Similaridades y diferencias entre los textos sagrados de las religiones asiáticas

Edwin David Narváez Miranda. Autor

Abstract— A lo largo de la historia de la humanidad han surgido varias religiones que han modelo el carácter de las personas a través de los años, de este modo una dinámica interesante es encontrar que similitudes o diferencias se encuentra en los textos sagrados de las religiones. Este proyecto busca despejar estas interrogantes a través de técnicas de minería de textos. Para ello se cuenta con un corpus de texto de cuatro religiones originarias de Asia. Se realiza un pre procesamiento de datos, se representa de forma vectorial los documentos y se emplea técnicas de aprendizaje no supervisado: K means y clusterización jerárquica. Al final se encuentra claras diferencias entre dos grupos de religiones: el cristianismo y las religiones del este asiático, así también se explica cuáles son los textos más semejantes.

Index Terms—Minería de textos, textos sagrados, religiones de Asia, clúster, k means, cluster jerárquicos.

1 Introducción

El continente asiático fue el lugar donde nació varias de las religiones más prácticas en el mundo moderno, estas religiones tienen varios propósitos que han ido evolucionando junto con la humanidad, destacan los valores morales que están conllevan: como el amor, la compasión la paciencia, la humildad y el perdón. Todos estos valores se pueden apreciar en los textos sagrados que a su vez sirven como medio de preservación durante el tiempo y medio de enseñanza a las demás generaciones. [1]

En este trabajo se ha utilizado una recopilación de textos sagrados de diferentes religiones de Asia como los Yogasutra, Upanishad del hinduismo, las cuatro nobles verdades del budismo, el Tao Te Ching del taoísmo y los libros de Proverbios, Eclesiastés, Eclesiásticos y de la Sabiduría del cristianismo. Estos textos han sido escritos en diferentes épocas por distintas civilizaciones de Asia y han sido traducidos al inglés y recopilados por el proyecto Gutenberg.

Analizar las religiones de cada civilización constituye en una manera compleja para entender las sociedades modernas, su origen y sus principales diferenciares. [2]. Varios autores que sentaron las bases de la sociología orientada a la religión resaltan la importancia de la religión en el entendimiento histórico-social de nuestras sociedades. [2]

El objetivo de este proyecto es identificar si es que existen algunas semejanzas entre los textos sagrados asiáticos y en el caso de haberlo cuales serían; para ello se emplea técnicas de Minería de textos para el pre procesamiento del corpus de datos, representación vectorial de términos y algoritmos no supervisados para encontrar agrupaciones.

Como entradas de este análisis se contará con el dataset de los textos sagrados de distintas religiones de Asia, y como salidas se tiene las similaridades y diferencias de los textos sagrados gracias al conjunto de cluster obtenidos.

2 Trabajos relacionados

Existen un trabajo relacionado, que utiliza el mismo conjunto de datos. Dicho trabajo trata de encontrar que tienen en común las religiones asiáticas a través de un enfoque de aprendizaje no supervisado.

Los autores este trabajo utilizan una metodología común de minería de textos, la cual consiste en representar el corpus a través de Document Term Frecuency (DTM), normalizar esta representación, encontrar matrices de similaridad usando varias métricas, por último emplearon 2 enfoques: uno aplicando métodos de aprendizaje supervisado como K NN, SVM y Random Forest para predecir con precisión si cualquier capitulo en efecto pertenece a la etiqueta y otro aplicando aprendizaje no supervisado con K means para revelar la similaridad entre estos textos sagrado.

Al final obtuvieron los siguientes resultados:

- La distancia euclidiana fue la que permitió separar mejor los textos.
- La representación Bag of Words es poderosa para encontrar patrones de fuerte cercanía entre el budismo, taoísmo e hinduismo cuyo lugar de origen es geográficamente cercano.
- Los dos libros más similares son el Tao Te Ching y el Upanishad.

