# Universidad de Cuenca

# Reporte Final del Trabajo de Minería de Textos

Abstract (1 párrafo)

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.

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.

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.

1. 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.

1. 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 1.

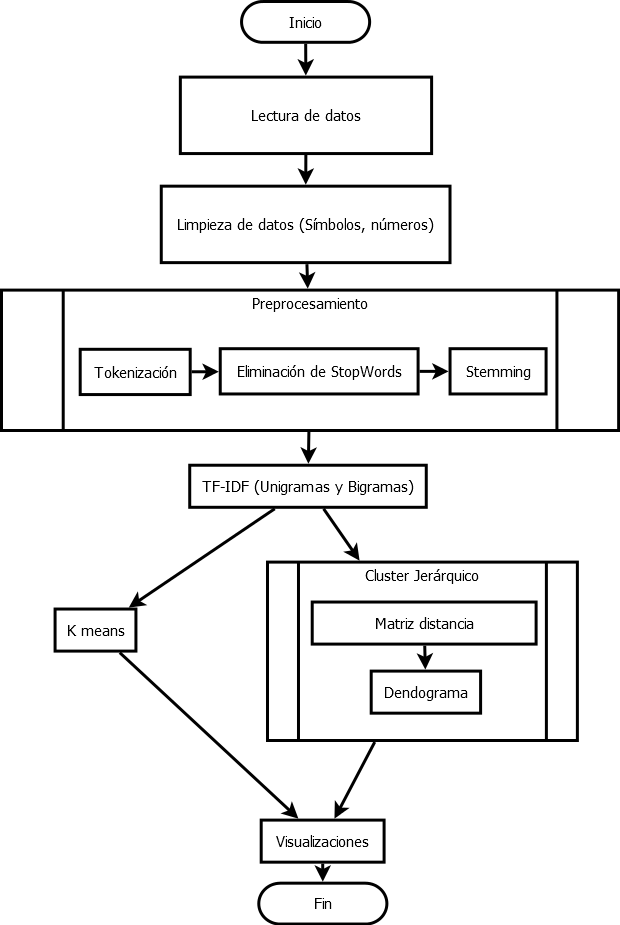


Figura 1. Metodología empleada

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.

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 [2]. Llevado esto al campo de la clusterización 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.

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.

1. Descripción del dataset

El conjunto de datos que se empleó constaba de 590 textos sagrados de las distintas religiones, repartidos como se muestra en la figura 2. Teniendo a YogaSutra como el texto sagrado que tiene mayores documentos y el libro de Eclesiastés con el menor número de documentos.

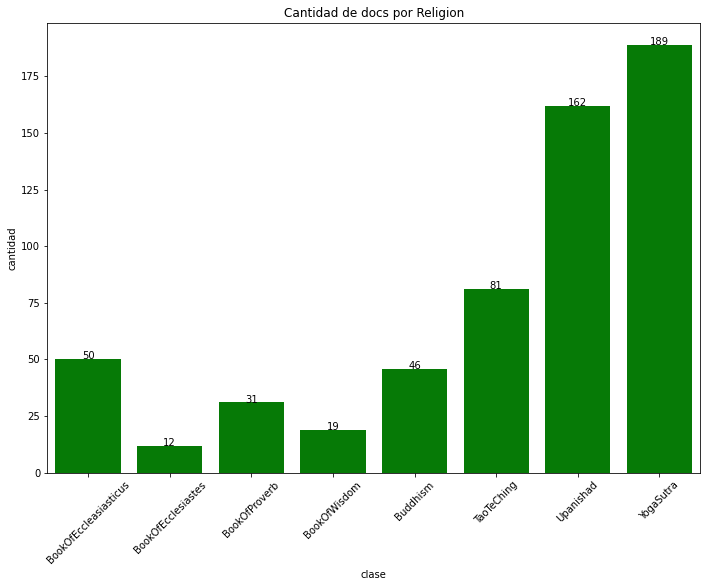


Figura 2. Textos por Religión

Como se mencionó en la metodología se realizó la limpieza de datos, eliminando símbolos y números de los textos, luego se pre procesó el texto eliminando 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.

1. 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 = 5, en la figura 3 se aprecia. Como se puede observar el valor promedio de silueta es bajo a pesar de ser el mejor obtenido.

