ISSN: 1692-7257 - Volumen 2 – Número 38 - 2021

Revista Colombiana de
Tecnologías de Avanzada

Recibido: 7 de agosto de 2021
Aceptado: 28 de septiembre de 2021

APLICACIÓN DE TÉCNICAS DE MINERÍA DE TEXTO PARA EL DESCUBRIMIENTO DE RELACIONES CONCEPTUALES ENTRE TRABAJOS DE GRADO

APPLICATION OF TEXT MINING TECHNIQUES FOR THE DISCOVERY OF CONCEPTUAL RELATIONSHIPS BETWEEN DEGREE PROJECTS

MSc. Jimmy Mateo Guerrero Restrepo [*], PhD. Ricardo Timarán Pereira [*]

[*] **Universidad de Nariño,** Facultad de Ingeniería, Programa de Ingeniería de sistemas
Grupo de investigación aplicada en sistemas Grias.
Torobajo - Calle 18 Carrera 50, Pasto, Nariño, Colombia.
Teléfono y Fax, con indicativos internacional y nacional.
E-mail: {jimaguere, ritimar}@udenar.edu.co.

Resumen: En este artículo se presentan uno de los resultados del proyecto de investigación cuyo objetivo fue descubrir relaciones conceptuales entre los trabajos de grado de la Universidad de Nariño utilizando técnicas de minería de texto que faciliten la recuperación de trabajos de grado relacionados con la temática de la búsqueda identificando similitudes y diferencias entre ellos. Se utilizó CRISP-DM como metodología. Usando técnicas de minería de texto se estructuraron los documentos del repositorio de trabajos de grado de la Universidad de Nariño. Se utilizaron técnicas de aprendizaje no supervisado para encontrar relaciones taxonómicas. Se entrenó el modelo Word2vec para encontrar relaciones temáticas. Se encontró el número óptimo de categorías, se logró interpretar los conceptos de cada categoría y sus relaciones.

Palabras clave: Minería de texto, Aprendizaje automático, Relaciones conceptuales, Doc2vec, Word2vec.

Abstract: This article presents one of the results of the research project whose objective was to discover conceptual relationships between the degree projects of the University of Nariño using text-mining techniques that facilitate the recovery of degree projects related to the topic of the search identifying similarities and differences between them. CRISP-DM was used as a methodology. Using text-mining techniques, the documents from the repository of degree works of the University of Nariño were structured. Unsupervised learning techniques were used to find taxonomic relationships. The Word2vec model was trained to find thematic relationships. The optimal number of categories was found, it was possible to interpret the concepts of each category and their relationships.

Keywords: Text mining, Machine learning, Conceptual relationships Word2vec, Doc2vec

1. INTRODUCCIÓN

La minería de texto ha despertado un enorme interés en la comunidad científica, debido a la creciente cantidad de documentos disponibles en formato digital, y también a la necesidad de organizar y obtener el conocimiento contenido en textos. La literatura en el campo de la minería de textos ofrece una serie amplia de aplicaciones tales como la clasificación supervisada, la recuperación de información, la clasificación no supervisada (CNS), extracción de entidades con nombre (NER), encontrar tendencias mediante nubes de palabras, extraer resúmenes de grandes volúmenes de texto, La mayoría de las técnicas propuestas en este ´ámbito se basan en el paradigma del aprendizaje artificial (Sebastiani, 2002). Un pilar importante de la minería de textos es la representación de documentos no estructurados de tal forma que reflejen los distintos rasgos de su contenido de la mejor manera posible. Esto es sumamente importante cuando se trabaja con colecciones de documentos no etiquetados mediante una clase, como los que se trata en este proyecto denominados problemas de CNS (Jain, 2010).

