UNIVERSIDAD DE SANTIAGO DE CHILE

FACULTAD DE CIENCIA

Departamento de Matemática y Ciencia de la Computación

Predicción de materiales bioplásticos por medio de aprendizaje no supervisado

Fernanda Abril Véliz Durán

Profesor Guía: Felipe Herrera Urbina

Tesis para optar al título de Analista en Computación Científica

Santiago - Chile

2023

RESUMEN

En el estudio de polímeros amigables con el medio ambiente, los principales ficocoloides como alginato, agar y carragenina, han sido prometedores para la realización de films bioplásticos, no obstante, para cumplir con propiedades similares al plástico tradicional, es necesario la incorporación de otros ingredientes. La literatura científica ofrece una amplia gama de experimentos con diversos aditivos que pueden contribuir a la obtención de resultados deseados. Sin embargo, los artículos publicados en ciencia de materiales han aumentado exponencialmente en las últimas décadas, lo cual restringe la formulación de nuevas hipótesis que se puedan proponer.

Este estudio pretende asociar ingredientes que den origen a bioplásticos por medio de aprendizaje no supervisado en contexto de polímeros a base de algas. Se utilizó un conjunto de datos de dos mil abstracts relacionados a polímeros naturales. A través del uso de métodos de procesamiento de lenguaje natural y el modelo Word2vec, se generaron representaciones vectoriales entrenadas para cada palabra del corpus y, mediante el uso de t-SNE, se visualizaron las predicciones correspondientes a cada propiedad y aplicación. El enfoque metodológico es cualitativo y se basa en técnicas de análisis textual.

Los resultados obtenidos indican que algunos ingredientes presentan ciertas tendencias, proporcionando información relevante en cuanto a su comportamiento, tanto para propiedades como aplicaciones aunque esto no siempre se traduce en una consideración óptima. Es importante destacar que, pese a los avances realizados, aún existen ciertos niveles de ruido en los datos, lo que implica que el modelo no es capaz de distinguir adecuadamente entre palabras ambiguas.

Palabras clave: procesamiento de lenguaje natural, aprendizaje no supervisado, bioplásticos

ABSTRACT

In the study of environmentally friendly polymers, the main phycocolloids such as alginate, agar, and carrageenan have been promising for the production of bioplastic films. However, to achieve properties similar to traditional plastic, the incorporation of other ingredients is necessary. Scientific literature offers a wide range of experiments with various additives that can contribute to the desired results. However, articles published in materials science have exponentially increased in recent decades, which restricts the formulation of new hypotheses that can be proposed.

This study aims to associate ingredients that give rise to bioplastics through unsupervised learning in the context of algae-based polymers. A dataset of two thousand abstracts related to natural polymers was used. Through the use of natural language processing methods and the Word2vec model, vector representations trained for each word in the corpus were generated, and predictions corresponding to each property and application were visualized using t-SNE. The methodological approach is qualitative and based on textual analysis techniques.

The results obtained indicate that certain ingredients exhibit certain trends, providing relevant information regarding their behavior, both for properties and applications, although this does not always translate into optimal consideration. It is important to note that, despite the advances made, there are still certain levels of noise in the data, which implies that the model is not able to distinguish adequately between ambiguous words.

Keywords: natural language processing, unsupervised learning, bioplastics

Tabla de contenidos

Índice de tablas

Índice de figuras

Introducción

Los materiales biodegradables son las alternativas más atractivas para la sustitución de polímeros de origen fósil o plásticos tradicionales (Bartolo *et al.*, 2021). Además, estos pueden reemplazar el segmento de empaques, donde la vida útil del empaque es corta y la cantidad de desechos posconsumo es grande y genera problemas significativos con su uso (Izdebska-Podsiadły, 2019). Las algas marinas han sido reconocidas como materia prima sostenible para la producción de bioplásticos dado que no requieren agua dulce, tierra cultivable o fertilizantes para crecer, absorben el exceso de nutrientes del agua de mar y actúan como sumideros de carbono, lo que tiene un efecto mitigador sobre el cambio climático (Radulovich *et al.*, 2015). Sin embargo, la escalabilidad de los films de biopolímeros sigue siendo limitada debido a sus escasas propiedades físicas, mecánicas y de barrera en comparación con los materiales de envasado derivados del petróleo (Blanco-Pascual *et al.*, 2014) (Gomaa *et al.*, 2018). Para abordar este problema, existen diversos aditivos que afectan a distintas propiedades funcionales de los compuestos a base de algas, como las técnicas de incorporación y las aplicaciones, con especial atención en el ámbito alimentario y farmacéutico (Khalil *et al.*, 2017). Sin embargo, es importante tener en cuenta que este campo está en constante evolución y requiere un conocimiento completo y actualizado de la ciencia de los materiales.

