ULPGC
Escuela de Ingeniería Informática

Práctica BOW

José Juan Hernández Gálvez[1] [1]*jose.hernandez219@alu.ulpgc.es*
Jorge Lang-Lenton Ferreiro[2] [2]*jorge.lang101@alu.ulpgc.es*

Resumen

Este estudio aborda la identificación de agrupaciones temáticas en conjuntos de textos mediante técnicas de procesamiento de lenguaje natural y aprendizaje automático. Utilizando la representación numérica proporcionada por el método Bag of Words (BOW), aplicamos el algoritmo de clustering k-means para segmentar los textos en grupos basados en su similitud temática. Evaluamos la coherencia y pertinencia de las configuraciones de clustering utilizando el coeficiente de silueta, una métrica cuantitativa. Nuestros hallazgos sugieren que una configuración de 6 clusters ofrece la segmentación más detallada y relevante, aunque otras configuraciones presentan méritos significativos. El estudio subraya la eficacia de combinar técnicas tradicionales con métodos modernos para analizar y categorizar grandes volúmenes de datos textuales.

1. Introducción

No es algo nuevo que el ser humano tienda a agrupar elementos que se vean similares entre sí, no sólo por seguir un cierto orden en el día a día, sino por añadir simplicidad en el proceso de búsqueda de dichos elementos.

Conociendo ciertas características de esos objetos, podemos ir *redirigiendo nuestra búsqueda* hasta llegar hasta el objeto en concreto. Es lo que conocemos, a día de hoy, como **indexación**.

Por poner ejemplos, empezando por algo grande podríamos destacar una biblioteca. En ella, los libros se encuentran organizados por estanterías, en las cuáles los libros se agrupan por la temática de los mismos (ciencia ficción, novela negra, romance, etc). De hecho, posiblemente estas clasificaciones se puedan redirigir todavía más en base a otro criterio. Por ejemplo, la letra inicial del libro en cuestión.

También un libro pudiera tener un índice según los temas que se traten en el mismo (por ejemplo, en una enciclopedia), lo cual permite a un usuario acceder directamente a la parte del libro que le interese.

Aplicado a un ámbito un poco más cercano, encontramos Google. Su motor de búsqueda pretende buscar una serie de páginas relacionadas con la búsqueda realizada, además de encontrar documentos similares entre sí.

Es decir, buscamos la **similitud** entre elementos. Evidentemente, dicha similitud entre objetos tiene muchas otras utilidades, como pudiera ser la detección de plagio, resumen automático de textos o la recomendación de contenidos.

Por tanto, no podemos negar que la indexación de elementos es crucial, ya que permite ahorrarnos mucho tiempo.

El fin de este paper es dar a conocer uno de los métodos más utilizados en el procesamiento del lenguaje natural para convertir un objeto, en este caso un documento, a un formato que pueda ser utilizado para comprender, organizar y clasificar su información de forma eficaz por parte de una inteligencia artificial. Todo ello nos ayudará, finalmente, a calcular la similitud de un documento con otros, para darle alguna de las utilidades nombradas anteriormente.

2. Preparación y Preprocesamiento de los textos

Antes de adentrarnos en técnicas y métodos avanzados de preprocesamiento, es esencial tener una clara idea del contenido y estructura de nuestros datos. Para ilustrar esto, tenemos un conjunto de 12 archivos de texto. Cada uno de estos documentos tiene una estructura y características específicas que lo distinguen:

doc01.txt : 102 palabras, 5 oraciones, 75 palabras únicas
doc02.txt : 95 palabras, 4 oraciones, 69 palabras únicas
doc03.txt : 84 palabras, 4 oraciones, 63 palabras únicas
doc04.txt : 86 palabras, 5 oraciones, 67 palabras únicas
doc05.txt : 76 palabras, 4 oraciones, 59 palabras únicas
doc06.txt : 81 palabras, 4 oraciones, 63 palabras únicas
doc07.txt : 71 palabras, 3 oraciones, 57 palabras únicas
doc10.txt : 81 palabras, 3 oraciones, 58 palabras únicas
doc08.txt : 75 palabras, 4 oraciones, 60 palabras únicas
doc09.txt : 83 palabras, 4 oraciones, 61 palabras únicas
doc11.txt : 137 palabras, 6 oraciones, 90 palabras únicas
doc12.txt : 143 palabras, 6 oraciones, 93 palabras únicas