• *Edwin David Narváez Miranda Universidad de Cuenca. E-mail: edwin.narvaez@ucuenca.edu.ec*

Otros autores han abordado el estudio de textos sagrados con técnicas de minería de textos, pero desde una perspectiva del parte del habla (POS), detectaba el número de sustantivos, verbos, adjetivos, adverbios, etc. Además, incorporaban algortimos de calculo de densidad léxicas (LD) para determinar el procentaje de las palabras léxicas y variación léxica (LV) para determinar el radio en procentajes entre las diferentes palabras en el texto y el total de palabras. [3]

3 DESCRIPCIÓN DEL DATASET

El conjunto de datos disponible consitía en 3 archivos, uno tipo texto y dos csv. El archivo de tipo texto tenían párrafos de los textos sagrados de las religiones, estos textos estaban sin etiquetar.

En cambio, los csv consistían en el corpus en representación DTM, la diferencia entre estos csv es que el uno tenía las etiquetas y el otro no.

El conjunto de datos que se empleó constaba de 590 textos sagrados de las distintas religiones, repartidos como se muestra en la figura 1. Teniendo a YogaSutra como el texto sagrado que tiene mayores documentos y el libro de Eclesiastés con el menor número de documentos.

Como se mencionará mas adelante en la metodología se realizó la limpieza de datos, eliminando símbolos y números de los textos, luego se pre procesó el texto removiendo las stopwords, tokenizando y stemming, para finalizar se representó los textos con TF-IDF usando unigramas y bigramas.

Para visualizar los textos en dos dimensiones se utilizó MDS, cabe recalcar que también se probó con SVD, pero las visualizaciones parecían no ser mejores a las obtenidas con MDS.

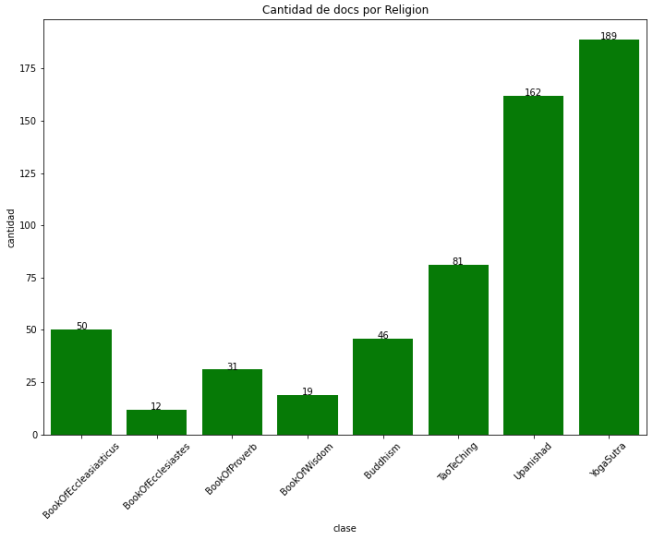

Figura 1.Cantidad de textos por documentos

3 METODOLOGÍA

La metodología empleada en este proyecto consiste en una especificación de la metodología común en minería de textos y se encuentra representada en la figura 2.

Los componentes de la metodología son similares a los vistos en clase y al revisado en trabajo relacionado, la principal diferencia es la utilización de TF-IDF con unigramas y bigramas en lugar de DTM.

Componentes de la metodología

Lectura de datos

Se leyó el archivo de texto con el fin de obtener el corpus del análisis, del csv solo se obtuvo las etiquetas de los textos con fines orientativos para poder representar los textos y conocer de que texto sagrado se trata. Más no como entrada de los modelos.

Librerías: pandas, numpy

Limpieza de datos

La limpieza de los datos consistió en la eliminación símbolos, números y espacios vacíos en los textos.

Librerías: re

Prepocesamiento

En esta etapa se considera tres procesos fundamentales: la tokenización, eliminación de las stopwords y el stemming.

Librerias: ntlk

Representación de documentos

Contrario a la metodología planteada por los autores del trabajo relacionado, en esta metología se utiliza una representación TF-IDF (unigramas y bigramas).

