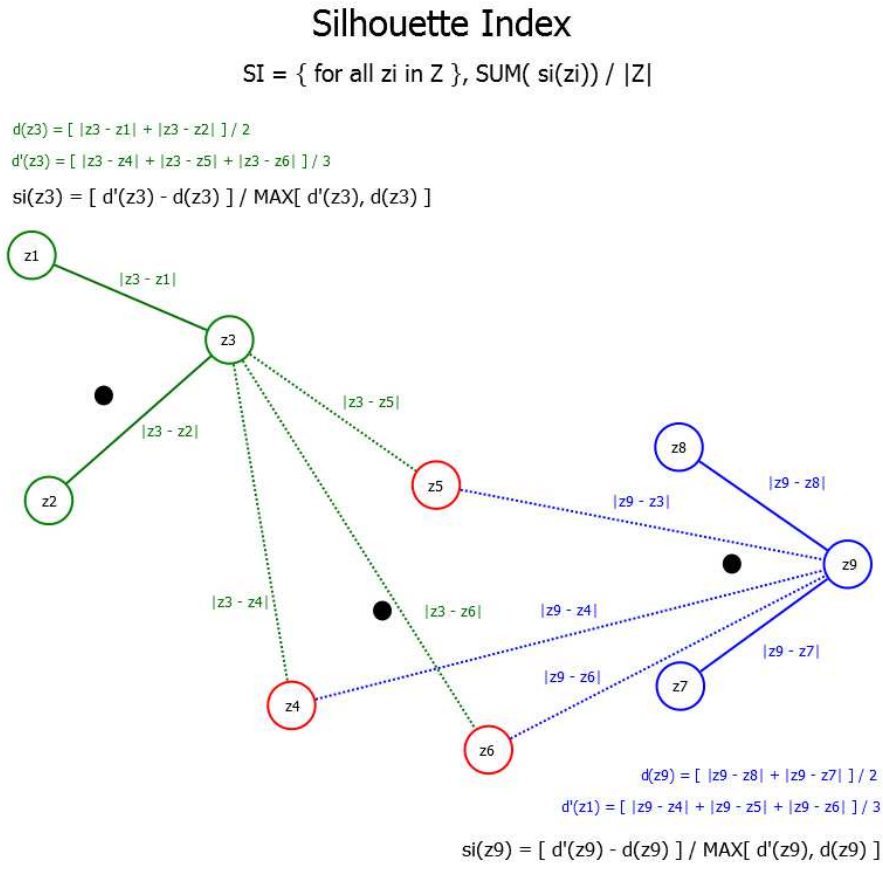


Figura 2-6.: Descripción del calculo para el coeficiente Silueta.

de agrupación de los mismos. En las siguientes secciones se desarrolló el marco teórico correspondiente al objetivo número 2, para lograr determinar un método de agrupamiento no supervisado que utilice la representación construida para un determinado conjunto de textos de extensión corta, en esta segunda parte se presentaron los diferentes métodos disponibles y también las medidas de validez internas y externas que están disponibles en el estado del arte para evaluar el desempeño de los algoritmos de agrupación.

3. Metodología propuesta

El objetivo de este capítulo es dar la descripción detallada de la propuesta metodológica para la agrupación de textos cortos, haciendo uso de las representaciones distribucionales de los términos. Se trata, entonces, del método que se utilizará en los experimentos para obtener los grupos objetos de estudio de esta investigación.
La fígura **3-1** presenta la visión general del método en sus diferentes etapas y procesos. El método consta de cuatro etapas principales.

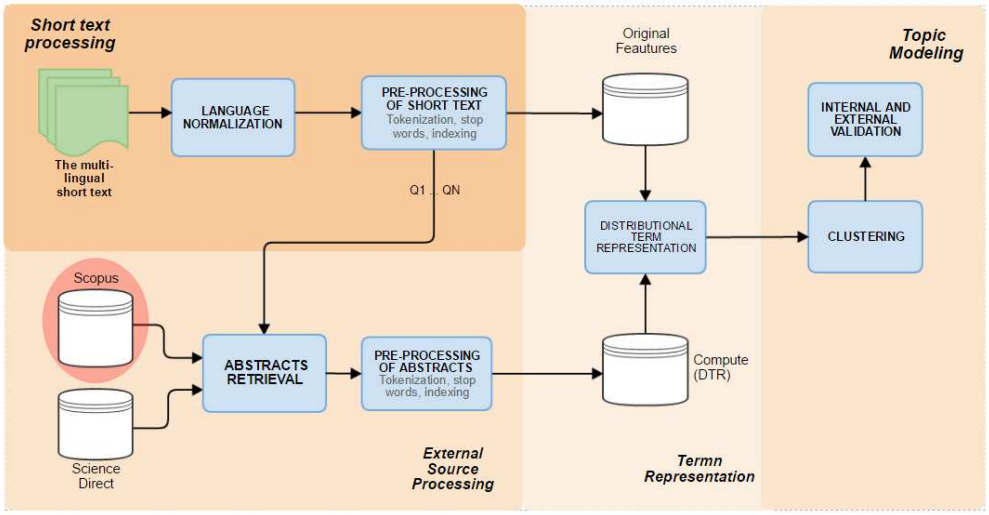


Figura 3-1.: Arquitectura general del método de agrupación de textos cortos.

En la etapa *Procesamiento inicial de textos cortos* se hace el pre-procesamiento de los textos originales, lo cual consiste en la depuración y normalización de los términos de los textos. La siguiente etapa del proceso se denomina *Expansión de consultas*, en la cual partiendo de los textos depurados se construyen consultas, que se utilizan para la expansión de los términos, esto es, para encontrar elementos textuales de mayor extensión, por ejemplo, resúmenes de artículos, títulos, citas, y/u otros elementos que pueden estar asociados a la temática de los textos originales. En la etapa *Expansión y representación de términos* se extrae la representación distribucional de los textos a partir de los elementos recuperados en la etapa anterior, dicha representación se usa para encontrar el sentido semántico de los términos y poder encontrar relaciones entre los textos que componen la colección. En la última etapa, *Moldeamiento de tópicos*, se aplican los diferentes métodos de agrupamiento sobre la colección

de textos cortos después de aplicar la expansión de los términos.
Partiendo de las etapas del proceso propuesto para obtener los vectores característicos de los textos cortos y el posterior proceso de agrupación de los textos, el tratamiento temático de este capítulo se ha organizado de la siguiente manera, en la sección 3.1 se presentara los aspectos ligados a la etapa de *Procesamiento inicial de textos cortos*, por su parte la etapa de *Expansión y representación de términos*, se divide a su vez en dos secciones: sección 3.2 en donde se presentan los aspectos relacionados con la construcción de las consultas para el proceso de recuperación de información y la sección 3.3 en donde se describe el algoritmo para DTR. La sección 3.3 también presenta el esquema metodológico propuesto para la experimentación de los diferentes algoritmos de agrupación seleccionados, finalmente en la sección 3.4 presenta los componentes del diseño usado para la implementación de las etapas descritas con anterioridad.

3.1. Preprocesamiento inicial de textos cortos

Como podemos ver la figura **3-1**, el primer paso del proceso propuesto, dentro de la actividad de normalización de los textos, es la identificación del lenguaje, para lo cual se construye un clasificador capaz de detectar el idioma, esto resulta necesario porque los proyectos de investigación de la universidad pueden estar registrados tanto en inglés como en español.
El primer paso es entonces usar un clasificador para hacer la detección del idioma, para lo cual dentro de la investigación, se usaron dos enfoques diferentes: el primero consiste en usar las listas de palabras vacías (*stopwords*, en inglés) de los diferentes lenguajes para construir un clasificador. La idea general de este clasificador es encontrar la frecuencia de ocurrencia de estas listas y seleccionar el idioma en que aparezca un mayor número de palabras claves [2]. El otro enfoque utilizado fue usar la función *get_languages*[1], que usan los recursos web de Google, con el fin de identificar y devolver los códigos de idioma iso639-1, para los idiomas admitidos en la API de Google Translate. Para mejorar el desempeño y los tiempos de la tarea de clasificación se decidió utilizar el segundo enfoque, dado que permite de una manera más ágil la identificación del lenguaje del texto original.
El paso siguiente a la detección del idioma consiste en traducir los textos cortos a un lenguaje base, para tal fin se usa Google Translate API[2]. El lenguaje base que se ha elegido para el método es el inglés, y en este mismo lenguaje se construirá y aplicará lo faltante del proceso, debido a que muchos artículos han sido escritos y publicados en libros y revistas en inglés. Una vez traducidos los textos, en esta etapa eliminamos la lista tradicional de palabras de paro inglesas para formular la consulta; este proceso se presentará en la siguiente sección.
La segunda actividad en la clasificación de documentos es el pre-procesamiento. Consiste en transformar los documentos en una representación con la que un algoritmo de aprendizaje

[1] *get_languages* está disponible en el paquete goslate python (https://pypi.python.org/pypi/goslate).
[2] Google Translate API puede convertir dinámicamente los textos entre ciertos pares de idiomas, esta API está disponible en https://cloud.google.com/translate/docs/.

pueda llevar a cabo la tarea de clasificación. Regularmente, para esta tarea se realizan algunos de los siguiente procesos:

- **Limpieza de documentos:** consiste en remover todo aquello que se considere ruido, por ejemplo: etiquetas HTML o XML, que normalmente se usan para organizar las colecciones de documentos en distintas categorías o de alguna forma particular. Sin embargo, estas etiquetas se tienen que remover para que el algoritmo de aprendizaje únicamente tome en consideración la información del contenido del documento. También incluye la eliminación de encabezados, separadores, tablas, caracteres extraños, entre otros, al igual que las palabras vacías, se trata de palabras muy frecuentes que por lo general, no aportan información del contenido del documento (por ejemplo: artículos, pronombres, preposiciones y conjunciones).
- **Lematización de palabras**: un lematizador obtiene las raíces morfológicas de las palabras eliminando desinencias por conjugación, número, género. Ejemplo: doctor se obtiene como lema de doctora o de doctores.

3.2. Expansión de consultas

La siguiente etapa del proceso, que forma parte de la fase de recuperación de información, es la formulación de consultas para enviar al motor de búsquedas y así poder obtener documentos que estén relacionados con el texto original, en nuestro caso, resúmenes de artículos científicos. Para ello se parte de un texto particular una vez se ha sometido a la depuración presentada en la sección anterior, dicho texto corresponde, como se ha mencionado, al título de un proyecto de investigación en el idioma seleccionado como base, $(d_i \in D = \{d_1, \ldots, d_N\})$. La mejor manera de ver esto es presentar un ejemplo que ilustra qué consultas consideramos para expandir la representación. La siguiente figura presenta un texto de ejemplo del conjunto original de datos después de su depuración, así como las consultas construidas a partir del texto:

$\mathbf{t_j}$ = "Multi dynamics algorithm for global optimization" $\Rightarrow$

$\mathbf{Q_1}(\mathbf{t_j})$ = "TITLE-ABS-KEY(Multi + dynamics + algorithm + global + optimization)"

$\mathbf{Q_2}(\mathbf{t_j})$ = "TITLE-ABS-KEY(Multi + OR+ dynamics +OR+ algorithm + OR+ global +OR+ optimization)"

$\mathbf{Q_3}(\mathbf{t_j})$ = "AUTHKEY(Multi + dynamics + algorithm global + optimization)"

Tal como muestra el ejemplo anterior, se realizaron tres consultas diferentes. Estas se lanzan en el motor de búsqueda de la siguiente manera: $\mathbf{Q_1}(.)$ realiza la búsqueda booleana que devuelve los documentos donde aparecen simultáneamente todos los términos de la consulta.

El motor de búsqueda, en nuestro caso Scopus o Sciencedirect, busca estas palabras dentro del título, dentro de las palabras clave y dentro del resumen. Se considera que esta consulta es exitosa cuando recupera al menos un documento, es decir, si en alguna de las partes del documento se encuentran simultáneamente todos los términos que componen la consulta. Si esto ocurre, ya no habría necesidad de lanzar consultas adicionales. Si no se recupera ni un documento relevante, se realiza una segunda consulta, esto es:$\mathbf{Q_2}$ (.), la cual está compuesta por los mismos términos de la consulta anterior, excepto que devuelve los documentos que contienen en el título, en el resumen o en las palabras clave al menos uno de los términos que conforma la consulta (se le agrega el operador "OR"). Otro tipo de consulta realizada es $\mathbf{Q_3}$ (.), que fue utilizada en la construcción para la representación Word2Vec y la construcción de otro conjunto de validación adicional. El detalle de la base de datos creada a partir de los documentos recuperados se describe en la sección 4.1, y los detalles del modelo de representación fueron presentados ya en la sección 2.2.
Para lanzar las consultas previamente descritas, se diseñó un programa empaquetador que se conecta con dos interfaces de programación de aplicaciones (API, por sus siglas en inglés). Estas interfaces están en línea, y se encargan de la integración del contenido y datos de productos de Elsevier[3]. De estas herramienta se seleccionaron:

- La API de búsqueda de ScienceDirect. Este recurso es la mayor base de datos de investigación científica y médica primaria revisada por pares, tiene 14 millones de usuarios mensuales [39].

- la API de Elsevier Scopus. Scopus es la base de datos de citas y estudios bibliográficos más extensa del mundo, con más de 65 millones de registros de información de 5000 editores diferentes [39].