Las representaciones distribuidas de palabras en un espacio vectorial ayudan a los algoritmos de aprendizaje a lograr un mejor rendimiento en tareas de procesamiento del lenguaje natural agrupando palabras similares. El algoritmo de Word2vec, especialmente el modelo Skip-gram, es un método eficiente para aprender representaciones vectoriales de alta calidad de palabras a partir de grandes cantidades de datos de texto no estructurados (Mikolov, Chen *et al.*, 2013a). Estos vectores se pueden utilizar para analizar y comparar las similitudes entre los términos, lo que permite una mejor comprensión de las relaciones

y patrones en los datos permitiendo capturar conceptos de polímeros como las relaciones de propiedad en los ingredientes estudiados. Además, este método no supervisado puede recomendar ingredientes para aplicaciones, ya que tiene la capacidad de extraer conocimiento y relaciones del cuerpo masivo de literatura científica de manera colectiva (Tshitoyan *et al.*, 2019).

Objetivo general

El objetivo de este trabajo de titulación es predecir por aprendizaje no supervisado set de propiedades que se puedan asociar a un set de ingredientes que den origen a bioplásticos. Este trabajo muestra una perspectiva desde la ciencia de la computación de cómo resolver un problema multidisciplinario como lo es la ciencia de materiales a través del aprendizaje no supervisado, analizando datos e implementando métodos y algoritmos capaces de comprender el lenguaje natural.

Objetivos específicos

- Analizar las técnicas de aprendizaje no supervisado
- Caracterizar el material bioplástico
- Implementar modelos para entrenar Word Embeddings
- Analizar resultados de la aplicación del modelo

Capítulo 1

Marco Teórico

En este capítulo se expone información sobre el Procesamiento de Lenguaje Natural y las ramas que lo componen. También se mencionarán distintas representaciones de palabras para trabajar con Word2Vec junto con su modelo Skip-gram para la predicción de palabras contexto y finalmente su aplicación en literatura científica.

1.1. Procesamiento de Lenguaje Natural

El procesamiento de lenguaje natural (desde ahora PLN) es un área de estudio donde se procesa el lenguaje humano utilizando inteligencia artificial y la lingüística diseñando métodos para que el computador procese y entienda el lenguaje humano. Hoy en día, el PLN es utilizado en varias aplicaciones como la computación lingüística, modelos de clasificación, traducción automática y sistemas de búsqueda de información. El PLN se enfoca en inspeccionar modelos e interpretar los vínculos de texto para examinar de forma efectiva grandes volúmenes de datos (Moreira *et al.*, s.f.).

El lenguaje natural está compuesto por caracteres, los cuales forman palabras y a su vez un conjunto de palabras forman conceptos, eventos, ideas y acciones. Los caracteres y las palabras son símbolos discretos que representan una imagen mental pero a su vez pueden ser símbolos distintos, como por ejemplo las palabras pizza y hamburguesa (Goldberg, 2017, p. 1). Además las palabras son de carácter compositivo ya que un conjunto de palabras puede formar infinitas frases y oraciones que tienen un mayor significado que simples palabras individuales. Existen métodos de aprendizaje supervisado donde el algoritmo intenta inferir patrones y regularidades a partir de un

documento, por ejemplo, tareas de clasificación de texto según su categoría identificando palabras claves y creando patrones (Vajjala *et al.*, 2020, p. 6). Algunas tareas principales de PLN son:

- *Language modelling*: Esta tarea se encarga de predecir qué palabra es la siguiente en una oración basándose en las palabras anteriores, donde el objetivo es aprender dependiendo de la probabilidad en que las palabras aparecen de acuerdo a un texto. Esta tarea es utilizada para solucionar problemas como el reconocimiento de voz y escritura de caracteres, traducción y corrección de ortografía.

- *Text Classification*: Esta tarea agrupa el texto en conjuntos de categorías dependiendo del contenido que este tenga, por ejemplo, clasificar si un email es de categoría spam o no.

- *Sistemas de PLN especializados*: Existen sistemas de conocimiento y reglas contextuales para identificar frases específicas en artículos científicos. Este es un análisis sintáctico, el cual implica la simplificación de un conjunto de oraciones de muestra que contienen verbos de algún área de interés (Friedman *et al.*, 2001).