Librerias: Sklearn

Modelos de clusterización

En el componente K means, se emplea el método de la silueta para encontrar el mejor K para particionar los clústeres.

En cuanto al clúster jerárquico, se calcula la matriz de distancia utilizando la métrica de coseno, a ella se aplica el método de linkage “average” ya que con este método se halló el mayor valor de Cophenet. Cophenet es una medida de cuán fielmente un dendograma conserva las distancias por pares entre los puntos de datos originales sin modelar [4]. Llevado esto al campo de la clusteriza-

ción compara (correlaciona) las distancias reales por pares de todas sus muestras con las implicadas por la agrupación jerárquica. Cuanto más cerca esté el valor de 1, mejor la agrupación conserva las distancias originales.

Librerias: Sklearn, Scipy

Visualizaciones de los datos

Para finalizar las visualizaciones que se consideran son:

- la visualización de los documentos en dos dimensiones utilizando MDS,
- una nube de palabras con los términos más comunes por clúster,
- una visualización geográfica de los clústeres en un mapa del continente asiático, y
- una visualización de la relación de los términos más representativos de los clústeres.

Librerias: matplotlib, seaborn, ploty
Software: VOSviewer.

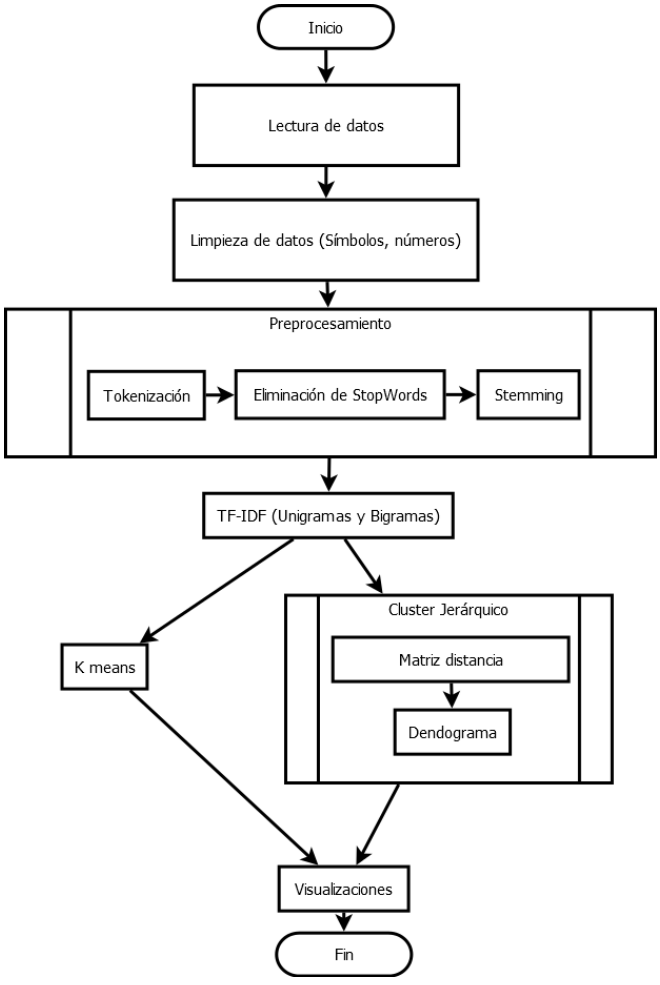

Figura 2. Metodología empleada

Las librerías empleadas en este análisis fueron las que se utilizaron a lo largo del curso, listadas de la siguiente manera:

- **Pandas:** para el manejo de estructuras de datos como dataframes, series, etc.
- **Numpy:** para el manejo de matrices unidimensionales y bidimensionales.
- **Sklearn:** para el uso de los modelos TF-IDF, K means, métricas de silueta, similitudes entre documentos, MDS.
- **Nltk:** para los algoritmos de minería de textos como Tokenización, Stemming, eliminación de StopWords.
- **Re:** para el procesamiento de texto.
- **Matplotlib, Seaborn:** para la generación de gráficas.
- **WorldCloud:** para mostrar las palabras más comunes en un gráfico.
- **Ploty:** para los gráficos interactivos.
- **Scipy:** para los modelos de clustering aglomerativos, dedongramas.