Estas dos APIs fueron seleccionadas por encima de los recursos de Web of Science, debido a que daban la opción de almacenar los resúmenes de los artículos científicos, los cuales consideramos ser una importante información textual para la ampliación semántica de los términos originales [39].
Se ha desarrollado un programa para almacenar una estructura de datos predefinida, la cual se muestra en la figura **3-2**. En dicha imagen podemos apreciar la información considerada relevante en las consultas realizadas: el localizador uniforme de recursos (*URL*, por sus siglas en inglés) de la consulta utilizada en la web. Adicionalmente, vemos que se guarda el texto de la consulta que fue utilizado en el motor de búsqueda (*searchTerms*) y el lenguaje según la clasificación Scopus (*name*); en el campo *entry* se almacenan todos los resúmenes de trabajos que están relacionados con el texto original, un ejemplo de la información recuperada y guardada en este campo se muestra en la Figura **3-3**. Sin embargo, en las API de Elsevier tenemos una restricción respecto de este punto solo podemos descargar hasta 100 documentos relevantes.

[3] Las APIs de Elsevier están disponibles en https://dev.elsevier.com/index.html.

```
object {
    "url_mia": string "http://api.elsevier.com/content/search/scopus?query=TITLE-ABS-
    KEY%28cooperation+OR+agreement+OR+alcaravan%29&count=100&facets=language%28count%3D
    1%29&apikey=3f8eebe2fd170110dc0c74a072238d9f&view=COMPLETE",
    "search-results": object {
        "opensearch:Query": object {
            "@searchTerms": string "TITLE-ABS-
            KEY%28cooperation+OR+agreement+OR+alcaravan%29",
            "@role": string "request",
            "@startPage": string "0"
        },
        "opensearch:itemsPerPage": string "100",
        "opensearch:totalResults": string "1191854",
        "facet": object {…},
        "link": array [4],
        "opensearch:startIndex": string "0",
        "entry": array [1]
    }
}
```


Figura 3-2.: Ejemplo de la estructura de almacenamiento del procedimiento para una consulta en Scopus o Sciendirect.

```
            "@startPage": string "0"
        },
        "opensearch:itemsPerPage": string "100",
        "opensearch:totalResults": string "1191854",
        "facet": object {…},
        "link": array [4],
        "opensearch:startIndex": string "0",
        "entry": array [
            object {
                "pii": string "S0260877415003416",
                "author": array [3],
                "prism:volume": string "168",
                "dc:title": string "Detection of fluorescence signals from ATP in the
                second derivative excitation-emission matrix of a pork meat surface
                for cleanliness evaluation",
                "affiliation": array [1],
                "prism:publicationName": string "Journal of Food Engineering",
                "prism:issn": string "02608774",
                "dc:identifier": string "SCOPUS_ID:84939455031",
                "subtypeDescription": string "Article",
                "authkeywords": string "ATP | Excitation-emission matrix | Food safety
                | Hygiene monitoring | Second derivative",
                "citedby-count": string "0",
                "dc:description": string "We investigated the potential application of
                excitation-emission matrix (EEM) spectroscopy in the rapid, non-
                destructive evaluation of cleanliness in meat processing plants....",
                "source-id": string "20586",
                "author-count": object {
                    "$": string "3",
```


Figura 3-3.: Ejemplo de la estructura de almacenamiento del procedimiento, de los artículos, para una consulta en Scopus o Sciendirect.

En la tabla **3-1** resumimos la secuencia de pasos que se toman para abordar la construcción de una nueva representación a partir de documentos recuperados de nuestro método, inclu-

yendo los procesos de las actividades relacionadas con el preprocesamiento de los textos y la recuperación de la colección de documentos, esta puede ser considera la ruta critica para la construcción de la representación distribucional de los textos.

Tabla 3-1: Procedimiento principal o crítico utilizado para lanzar una consulta y depurar los textos.

1. Búsqueda de resúmenes científicos
   - Lanzar la consulta $\mathbf{Q_1}$ (.)
   - Lanzar $\mathbf{Q_2}$ (.), si $\mathbf{Q_1}$ no encuentra resultados.
   - Lanzar $\mathbf{Q_3}$ (.), se usa únicamente en la construcción de Word2Vec y en la construcción de la base de datos de Scopus 4.2.
2. Etapa de Pre-procesamiento
   - Estructuración de la información.
   - Quitar caracteres especiales de las cadenas.
   - Quitar palabras vacías en el campo de *resumen*.
   - Lematización de palabras.
   - Detección de palabras con alta frecuencia de aparición.
   - Construir nueva representación de acuerdo con la metodología descrita en la sección 2.2.

3.3. Representación DTR y agrupación

El siguiente paso en el método propuesto es la representación distribucional de los términos y la ejecución de los algoritmos de agrupamiento. En esta sección se presenta toda una nueva clase de métodos de agrupación para textos de extensión corta, basados en representaciones distribucionales de los términos. De acuerdo con las definiciones de la sección 2, sea $\mathbf{W} = (\mathbf{w}_1, ..., \mathbf{w}_d)$ la matriz de representación de los documentos originales, con base en la nueva colección de documentos (usando DOR, TCOR o word2Vec); sea A la matriz de representación de los términos en la colección original (usualmente tf-idf). El algoritmo 3.2 define todo un nuevo conjunto de métodos de agrupación para textos de longitud corta, como se puede ver, en el algoritmo los resultados dependen de la representación distribucional seleccionada y del método de agrupación seleccionado.

Algoritmo 3.2: Agrupación de textos cortos usando DTR.

Datos: Matriz de pesos A
Datos: Matriz de pesos W de acuerdo con las expresiones (2-2), (2-4) o (2-5)
Resultado: Los subconjuntos $C_i \neq \varnothing$ tal que $\bigcup_{i=1}^{k} C_i = \mathbf{X}$ y $C_i \cap C_j = \varnothing$ para $i \neq j$.

```
1 inicio
2     Calcular representación DTR de los textos usando la ecuación 2-1.
3     Calcular matriz de similaridad de coseno S_cos(x_i, x_j) = x_i^t x_j / (||x_i|| ||x_j||)
4     para 1 ∈ {1, ..., K} hacer
5         Usar el Algoritmo 2.1 o el Algoritmo 2.2
6         Validez Interna sección 2.4.2, calcular φ_DB, φ_QE y φ_Silueta
7         si κ vector de etiquetas de las clases "verdaderas". entonces
8             Validez Externa sección 2.4.1, calcular φ_ARI, φ_V, φ_AMI, φ_MI y
              φ_pureza
9     Selección del mejor algoritmo de acuerdo con φ.
```

En particular, para la experimentación, se va a estudiar los clasificadores que surgen de las combinaciones de las propuestas para representar los textos cortos, descritos en la sección 2.2 y de los métodos de agrupación presentados en esta disertación y descritos en la sección 2.3. La tabla **3-3** presenta los diferentes combinaciones seleccionados y las abreviaturas o etiquetas con la cual serán presentados, mas adelante, en la sección de resultados experimentales.

Tabla 3-3.: Métodos evaluados en la experimentación

Etiqueta	Representación Distribucional	Método de agrupación
KK (tf_idf)	tf-idf	K-medias con Kernel
SPEC (tf_idf)		Agrupación Espectral
KK (DOR)	DOR	K-medias con Kernel
SPEC (DOR)		Agrupación Espectral
NMF (DOR)		Factorización de matrices no negativas
KK (TCOR)	TCOR	K-medias con Kernel
SPEC (TCOR)		Agrupación Espectral
NMF (TCOR)		Factorización de matrices no negativas

Continúa en la siguiente página

Tabla 3-3 – Continúa de la página anterior

Etiqueta	Representación Distribucional	Método de agrupación
KK (W2V_G)	word2Vec entrenado con Google News	K-medias con Kernel
SPEC (W2V_G)		Agrupación Espectral
KK (W2V_S)	word2Vec entrenado con Scopus	K-medias con Kernel
SPEC (W2V_S)		Agrupación Espectral
KK (W2V_W)	word2Vec entrenado con Wikipedia	K-medias con Kernel
SPEC (W2V_W)		Agrupación Espectral
KK (TCOR * W2V_S)	TCOR $\otimes$ word2Vec entrenado con Scopus	K-medias con Kernel
SPEC (TCOR * W2V_S)		Agrupación Espectral
NMF (TCOR * W2V_S)		Factorización de matrices no negativas

Para aprovechar las múltiples características exhibidas por diferentes medidas de similitud, obtenidas a partir de diferentes representaciones distribucionales de los términos que componen los textos, existen propuestas para asignar pesos a la similitud medida de un par de documentos por otra medida de similitud. Dadas dos matrices de similitud S_1 y S_2, se define una nueva matriz de similitud como:

$Y = S_1 \otimes S_2$

Donde $\otimes$ es la multiplicación elemento-elemento de las dos matrices. El enfoque propuesto por[42] y depues usar algún metido de agrupación basada en la nueva matriz de similitud de los elementos.

3.4. Diseño e implementación

En el desarrollo de esta investigación fue necesaria la estructuración del código de las diferentes etapas del proceso. Se necesita de una herramienta con la que se puedan implementar y/o adaptar los diferentes algoritmos de agrupación seleccionados, así como las representaciones semánticas seleccionadas, las medidas de evaluación que se van a utilizar y, finalmente, los reportes de resultados de las experimentaciones. Con el fin de implementar correctamente el esquema presentado en la sección 3.3, tener un diseño de los procesos involucrados es primordial para una buena ejecución. En esta sección, explicamos el diseño y la implementación de dicho proceso.

La implementación de la solución y el código fuente está escrito en el lenguaje de programación Python [16], este lenguaje de programación de uso general, orientado a objetos e interpretado, se seleccionó como herramienta, dado que tiene muchas facilidades para la

computación científica. Como se puede ver en la figura **3-4** el programa se compone de seis clases principales:

- *searchScopus* es una clase que contiene las funciones necesarias para construir y lanzar las consultas a la API de Elsevier Scopus. Los principales componentes de las clase son: *get_search_url* y *normalize_query* que permite la construcción de la Url que sera utilizada para hacer la recuperación de los documentos, dicha url tiene incluida el atributo *scopusApyKey*, suministrada por el servicio web de Elsevier; *get_abstract* es la encargada de enviar las consultas y guardarlas en objetos de la clase *elsevierResult*.

- La clase *searchScience*, hereda los atributos y los métodos de la clase *searchScopus*, se hace la redefinición de los métodos *get_search_url* y *normalize_query*, para construir adecuadamente las consultas que serán enviadas al sistemas de Elseivir. En esta clase todas las funciones están orientadas a obtener los resultados de la API de búsqueda de ScienceDirect.

- La clase *elsevierResult* está diseñada para capturar la información relevante de una consulta generada por *searchScience* o *searchScopus*. Los campos almacenados según lo descrito en la figura **3-2** tienen dos métodos construidos *limit_str_size* y *_repl_*, los cuales son métodos programados con validaciones internas para excluir aquellos elementos de las consultas que no contienen texto y/o no tienen la información minima para identificar un articulo (Elsevier_ID, autores, palabras claves, etc.).

- Una vez se tengan el resultado de las consultas y el texto almacenado en el campo *abstractArticle*, *computeDTR* es la clase encargada de calcular las representaciones semánticas de los textos en el conjunto original, utilizando los métodos *weight_DTR* y *compute_DTR*, con base en los textos recuperados y las formulaciones descritas en la sección 2.2. Existen funciones adicionales para validar la frecuencia de ocurrencia de los términos y poder depurar el diccionario de palabras con el que se calcularán las representaciones.

- *resultCluster* es la clase creada, a partir de una matriz de representación semántica DTR encontrada usando la clase *abstractArticle*, descrita con anterioridad, esta funcionalidad ejecuta los algoritmos programados como métodos implementados al interior de la clase y descritos en la sección 2.3.

- La última clase del procedimiento realiza la validación de los grupos encontrados en los pasos anteriores, se trata de la clase *oneCluster*, que fue diseñada para calcular y almacenar los resultados de las medidas de validez interna y externa. Al interior de esta clase se encuentra la implementación de las diferentes validaciones descritas en la sección 2.4.

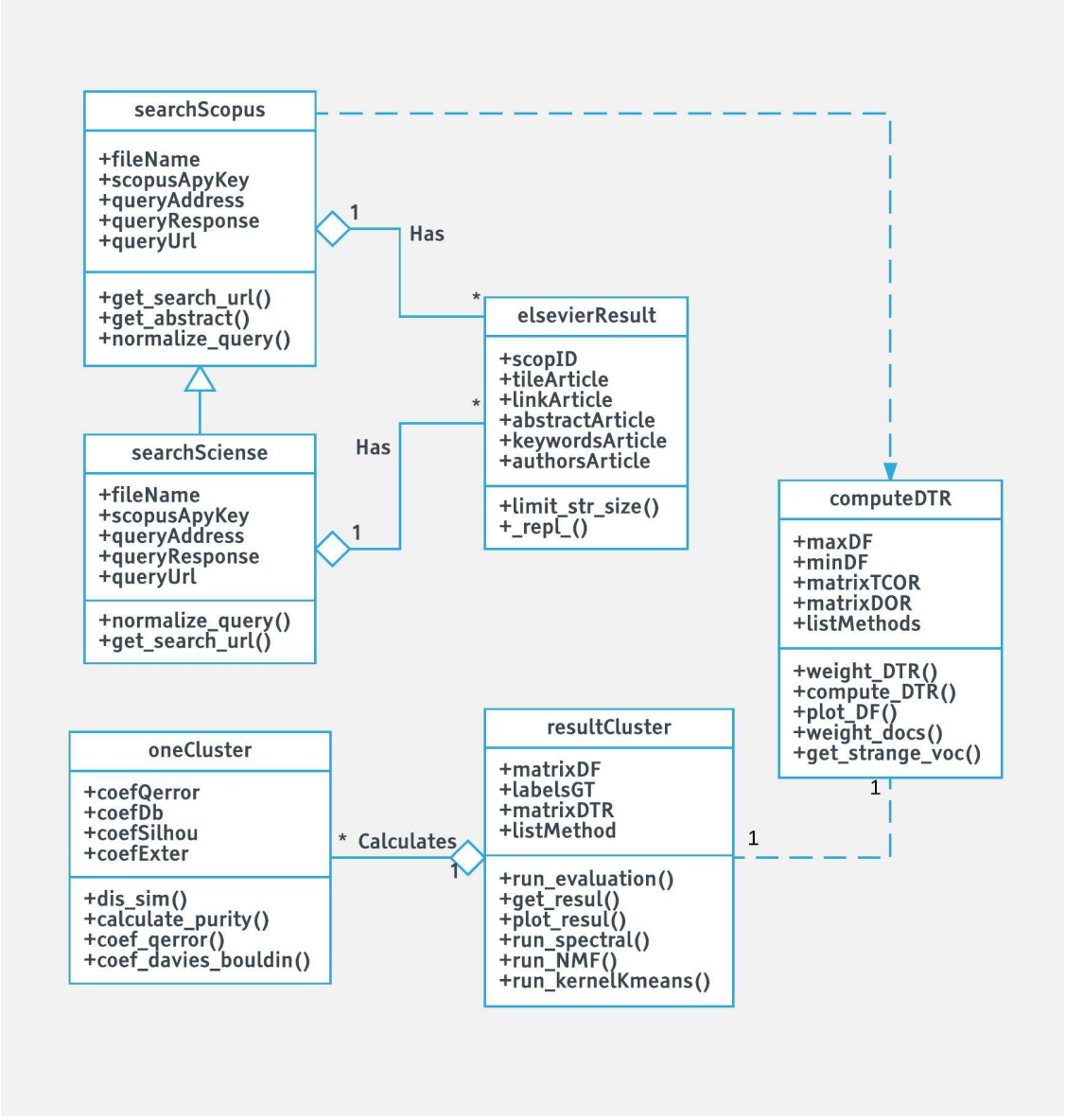


Figura 3-4.: Diagrama UML de clases utilizado para la implementación de la estrategia agrupación de textos cortos.