Para optimizar la calidad de los datos de texto y facilitar su posterior análisis, es esencial realizar un preprocesamiento adecuado. Los archivos que poseemos presentan varios elementos que podrían interferir en la extracción de información relevante, como signos de puntuación, stopwords, diferencias en mayúsculas y minúsculas, y números. A continuación, describimos el proceso que llevaremos a cabo para cada uno de estos archivos de texto:

1. **Conversión a minúsculas**: La primera etapa del proceso implica convertir todo el texto a minúsculas. Esta acción es crucial para evitar que el sistema reconozca como diferentes palabras que son esencialmente las mismas pero con diferentes formatos, como “Palabra” y “palabra”.

2. **Eliminación de los signos de puntuación**: Los signos de puntuación no aportan un valor semántico significativo cuando se analiza el contenido de un texto en términos de palabras clave o temas. Por tanto, serán eliminados para simplificar el contenido y reducir el ruido.

3. **Tokenización**: Este proceso implica dividir el texto en unidades más pequeñas, llamadas tokens. En el caso de nuestros archivos, tokenizaremos por espacios, convirtiendo cada palabra en un token independiente. Esto facilita la identificación y el tratamiento de palabras específicas.

4. **Eliminación de las stopwords**: Las stopwords son palabras comunes que, aunque necesarias para la estructura gramatical, suelen carecer de significado por sí mismas (por ejemplo, “y”, “de”, “la”, “que” etc.). Estas palabras pueden afectar el análisis posterior, ya que su alta frecuencia podría oscurecer las palabras verdaderamente significativas. Por ello, serán eliminadas de nuestros archivos.

5. **Eliminación de los números**: A menos que los números sean esenciales para el contexto del análisis, a menudo se consideran ruido en los análisis de texto. En este caso, optaremos por eliminar cualquier número presente en los archivos, centrándonos exclusivamente en el contenido textual.

Una vez completado este proceso de preprocesamiento, los textos estarán limpios, simplificados y listos para un análisis más efectivo y eficiente.

2.1. Bag of Words (BOW)

Bag of Words (BOW) es una técnica comúnmente utilizada en el procesamiento de lenguaje natural y la minería de textos para representar documentos o frases como vectores numéricos. En esencia, BOW trata cada documento como una ”bolsa”de palabras, sin tener en cuenta el orden o la estructura gramatical, pero preservando la multiplicidad. El proceso de creación de una representación BOW generalmente implica los siguientes pasos:

- **Construir un Vocabulario**: Se crea un vocabulario de todas las palabras únicas presentes en todos los documentos o frases del conjunto de datos.
- **Codificación One-hot**: Cada palabra del vocabulario se representa mediante un vector en el que todos los elementos son cero, excepto en la posición que corresponde a esa palabra, que es uno.
- **Representar Documentos**: Para cada documento, se suman los vectores one-hot de todas las palabras presentes en ese documento.

Bag of Words (BOW) Ejemplo:

Consideremos las siguientes dos frases ya preprocesadas:
Frase 1: "gato juega"
Frase 2: "perro corre"

El vocabulario (conjunto de palabras únicas) sería:

$$\{\textbf{gato}, \textbf{juega}, \textbf{perro}, \textbf{corre}\}$$

Ahora, representaremos cada palabra del vocabulario con un vector *one-hot*:

$$\textbf{gato} = \begin{bmatrix}1\\0\\0\\0\end{bmatrix}, \quad \textbf{juega} = \begin{bmatrix}0\\1\\0\\0\end{bmatrix}, \quad \textbf{perro} = \begin{bmatrix}0\\0\\1\\0\end{bmatrix}, \quad \textbf{corre} = \begin{bmatrix}0\\0\\0\\1\end{bmatrix}$$