5 Resultados y discusiones

K means

Se probó con varios valores de K para el agrupamiento de clúster, pero el mejor valor de silueta se obtuvo con K = 3, en la figura 3 se aprecia la gráfica de la silueta, como se puede observar el valor promedio de silueta es bajo a pesar de ser el mejor obtenido.

Recordemos que el valor de la silueta es orientativo e indica que tan buena es la asignación de una observación de su cluster comparando su similitud con el resto de documentos pertenecientes a ese cluster.

El score mas alto de la silueta lo obtiene el cluster 1, que además es el posee más documentos, 357 documentos contrarios a los 110 para el cluster 0 y 123 para el cluster 2.

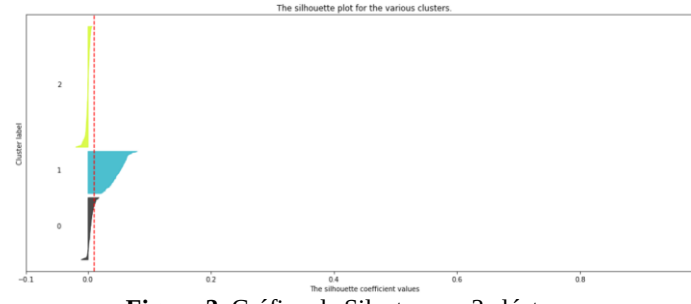

Figura 3. Gráfico de Silueta para 3 clúster

Graficando los textos en un espacio de dos dimensiones figura 4 (a) intuitivamente se puede verificar que hay dos grupos definidos: el grupo ubicado en la parte superior conformada por documentos del taoísmo, budismo

e hinduismo y otro grupo en la parte inferior conformada por documentos de la religión cristiana.

La figura 4 (b) muestra los 5 clúster obtenidos con Kmeans, donde se puede confirmar que el algortimo toma a los documentos cristianos como un solo clúster, en cambio a las otras religiones intenta dividirlos.

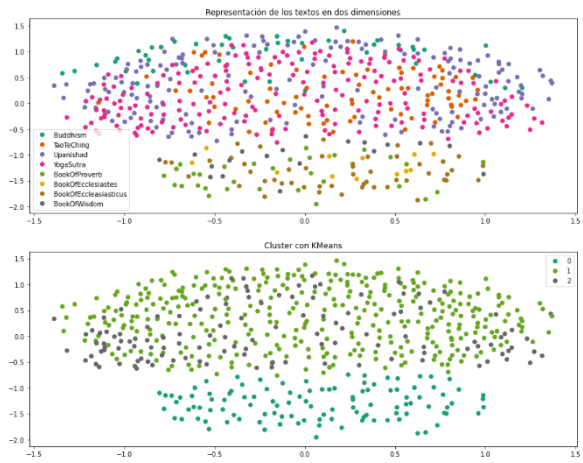

Figura 4. (a) Representación de los textos en dos dimensiones (b) Clúster con K means

El algoritmo de K means se comporta de manera aleatoria al momento de elegir sus centroides, es por ello que en cada ejecución se presenta resultados diferentes, sin embargo, esta diferencia no es tan notable.

Clúster jerárquico

Al aplicar el clúster jerárquico, se obtiene resultados similares a los obtenidos con K means. Con el método de linkage "average" se obtiene el mejor valor de Cophenet, luego de probar con varios valores de K se obtiene que en todas las ocasiones un clúster que no se separa, este clúster es el que está conformado por documentos de la religión cristiana, en cambio los documentos de las otras religiones si son separados, solo el Tao Te Ching y el Upanishad se mantenían como otro custer, sin embargo, los demás nuevos clústers se encuentran muy desbalanceados.