1.2. Machine Learning

La inteligencia artificial (IA) es una rama de la ciencia de la computación donde se desarrollan sistemas que puedan realizar tareas que requieran inteligencia humana. Machine Learning es una rama de la IA donde se utilizan algoritmos que pueden aprender a realizar tareas automáticamente basado en un gran número de ejemplos utilizando una representación numérica o característica para entrenar los datos y aprender patrones (Vajjala *et al.*, 2020, p. 14). Esta área está dividida en varias subcategorías. Las categorías más utilizadas son el aprendizaje no supervisado y el aprendizaje supervisado,

- *Aprendizaje no supervisado*: Como su nombre lo indica, este aprendizaje no necesita de supervisión, es decir, no necesita datos etiquetados sino que sólo datos, donde se identifican qué patrones ocurren con más frecuencia que otros en una estructura dada, de ello ver qué sucede y que no. Una de las aplicaciones que más utiliza este aprendizaje es el *clústering* , donde su objetivo es encontrar ciertas agrupaciones o

grupos en el input según sus atributos que tengan mayor similitud (Alpaydin & Bach, 2014).

- *Aprendizaje supervisado*: En esta categoría los datos consisten en ser características (*features*) y tener un objetivo (*target*) el cual puede ser de forma cuantitativa (de regresión) o categórica (clasificación). Bajo estas circunstancias, llamamos al conjunto de datos un conjunto de datos etiquetado (Kyriakides & Margaritis, 2019)

Para resolver tareas de aprendizaje supervisado o no supervisado, se necesita una serie de pasos para procesar los datos de textos, estos son la extracción de características del texto, usar la representación de características para aprender un modelo, y evaluar y mejorar el modelo.

1.2.1. Machine learning en PLN

Las técnicas de Machine learning (ML) se aplican para varias formas de datos como las imágenes, audios y textos. En este último caso, existen varias técnicas generales disponibles, como la traducción de texto. La traducción de texto consiste en convertir una secuencia de tokens discretos de un vocabulario a otra secuencia de tokens discretos en otro vocabulario, debido a que el PLN tiene la característica de ser datos discretos. El Machine learning al ser una rama de la inteligencia artificial busca como función principal desarrollar computacionalmente aplicaciones con el mismo rango de las habilidades humanas (Russell Stuart & Norvig, 2009).

La habilidad del lenguaje en el ser humano es una de las características de inteligencia más básicas que posee, por lo tanto ya es considerado como un prerrequisito para la inteligencia artificial. Sin embargo al procesar el lenguaje se debe considerar el razonamiento que existe en las frases, el esquema de Winograd hace un desafío de razonamiento utilizando un par de oraciones que difieren en una o dos palabras con un pronombre muy ambiguo y para resolver correctamente el significado se necesita un conocimiento previo y por lo tanto la máquina deben tener una comprensión profunda del contenido del texto, este desafío se presenta como una alternativa al Test de Turing (Kocijan *et al.*, 2020).

Existen dos tipos de evaluación, estas son: evaluación intrínseca y evaluación extrínseca.

- *Evaluación Intrínseca*: Esta forma de evaluación es ampliamente utilizada en sistemas de Procesamiento del Lenguaje Natural (PLN) y se centra en medir el rendimiento del sistema utilizando medidas como precisión y recuperación. Uno de los indicadores más importantes es la *exactitud*, que indica la fracción de veces que el modelo realiza predicciones correctas en comparación con el total de predicciones realizadas.

- *Evaluación Extrínseca*: Está centrada en la evaluación del desempeño del modelo en el objetivo final, es decir, en el resultado real. Este tipo de evaluación son las pruebas definitivas para garantizar que el modelo realmente es útil.

1.3. Deep Learning

Deep learning es una rama del Machine Learning el cual se caracteriza por aprender predicciones basadas en observaciones pasadas y representar datos. Este aprendizaje funciona alimentando a una red que produce transformaciones sucesivas de datos de entrada hasta que la transformación final predice una salida, estas redes se llaman redes neuronales (Russell Stuart & Norvig, 2009).

1.3.1. Redes neuronales

Las redes neuronales son un tipo de aprendizaje inspirado en la forma en que funciona la computación en el cerebro. Este tipo de aprendizaje se basa en una red de perceptrones interconectados (Prieto *et al.*, 2016). El modelo de perceptrón es la red neuronal más simple, la cual aprende realizando un mapeo lineal (suma de productos de pesos o características) basado en la entrada y salida cuando se entrena en un conjunto de datos etiquetados. Para entrenar las redes neuronales, se utiliza el algoritmo *Backpropagation* y el conjunto inicial de pares de vectores (input, output). La figura 1.1 muestra cómo cada entrada x tiene asociado un peso W, y la neurona multiplica cada entrada por sus pesos y luego realiza una suma para producir la salida.