A continuación, representaremos cada frase sumando los vectores *one-hot* de las palabras que contiene:
Frase 1 (gato juega):

$$\textbf{Frase 1 (gato juega)} = \begin{bmatrix}1\\1\\0\\0\end{bmatrix}, \quad \textbf{Frase 2 (perro corre):} = \begin{bmatrix}0\\0\\1\\1\end{bmatrix}$$

3. Análisis de los resultados

Tras concluir la fase de preparación y preprocesamiento de los textos, donde se han eliminado las stopwords, tokenizado las palabras y depurado los signos de puntuación, procederemos a realizar un análisis de los datos.

Para representar los textos en un formato que permita su comparación cuantitativa, utlizaremos el método Bag of Words (**BOW**). Esta técnica permite convertir un texto a una representación numérica, donde cada palabra única del texto se representa por un número, y su frecuencia en el texto determina el valor de ese número en el vector.

Con el objetivo de detectar similitudes entre los textos, empleamos la diferencia del coseno. Esta métrica nos proporciona un valor entre 0 y 1, donde 0 indica que los textos son idénticos y 1 que son completamente diferentes. Para determinar la mejor configuración de agrupación de textos basada en sus similitudes, recurrimos al método de la silueta. Este método evalúa la cohesión y separación de los clusters generados, proporcionando un indicador de la calidad de la agrupación.

3.1. Cálculo de la Similud del Coseno

En el análisis de textos, especialmente cuando se busca determinar similitudes o diferencias entre documentos, la medida de la similitud del coseno es una herramienta esencial. Esta métrica cuantifica la similitud entre dos vectores, en nuestro caso, los vectores que representan los textos en el espacio multidimensional generado por el modelo Bag of Words (**BOW**). La similiutd del coseno se basa en el cálculo del coseno del ángulo entre dos vectores. Un valor cercano a 1 indica que los vectores (o textos) son muy similares, mientras que un valor cercano a 0 sugiere que son muy diferentes. La fórmula para calcular la similitud del coseno entre dos vectores $\mathbf{A}$ y $\mathbf{B}$ es:

$$\text{similitud}(\mathbf{A}, \mathbf{B}) = \frac{\mathbf{A} \cdot \mathbf{B}}{||\mathbf{A}|| \times ||\mathbf{B}||}$$

Donde $\mathbf{A} \cdot \mathbf{B}$ es el producto punto de los vectores y $||\mathbf{A}||$ y $||\mathbf{B}||$ son las normas L2 (o magnitudes) de los vectores $\mathbf{A}$ y $\mathbf{B}$, respectivamente.

Una vez hemos empleado el método de Bag-Of-Words para transformar nuestros documentos en "bolsas de palabras", pasando por los procesos de tokenización, construcción del vocabulario y vectorización de cada uno de los documentos, es hora de utilizar sus resultados para consultar la similitud entre los documentos de los que disponemos, mediante la similitud coseno previamente nombrada.

Podemos ver los resultados a continuación, en la Figura 1.