En el proceso investigativo realizado en el Grupo de Investigación Aplicada en Sistemas - GRIAS, en la línea de investigación de Herramientas y Sistemas de Gestión de Conocimiento y Recuperación de Información, se han desarrollado dos proyectos de investigación financiados por el sistema de investigaciones de la Universidad de Nariño: uno por la convocatoria estudiantil denominado ”Construcción de una Ontología de Aplicación que Soporte la Búsqueda Inteligente sobre los Trabajos de Grado de la Universidad de Nariño denominada SAWA, utilizando la herramienta de software libre Protégé” (Cabrera et al., 2015) y otro en la convocatoria de trabajos de grado denominado ”UMAYUX: Un Modelo de Software de Gestión de Conocimiento Soportado en una Ontología Dinámica Débilmente Acoplado con un Gestor de Bases de Datos para la Universidad de Nariño” (Benavides y RESTREPO , 2014). Estos proyectos fueron delimitados a los trabajos de grado del programa de Ingeniería de Sistemas de la Universidad de Nariño. Como resultado de estos proyectos se cuenta con “MASKANA” (Restrepo y Pereira, 2015), un prototipo de gestor documental para recuperación de información relacionada con los trabajos de grado del programa de Ingeniería de Sistemas almacenados en formato digital. En estos proyectos se dispone de un repositorio textual de documentos no estructurados sin etiqueta de clase, el cual se limitó a encontrar relaciones semánticas sin tener en cuenta los conceptos y NER.

Según (Vivas y Coni, 2013),” los conceptos son elementos esenciales para el reconocimiento del mundo que nos rodea. Ellos constituyen una representación de una clase de cosas. Frecuentemente, se suelen confundir o utilizar indistintamente los términos concepto y palabra. El concepto [escuela], por ejemplo, debe ser distinguido de la palabra ’escuela’. [Escuela] es un tipo de [institución educativa]. El concepto [escuela] puede, por ejemplo, ser expresado por las palabras ’escuela’, ’lugar para educar’, ’institución educativa’. Los conceptos están profundamente relacionados unos con otros de manera que la activación de unos genera la activación de otros. Los vínculos que los interconectan se denominan relaciones conceptuales. Este tipo de relaciones no debe ser confundido con las relaciones entre términos o palabras. Mientras que a las primeras se las suele denominar relaciones conceptuales, a las segundas se las suele denominar relaciones semánticas. Por ejemplo, las relaciones de sinonimia o de homonimia son relaciones semánticas, mientras que las relaciones taxonómicas y temáticas son relaciones primordialmente entre conceptos.”

Para (Estes et al., 2011), “las relaciones taxonómicas (también llamadas relaciones categoriales) son las que vinculan a conceptos de una misma categoría semántica. Los objetos de la misma categoría taxonómica usualmente comparten el nombre genérico. Dado que los componentes de este tipo de relaciones tienen rasgos comunes, las vinculaciones se establecen principalmente mediante mecanismos de detección de similitudes, es decir, comparando las propiedades de ambos conceptos. Este tipo de relaciones permiten, agrupar los conceptos de una misma categoría, anticipar, mediante procesos de deducción e inferencia, las propiedades que tendrá un nuevo elemento que se incluya dentro de la categoría.”

De acuerdo a (Golonka y Estes, 2009), “las relaciones temáticas, son relaciones contextuales entre objetos que no son del mismo tipo pero que pueden ser encontrados en los mismos esquemas. Específicamente, una cosa esta temáticamente relacionada con otra cuando ambas desempeñan roles complementarios en el mismo escenario o situación”.

Según (Barsalou et al., 2003), "las relaciones temáticas permiten organizar contextualmente la experiencia, así como establecer predicciones frente a situaciones futuras similares mediante el mecanismo de inferencia de completamiento de patrones."

Se han propuesto diferentes técnicas de minería de textos, en (Troyano et al., 2003) describen los enfoques de extracción de NER y conceptos ligados al conocimiento, en (Figuerola et al., 2004) aplican técnicas para extraer palabras clave de documentos textuales. Los siguientes autores (Llorens et al., 1998), (Santana Mansilla et al., 2014), (Barrera, 2016) y (Rodríguez-Tapia y Camacho-Cañamón, 2018) proponen aplicar técnicas de minería de textos para representar documentos no estructurados, en (Montes y Gómez et al., 2005) y (Muñoz y Otón ,2010) usan grafos conceptuales como representación del contenido de los textos, y obtiene algunos patrones descriptivos de los documentos aplicando varios tipos de operaciones sobre estos grafos. Estos antecedentes proponen diferentes alternativas de minería de textos, pero ninguno de ellos aplicado al dominio de trabajos de grado. Esto implico investigar diferentes técnicas de minería de textos y minería de datos, aplicarlas en el repositorio, evaluar su correcto funcionamiento e interpretar los patrones obtenidos generando conocimiento útil para el repositorio de la biblioteca Alberto Quijano Guerrero de la universidad de Nariño.