En este punto ya se encontrar similitudes entre algunos textos sagrados, por un lado y un tanto más obvio todos los textos de la religión cristiana van y deben parecerse entre sí, razón por la cual no son separados al momento de aumentar el número de cluster. De igual manera, como se describe en el párrafo anterior, los textos del Tao Te Ching y el Upanishad también tienen semejanzas es por ello que no son separados tan fácilmente, sólo cuando ya se llega a un número K mayor a 7 empieza a formarse nuevos cluster separándo los documentos pertenecientes a estos dos libros sagrados.

Para continuar con el proceso de inrpretación y análisis de cluster se cree conveniente llevar el estudio al analisis dos clústers únicamente, los cuales se muestran en la figura 5.

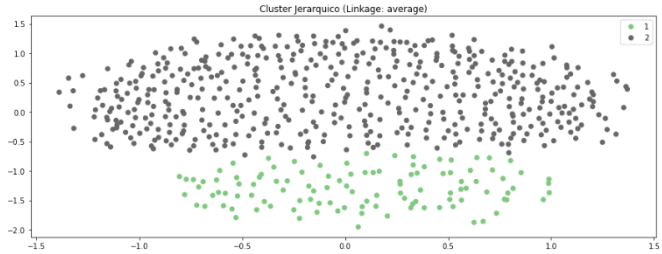

Figura 5. Clúster jerárquico k = 2

De cada uno de los clústers se encontró los términos más comunes y se las represento por mediante una nube de palabras, figura 6.

Por un lado, tenemos al Cluster 1 en cuyos términos principales se encuentra palabras como *Dios, señor, mal, malvado, alma, gloria, hijo, palabra.* Que hace referencia a la presencia de un ente superior que maneja el bien y el mal. Estos términos están presentes en la religión cristiana y son la base de sus enseñanzas.

Por otro lado, en el Cluster 2 tenemos términos como: *poder, conciencia, mente, conocimiento, uno mismo, saber, sentido, cuerpo, espiritual, cosas.* Se nota ausencia de un ente superior y se centra más en el ser humano y sus virtudes, propias de las religiones del este de Asia.

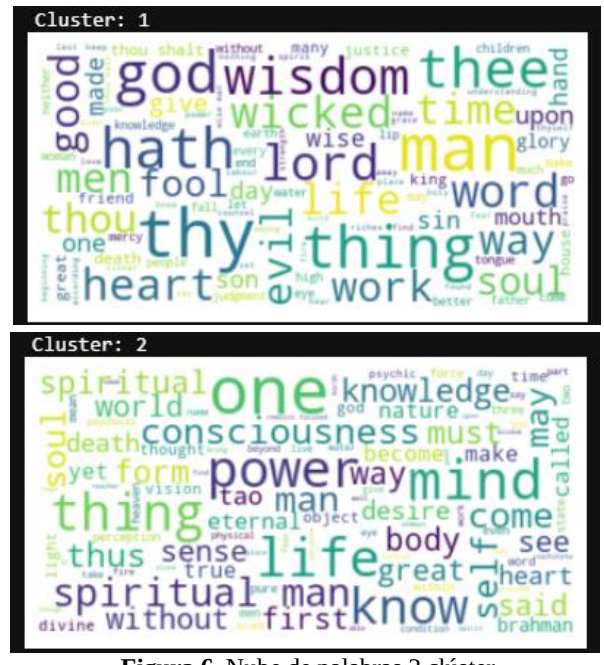

Figura 6. Nube de palabras 2 clúster

Entonces, sin necesidad de las etiquetas de cada texto ya se puede hablar de una clara diferenciación entre estos

dos grupos de cluster, y es el enfoque de los textos sagrados, Cluster 1: eje central un Dios, Cluster 2: eje central Ser Humano.