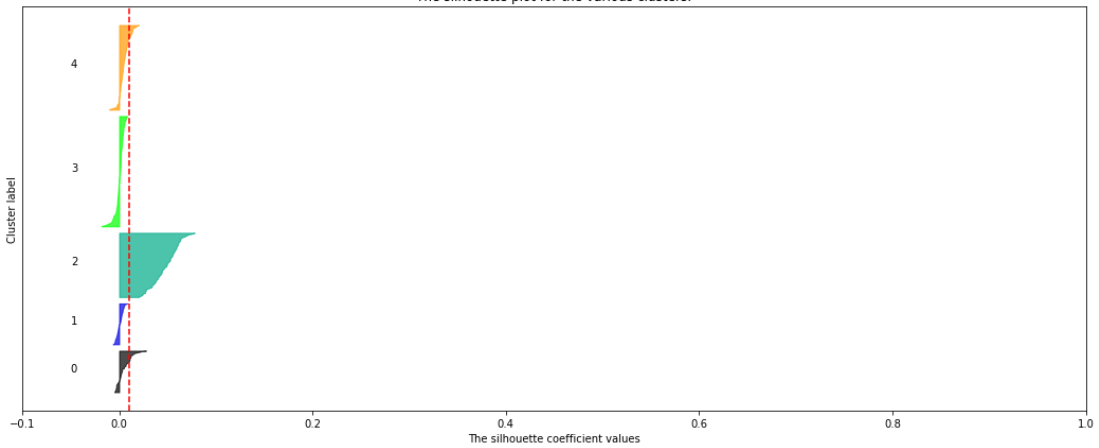


Figura 3. Gráfico de Silueta para 5 clúster

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 K means, donde se puede confirmar que toma a los documentos cristianos como un solo clúster, en cambio a las otras religiones intenta dividirlos.

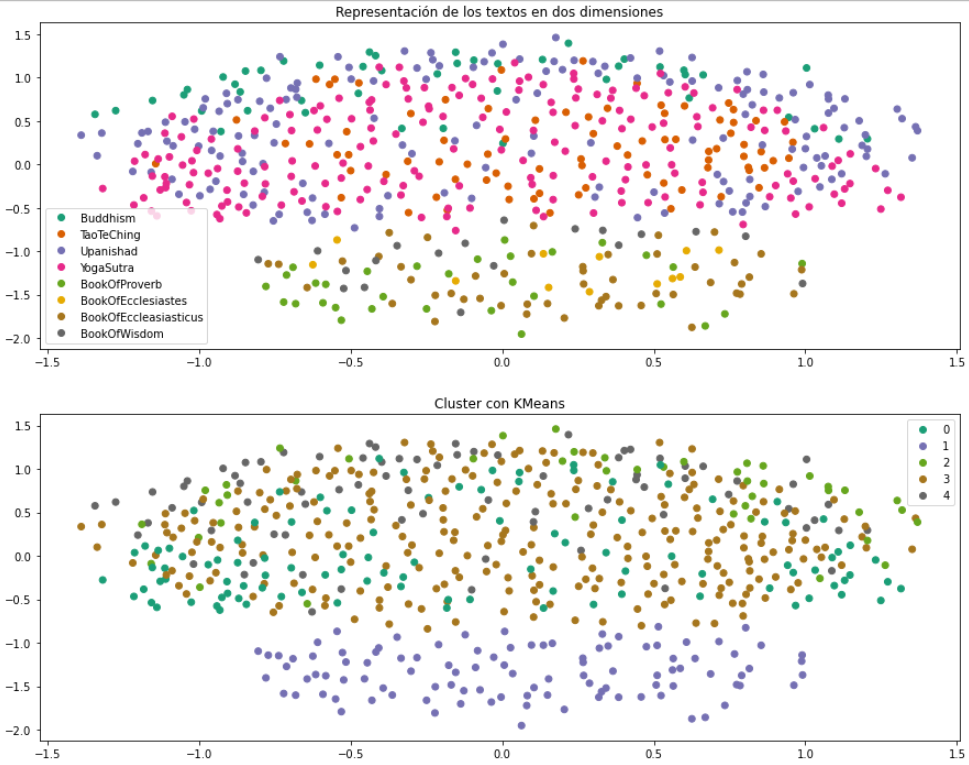


Figura 4. (a) Representación de los textos en dos dimensiones (b) Clúster con K means

**Clúster jerárquico**

Resultados similares se obtiene al aplicar el clúster jerárquico, con un 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 existe un clúster que no se separa, este clúster es el que está conformado por documentos de la religión cristiana, en cambio al ir modificando los valores de K los documentos de las otras religiones si son separados, sin embargo, estos nuevos clústers estaban muy desbalanceados.

Es por eso que se cree conveniente analizar únicamente dos clústers los cuales se muestran en la figura 5.