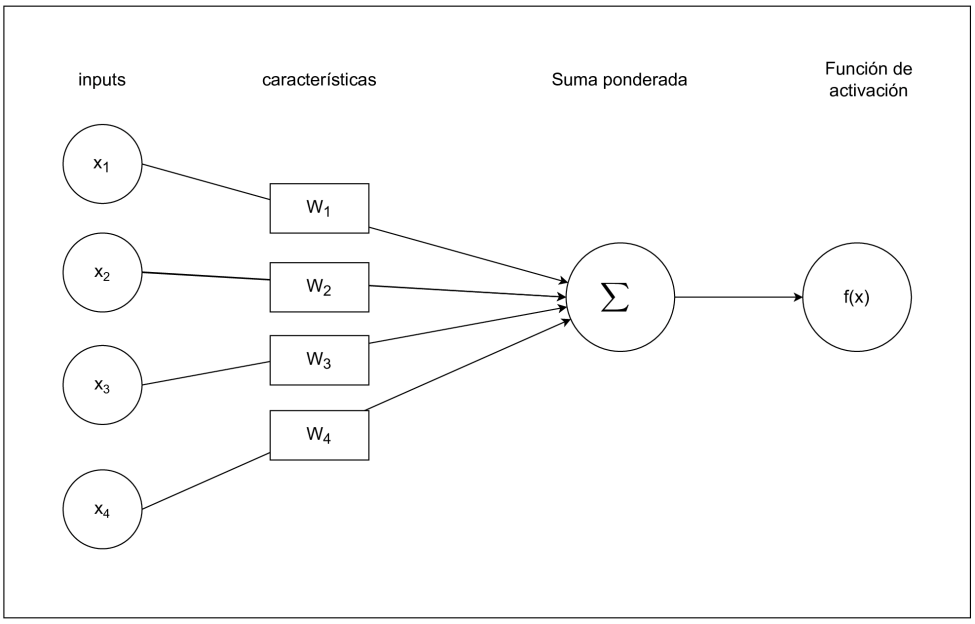

Figura 1.1: Representación de un perceptrón. Fuente: Adaptado de Arumugam y Shanmugamani (2018)

El mapeo final pasa a través de una función de activación $f(x)$ la cual transforma la entrada de un nodo de red neuronal a una salida no lineal. Esto permite que la red aprenda mapeos no lineales a partir de los datos (Arumugam & Shanmugamani, 2018). Estos mapeos producen que los datos tengan transformaciones donde facilita la relación de los datos con la etiqueta deseada.

Las redes neuronales (NN) *Feed-Forward* tiene una arquitectura donde las neuronas están conectadas a cada una formando una red y cada output de la neurona alimenta a cada input.

Como se muestra en la figura 1.2, cada círculo representa una neurona. La capa inferior no tiene flechas entrantes y es el input de la red (Goldberg, 2016). La capa superior no tiene flechas salientes el cual es el output de la red. El input es un vector de 4 dimensiones x, la siguiente capa es una transformación lineal de 4 dimensiones a 6 dimensiones (h). Formalmente las capas conectadas realizan una multiplicación vectorial $xW = h$, donde el peso de la conexión desde la i-ésima neurona en la fila de entrada hasta la j-ésima neurona en la fila de salida es w_{ij}. Los valores de h son:

$$h_j = \sum_{i=1}^{4} x_i \cdot w_{ij}$$

Luego estos valores son transformados por una función no lineal g que son aplicados a cada valor antes de pasar por al siguiente input. El resultado completo de esta red se escribe como:

$$(g(W^1))W^2$$

Donde W^1 son los pesos de la primera capa y W^2 son los pesos de la segunda capa.

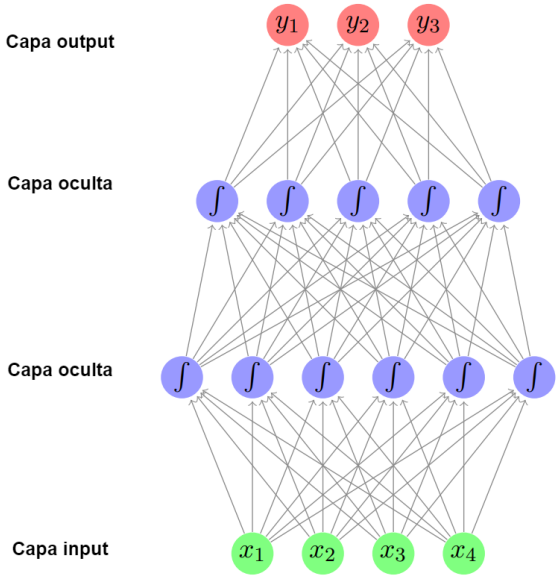

Figura 1.2: Representación de una red neuronal Feed-forward con dos capas ocultas. Fuente: Goldberg (2016)