Breve síntesis del capítulo En este capítulo, se describió el proceso metodológico usado en la representación distribucional de los términos que componen los textos objeto de estudio. En la primera parte de la sección se explica en detalle la metodología para la búsqueda y la recuperación de los textos relacionados por medio de la construcción de consultas, adicionalmente se describe cual fue el preprocesamiento de los textos, esta sección aporta a la resolución del objetivo general de esta investigación, describiendo en detalle la estrategia diseñada para la detección de grupos o temas latentes en los textos de extensión corta. Adicionalmente apoya la realización del objetivo número 2, puesto que describe la selección de los métodos de agrupación y las combinaciones propuestas en para la experimentación. Posteriormente se hace énfasis en el proceso para obtener los vectores característicos de los textos cortos, así como las fuentes utilizadas para la construcción de esos vectores característicos. Al igual, se explica el diseño metodológico de la implementación del proceso con la información, atributos y métodos que han sido implementados para el almacenamiento de la información obtenida de las consultas. Esta información permitirá la ampliación semántica de los textos de acuerdo con las estrategias discutidas en el capítulo 2.

4. Evaluación Experimental

En esta investigación se hicieron diversas pruebas para evaluar los algoritmos de agrupación seleccionados en todos los aspectos de interés: validez en la agrupación de los objetos, tiempo de procesamiento y número de grupos resultantes. En este capítulo se explicara en detalle algunos experimentos para evaluar el desempeño y la precisión de la metodología propuesta, comparándola con los diferentes métodos del estado del arte. El presente capítulo está organizado de la siguiente forma: En la primera parte (sección 4.1), se describe de manera general los dos conjuntos de datos seleccionados para validar el método propuesto. En la segunda parte (sección 4.2), mencionamos las características del conjunto de datos construído con Scopus y la evaluación de los diferentes experimentos realizados con este conjunto. Para finalizar, en la tercera y última parte (sección 4.3), se presentan varios experimentos adicionales que están relacionados con los textos cortos de las investigaciones que están en curso en la Universidad Nacional de Colombia.

4.1. Conjuntos de datos utilizados para experimentación

Conjunto de datos	Número de textos	Número de palabras	Promedio de Palabras	Número de clases	Artículos recuperados
Scopus	1696	7373	13.43	20	22267
UN Títulos	2138	7307	20.91	No disponible	20279

Tabla 4-1.: Descripción de los conjuntos de datos y número de textos utilizados en los experimentos.

Para probar nuestro algoritmo realizamos experimentos en dos conjuntos de datos. La descripción de estos conjuntos de datos se puede ver en la tabla **4-1**. 1) Los títulos de proyectos de investigación de la Universidad Nacional de Colombia: Este conjunto de datos contiene (3718) documentos de texto no etiquetados almacenados en formato "XLSX", y como se mencionó anteriormente no tienen las mismas características del conjunto de datos de referencia. 2) Para la construcción del segundo conjunto de datos, se descargaron alrededor de (1696) títulos de artículos con sus respectivos resúmenes y palabras clave de la base de datos Scopus. Estas palabras clave están categorizadas en la tabla **4-2** cuya búsqueda recupera los

100 documentos más relevantes que contienen todos los términos de la consulta en el campo *keywords*, utilizando el mismo procedimiento de descarga de la representación de texto propuesta y descrita en la sección 3.2.

4.2. Conjunto de prueba scopus (Scopus Títulos)

La tabla **4-2**, muestra los resultados de las 20 consultas enviadas a la API de Elsevier Scopus. Para recuperar nuevos artículos, se usó la consulta $\mathbf{Q_3}$ (.) descrita en la sección 3.2. Las palabras clave con las que se construyeron las consultas, fueron seleccionadas entre las palabras clave de los artículos recuperados en el proceso de agrupamiento de los proyectos de investigación de la universidad, adicional a esto, se seleccionaron las palabras de acuerdo con la categorización de las áreas mencionadas por Bellotti et al. [3]. La palabra clave con la cual se recuperó el artículo fue considerada la clase de los textos, y esta clase es la que se quiere recuperar con la ampliación distribucional propuesta.

Tabla 4-2: Categorización de palabras claves y número de documentos recuperados

Palabra Clave	Número de Artículos
Academic and achievement	78
Artificial and neural and network	79
Asphalt and mixture	98
Cultural and heritage	73
Ecological and compensation	93
Electrophysiology	89
Energy and savings	83
Image and analysis	78
Intracellular and trafficking	97
Ionotropic and glutamate and receptors	85
Logistics and clusters	93
Motivation	57
Mycobacterium and tuberculosis	97
Parallel and robot	90
Periodontal and diseases	83
Pharmacology	81
Rural and development	75
Social and capital	86
Thermal and cracking	89
Violent and behavior	92
Número total de títulos en la colección	**1696**

4.2.1. Resultados representaciones distribucionales de documentos Word2Vec

Para entrenar correctamente el modelo Word2Vec (descrito en la sección 2.3), se necesita de mucha información textual con el fin de obtener una representación adecuada de las palabras. En este experimento, utilizamos una implementación ya construída para el entrenamiento word2vec, dicha implementación se encuentra en python disponible en el paquete *gensim* [38]. Además, se utilizó los siguientes recursos para calcular la representación distribucional de los términos con base en el modelo word2Vec:

- Se construyó un *CBOW* word2vec usando la mayoría de los parámetros establecidos por Mikolov et al. [29]. Esta representación fue entrenada con un conjunto de datos de *Google News*[1], el cual contiene 3 millones de palabras en inglés. El modelo final tendrá alrededor de 3 millones de vectores de una dimensión 300.

- Otra de las fuentes disponibles en la web son los textos de Wikipedia. Estos han sido una fuente confiable para la representación de palabras en otros experimentos. Los modelos preconfigurados de word2vec, entrenados con Wikipedia, están disponibles en el repositorio como wiki2Vec [2]. Este repositorio tiene un total de 1,151,090 vectores de palabras representados en un espacio de 1000 dimensiones.

- Entrenamos otro *CBOW* mediante el uso de resúmenes recuperados de las diferentes APIs de Elsevier. Este conjunto de datos se construyó utilizando un índice para almacenar todas las *keywords* del repositorio de artículos recuperados a lo largo de los experimentos realizados. El repositorio fue compuesto por la búsqueda de tantas consultas como fuera posible, haciendo uso de la formulación $\mathbf{Q_3}$ presentada en la sección 3.2. En total, la API recuperó 1,045,144 diferentes artículos, junto con la información recuperada de acuerdo con lo descrito en la figuras **3-2** y **3-3**.

  Los textos que conforman los resúmenes de los artículos recuperados se utilizaron como corpus, con un mínimo de preprocesamiento (en este caso solo se eligió la conversión de los textos a minúsculas). En el entrenamiento del modelo se seleccionó un tamaño de ventana de 5 palabras, con el fin de la definición del largo de los contextos usados en el entrenamiento. Para la generación de los *embeddings* o *CBOW*, en este experimento, se probaron diferentes dimensiones del espacio de representación de salida con tamaños de 200, 300 y 500.

Adicionalmente, en este ejercicio se contaban con dos conjuntos diferentes de documentos recuperados: una primer conjunto de entrenamiento fue construído con resultados parciales de la búsqueda, con alrededor de 400,000 artículos recuperados. El segundo conjunto de

[1] https://github.com/mmihaltz/word2vec- GoogleNews-vectors
[2] https://github.com/idio/wiki2vec

datos utilizados en el entrenamiento fue entrenando el modelo con 1,045,144 artículos, es decir la totalidad de los documentos. Los resultados de los experimentos con las diferentes

Método	ϕ_{DB}	ϕ_{QE}	$\phi_{Silueta}$	ϕ_{ARI}	ϕ_{HOM}	ϕ_{V}	ϕ_{AMI}	ϕ_{MI}	ϕ_{pureza}
KK (W2V_SV400_200)	0.381	2.789	0.041	0.207	0.392	0.403	0.371	1.179	0.398
KK (W2V_SV400_300)	0.383	2.867	0.029	0.205	0.380	0.394	0.358	1.136	0.384
KK (W2V_SV1000_300)	0.379	2.934	0.040	0.212	**0.394**	**0.401**	**0.370**	**1.179**	**0.407**
SPEC (W2V_SV400_200)	0.386	2.957	**0.048**	0.192	0.369	0.371	0.344	1.103	0.378
SPEC (W2V_SV400_500)	0.384	3.231	0.040	0.203	0.379	0.380	0.354	1.132	0.398
SPEC (W2V_SV1000_500)	0.381	3.282	0.038	0.209	0.390	0.391	0.365	1.164	0.402

Tabla 4-3.: Resultados obtenidos en los diferentes métodos experimentos propuestos word2vec (sección 3.3), construídos con la base de datos *Scopus Títulos*. Se presentan los indices de validez interna (ϕ_{DB}, ϕ_{QE} y $\phi_{Silueta}$) y los indices de validez externa (ϕ_{ARI}, ϕ_{HOM}, ϕ_{V}, ϕ_{AMI}, ϕ_{MI} y ϕ_{pureza}).

representaciones distribucionales descritas con anterioridad, se presentan en la tabla **4-3**. Tal como se observa por un pequeño margen, los mejores resultados se alcanzaron usando un espacio vectorial de 300 dimensiones y usando la base de datos que contenía los 1,045,144 recuperados.

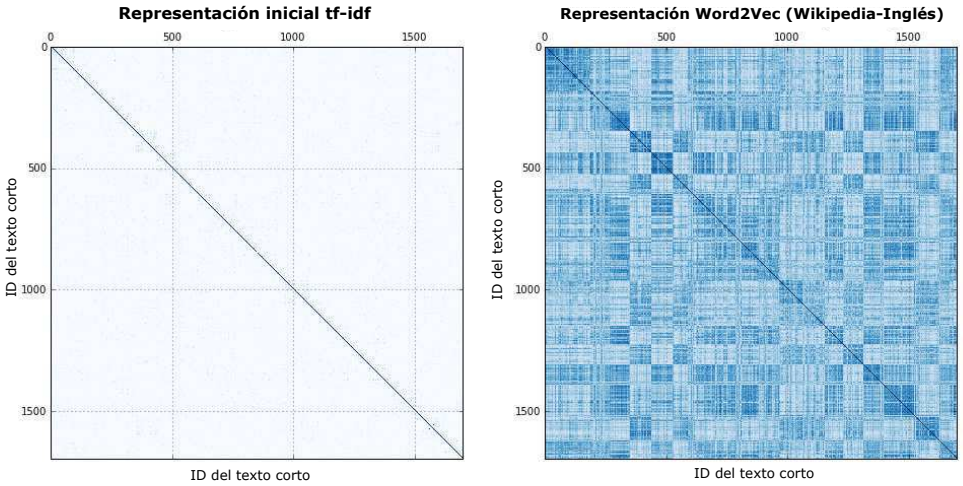


Figura 4-1.: Matriz de correlación usando tf-id (izquierda) y después de usar la representación distribucional word2Vec(derecha).

Los algoritmos de agrupación que fueron seleccionados para la experimentación, se basan en una medida de distancia que es calculada entre las representaciones distribucionales de los diferentes textos que componen el conjunto de datos. A manera de ejemplo la figura **4-1**, presenta la medida de asociación antes y después de aplicar una representación distribucional usando Word2Vec. En el gráfico entre más cercano sea un texto de otro mayor será la intensidad del color. Tal como se puede observar en la gráfica (izquierda), usar los textos originales no arroja suficiente información estadística, es decir, se puede concluír que no existe asociaciones entre los textos. Contrario a esto(gráfica derecha), después de hacer la representación semántica de los textos, se observan algunos patrones de asociación. Esto debido a que la representación semántica mejora, brindando mayor información para calcular la similaridad entre los textos.