En este artículo se presentan los resultados de aplicar minería de textos para descubrir relaciones conceptuales entre los trabajos de grado de la universidad de Nariño que faciliten la recuperación de trabajos de grado relacionados con la temática de búsqueda identificando similitudes y diferencias entre ellos.

2. MATERIALES Y MÉTODOS

Como fuentes de información se utilizó el repositorio de datos de los trabajos de grado de la biblioteca Alberto Quijano Guerrero de la universidad de Nariño.

En este capítulo se describen los procesos de construcción, limpieza y transformación del corpus de documentos de los trabajos de grado. Luego se detallan los experimentos realizados; con el objetivo de descubrir relaciones conceptuales en el corpus de trabajos de grado de la Universidad de Nariño utilizando la metodología CRISP-DM.

En la fase comprensión del problema se realizaron las actividades que permitieron profundizar y apropiar de una manera completa el problema objeto de estudio, los objetivos y los requisitos de la investigación, que posibilitaron la recolección de los datos correctos para interpretar adecuadamente los resultados. En esta fase, descubrir las relaciones conceptuales del repositorio de documentos no estructurados de trabajos de grado de la universidad de Nariño, se convirtió en un problema a resolver con minería de textos.

En la fase de comprensión de los datos se identificó, recopiló y familiarizó con el corpus de trabajos de grado, disponible en el repositorio de la biblioteca Alberto Quijano Guerrero de la universidad de Nariño.
Se construyó un repositorio inicial donde se integraron todos los trabajos de grado de los diferentes programas de la universidad de Nariño. Dando como resultado un repositorio compuesto por 8.076 documentos no estructurados, el cual sirvió de base para las subsiguientes fases.

En la fase de preparación de los datos se realizó el pre procesamiento del corpus; primero que todo se descartó de los textos la sección de agradecimientos, utilizando expresiones regulares; se obtuvo las secciones de resumen, introducción, marco teórico, metodología, resultados y conclusiones. Después de eliminar la sección de agradecimientos se eliminaron las palabras muertas de la colección usando la librería Spacy de Python, seguidamente se obtuvieron los lemas de las palabras del corpus, ejemplo desarrollo por desarrollar. Obteniendo como resultado un corpus limpio y listo para el desarrollo del trabajo. La figura 1 detalla el proceso de pre procesamiento del corpus.
En la fase de modelado una vez organizado el corpus se implementaron diferentes técnicas de minería de texto; para obtener diferentes conjuntos de datos estructurados. Para estructurar documentos se usaron tres técnicas diferentes, las dos primeras provenientes de la librería Sklearn; CountVectorizer (Bow) y TfidfVectorizer (TF-IDF); la tercera técnica utilizada fue Doc2vec proveniente de la librería Gensim.

Para iniciar el análisis, se verifica si existe tendencia al agrupamiento en cada uno de los conjuntos de datos estructurados por los modelos

generados anteriormente. Utilizamos el estadístico de Hopkins para establecerlo, recurrimos al paquete RANN de R. Seleccionamos 1000 puntos al azar de los datos de muestra y luego se los compara con 1000 puntos creados al azar.
Luego se calcula la distancia al vecino más cercano de ambas muestras en la cual se obtiene u=conjunto de distancias obtenidas de los objetos distribuidos al azar y w=conjunto de distancias obtenidos de los objetos extraídos del dataset original, finalmente se establece el estadístico para cada conjunto de datos de acuerdo a la ecuación 1.

$$H = \frac{\sum_{i=1}^{p} wi}{\sum_{i=1}^{p} ui + \sum_{i=1}^{p} wi} \quad (1)$$

Visto en "Introduction to Data Mining" (Tan et al., 2016), p547.