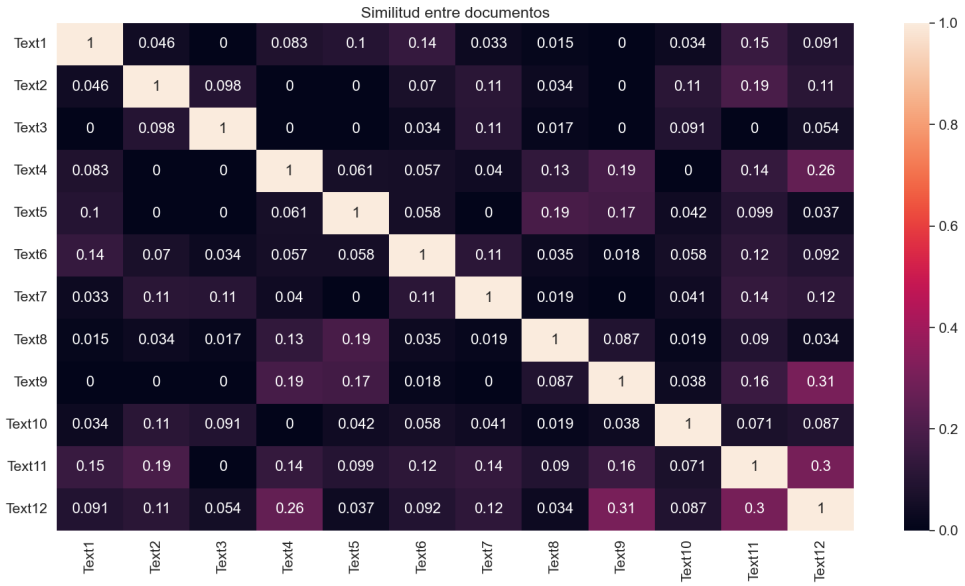

Figura 1: Similitud entre los documentos otorgados

Viendo los resultados de la matriz resultante, la cual proporciona la similitud para cada par de documentos, podemos extraer varias conclusiones.
Teniendo en cuenta que son 12 documentos, la media de similitud entre ellos debería ser de $1/12 = 0.083$. Por tanto, vamos a destacar aquellas similitudes relativamente más grandes con respecto a dicha cantidad.

1. **Similitud entre documento 9 y 12**: El documento 9 "Los supercargadores: carga rápida para coches eléctricos" y el documento 12 "App móvil revoluciona la carga de vehículos eléctricos en la ciudad" tienen una similitud del 31 %.

2. **Similitud entre documento 11 y 12**: El documento 11 "Integración avanzada de smartphones en vehículos eléctricos: El futuro de la conducción conectada" y el documento 12 "App móvil revoluciona la carga de vehículos eléctricos en la ciudad" tienen una similitud del 30 %.

3. **Similitud entre documento 4 y 12**: El documento 4 "El futuro de los coches eléctricos brilla más que nunca" y el documento 12 "App móvil revoluciona la carga de vehículos eléctricos en la ciudad" tienen una similitud del 26 %.

4. **Similitud entre el documento 2 y 11**: El documento 2 "La duración de la batería en smartphones: un desafío constante" tiene una similitud del 19 % con el documento 11 "Integración avanzada de smartphones en vehículos eléctricos": El futuro de la conducción conectada".

5. **Similitud entre documento 5 y 8**: El documento 5 "Subvenciones estatales impulsan la adopción de vehículos eléctricos" y el documento 8 "EcoDrive: el coche eléctrico económico para todos" tienen una similitud, también, del 19 %.

En general, podemos decir que el **documento 12 es uno que tiene bastante similaridad con los demás**, puesto que trata temas tanto sobre los smartphones como sobre los coches eléctricos. **El documento 11 también tiene cierta presencia de la misma manera**, pero no tanto como el anterior documento.

Por otro lado, también **cabe destacar la presencia de los documentos 4, 5, 8 y 9**, ya que presentan similitudes relativamente altas entre sí, alcanzando porcentajes del 13, 17 y 19 %, lo cual tiene bastante sentido teniendo en cuenta que tratan el tema de los coches eléctricos.

La matriz de correlación es simétrica con respecto a su diagonal principal. Esto implica que la similitud del coseno entre las variables a y b es la misma que entre b y a. Matemáticamente, si denotamos a la matriz de correlación como M, entonces para cualquier elemento $M_{i,j}$ (correlación entre la variable i y j), es cierto que $M_{i,j} = M_{j,i}$.

Por último, ya que estamos intentando descubrir similitudes entre documentos y, en definitiva, agrupándolos en cierta medida, vamos a observar si podríamos clasificarlos en **clústers**.