4.2.2. Tiempo de ejecución y exploración del número de artículos recuperados

Uno de los parámetros que pueden impactar el desempeño de las representaciones distribucionales propuestas, es el número de documentos relevantes (en adelante, denotado por M) que son recuperados en la búsqueda de la fuente externa (en este caso Scopus). El número de elementos recuperados puede generar un incremento en el tiempo de procesamiento de los textos. Adicionalmente, puede introducir una reducción en las frecuencias relativas de los términos, influyendo directamente en las representación semánticas de los mismos que componen un determinado documento. Otro de los parámetros que se debe tener en cuenta, es la frecuencia máxima de documentos (DF, por sus siglas en inglés). Esta identificara algunos términos que son muy frecuentes en la colección de documentos recuperados y que en la representación podrían, en alguna medida, sobrestimar la similitud de los documentos. Para esto se debe explorar la máxima DF permitida y eliminar los términos que tienen una DF mayor a un determinado valor.

Para determinar estos parámetros, se iteró entre varias posibilidades. Para el parámetro M, se exploraron los resultados en el conjunto $\Omega_M = \{5, 10, 15, 20, 30, 40, 50, 75, 100\}$ y para cada uno de estos valores, se exploró el punto de corte de la DF de los términos, teniendo en cuenta su distribución a lo largo de la colección recuperada con base en μ_{DF_M} y σ_{DF_M}, que son respectivamente la media y la desviación estándar de la frecuencia de documentos para la colección recuperada tomando M documentos relevantes en la búsqueda. Se construye el conjunto $\Omega_{DF_M} = \{\mu_{DF_M} + 3\sigma_{DF_M}, ..., \text{máx}(DF_M)\}$, restringido a $|\Omega_{DF_M}| = 10$ como el conjunto para explorar el corte de DF.

Para la evaluación de los parámetros se usaron dos criterios principalmente: el primero corresponde al tiempo de ejecución del proceso, entendiéndose este, como el tiempo de la búsqueda y recuperación de documentos relacionados, más el tiempo de la representación distribucional de los términos, y adicional se tiene en cuenta el tiempo de procesamiento de los grupos. El segundo criterio de evaluación, es la pureza de los grupos encontrados por el

algoritmo. En este ejercicio se utilizó la agrupación espectral debido a que era el método que mejores resultados arrojaba en experimentaciones previas.
En la figura **4-2** se muestra la matriz de dispersión de los tiempos de ejecución medidos frente al número de elementos recuperados. En esta, se puede evidenciar los gráficos de dispersión bivariados para cada par de variables, donde se puede ver una clara relacion lineal entre el tiempo de construcción de la representación DTR y el número de resúmenes recuperados. Igualmente, existe una relación lineal positiva, entre el número de elementos recuperados y el tiempo de ejecución de los algoritmos de agrupación presentados con anterioridad.

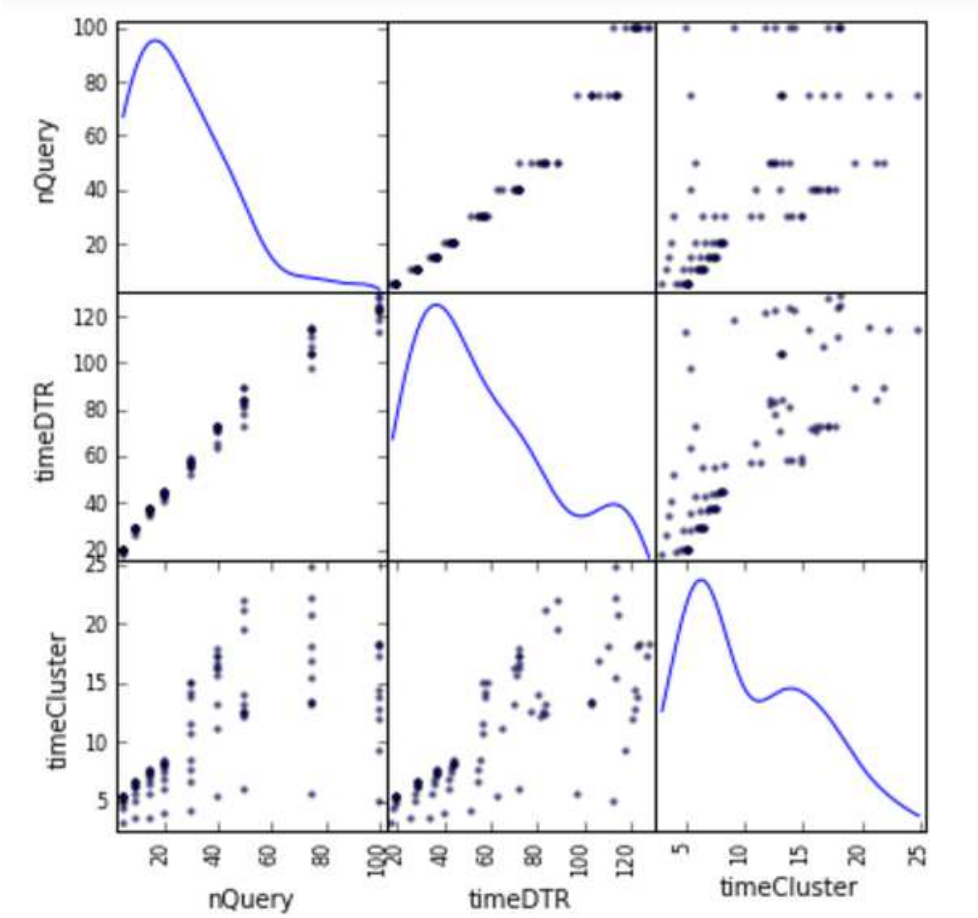


Figura 4-2.: Distribución de tiempos de cálculo DTR (*timeDTR*), tiempos de ejecución del algoritmo de agrupación(*timeCluter*) y número de documentos recuperados (*nQuery*).

El número de elementos recuperados de la consulta puede afectar el tiempo de ejecución de los procesos, si aumenta el número de documentos relacionados, mayor será el número de términos en la representación distribucional, por ende, mayor será el tiempo de construcción de la nueva representación y el tiempo de ejecución de los algoritmos. Evidencia de lo anterior se puede ver en la figura **4-3**, en donde a mayor número de resúmenes recuperados por consulta, mayor fueron los tiempos de ejecución de los diferentes experimentos.

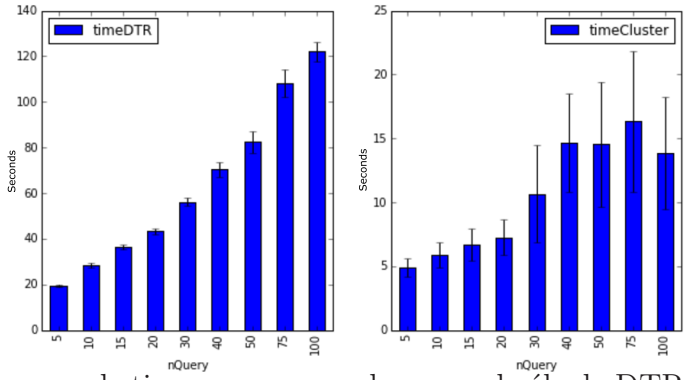


Figura 4-3.: Histograma de tiempos en segundos para el cálculo DTR (*timeDTR*) y tiempos del algoritmo de agrupación (*timeCluster*), dado el número de documentos recuperados (*nQuery*)

Con relación al segundo criterio de evaluación en los experimentos, la figura **4-4** presenta el índice de pureza encontrado para las diferentes configuraciones experimentales, iterando sobre los conjuntos Ω_M y Ω_{DF_M}, explicados con anterioridad. En la gráfica, podemos observar que la efectividad de los grupos decrece a medida que se recuperan más elementos por consulta (al incrementar M). También se puede observar que los mejores resultados, se obtienen al disminuír la máxima DF permitida.

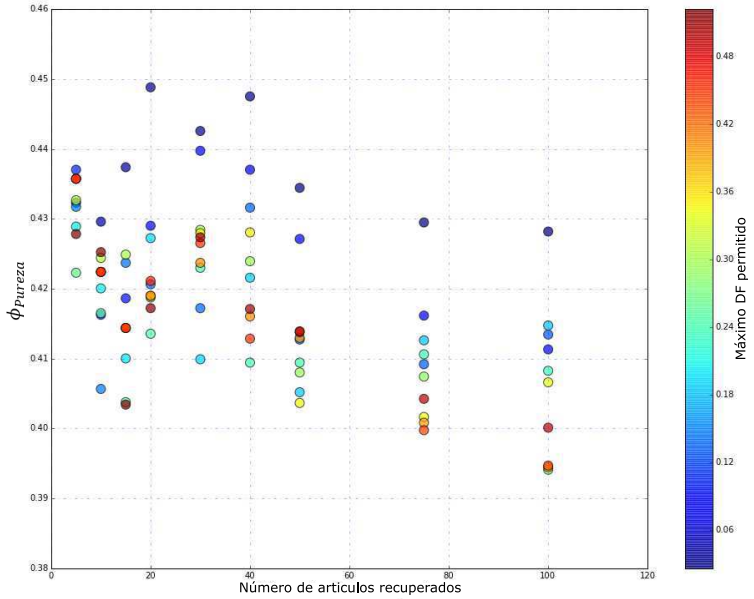


Figura 4-4.: Evaluación del índice de pureza de los grupos usando la representación TCOR, explorando diferentes números de artículos recuperados por consulta y diferentes puntos de corte DF para los términos. En azul, se encuentran los resultados usando los términos tales que $DF(t, D) \leq \mu_{DF_M} + 3\sigma_{DF_M}$, que implica eliminar un número mayor de términos; en rojo, se encuentran los resultados utilizando la cota $DF(t, D) \leq \text{máx}_{DF_M}$, lo cual implica eliminar un número menor de términos.

La disminución del índice de pureza al aumentar el número de elementos recuperados, podría deberse al efecto de la organización de los artículos en los sistemas de búsqueda Elsevier. En estos, los elementos son organizados por la relevancia del contenido con relación a la consulta lanzada. Por lo tanto, incluir un gran número de artículos en cada consulta, podría recuperar textos que no tengan mucha relación con el dominio/tema de la consulta e introducir sesgo en la representaciones semánticas de algunos términos.
Por otra parte, el aumento del índice causado por la disminución del umbral de DF, podría explicarse por la exclusión de algunos términos que son comunes en los resúmenes de los artículos científicos, como lo son: *addit, affect, aim, analysi, applic, approach, area, assess, associ, etc.*

4.2.3. Comparación métodos de agrupación

En la tabla **4-3** se presentan los resultados obtenidos para el conjunto de datos "*Scopus Títulos*", con los diferentes algoritmos de agrupación seleccionados, descritos en la tabla **3-3**. Los resultados en negrilla representan el mejor desempeño de la agrupación, relativo a cada uno de los índices utilizados. Tal como se mencionó en la sección 2.4.1, se espera que entre mejor sea la agrupación encontrada, lo índices de validez externa (ϕ_{ARI}, ϕ_{HOM}, ϕ_V, ϕ_{AMI} y ϕ_{MI}) incrementen. Estos índices, en su gran mayoría, se encuentran acotados entre 0,0 y 1,0, en donde 1 significa que la agrupación concuerda con las clases reales de los objetos.

Método	ϕ_{DB}	ϕ_{QE}	$\phi_{Silueta}$	ϕ_{ARI}	ϕ_{HOM}	ϕ_V	ϕ_{AMI}	ϕ_{MI}	ϕ_{pureza}
SPEC (tf_idf)	0.974	7.513	0.010	0.046	0.207	0.223	0.174	0.617	0.253
SPEC (TCOR)	0.405	4.540	-0.028	**0.226**	**0.408**	**0.415**	**0.385**	**1.221**	**0.448**
SPEC (W2V_G)	3.033	4.127	0.014	0.147	0.305	0.306	0.277	0.912	0.327
SPEC (W2V_W)	12.336	3.573	0.001	0.167	0.341	0.342	0.314	1.018	0.380
SPEC (TCOR *W2V_S)	0.383	3.014	0.049	0.200	0.397	0.404	0.372	1.185	0.392
KK (tf_idf)	0.979	9.704	0.007	0.064	0.176	0.177	0.142	0.524	0.238
KK (TCOR)	0.406	5.002	-0.041	0.197	0.383	0.393	0.358	1.145	0.391
KK (W2V_G)	3.006	4.025	-0.015	0.148	0.319	0.322	0.292	0.953	0.318
KK (W2V_S)	0.379	**2.934**	0.041	0.212	0.395	0.402	0.371	1.179	0.407
KK (W2V_W)	12.265	3.646	0.003	0.169	0.352	0.354	0.326	1.052	0.364
KK (TCOR * W2V_S)	0.379	**2.934**	0.041	0.212	0.395	0.402	0.371	1.179	0.407
NMF (TCOR)	0.338	13.628	-0.038	-0.001	0.033	0.033	-0.005	0.10	0.097
NMF (TCOR * W2V_S)	0.338	13.451	-0.041	0.001	0.040	0.040	0.002	0.12	0.101

Tabla 4-4.: Resultados obtenidos mediante el uso de TF-IDF original basado en el conjunto de datos de títulos Scopus, mejor entrenamiento TCOR, y los mejores resultados utilizando word2vec basado en la representación construída a partir de otras fuentes externas de información, presentadas en la sección 3.3(Wikipedia, Google News y artículos de Scopus)

Por otra parte, los índices de validez interna ϕ_{DB} y ϕ_{QE}, presentados en la sección 2.4.2, se basan en la medición de las similaridades intergrupos y/o intragrupos, la mejor partición de los objetos de análisis será aquella que logre minimizar las distancias intragrupos y por ende logren índices más bajos (cercanos a cero). Contrario a estos dos índices, el índice $\phi_{Silueta}$, toma valores cercanos a uno, cuando los grupos encontrados por el algoritmo están mejor definidos, es decir mayor diferencia promedio entre grupos y menores diferencias promedios intragrupo (homogeneidad dentro del grupo), por otra parte, toma valores cercanos a cero cuando se presentan mayores diferencias promedio intragrupo y menores diferencias promedios entre grupos.
Teniendo en cuenta la descripción de los índices, se puede observar que el rendimiento de los algoritmos usando la representación semántica W2V_S (representación Word2vec entrenado con las entradas de Scopus), presenta mejores resultados en comparación con los otros dos *CBOW* utilizados (Word2vec entrenado con Wikipedia y otro entrenando con entradas de Google News).
Según lo descrito en tabla **4-3**, la representación de TCOR tiene un rendimiento competitivo en comparación con W2V_S, de hecho, obtuvo el mejor rendimiento para dos de las tres medidas de validez interna. Podemos concluír que el desempeño de los métodos usando la representación distribucional, a partir de una colección de textos externos relacionados con la colección de documentos, obtiene mejor desempeño que los métodos tradicionales.