La figura 1 detalla el proceso.

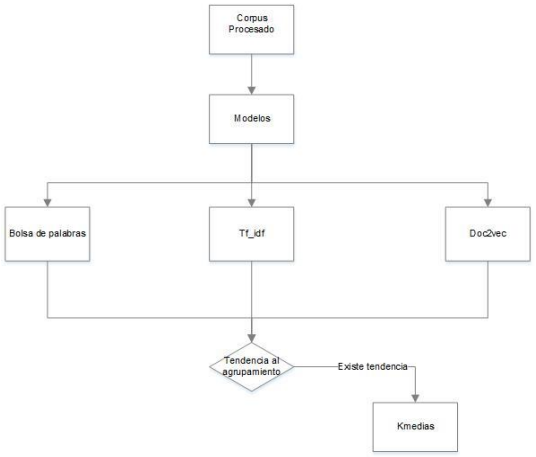

Fig. 1. Flujo de trabajo para la generación de un conjunto de datos estructurado

La agrupación es un ejercicio popular de aprendizaje automático, y las técnicas utilizadas en las tareas de clustering clásicas pueden utilizarse para texto una vez estructurado, con la idea de formar grupos de trabajos de grado de acuerdo a su dominio de conocimiento. Para determinar la cantidad de grupos óptimos (K) se corrió el algoritmo kmedias para diferentes K, variando desde 20 hasta 44. Se calculó la suma de errores al cuadrado (SSE) y el coeficiente de Silhouette para cada valor de K. SSE mide la distancia de los puntos al centroide por lo cual esperamos que SSE sea lo más bajo posible como se lo describe en la ecuación 2.

$$SSE = \sum_{i=1}^{k} \sum_{xeci} dist(ci, x)^2 \quad (2)$$

Donde k es el número de clústeres y ci es el centroide del clúster Ci
Visto en "Introduction to Data Mining" (Tan et al., 2016).

El coeficiente de Silhouette es útil para analizar la cohesión y la separación del agrupamiento, valores cercanos a 1 indican que el agrupamiento es bueno, valores alrededor de cero o hasta negativos indican que el k no es el óptimo o la tendencia al agrupamiento no es tan buena, su fórmula se describe en la ecuación 3.

$$si = \frac{(bi - a1)}{max(ai, bi)} \quad (3)$$

Visto en "Introduction to Data Mining" (Tan et al., 2016).
Para la tarea de agrupación se usó la librería sklearn de Python y el algoritmo kmeans en los conjuntos de datos estructurados con tendencia al agrupamiento.

Por cada grupo encontrado se entrenó el modelo Word2vec, con el fin de encontrar relaciones conceptuales entre los términos de dominio presentes en los documentos. Para este proceso se configuró un clúster con Apache Spark con el fin de paralelizar el entrenamiento y acelerar el tiempo de procesamiento, se utilizó la librería Pyspark para la programación en paralelo y la librería Gensim para el entrenamiento del modelo Word2vec y como entrada el corpus de lemas clasificado por su respectivo grupo encontrado.

La fase de evaluación de los modelos se describe en los capítulos de Resultados y Discusión.

3. RESULTADOS

3.1 Compresión de los datos

Se obtuvo un repositorio inicial donde se integraron todos los trabajos de grado de los diferentes programas de la universidad de Nariño. Dando como resultado un repositorio compuesto por 8.076 documentos no estructurados. La tabla 1 indica las medidas de resumen la cantidad de palabras o tokens del repositorio inicial.

Tabla 1: Estadísticos descriptivos en cuanto a la cantidad tokens del repositorio inicial

Media	36.120
STD	25.879
Mínimo	15.423
Q1	20.695
Q2	31.219
Q3	45.480
Máximo	672.312

3.2 Preparación de los datos

Se obtuvo un repositorio limpio y listo para la aplicación de técnicas de minería de texto para estructurarlo. La tabla 2 indica las medidas de resumen en cuanto a la cantidad de palabras o tokens del repositorio procesado.