En la figura 7, se puede apreciar los clústers de manera geográfica, los cuales fueron localizados gracias a las etiquetas de los textos, cabe reclacar que en esta ocasión las etiquetas fueron usadas sólo con fines explicativos, no se utilizó como entrada en ningún modelo. En el cluster 1 se encuentran los textos originarios del oeste de Asia y el Cluster 2 se encuentra los originados al este.

Figura 7. Ubicación geográfica de los clústers

Para finalizar, con la ayuda del software VOSviewer se obtuvo la figura 8, la cual muestra las relaciones entre los términos que componen cada uno de los dos cluster, el cluster de la izquierda, el de color rojo puede ser interpretado como el cluster 1 y el cluster de la derecha, el de color rojo se puede interpretar como el cluster 2.

Donde se puede identificar que muchos términos que están asociados al clúster 2, pertenecen exclusivamente a este clúster, en cambio los términos del clúster 1 pueden relacionarse con el clúster 2.

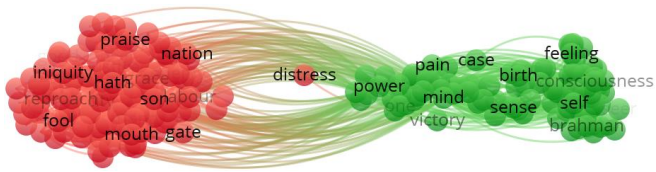

Figura 8. Relaciones entre términos

6 Conclusiones

Se puede concluir que los textos que pertenecen a las religiones del este de Asia: taoísmo, budismo e hinduismo tienen muchas similitudes entre sí, porque a pesar de no ser las mismas religiones los términos que componen sus textos sagrados se relacionan en mayormente entre los miembros del mismo clúster en mayor medida que los textos de la religión cristiana. Esto también se pudo evidenciar cuando se clusterizaba a través de clúster jerárquicos, a medida que K aumentaba se dividía el clúster de religiones del este de Asia, pero se encontraban grupos muy desbalanceados, es decir, el Tao Te Ching y el Upanishads no se separaban y algunos pocos documentos de los textos religiosos formaban pequeños grupos. Por lado de los textos cristianos siempre permanecían juntos ya que pertenecen a la misma religión.

También se pudo diferenciar claramente las posturas de las religiones predominantes en Asia. La religión cristiana que se basa en la presencia de un Dios más poderoso que el humano, quien está en una escala superior, que controla todo y puede castigar o premiar en una vida posterior a la muerte y por otro lado las religiones del este de Asia que cultivan las virtudes de cada persona con el fin de hacer que la estadía en la vida terrenal sea más placentera.

Un potencial trabajo futuro se podría extraer de este análisis direccionándolo a la extracción de características de los textos con LDA. También se podría direccionar el análisis con otros modelos como DBSCAN para hallar posibles nuevos clúster en los documentos con alta densidad.

7 Referencias

[1] 2021. [Online]. Available: https://www.researchgate.net/publication/338137955_What_do_Asian_Religions_Have_in_Common_An_Unsupervised_Text_Analytics_Exploration. [Accessed: 18- Feb- 2021].

[2] "Vista de Religión y sociedad", Sociologicamexico.azc.uam.mx, 2021. [Online]. Available: http://sociologicamexico.azc.uam.mx/index.php/Sociologica/article/view/1548/1593. [Accessed: 18- Feb- 2021].

[3] 2021. [Online]. Available: https://www.researchgate.net/profile/Mayuri_Verma/publication/317608712_Lexical_Analysis_of_Religious_Texts_using_Text_Mining_and_Machine_Learning_Tools/links/5943810a0f7e9b6910eb0376/Lexical-Analysis-of-Religious-Texts-using-Text-Mining-and-Machine-Learning-Tools.pdf. [Accessed: 18- Feb- 2021].

[4] "Cophenetic correlation", En.wikipedia.org, 2021. [Online]. Available: https://en.wikipedia.org/wiki/Cophenetic_correlation. [Accessed: 17- Feb- 2021].