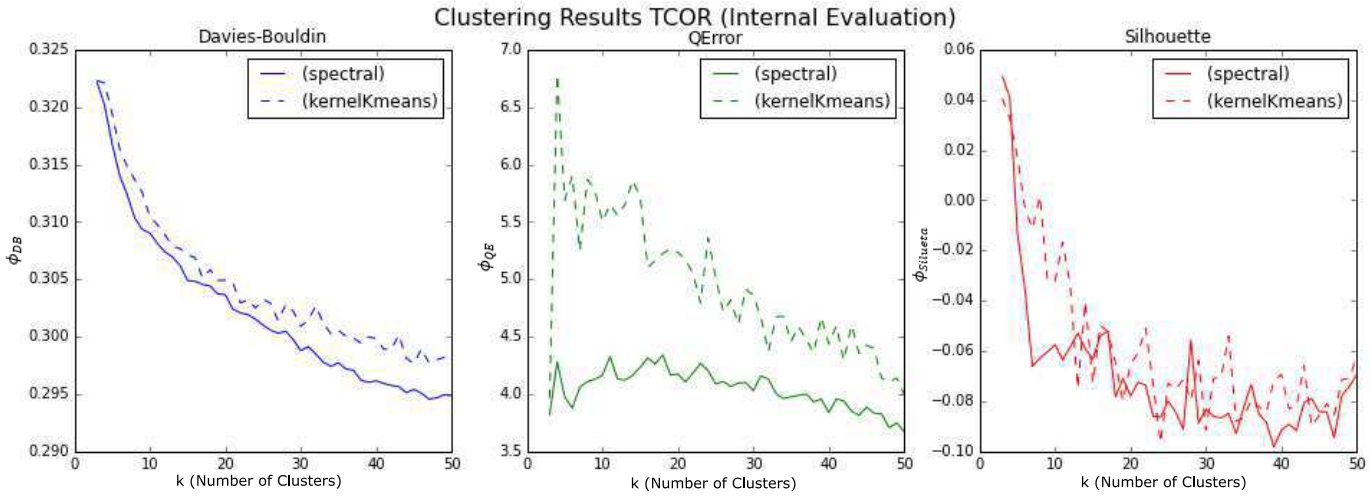


Figura 4-5.: Exploración del número de grupos (k) para *Scopus Títulos*, usando las medidas de validez interna: Índice Davies-Bouldin, Índice de cuantificación de error (*QError*), Coeficiente Silueta (*Silhouette*); y utilizando los métodos de agrupación seleccionados: espectral (*spectral*) y K-medias con Kernel (*kernelK-means*).

Los resultados previos se obtuvieron bajo una condición específica, comprobar la calidad de los grupos asumiendo el número de grupos encontrados igual al número existente de categorías, en nuestro caso 20 clases. A modo de ejemplo, las figuras **4-5** y **4-6**, representan la exploración de las medidas de validez internas y externas respectivamente, investigando si la calidad de los grupos permanece igual bajo diferentes condiciones, con diferentes algoritmos de agrupación y cuando se buscan estructuras con un número diferente de grupos, en nuestro caso explorando $K = \{1, \ldots, 50\}$ y asumiendo la mejor representación distribucional TCOR.

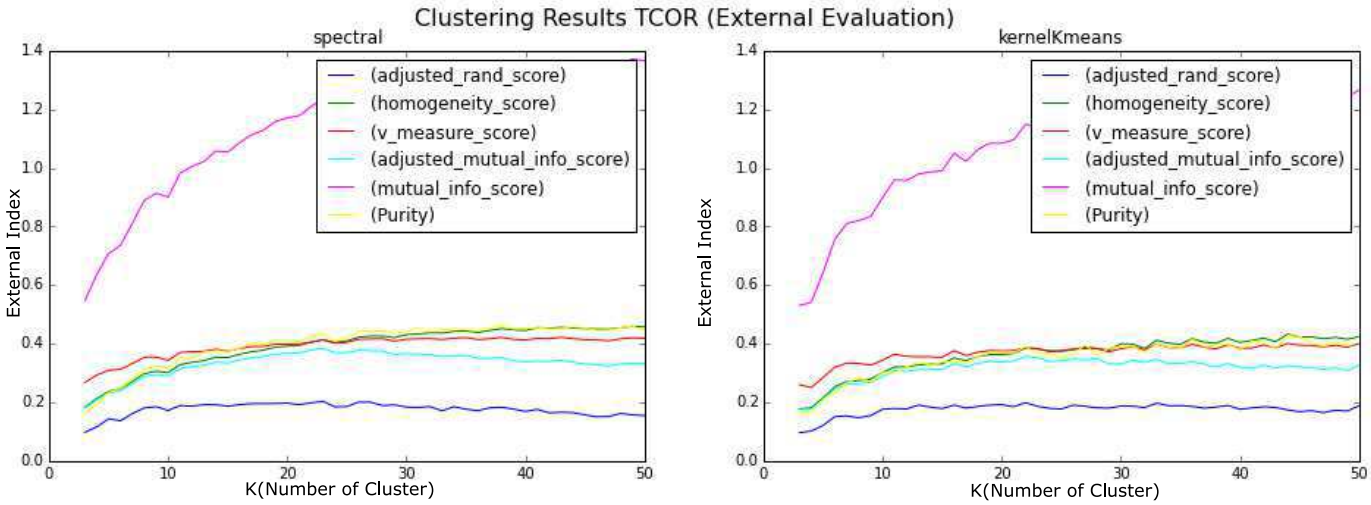


Figura 4-6.: Exploración del número de grupos (k) para *Scopus Títulos*, usando los índices de validez externos seleccionados (pureza, Rand ajustado, información mutua, información mutua ajustada y índice V) y utilizando los métodos de agrupación seleccionados: espectral (*spectral*) y K-medias con Kernel (*kernelKmeans*).

En la figura **4-5**, se puede observar que los índices ϕ_{DB} y ϕ_{QE}, disminuyen cuando mayor es el número de grupos explorados. Esto tiene sentido, dado que a mayor número de grupos pueden resultar particiones más homogéneas al interior de los mismos. Sin embargo, se puede observar una inconsistencia con los resultados del índice $\phi_{Silueta}$ el cual, contrario a lo esperado, decrece de forma monótona cuando crece el número de grupos explorados, idealmente este índice debe ser máximo al acercase al número de particiones reales o clases del conjunto de datos. Este comportamiento errático puede deberse en parte, a la alta dimensional del espacio de características, resultante de la representación distribucional aplicada, o deberse a la medida de similaridad seleccionada para el cálculo del índice.

Desde otro ángulo, en la figura **4-6** se puede ver el comportamiento de los índices de validez externa, como es de esperarse, se observa un leve incremento al acercarse al número de clases reales de los datos, pero se mantiene constante o decrece al aumentar el número de grupos en los algoritmos de agrupación, lo que podría indicar que un número mayor de grupos, arrojaría peores particiones, es decir agrupaciones que no concuerdan con las clases reales de los objetos.

4.3. Proyectos de investigación de la Universidad Nacional de Colombia (UN Títulos)

Este conjunto de datos contiene títulos originales de proyectos de investigación de la Universidad Nacional de Colombia, extraídos a partir de la base de datos de Vicerrectoría de Investigación para los periodos 2014-2015. Se cuenta con descripciones muy cortas de cada tema de investigación, la mayoría de ellos corresponden a los títulos de proyectos de investigación que adelanta la universidad, estos textos tienen una longitud promedio de 20,98 términos. La base contiene en total 3718 textos no etiquetados almacenados en formato "XLSX".

Expresión regular	Expresión regular	Expresión regular
GRUPO DE INVESTIGACI(Ó|O)N(EN)?:?	TESIS DE MAESTR(I|Í)A:?	TESIS DE DOCTORADO:?
TRABAJO DE GRADO:?	CONTRAPARTIDA DIPAL:?	COLCIENCIAS(\\s+\\d+)?:? (\\s+-\\s+\\d{4})?
PROGRAMA DE FORTALECIMIENTO (PARA EL?|DEL|EN EL ÁREA DE|EN)?	PROGRAMA DE INVESTIGACI(Ó|Ó)N:?	FORTALECIMIENTO DE LA INVESTIGACI(Ó|Ó)N(:| EN|EN EL (Á|Á)REA DE)?
(PROGRAMA)?\\s?(DE)?J(Ó|O)VENES INVESTIGADORES\\s?(E INNOVADORES)?(VIRGINIA GUTI(E|É)RREZ DE PINEDA)?	CONTRAPARTIDA VICERRECTORIA DE INVESTIGACI(Ó|O)N:?	- SEDE ORINOQUIA
CTO\\. \\d+	- FUENTE INTERNA	PROCESOS DE INVESTIGACI(Ó|O)N(-|:)?
CONV\\. NACIONAL	MOD\\. \\d	FORTALECIMIENTO DE LA VISIBILIDAD DE LA PRODUCCI(Ó|O)N ACAD(É|E)MICA

Tabla 4-5.: Primer grupo de expresiones regulares para la depuración de la base *UN Títulos* (Parte I).

Adicional a las actividades de pre-procesamiento de texto presentadas en la sección 3.1, se han realizado esfuerzos sistemáticos para depurar los títulos de los proyectos de investigación, eliminando algunos registros relacionados con la asistencia a eventos o conferencias de algunos investigadores, así como también se eliminan términos específicos que se repiten a lo largo de los textos, mediante el uso de expresiones regulares utilizando operaciones de búsqueda y reemplazo. En la tabla **4-5**, se encuentran enunciadas algunas de las expresiones regulares que fueron anotadas y extraídas de una revisión del conjunto de datos. Las demás expresiones regulares son descritas en más detalle en el Anexo A.
Eliminar del texto estas expresiones es relevante, debido a que pueden introducir relaciones

en los textos que no son propiamente debido al campo del conocimiento que abarca la investigación. Las relaciones encontradas pueden ser espurias y podrían ser causadas por fragmentos de los textos que se repiten en los títulos de investigación (por ejemplo, tesis de maestría, tesis de doctorado, trabajo de grado, etc.).
Después de la etapa de depuración de los textos por medio de las expresiones regulares mencionadas con anterioridad, se procedió, según lo descrito en la sección 3.1, con la traducción de los textos cortos a un lenguaje base, en este caso inglés, usando Google Translate API[3]. Luego de la traducción de los textos, se elimina las palabras vacías del idioma inglés y posteriormente se realiza las consultas a la base de datos de Elsevier, de acuerdo con lo descrito en la sección 3.2.
Después de los filtros y la depuración de los datos se eliminaron 1580 textos, debido a que no contenían ningún termino después de la depuración, es decir, la base final está compuesta por 2138 textos. La tabla **4-6** relaciona la cantidad de documentos recuperados de consultas realizadas en el sistemas Scopus, así como una breve descripción del conjunto de datos que se recuperó luego de lanzar las consultas construidas.

Característica	Conteo
No. de documentos recuperados	24945
No. de términos	103174
Promedio del No. de documentos	73,81
Desviación del No. de documentos	30,32

Tabla 4-6.: Descripción de artículos recuperados en la colección *UN Títulos*.

4.3.1. Depuración de términos

Como se describió en la sección 4.2.2, el número de documentos que son recuperados de las consultas y la *frecuencia máxima de documentos* de los términos (DF, por sus siglas en ingles), resultan ser dos parámetros relevantes en la representación distribucional del conjunto de datos, cuando se usan las representaciones TCOR o DOR. Para evaluar estos parámetros en el conjunto de datos *UN Títulos*, el gráfico **4-7** presenta en la parte superior la distribución en escala logarítmica del número de términos de acuerdo con el número de documentos en el que el término aparece, por otra parte en la figura inferior se presenta la frecuencia de términos de acuerdo con su DF. Estas gráficas son utilizadas para identificar algunos términos que son muy frecuentes en la colección de documentos recuperados y que

[3] Google Translate API puede convertir dinámicamente los textos entre ciertos pares de idiomas, esta API está disponible en https://cloud.google.com/translate/docs/

en la representación podrían, en alguna medida, sobrestimar la similitud de los documentos, para esto se debe explorar máximo DF permitido y eliminar los términos que tienen un DF mayor a un determinado valor.