Tabla 2: Estadísticos descriptivos en cuanto a la cantidad tokens del repositorio procesado

Media	6.672
STD	4.768
Mínimo	1.500
Q1	3.583
Q2	5.620
Q3	8.622
Máximo	66.538

3.3 Modelado y evaluación

En esta fase se obtuvieron modelos para estructurar los documentos y modelos de aprendizaje no supervisado para identificar áreas de conocimiento en los trabajos de grado validados con sus respectivas métricas.

3.3.1 Modelos para estructurar el corpus

La tabla 3 muestra el estadístico de Hopkins para cada conjunto de datos generado.

Con un umbral establecido en 0.5 podemos confirmar que existe tendencia al agrupamiento en los conjuntos de datos generados por los modelos Doc2vec bolsa de palabras distribuida (PV-DBOW) y Doc2vec memoria distribuida (PV-DM). Los modelos Doc2vec logran captar relaciones conceptuales mediante el contexto que representa cada documento, usando esta representación se podrá medir qué tan relacionado está un documento con respecto a los demás.

Tabla 3: Descripción de la tendencia al agrupamiento en los conjuntos de datos estructurados.

Modelo	Estadístico Hopkins	Descripción
Bow	0.61	Conjunto Bow dimensión (8.076,10.000)
Bow	0.58	Conjunto Bow dimensión (8.076,20.000)
Tf-idf	0.54	Conjunto Tf-idf dimensión (8.076,20.000)
Tf-idf	0.52	Conjunto Tf-idf dimensión (8.076,10.000)
Doc2vec Pv-Dbow bolsa de palabras distribuida	0.38	Conjunto PV-DBOW dimensión (8.076,20)
Doc2vec Dm memoria distribuida	0.42	Conjunto DM (8.076,20)

3.3.2 Modelos de aprendizaje automático no supervisados

Con el fin de descubrir grupos de conocimiento de acuerdo al contexto de los diferentes trabajos de grado se corrió el algoritmo k-medias. Para la selección del k óptimo generamos diferentes grupos iterando k desde 20 a 44 evaluando el coeficiente de Silhouette y el error cuadrático en los 2 conjuntos de datos generados con el modelo Doc2vec ya que tienen tendencia al agrupamiento.

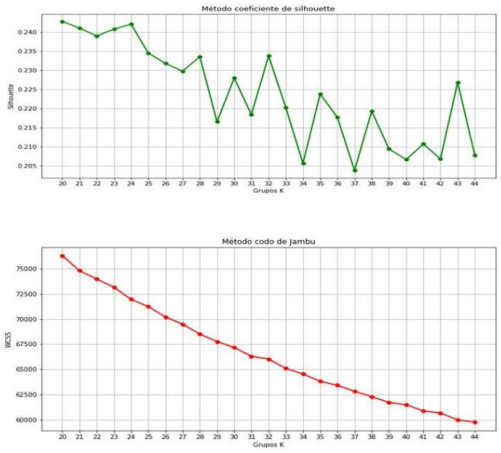

Fig. 2 Métodos para encontrar el número de grupos optimo

En la figura 2 se visualiza el método de coeficiente de Silhouette; el cual sugiere que el número adecuado de grupos es de 24, también se visualiza el método del codo de Jambu; para el cual miramos que a medida que k se incrementa la distancia de los puntos al centroide va disminuyendo. Para la selección del K óptimo se contrasto los dos métodos y se eligió el k en 32; ya que se observa que su coeficiente de Silhouette está entre los más altos y a la vez la distancia de los puntos al centroide es baja. Por lo tanto, se corrió el algoritmo k medias con k =32, con el fin de clasificar los documentos en 32 áreas de conocimiento de acuerdo a su temática de contexto.

3.3.3 Interpretación de las categorías mediante grafos conceptuales.

Para interpretar las temáticas de las 32 categorías encontradas por el modelo k medias con las representaciones generadas con Doc2vec se entrenó el modelo Word2vec y se elaboró el algoritmo visto en la figura 3.