Para decidir cuál es la mejor configuración de clusters, consideraremos tanto la interpretación temática de cada agrupación como el coeficiente de silueta.

3.2. Análisis de Clusters

El análisis de clusters tiene como finalidad descubrir agrupaciones temáticas inherentes en los textos. Aunque existen varios métodos para llevar a cabo esta tarea, como k-means y knn, hemos elegido k-means para evaluar el poder descriptivo del modelo Bag of Words (BOW).

El algoritmo de K-means es una técnica popular de clustering que tiene como objetivo dividir un conjunto de observaciones en k grupos, donde cada observación pertenece al grupo cuyo valor medio es más cercano. Es importante mencionar que es necesario definir el número k de clusters con anticipación.

A continuación se presenta una descripción generalizada del método: Las diferentes configuraciones de clustering exploradas se basan en la similaridad de contenido entre los textos:

- **K-means con 2 Clusters**: Una segmentación general, donde el primero engloba la mayoría de los textos y el segundo destaca la tecnología de carga rápida.
- **K-means con 3 Clusters**: Ofrece un balance al categorizar los textos en smartphones, vehículos eléctricos y su adopción.
- **K-means con 5 Clusters**: Brinda una segmentación más granular con temáticas claras y bien definidas.
- **K-means con 6 Clusters**: Presenta un mayor nivel de detalle, con clusters temáticamente específicos y distintivos.

Para evaluar la calidad de estas agrupaciones, se ha empleado **el coeficiente de silueta**. Este coeficiente mide cuán similares son los objetos dentro de su propio cluster comparado con otros clusters. Valores cercanos a 1 indican agrupaciones adecuadas, mientras que valores cercanos a -1 sugieren que las agrupaciones no son óptimas. Los valores obtenidos para cada configuración son:

- **2 clusters**: 0.2917
- **3 clusters**: 0.3416
- **5 clusters**: 0.4046
- **6 clusters**: 0.4641

Aunque la configuración de 6 clusters obtuvo el coeficiente de silueta más elevado, la elección óptima debe alinear las métricas cuantitativas con los objetivos cualitativos del análisis.

4. Conclusión

A lo largo de este estudio, hemos explorado diversas técnicas y enfoques con el propósito de identificar agrupaciones temáticas en un conjunto de textos. La elección del método Bag of Words (BOW) nos ha permitido convertir el contenido textual en una representación numérica, facilitando la aplicación de algoritmos de clustering. Específicamente, hemos empleado el algoritmo k-means, un método ampliamente reconocido por su capacidad de dividir conjuntos de datos en grupos basándose en su similitud.

El análisis de clusters se centró en distintas configuraciones, con el fin de determinar cuál proporciona una representación más adecuada y coherente de los temas presentes en el conjunto de textos. Además, se ha incorporado el coeficiente de silueta como métrica cuantitativa, brindando un criterio adicional para evaluar la calidad de los clusters generados.

Al considerar tanto las métricas cuantitativas como las interpretaciones cualitativas de los resultados, hemos concluido que, aunque existen varias configuraciones de clustering con mérito, la configuración de 6 clusters parece ofrecer la segmentación más detallada y pertinente. Sin embargo, es esencial recordar que la elección final de la configuración debe basarse en la relevancia y utilidad específicas para el propósito del análisis.

El método utilizado y los hallazgos de este estudio demuestran la importancia y la potencia de combinar técnicas tradicionales de procesamiento de lenguaje natural con algoritmos de aprendizaje automático para desentrañar patrones y tendencias ocultas en grandes conjuntos de datos textuales.

Referencias

[1] Harris, Z. S. (1954). *Distributional structure*. Word, 10(2-3), 146-162.

[2] MacQueen, J. (1967, June). *Some methods for classification and analysis of multivariate observations*. In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability (Vol. 1, No. 14, p. 281-297).

[3] Rousseeuw, P. J. (1987). *Silhouettes: a graphical aid to the interpretation and validation of cluster analysis*. Journal of computational and applied mathematics, 20, 53-65.