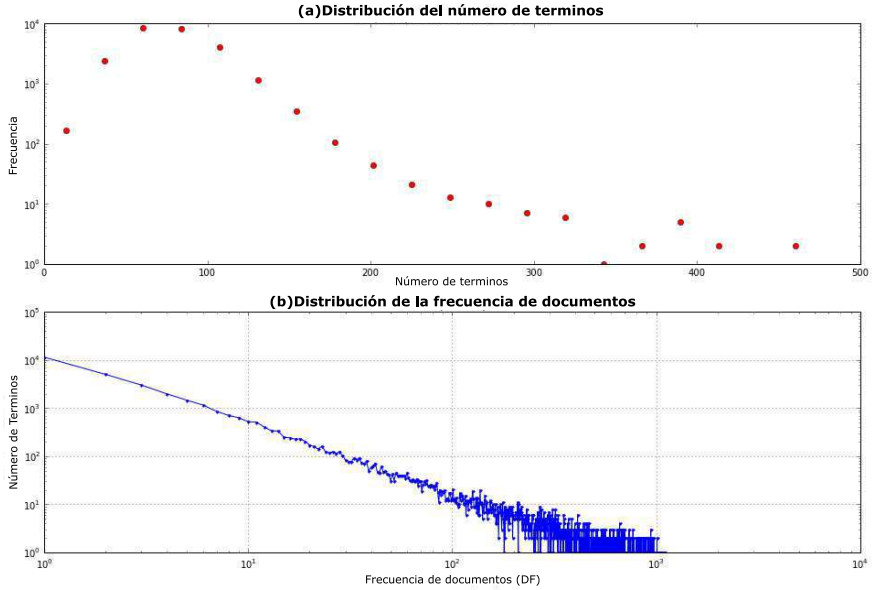


Figura 4-7.: (a): Presenta la distribución del número términos de un documento. (b): Muestra la distribución de los terminos de acuerdo a las frecuencias de documentos (DF, por sus siglas en inglés).

La gráfica 4-8 presenta el histograma de DF y el histograma del número de documentos, de los términos que componen la colección de resúmenes del proceso de recuperación de información, en esta se observa que la mayoría de los términos que componen la colección tienen una frecuencia de documentos muy baja, de al rededor de un solo documento. Con base en estos gráficos y con los hallazgos encontrados en la sección 4.2.2 se decido utilizar un punto de corte máximo de DF igual 0,07, eliminando así al rededor de 230 palabras que tienen un DF mayor a ese criterio y que son palabras comunes en los resumes capturados en el proceso de recuperación de información.

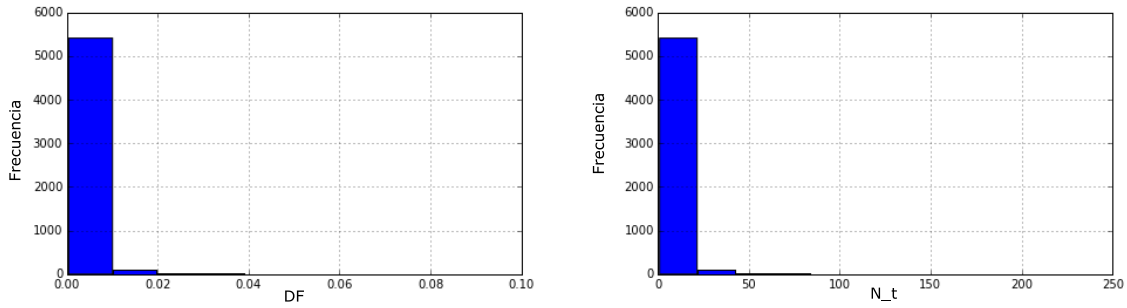


Figura 4-8.: (Izquierda) Histograma de la *frecuencia de documentos "DF"*. (Derecha) Histograma del *número de términos en un documento "N_t"*.

4.3.2. Comparación técnicas de agrupamiento

A manera de ejemplo, la figura **4-9** presenta la exploración de las medidas de validez interna del conjunto *UN Títulos*, investigando si la calidad de los grupos permanece igual bajo diferentes condiciones de experimentación: con diferentes algoritmos de agrupación y cuando se buscan estructuras con un número diferente de grupos, en nuestro caso explorando $K = \{1, \ldots, 50\}$ y asumiendo la representación distribucional TCOR.

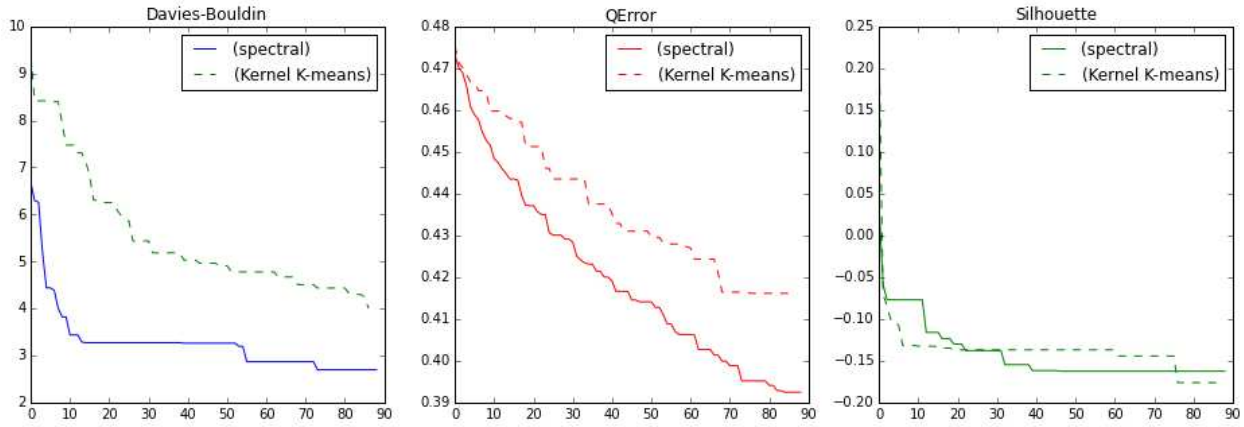


Figura 4-9.: Exploración del número de grupos (k) conjunto de datos *UN Títulos*, usando las medidas de validez interna: Índice Davies-Bouldin(ϕ_{DB}), Índice de cuantificación de error(ϕ_{QError}), Coeficiente Silueta($\phi_{Silhouette}$); y utilizando los métodos de agrupación seleccionados: espectral (*spectral*) y K-medias con Kernel (*kernelKmeans*).

En la figura **4-9**, el índices Davies-Bouldin (ϕ_{DB}) y el índice de cuantificación de error (ϕ_{QE}), disminuyen cuando mayor es el número de grupos explorados, esto tiene sentido, dado que, a mayor número de grupos, el método de agrupación puede resultar particiones más homogéneas al interior de los grupos, es decir, textos más parecidos entre sí.

Método	ϕ_{DB}	ϕ_{QE}	$\phi_{Silueta}$
SPEC (tf_idf)	0.405	4.386	-0.008
SPEC (TCOR)	0.974	7.513	0.010
SPEC (DOR)	0.405	4.386	-0.008
KK (W2V_S)	0.413	4.050	0.007
KK (W2V_G)	0.433	4.159	-0.0004
KK (W2V_W)	0.418	4.406	-0.002

Tabla 4-7.: Resultados obtenidos en la base *UN Títulos*, mejor entrenamiento TCOR, y el mejor resultado utilizando word2vec basado en la representación construida de otras fuentes externas (Wikipedia, Google News y artículos de Scopus).

Sin embargo, existe un comportamiento errático con el tercer índice de validez interna, el índice Silueta el cual decrece al aumentar el número de grupos explorados, contrario al comportamiento esperado, el cual siguiere que el índice aumenta para así poder encontrar el número de grupos existentes en el conjunto de datos. El anterior comportamiento puede ser causado por la alta dimensionalidad del espacio de características o puede ser causado por la medida de similitud seleccionada para el cálculo de este índice.
La figura **4-9** apoya en gran medida la determinación del número de grupos que se pueden seleccionar, en esta figura se observa un codo al acercarse a una cantidad de 20 grupos, que supondríamos son el número de agrupaciones latentes que existen en el conjunto de datos, y por esto se selecciona una cantidad de 20 grupos como insumo en los algoritmos de agrupación. Cabe anotar que se encontraron comportamientos equivalentes, cuando se probaron los diferentes métodos de representación distribucional seleccionados y descritos en la sección 2.2.
Como se observa en la tabla **4-7**, se encontraron conclusiones similares a los experimentos realizados con el conjunto de datos *Scopus Títulos*. En el caso de la generación de clústeres temáticos utilizando la base de datos *UN Títulos*, se demuestra que los algoritmos propuestos basados en las representaciones distribucionales de términos (TCOR), presentan un buen desempeño en las agrupaciones de textos cortos, incluso frente al enfoque de bolsa de palabras continuas o comúnmente conocido como Word2Vec. Este experimento también arroja conclusiones sobre el mejor algoritmo de agrupación, el mejor escenario de los resultados obtenidos fue utilizando la representación TCOR y usando el algoritmo de agrupación espectral.

Breve síntesis del capítulo En este capítulo se presentaron los resultados experimentales que validan el método de agrupación para textos con extensión corta propuesto en este trabajo de investigación. Se presentaron dos escenarios de experimentación, el primero en el cual se construyó un conjunto de datos en el que se conoce la verdadera categoría de cada texto. Con este conjunto se evaluó principalmente la consistencia externa de la agrupación, es decir, que los grupos encontrados correspondieran a las clases originales de los textos. El segundo conjunto de datos corresponde a los títulos de proyectos de investigación de la Universidad Nacional de Colombia. Los experimentos y los resultados de la validación en la agrupación de los objetos, tiempo de procesamiento y número de grupos resultantes, que fueron presentados en este capítulo, se concentraron principalmente en apoyar el objetivo número 3 de evaluar el método desarrollado en un conjunto de datos concreto relacionado con la tarea de detección de temas latentes en una colección de descripciones cortas de proyectos de investigación de la Universidad Nacional de Colombia.

5. Conclusiones y Trabajo Futuro

5.1. Conclusiones

En esta investigación, hemos propuesto un método para la agrupación de texto con extensión muy corta, basado en la representación distribucional de los términos (DTR), utilizando una fuente de información externa, en nuestro caso, utilizando las bases de datos de Elsevier. Dicha representación distribucional tiene como objetivo la obtención de conceptos/términos que pueda ampliar la representación semántica de los términos que conforman los textos originales, para lograr capturar mejor la similitud de los textos, y de esa manera poder mejorar los resultados de las técnicas de agrupación.

Se realizó la evaluación experimental con dos conjuntos de datos diferentes: Uno en donde se conoce las etiquetas reales de los datos, y el cual fue construído a partir de consultas enviadas a la API de Elsevier Scopus; el otro conjunto de datos, consiste en la colección de títulos depurados de los proyectos de investigación de la Universidad Nacional de Colombia. Las principales conclusiones que se obtuvieron de los experimentos realizados son:

- Tal como se mencionó en el capítulo 4, como estudio experimental en los dos conjuntos de datos, se concluyo que es posible lograr un rendimiento significativamente mejor que los métodos tradicionales para la agrupación de textos. En comparación con la representación *tf-idf* (conocida como BOW, Bolsa de palabras por sus siglas en inglés), se logra una mejora de alrededor del 70 % en los índices de validez externa de los grupos. También se puede concluir que existe una mejora relativa de alrededor del 10 % con relación a los métodos de bolsa de palabras continuas (CBOW, por sus siglas en inglés), entrenados con el popular método word2Vec, el cual genera representaciones usadas en métodos de vanguardia y que han mostrado buenos resultados en diferentes problemáticas relacionadas con minería de textos.

- Se demostró que el uso de DTRs (presentados en la sección 2.2) para la ampliación semántica de los textos, permite obtener resultados aceptables, considerando que encuentra relaciones entre los textos dada una representación más amplia de los términos que los componen resolviendo el problema de baja frecuencia de los términos y las representaciones dispersas de los mismos.

- En la base de datos de los textos de la Universidad Nacional, se encontró que los

métodos propuestos en este trabajo de investigación, sobre el agrupamiento de textos cortos basados en el uso de DTRs, mostraron resultados igualmente útiles que en los conjuntos de prueba que fueron construidos, permitiendo encontrar grupos temáticos en una colección de descripciones cortas de proyectos de investigación. No obstante las reglas de depuración establecidas y descritas en las sección 4.3 y A resultan ser muy estrictas eliminando al rededor del 40 % de los registros, esto podría proponer una acción de mejora al proceso de recolección de información de la Vicerrectoría de Investigación sobre las temáticas investigadas en la Universidad.

- La metodología que se presento en la sección 3, brinda una nueva herramienta para las consultas sobre los recursos web de Elsevier, específicamente haciendo uso de la API de Elsevier Scopus y la API de búsqueda de ScienceDirect. La construcción de este programa en Python permite la captura de la información bibliográfica de muchos artículos disponibles en estas dos bases de datos. Una estructuración de esta información podría apoyar posteriores investigaciones sobre análisis bibliométricos de artículos científicos.

- Otro resultado relevante de esta investigación, fue la construcción de nuestra propia representación de bolsa de palabras continuas (CBOW), la cual se construye mediante el uso de resúmenes recuperados de las diferentes APIs de Elsevier. Este conjunto de datos compuesto por 1,045,144 diferentes artículos científicos, junto con la información recuperada de acuerdo con lo descrito en la sección 3.2 puede ser utilizada en la representación semantica de los términos de documentos científicos en futuros problemas de minería de textos.

Entre las limitaciones encontradas en el estudio se puede indicar que, si bien, el método propuesto permite lograr los propósitos de la investigación encontrando grupos temáticos de textos con extensión muy corta, al ser una primera aproximación al problema de estudio, no cuenta con bases de datos del todo estandarizadas y con el conocimiento de cuáles son las verdaderas clases, es decir, los temas latentes que se encuentran dentro de los conjuntos de experimentación.

El idioma resulta ser una de las limitaciones del estudio, debido a que la mayoría de los textos con los que cuenta la base de datos de la vicerrectoría de investigación son en español, a diferencia de los recursos externos que están disponibles en la web: como las entradas de Wikipedia, la información de los artículos de las bases de datos de Elsevier y las descargas de Google News, las cuales se encuentran en inglés, este aspecto limita en algunas ocasiones la expansión distribucional de los términos en español.