Input: T: contenido textual documentos relacionados
Output: G: Grafo Conceptual
1: *conceptos_ner* ← ∅
2: *G* ← ∅
3: spacy.load('es_core_news_sm')
4: **for** each token ∈ T **do**
5: **if** spacy.isNer(token) **then**
6: *conceptos_ner*←adicionar(conceptos_ner,token)
7: **end if**
8: **end for**
9: **for** each c ∈ conceptos_ner **do**
10: **for** rl ∈ word2vec.sim(c) **do**
11: *n1*←crearNodo(c)
12: *n2*←crearNodo(rl)
13: *G*←relacionarNodos(n1,n2)
14: **end for**
15: **end for**

Fig. 3 Algoritmo Maskanita relaciones conceptuales de trabajos de grado.

El primer paso toma como entrada los documentos de cada categoría, En segundo lugar, se obtiene conceptos relevantes de los documentos de entrada mediante la tarea de reconocimiento de entidades nombradas de la librería Spacy de Python, Para cada entidad o concepto reconocido se aplica el modelo Word2vec para conocer sus relaciones, finalmente se construye el grafo conceptual usando la librería D3.

En la figura 4 se muestra el grafo conceptual generado por el modelo Word2vec de los trabajos de grado del grupo 30. Donde se pueden establecer vínculos entre conceptos dentro de minería de datos, descubrimiento de conocimiento, algoritmos de clasificación y asociación tales como equipasso, fpgrowth, árboles de decisión, a priori e itemsets frecuentes.

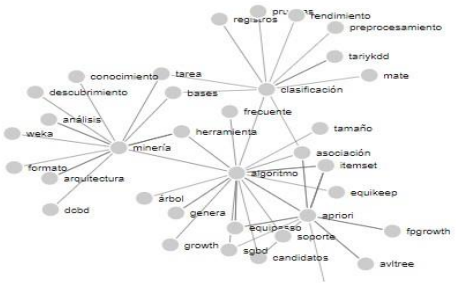

Fig. 4 Grafo relaciones temáticas grupo 30

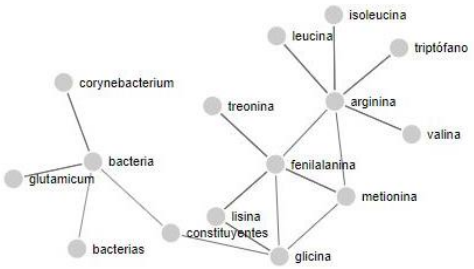

Fig. 5 Grafo relaciones temáticas grupo 17

El grupo 17 contiene tópicos relacionados con bacterias, microorganismos, compuestos antioxidantes, proteínas y aminoácidos como se describe en la figura 5.

4. DISCUSIÓN

En este capítulo se evalúa las relaciones conceptuales taxonómicas o relaciones categoriales de los métodos propuestos contrastando los resultados con otros trabajos relacionados. Dado que los trabajos de grado de este tipo de relaciones tienen rasgos comunes, las vinculaciones se establecen principalmente mediante mecanismos de detección de similitudes. La medida de validación interna del agrupamiento, que se utiliza en este trabajo, consiste en determinar qué tan cohesionados están los grupos entre sí, se busca que los documentos relacionados de un mismo grupo se parezcan contextualmente más entre ellos que con los documentos de otros grupos; y por otra parte determinar qué tan separados son los documentos de un grupo con respecto a todos los documentos de otros grupos. Por tanto, la métrica que permite evaluar la agrupación de los dominios obtenidos es el coeficiente de Silhouette. El coeficiente de Silhouette muestra qué documentos se encuentran completamente dentro de un grupo y cuáles han sido mal clasificados. Esta métrica tiene

un rango de [-1, 1], un valor mayor a 0 y cercano a 1 indica que el documento está lejos de los grupos vecinos, 0 indica que el documento está en o muy cerca de la frontera de decisión entre dos grupos vecinos, y valores negativos indican que el documento podría haber sido mal asignado al grupo.