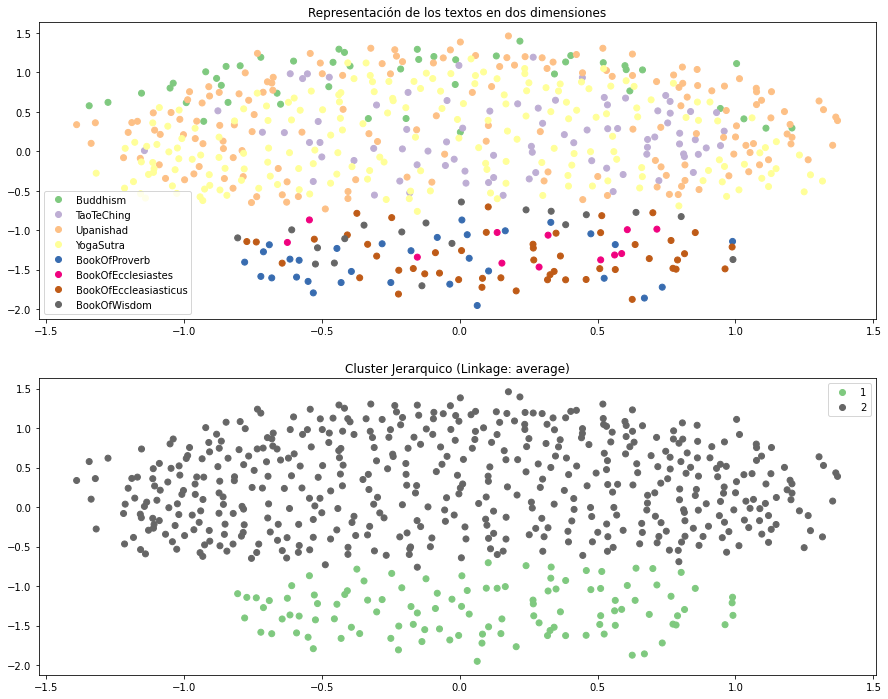


Figura 5. Clúster jerárquico k = 2

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 alguien superior que maneja el bien y el mal. Estos términos están presentes en la religión cristiana. Por otro lado, tenemos términos como: *poder, conciencia, mente, conocimiento uno mismo, saber, sentido, cuerpo, espiritual, primero, cosas.* Se nota ausencia de un ente superior y se centra en el ser humano y sus virtudes, propias de las religiones del este de Asia.



Figura 6. Nube de palabras 2 clúster

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 las etiquetas fueron usadas sólo con fines explicativos, no se utilizó como entrada en ningún modelo.

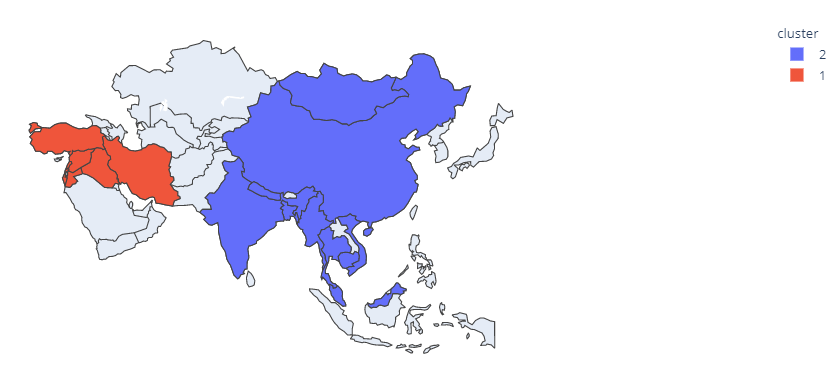


Figura 7. Ubicación geográfica de los clústers

Para finalizar, con la ayuda del software VOSviewwer se obtuvo la figura 8, la cual muestra las relaciones entre los términos, donde se 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.

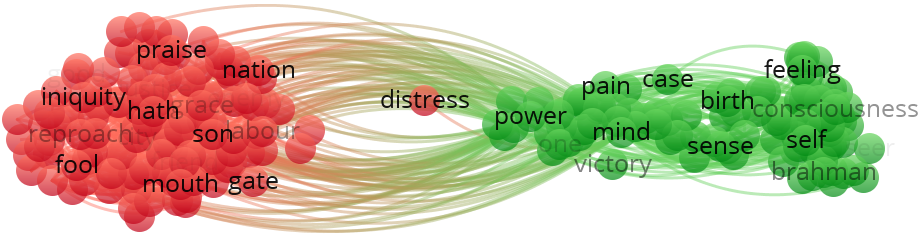


Figura 8. Relaciones entre términos

1. Conclusión

Se puede concluir que los textos que pertenecen a las religiones del este de Asia: taoísmo, budismo e hinduismo tienen muchas similaridades entre sí, porque a pesar de no ser las mismas religiones los términos que componen sus textos sagrados se relacionan 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.

1. Contribuciones

Este proyecto fue realizado por Edwin Narváez en su totalidad.

1. 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]"Cophenetic correlation", *En.wikipedia.org*, 2021. [Online]. Available: https://en.wikipedia.org/wiki/Cophenetic\_correlation. [Accessed: 17- Feb- 2021].