Otra limitación del estudio se encuentra en el tipo de textos que se obtienen en el proceso de recuperación de información (descrito en la sección 3.2), en vista de que la API de Elsevier Scopus y la API de búsqueda de ScienceDirect, solo permiten recuperar como máximo los resúmenes de los artículos. Esto implica que las representaciones distribucionales de los

términos se restringen a las relaciones entre las palabras que se encuentran en los resúmenes científicos, perdiendo información importante contenida en los textos completos de los artículos. Estas cuestiones de interés precisarían un tratamiento complementario a los textos y/o utilizar una metodología de recuperación de información diferente a la que fue propuesta en esta investigación.

5.2. Trabajo Futuro

Existen varios componentes relacionados con el método propuesto que podrían mejorar los resultados encontrados, para esto es necesario realizar otras exploraciones e incluír otro tipo de información externa, que podría capturar de una mejor forma el significado semántico y sintáctico de los textos de extensión corta. Los siguientes son algunos componentes claves detectados en el desarrollo de la investigación, considerados como trabajo futuro:

- Se propone la reorganización de los campos capturados en la base de registro de la Vicerrectoría de Investigación de la Universidad Nacional de Colombia y la ampliación de esta base con información como: el tipo de apoyo (si es participación a un conferencia, congreso, u otro tipo de actividad académica), el dominio de la investigación (entendido como el campo de la ciencia con el que se puede caracterizar la investigación) y la descripción del proyecto o actividad (la cual debería contener una cantidad mínima de palabras). La inclusión de una descripción de los proyectos facilitaría posteriores análisis con la información textual contenida en esta base de datos.

- Como se mencionó en la sección 3.1, una de las decisiones para construír la representación distribucional, fue la traducción de los títulos de los proyectos de investigación de español a inglés. La traducción de los textos podría introducir un sesgo en la representación de las palabras cuya traducción no fue la correcta y, por ende, la representación distribucional construída estaría sesgada. La construcción de la expansión distribucional (DOR, TCOR o Word2Vec) con otras fuentes textuales en español, podría generar una mejor representación de los textos y por consiguiente una mejor agrupación de los mismos. Entre las fuentes de información identificadas en español se encuentran: los textos completos de los trabajos de investigación de la universidad, como lo son monografías, tesis, revistas, etc.; y la información bibliográfica de artículos en español, título, resumen, palabras claves, entre otros elementos.

- Estudiar el impacto de la inclusión de otras distancias o distribuciones, para relacionar los textos de extensión corta, incluyendo por ejemplo divergencias (como la distancia de Kullback-Leibler, la distancia de Mahalanobis) o medidas de similitud de los textos, diferentes a la similitud coseno ampliamente usada en esta experimentación. La utilización de estas medidas podrían generar mejores resultados en los métodos de agrupación seleccionados.

- Con relación al método de k-medias usando la transformación kernel, la exploración de diferentes funciones kernel, las cuales proporcionan una forma de obtener relaciones de los datos, podría requerir una exploración y un análisis más profundo, y así seleccionar el mejor método kernel de acuerdo con las características de los textos que se están trabajando.

Bibliografía

[1] Bashar Al-Shboul and SH Myaeng. Wikipedia-based query phrase expansion in patent class search. *Information Retrieval*, 17(5-6):430–451, nov 2013. ISSN 1386-4564. doi: 10.1007/s10791-013-9233-4. URL http://link.springer.com/article/10.1007/s10791-013-9233-4.

[2] Timothy Baldwin and Marco Lui. Language identification: The long and the short of the matter. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 229–237. Association for Computational Linguistics, 2010.

[3] Elisa Bellotti, Luka Kronegger, and Luigi Guadalupi. The evolution of research collaboration within and across disciplines in italian academia. *Scientometrics*, 109(2):783–811, 2016.

[4] Pavel Berkhin. A survey of clustering data mining techniques. In *Grouping multidimensional data*, pages 25–71. Springer, 2006.

[5] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*, 2016.

[6] Juan Manuel Cabrera, Hugo Jair Escalante, and Manuel Montes-Y-Gómez. Distributional term representations for short-text categorization. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 7817 LNCS, pages 335–346, 2013. ISBN 9783642372551. doi: 10.1007/978-3-642-37256-8_28.

[7] Leticia Cagnina, Marcelo Errecalde, Diego Ingaramo, and Paolo Rosso. An efficient Particle Swarm Optimization approach to cluster short texts. *Information Sciences*, 265:36–49, may 2014. ISSN 00200255. doi: 10.1016/j.ins.2013.12.010. URL http://www.sciencedirect.com/science/article/pii/S0020025513008542.

[8] C. Carretero-Campos, P. Bernaola-Galván, a.V. Coronado, and P. Carpena. Improving statistical keyword detection in short texts: Entropic and clustering approaches. *Physica A: Statistical Mechanics and its Applications*, 392(6):1481–1492, mar 2013. ISSN 03784371. doi: 10.1016/j.physa.2012.11.052. URL http://linkinghub.elsevier.com/retrieve/pii/S0378437112010175.

[9] Xiaohui Cui, Thomas E Potok, and Paul Palathingal. Document clustering using particle swarm optimization. In *Swarm Intelligence Symposium, 2005. SIS 2005. Proceedings 2005 IEEE*, pages 185–191. IEEE, 2005.

[10] David L Davies and Donald W Bouldin. A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*, (2):224–227, 1979.

[11] Inderjit S Dhillon. Kernel k-means , Spectral Clustering and Normalized Cuts. 2004.

[12] Chris Ding, Xiaofeng He, and Horst D Simon. On the equivalence of nonnegative matrix factorization and spectral clustering. In *Proceedings of the 2005 SIAM International Conference on Data Mining*, pages 606–610. SIAM, 2005.

[13] Marcelo Errecalde, Diego Ingaramo, and Paolo Rosso. ITSA*: An effective iterative method for short-text clustering tasks. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 6096 LNAI, pages 550–559, 2010. ISBN 3642130216. doi: 10.1007/978-3-642-13022-9_55.

[14] Zhou Faguo, Zhang Fan, Yang Bingru, and Yu Xingang. Research on Short Text Classification Algorithm Based on Statistics and Rules. *2010 Third International Symposium on Electronic Commerce and Security*, (2):3–7, jul 2010. doi: 10.1109/ISECS.2010.9. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5557448.

[15] Adil Fahad, Najlaa Alshatri, Zahir Tari, Abdullah Alamri, Ibrahim Khalil, Albert Y Zomaya, Sebti Foufou, and Abdelaziz Bouras. A survey of clustering algorithms for big data: Taxonomy and empirical analysis. *IEEE transactions on emerging topics in computing*, 2(3):267–279, 2014.

[16] Python Software Foundation. Python language reference, version 2.7. http://www.python.org, 1990–2017.

[17] Joydeep Ghosh and Alexander Strehl. Similarity-based text clustering: a comparative study. In *Grouping Multidimensional Data*, pages 73–97. Springer, 2006.

[18] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. Unsupervised learning. In *The elements of statistical learning*, pages 485–585. Springer, 2009.

[19] Xia Hu, Nan Sun, Chao Zhang, and Tat-seng Chua. Exploiting Internal and External Semantics for the Clustering of Short Texts Using World Knowledge. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 919–928, 2009. ISBN 9781605585123. doi: http://doi.acm.org/10.1145/1645953.1646071. URL http://doi.acm.org/10.1145/1645953.1646071.

[20] Xia Hu, Nan Sun, Chao Zhang, and Tat-seng Chua. Exploiting Internal and External Semantics for the Clustering of Short Texts Using World Knowledge. *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 919–928, 2009. ISSN 1605585122. doi: http://doi.acm.org/10.1145/1645953.1646071. URL http://doi.acm.org/10.1145/1645953.1646071.

[21] Lan Huang. *Concept-based text clustering*. PhD thesis, University of Waikato, 2011.

[22] Jingu Kim and Haesun Park. Sparse Nonnegative Matrix Factorization for Clustering. *Science*, pages 1–15, 2006. URL http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.131.4302&rep=rep1&type=pdf.

[23] Da Kuang, Haesun Park, and Chris Ding. Symmetric nonnegative matrix factorization for graph clustering. *International Conference on Data Mining*, pages 494–505, 2012. URL http://siam.omnibooksonline.com/2012datamining/data/papers/130.pdf.

[24] Oh-Woog Kwon and Jong-Hyeok Lee. Text categorization based on k-nearest neighbor approach for Web site classification. *Information Processing & Management*, 39(1):25–44, jan 2003. ISSN 03064573. doi: 10.1016/S0306-4573(02)00022-5. URL http://www.sciencedirect.com/science/article/pii/S0306457302000225.

[25] Alberto Lavelli, Fabrizio Sebastiani, and Roberto Zanoli. Distributional term representations. In *Proceedings of 13th ACM*, page 615, 2004. ISBN 1581138741. doi: 10.1145/1031171.1031284. URL http://doi.acm.org/10.1145/1031171.1031284.

[26] Alberto Lavelli, Fabrizio Sebastiani, and Roberto Zanoli. Distributional term representations: an experimental comparison. In *Proceedings of the thirteenth ACM international conference on Information and knowledge management*, pages 615–624. ACM, 2004.

[27] Chien-Liang Liu, Tao-Hsing Chang, and Hsuan-Hsun Li. Clustering documents with labeled and unlabeled documents using fuzzy semi-Kmeans. *Fuzzy Sets and Systems*, 221:48–64, jun 2013. ISSN 01650114. doi: 10.1016/j.fss.2013.01.004. URL http://www.sciencedirect.com/science/article/pii/S0165011413000213.

[28] Chien-Liang Liu, Wen-Hoar Hsaio, Chia-Hoang Lee, and Chun-Hsien Chen. Clustering tagged documents with labeled and unlabeled documents. *Information Processing & Management*, 49(3):596–606, may 2013. ISSN 03064573. doi: 10.1016/j.ipm.2012.12.004. URL http://www.sciencedirect.com/science/article/pii/S0306457312001422.

[29] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[30] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In

C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc., 2013. URL http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf.

[31] Vikramjit Mitra, Chia-Jiu Wang, and Satarupa Banerjee. Text classification: A least square support vector machine approach. *Applied Soft Computing*, 7(3):908–914, jun 2007. ISSN 15684946. doi: 10.1016/j.asoc.2006.04.002. URL http://www.sciencedirect.com/science/article/pii/S156849460600038X.

[32] Yuan PING, Ya-jian ZHOU, Chao XUE, and Yi-xian YANG. Efficient representation of text with multiple perspectives. *The Journal of China Universities of Posts and Telecommunications*, 19(1):101–111, feb 2012. ISSN 10058885. doi: 10.1016/S1005-8885(11)60234-3. URL http://www.sciencedirect.com/science/article/pii/S1005888511602343.

[33] D Pinto. Analysis of narrow-domain short texts clustering. (September), 2007. URL http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Analysis+of+narrow-domain+short+texts+clustering#0.

[34] D. Pinto, P. Rosso, and H. Jimenez-Salazar. A Self-enriching Methodology for Clustering Narrow Domain Short Texts. *The Computer Journal*, 54(7):1148–1165, 2011. ISSN 0010-4620. doi: 10.1093/comjnl/bxq069. URL http://comjnl.oxfordjournals.org/cgi/doi/10.1093/comjnl/bxq069.

[35] DAVID EDUARDO PINTO AVENDAÑO. On Clustering and Evaluation of Narrow Domain Short-Test Corpora, jul 2008. URL http://www.tesisenred.net/handle/10803/22037.

[36] Yonggang Qiu and Hans-Peter Frei. Concept based query expansion. In *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, volume 11, pages 160–169, 1993. ISBN 0897916050. doi: 10.1145/160688.160713. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2873540&tool=pmcentrez&rendertype=abstract.

[37] Aniket Rangrej. Comparative Study of Clustering Techniques for Short Text Documents. *Media*, pages 111–112, 2011. doi: http://doi.acm.org/10.1145/1963192.1963249. URL http://portal.acm.org/citation.cfm?doid=1963192.1963249.

[38] Radim Rehurek and Petr Sojka. Software framework for topic modelling with large corpora. In *In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Citeseer, 2010.

[39] RELX Group. Relx group company reports, 2017. URL http://www.relx.com/investorcentre/reports%202007/Documents/2016/relxgroup_ar_2016.pdf. Accedido 01-10-2013.

[40] Andrew Rosenberg and Julia Hirschberg. V-measure: A conditional entropy-based external cluster evaluation measure. In *EMNLP-CoNLL*, volume 7, pages 410–420, 2007.

[41] Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

[42] Sepideh Seifzadeh, Ahmed K Farahat, Mohamed S Kamel, and Fakhri Karray. Short-Text Clustering using Statistical Semantics. 15:805–810, 2015. doi: 10.1145/2740908.2742474. URL http://dx.doi.org/10.1145/2740908.2742474.

[43] Prajol Shrestha, Christine Jacquin, and Béatrice Daille. Clustering short text and its evaluation. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 7182 LNCS(PART 2):169–180, 2012. ISSN 03029743. doi: 10.1007/978-3-642-28601-8_15.

[44] Robert Tibshirani, Guenther Walther, and Trevor Hastie. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):411–423, 2001.

[45] Mireya Tovar and V Blanca. An Iterative Clustering Method for the XML-Mining Task of the INEX 2010. *Lecture Notes in Computer Science*, pages 377–382, 2011.

[46] Krutika Verma, Mukesh K Jadon, and Arun K Pujari. Information Retrieval Technology: 9th Asia Information Retrieval Societies Conference, AIRS 2013, Singapore, December 9-11, 2013. Proceedings. chapter Clustering, pages 145–155. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013. ISBN 978-3-642-45068-6. doi: 10.1007/978-3-642-45068-6_13. URL http://dx.doi.org/10.1007/978-3-642-45068-6_13.