Tabla 4: Validación Coeficiente de Silhouette algoritmo k medias para los conjuntos de datos generados.

Modelo	Coeficiente de Silhouette	K óptimo (20 –44)
Doc2vec PV-DBOW	0.36	32
Doc2vec PV-DM	0.34	27
Tf-idf 20000 variables	0.04	38
Tf-idf 10000 variables	-0.009	43

En la tabla 4 se observa que el valor más alto de coeficiente de Silhouette está en los conjuntos generados por los modelos Doc2vec, por tanto, los algoritmos implementados en esta investigación se basaron en estos, particularmente en Do2vec PV-DBOW; ya que logra una mejor agrupación y separación de los dominós de conocimiento en el corpus de trabajos de grado de la universidad de Nariño.

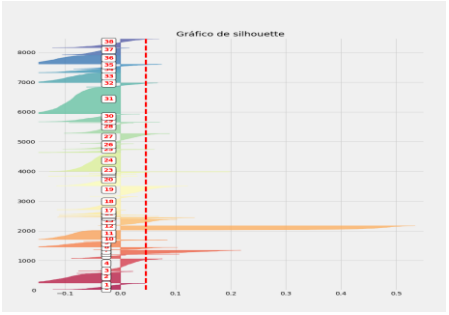

Fig. 6 Gráfico de Silhouette algoritmo K-medias con k=38, Tf-idf 20000 variables

En el gráfico de silhouette generado por el algoritmo K-medias con k=38, Tf-idf 20000 variables, figura 6 se observa que la mayoría de los documentos del corpus fueron mal clasificados a su grupo ya que tienen un valor negativo de silhouette.

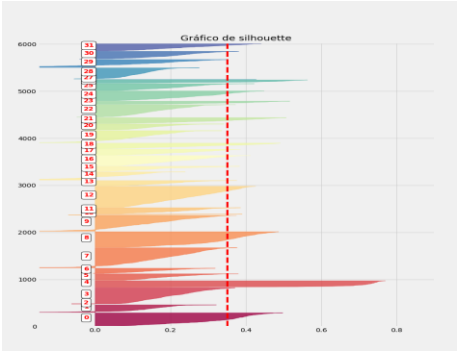

Fig. 7 Gráfico de Silhouette algoritmo K-medias con k=38, Tf-idf 20000 variables

Por otro parte, en el gráfico de silhouette generado por el algoritmo K-medias con k=32, método Doc2vec PVDBOW, figura 7 se observa que a pesar de existir documentos que quedaron mal clasificados; la gran mayoría de documentos están cohesionados de forma correcta a su respectivo grupo, Doc2vec tiene un rendimiento claramente superior a tf-idf por tanto es el método que se elige para relacionar trabajos de grado conceptualmente y generar recomendaciones.

Contrastando con (Kim et al., 2019) lograron obtener mejores resultados ensamblando los modelos TF-IDF, IDA Y Doc2Vec, los resultados de la presente investigación vistos en la tabla 3 demuestran que los conjuntos estructurados con los métodos TF-IDF no tienen tendencia al agrupamiento por tal motivo este método quedó descartado para el repositorio de trabajos de grado. En (Nandi et al., 2018) se observa que el método Doc2vec coincide en los resultados, reiterando que este tiene un mejor rendimiento y puede ser implementado en diferentes dominios de conocimiento.

5. CONCLUSIONES

Con la construcción, limpieza y transformación del repositorio de documentos de trabajos de grado de la universidad de Nariño; se logró estructurar este repositorio usando diferentes técnicas de minería de texto y se seleccionó la más adecuada que de acuerdo a los resultados fue Doc2vec, estructurar el corpus de trabajos permito utilizar el algoritmo k-medias, para encontrar relaciones categoriales y diferenciar los dominios de conocimiento del repositorio, los resultados mostraron que el número óptimo de categorías o grupos de conocimiento es 32. Finalmente, la tarea de Ner y el modelo Word2vec permitieron interpretar el conocimiento

relacionado a cada una de las categorías encontradas, estos 2 modelos fueron implementados en el algoritmo visto en la figura 3, el cual visualiza relaciones conceptuales temáticas entre los documentos del repositorio.