[47] Nguyen Xuan Vinh, Julien Epps, and James Bailey. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research*, 11(Oct):2837–2854, 2010.

[48] Peng Wang, Bo Xu, Jiaming Xu, Guanhua Tian, Cheng Lin Liu, and Hongwei Hao. Semantic expansion using word embedding clustering and convolutional neural network for improving short text classification. *Neurocomputing*, 174:806–814, 2016. ISSN 18728286. doi: 10.1016/j.neucom.2015.09.096.

[49] Rui Xu and Donald Wunsch. Survey of clustering algorithms. *IEEE Transactions on neural networks*, 16(3):645–678, 2005.

[50] Xiaohui Yan, Jiafeng Guo, Shenghua Liu, Xue-qi Cheng, and Yanfeng Wang. Clustering short text using ncut-weighted non-negative matrix factorization. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 2259–2262. ACM, 2012.

[51] Ka Yee Yeung and Walter L Ruzzo. Details of the adjusted rand index and clustering algorithms, supplement to the paper an empirical study on principal component analysis for clustering gene expression data. *Bioinformatics*, 17(9):763–774, 2001.

[52] Lingling Yuan. An effective Chinese short message texts clustering algorithm based on the ward's method. *2011 2nd International Conference on Artificial Intelligence, Management Science and Electronic Commerce (AIMSEC)*, pages 1897–1899, aug 2011. doi: 10.1109/AIMSEC.2011.6010901. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6010901.

[53] Wen Zhang, Taketoshi Yoshida, and Xijin Tang. Text classification based on multi-word with support vector machine. *Knowledge-Based Systems*, 21(8):879–886, dec 2008. ISSN 09507051. doi: 10.1016/j.knosys.2008.03.044. URL http://www.sciencedirect.com/science/article/pii/S0950705108000968.

[54] Wen Zhang, Taketoshi Yoshida, Xijin Tang, and Qing Wang. Text clustering using frequent itemsets. *Knowledge-Based Systems*, 23(5):379–388, jul 2010. ISSN 09507051. doi: 10.1016/j.knosys.2010.01.011. URL http://www.sciencedirect.com/science/article/pii/S0950705110000134.

A. Anexo: Lista de expresiones regulares

En este anexo se encuentran los diferentes grupos formados, que fueron consideradas para el proceso eliminación y depuración de las cadenas de texto, mediante operaciones de búsqueda y reemplazo. Los grupos fueron conformados por la cantidad de expresiones y paréntesis que se incluyen en las diferentes expresiones regulares.

Expresión regular	Expresión regular	Expresión regular
(EVENTO(-|))?SEMINARIO INTERNACIONAL (SOBRE)?	DIPLOMADO(EN)?:?	CURSO INTERNACIONAL (DE)?
MOVILIDAD DIB- \\d{4} M3	CONVENIO DE COOPERACI(Ó|O)N ENTRE	CURSO DE EDUCACI(Ó|O)N CONTINUA Y PERMANENTE:?
- BP	- AMVA	AÑO \\d{4}\\s+-?
^\\d{4}-EXT(-COLC)?	AÑO\\s+\\d{4}\\s?-\\s? (EXT|INV|COLC)	- SEMESTRE (\\w{2}|\\d|\\d{2})((DE)|-)\\d{4}
CURSO DE EDUCACI(Ó|O)N CONTINUA Y PERMANENTE:	CURSO DE ACTUALIZACI(Ó|O)N:?	(CONTRATO |COTRAPARTIDA)?(COLCIENCIAS)?RC\\.?(\\s+)?(NO\\.\\s+)? \\d+(\\s+)?-(\\s+)?\\d{4}
RC\\s+(NO\\.|N\\.)? (\\d{3}-\\d{4})?	(REFERENTE|REFERNTE)\\s+ AL\\s+(CONTRATO|CONVENIO) .+$	- INDER

```
(CONTRATO|CONVENIO)\\s+INTERADMINISTRATIVO(\\s+|:|,|\\.)(DE\\sCOOPERACI(Ó|O)N\\s
|DE\\sCOLSULTORIA)?((\\w+)?(\\s+)?(NO\\.|N\\.)?(\\s+)?(\\d+)?\\s(DEL|DE)?(\\s+)?
((\\w+\\s\\d{4})|\\d{4}|\\s\\d{2}\\/\\d{2}\\/(\\d{4}|\\d{2})))?
```

Tabla A-1.: Primer grupo de expresiones regulares para la depuración de la base UN Títulos (Parte II).

Expresión regular	Expresión regular	Expresión regular
(CONTRAPARTIDA)?DIB(\\s+)(\\d{4})? (((\\s+)?CONV.)|:?)	AÑO\\s\\d{4}(\\s+)? -(\\s+)?DIB	DIB(\\s+)?-(\\s+)?
^(\\s+)?DIB	'\\d{4}(-INV)?-DIB	CONVENIO ESPECIFICO(((\\s+)? (NO\\.|N\\.)(\\s+)?EP)|:| DE)?
-?(\\s+)?MOVILIDAD(\\s+)? (-|:)	CONVENIO DE COOPERACI(Ó|O)N\\s+ (NO\\.|N\\.)\\s+\\d+	-? FIDUCIARIA BOGOTÁ((\\s+)?\\d+ - \\d+)?
-?(\\s+)?UGI(\\s*-)? (\\s+)?\\d{4}(-\\d{4})?	^(\\d+)?(\\s+)?-(\\s+)?UGI (\\s+)?(-|\\.)	^UGI -|\\.
\\d{4}- DIB.(UGI)?	-?(\\s+)?UGI(\\s+)?-	UNIDAD DE GESTI(Ó|O)N DE INVESTIGACI(Ó|O)N
CONV\\.? VIC\\.? (INV|DE INV)?\\.?(\\s+Y\\s+\\.?EXT) ?(-|\\.)?(\\s+)?\\d{4} (\\w{1})?	CONV\\.\\s+NAL\\. \\s+INVEST\\.	CONV\\.(\\s+)?((COLC\\/UNAL) |EXTERNA|(ORLANDO FALS BORDA(\\s+)?(\\d{4})?))?
SUSCRITO ENTRE LA CORPORACI(Ó|OX)N PARA EL DESARROLLO	RES\\.?(\\s+)?\\d+(DE (\\d+|(VICESEDE\\.))\\.?)?	INVESTIGACI(Ó|O)N \\d{4}:
CONTRATO NO\\. \\d+	ORDEN DE SERVICIOS .+$	(CONVENIO)?CELEBRADO ENTRE COLCIENCIAS.+$
COD\\.(\\s+)?\\d+	^\\d{4}((-\\s?(EXT|INV |COLC|FIN))+)?	CONV\\. ORLANDO FALS BORDA \\d{4}
DID(\\.|-)(\\s+)?\\d{4}	CURSO \\d{4}:	U\\.?(\\s+)?G\\.?(\\s+)?I\\.?
- SEDE BOGOT(Á|A) (\\d{4})?	(EVENTO(-|))?SEMINARIO	

Tabla A-2.: Segundo grupo de expresiones regulares para la depuración de la base UN Títulos.

Expresión regular	Expresión regular	Expresión regular
(ESPECIAL DE COOPERACION CIENTIFICA Y TECNOLOGICA\\s+)?(ESPECIAL)?NO\\.\\s?PE\\.GDE\\.(\\d+ \\.)+\\d+\\s?-(\\d{4})?	CONVENIO DE ASOCIACION NO\\.?(\\s+)?\\d+(DE \\d{4})	(SUSCRITO)? ENTRE (EL|LA) .+Y LA UNIVERSIDAD NACIONAL DE COLOMBIA
(CONVENIO|CONTRATO) BANCO DE LA REPUBLICA NO. \\d+	(CONVENIO|CONTRATO)\\s+NO \\.\\s+\\d+- BANCO DE LA REPUBLICA	(CONVENIO|CONTRATO) NO\\.\\s+\\d+\\s+-\\s+BANCO DE LA REPUBLICA
(CONVENIO|CONTRATO) DE LA REPUBLICA NO\\.\\s+\\d+ -	-\\s+BANCO DE LA REPUBLICA$	ECOS-NORD \\d{4} - \\d{4}
CÓD\\. \\d+-\\d+-\\d+	ACUERDO DE SUB-PROYECTO NO\\.\\s+PO.(\\s+)?\\d+ - SECRETARIA GENERAL DE LA ORGANIZACION DE LOS ESTADOS AMERICANOS -	GINEBRA/UNAL \\d{4}
TESIS POSG\\. M\\d+(CORTE \\d)?	BIB COTRAPARTIDA	-INV-
CONTRAPARTIDA UNAL	CONVENIO ISAGEN NO\. \d+(DE \d+)?	CONTRATO(\s+-\s+)?(\s+NO\\.)? (\s+(\d|-)+)?
\\s+-(\\s+)?-\\s+	SUSCRITO ENTRE\s+Y LA UNIVERSIDAD NACIONAL DE COLOMBIA	INV\s+\d{4}:?
RES\s+\d+\/\d+	^\s?INV	PROYECTO COLCIENCIAS:?
VIC\. INV\. Y EXT\.	DE INVESTIGACION NO\.\s(IF(\d|-)+)?	CONTRAPARTIDA:
CONVENIO SDA-\d+-\d+	^-?EXT(|-\\d+)	FACARTES(\\s+)?-?(\\s+)? \\d{4}
^ POSGRADO	CONV COLC/UNAL	FIDUCIARIA BOGOTA

Tabla A-3.: Tercer grupo de expresiones regulares para la depuración de la base UN Títulos (Parte I).

Expresión regular	Expresión regular	Expresión regular
(SUSCRITO)? ENTRE (EL|LA) .+Y LA UN	\\s?NO \\d+	
MODALIDAD IIA.+$	NAL\\.(\\s+)?INVEST \\.(\\s+)?M\\d	
^\d+(-|\\s)	ACAC\\/UNAL\\s+\\d+	REFERENTE A LA CONVOCATORIA PARA EL ESTIMULO A LA INVESTIGACI(O|Ó) N(FACULTAD DE MEDICINA Y DIRECCION DE INVESTIGACION)?.+
REFERENTE A LOS GANADORES EN LA CONVOCATORIA PROYECTOS? DE INVESTIGACION.+	REFERENTE A LOS GANADORES EN LA CONVOCATORIA.+	^\\s*-\\s*\\d{4}\\s*-
\\(\\)\\s*-?\\d{4}	\\w+ JORNADAS DE INVESTIGACION DE LA FACULTAD DE MEDICINA DE LA UNIVERSIDAD NACIONAL DE COLOMBIA	MOVILIDAD M\\d \\w+ SIMPOSIO
MOVILIDAD M\\d	CAR-CONV \\d+\\/\\d+	CORTOLIMA \\d+\\/\\d+
INTERADMISTRATIVO INVIMA N\\.? \\d+ DE \\d+	^.+(CTO|CNT|CNV|CONT|CPS) \\.?\\s*\\d+(/\\d+)?	\\w+ CONVOCATORIA PROYECCION .+$
-PNUD \\d+-\\w+-\\d+	SEGÚN ACTA DE COMPROMISO N.+$	- CON \\d{3}-\\d{4} PFIZER
CONT \\d+ SDCRD	AÑO-\\d{4}-(EXT|INV|COLC)	ICTA\\s*-?\\s*\\d{4}
PEP\\s*-?\\s*\\d{4}	DIRIGIDO A FUNCIONARIOS (DE |DE LA|DEL).+$	MADR \\d+(\\s*-?\\s*\\d+)?’

Tabla A-4.: Tercer grupo de expresiones regulares para la depuración de la base UN Títulos (Parte II).

B. Anexo: Clasificador de idioma.

Después del proceso de lectura de los datos y de consolidación de la indicadora que determina el lenguaje en el que se encuentra la descripción corta del proyecto de investigación, se decidió construir un clasificador de idiomas basados en la frecuencia de aparición de las palabras vacías (stop words en inglés). Dado un documento j, se clasificara en 'Ingles' o 'Español', con base en la comparación de estas dos cantidades $|V_i \cap SW_I|$ y $|V_i \cap SW_E|$, siendo V_j el conjunto de tokens del texto j-esimo de la colección, SW_I el conjunto de stop words en inglés y SW_E el conjunto de stop words en Español [1].

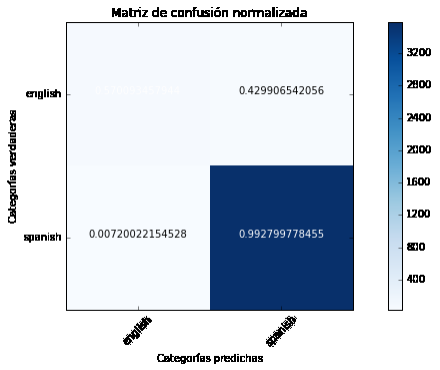


Figura B-1.: Matriz de confusión del clasificador de Idioma.

Como se puede ver en la matriz de confusión (Figura **B-1**), se obtuvieron buenos resultados para el clasificador construido, teniendo en cuenta la evaluación manual realizada de cada uno de los textos que conforman el conjunto de análisis.

[1] Las etiquetas para hacer la validación del lenguaje, es el resultado de la revisión manual del idioma original de cada una de las descripciones cortas del proyecto de investigación.

C. Anexo: Certificación de participación en Evento.

Certificate of Participation

Awarded to

Jorge Mario Carrasco

Oral Presentation - "*A very-short-text clustering method based on distributed representation to identify research capabilities of a Higher Education Institution*"

At the **6th *Global TechMining Conference*** September 13, 2016

Program Chairs

Figura C-1.: Anexo: Certificación de participación Global TechMining Conference