REFERENCIAS

Barrera, M.C. (2016). Minería de texto en la clasificación de documentos digitales. Biblios: Journal of Librarianship and Information Science, (64), 33–43.

Barsalou, L.W., Simmons, W.K., Barbey, A.K., and Wilson, C.D. (2003). Grounding conceptual knowledge in modality-specific systems. Trends in cognitive sciences, 7(2), 84–91.

Benavides, M. and RESTREPO, J.M.G. (2014). Umayux: un modelo de gestor de conocimiento soportado en una ontología dinámica débilmente acoplado con un gestor de base de datos.

Cabrera, O.E., Guerrero, J.M., Benavides, M.F., and and Pereira, R.T. (2015). Swa: ontología para la gestión de conocimiento sobre trabajos de grado. Revista Ontare, 1(2), 183–214.

Estes, Z., Golonka, S., and Jones, L.L. (2011). Thematic thinking: The apprehension and consequences of thematic relations. In Psychology of learning and motivation, volume 54, 249–294.

Elsevier. Figuerola, C.G., Berrocal, J.L.A., Rodríguez, A.F.Z., Rodríguez, E., and Reina, G. (2004). Algunas técnicas de clasificación automática de documentos. Cuadernos de documentación multimedia, ISSN-e, 1575–9733.

Golonka, S. and Estes, Z. (2009). Thematic relations affect similarity via commonalities. Journal of Experimental Psychology: Learning, Memory, and Cognition, 35(6), 1454.

Jain, A.K. (2010). Data clustering: 50 years beyond kmeans. Pattern recognition letters, 31(8), 651–666.

Kim, D., Seo, D., Cho, S., and Kang, P. (2019). Multico-training for document classification using various document representations: Tf–idf, lda, and doc2vec. Information Sciences, 477, 15–29.

Llorens, J., Velasco, M., Moreiro, J., and Morato, J. (1998). Características textuales como medida cualitativa de la información en la generación semiautomática de tesauros. Procesamiento del Lenguaje Natural, 23.

Montes y Gómez, M., Gelbukh, A., and López López, A. (2005). Minería de texto empleando la semejanza entre estructuras semánticas. Computación y Sistemas, 9(1), 63–81.

Muñoz, M.S. and Otón, E.M. (2010). Utilización de árboles semánticos para la comprensión de textos especializados a partir de su terminología. 18, 477–493.

Nandi, R.N., Zaman, M.A., Al Muntasir, T., Sumit, S.H., Sourov, T., and Rahman, M.J.U. (2018). Bangla news recommendation using doc2vec. In 2018 International Conference on Bangla Speech and Language Processing (ICBSLP), 1–5. IEEE.

Restrepo, J.G. and Pereira, R.T. (2015). Maskana: un gestor de conocimiento para recuperación y búsqueda inteligente de trabajos de grado en la universidad de Nariño. Revista Tecnológica-ESPOL, 28(5).

Rodríguez-Tapia, S. and Camacho-Cañamón, J. (2018). Los métodos de aprendizaje automático supervisado en la clasificación textual según el grado de especialización. Tonos Digital, 35(0).

Santana Mansilla, P.F., Costaguta, R.N., and Missio, D. (2014). Aplicación de algoritmos de clasificación de minería de textos para el reconocimiento de habilidades de e-tutores colaborativos.

Sebastiani, F. (2002). Machine learning in automated text categorization. ACM computing surveys (CSUR), 34(1), 1–47. Tan, P.N., Steinbach, M., and Kumar, V. (2016). Introduction to data mining. Pearson Education India.

Troyano, J.A., Díaz, V.J., Enríquez, F., Barroso, J., and Carrillo, V. (2003). Identificación de entidades con nombre basada en modelos de markov y árboles de decisión. Procesamiento del lenguaje natural, 31.

Vivas, L. and Coni, A.G. (2013). Relaciones conceptuales: Definición del constructo, bases neuroanatomías y formas de evaluación. Actualidades en psicología, 27(114), 1–18.