Existen varias funciones no lineales g, comúnmente se utilizan la sigmoide o también conocida como función logística la cual consiste en $\sigma(x) = \frac{1}{(1+e^{-x})}$, esta transforma cada valor x en un rango de [0,1]. Sin embargo, la función de activación que ha obtenido mejores resultados y ha demostrado ser más sencilla es ReLU. Esto se debe a que no involucra funciones costosas de calcular, además de ser la opción más adecuada para redes neuronales de múltiples capas (Glorot *et al.*, 2011).. La función ReLU se representa en la ecuación 1.1 donde recorta cada valor x <0 a 0.

$$ReLU(X) = max(0, x) = \begin{cases} 0 & si \quad x < 2 \\ x & si \quad \text{otro caso} \end{cases} \tag{1.1}$$

Para la predicción de palabras, es útil tener una distribución de probabilidades sobre posibles vectores de salida, por lo tanto se utiliza una transformación en la última

capa output. Comúnmente y para fines de esta tesis se utilizará la función *softmax*, la cual se define en la ecuación 1.2 como:

$$softmax(x_i) = \frac{e^{x_i}}{\sum_{j=1}^{k} e^{x_j}} \tag{1.2}$$

El resultado es un vector de números reales no negativos que suman uno.

1.3.2. Redes neuronales en PLN

Las redes neuronales es un aprendizaje útil para los problemas de lenguaje natural debido a la manera de manejar datos complejos y no estructurados como lo es el lenguaje. Se utilizan redes neuronales para maximizar el rendimiento de la traducción y para producir buenos resultados en oraciones más largas (Bahdanau *et al.*, 2016).

El uso de una capa de embedding es el componente principal para el procesamiento del lenguaje natural. A través de esta capa, se realiza un mapeo de símbolos discretos a vectores continuos, donde se asocia cada palabra o característica con un vector de *d-dimensiones*. Esta transformación convierte las palabras en objetos matemáticos representados por vectores, los cuales son aprendidos por la red neuronal durante el proceso de entrenamiento.

A medida que la red neuronal procesa los vectores de entrada, aprende a combinarlos de manera efectiva para producir un vector de salida, que puede ser utilizado para hacer predicciones.

Formalmente, los parámetros de una capa de embedding es una matriz $E \in R^{|v| \times d}$, la cual tiene filas del tamaño del vocabulario v y d columnas (Goldberg, 2017, p. 117). Cuando una palabra está representada como un vector *one-hot* $\vec{x}$ y que requiera localizar una capa de embedding, la matriz E se multiplica con el vector $\vec{x}$.

1.4. Representación de texto

Las representaciones de texto son muy importantes para realizar varias aplicaciones en el mundo real, por ejemplo: búsqueda, recomendaciones de anuncios, ranking y filtros de spam (Mikolov, 2016). Las palabras pueden ser representadas matemáticamente capturando la forma semántica de cada texto utilizando vectores numéricos de palabras.

El lenguaje natural representa características discretas como las palabras y letras, las cuales se pueden codificar para ser procesados en la tarea de predecir palabras que se encuentran en el vocabulario. A continuación se presentará el método One-Hot el cual será utilizado en el modelo Skip-Gram, además de la representación mediante a los word embeddings.

1.4.1. Reducción de vocabulario

El Vocabulario V se define como el conjunto de todos los términos distintos presentes en una colección de documentos, conocida como *corpus*. Durante el preprocesamiento de textos, se utilizan diversas técnicas para reducir el tamaño del vocabulario y eliminar términos no significativos. Por lo general, las palabras con mayor frecuencia suelen ser poco informativas, como los pronombres, preposiciones y conjunciones, que se consideran palabras vacías o *stop words*.

Asimismo, existen técnicas como el *stemming* y la *lematización* que permiten transformar los términos en un formato estándar. Estos procesos reducen las palabras a sus raíces léxicas, por ejemplo, todas las posibles conjugaciones de los verbos se transforman en una misma palabra (Perkins *et al.*, 2016).

1.4.2. Bag of Words

Una de las clásicas representaciones de texto utilizadas en varias técnicas en PLN son las Bag of Words. Es una colección de palabras sin considerar el orden y el contexto en que aparecen. Si dos piezas de texto tienen casi las mismas palabras, entonces pertenecen a la misma bag of Words. Cada documento del corpus es convertido en un vector de $|V|$ dimensiones, donde cada i-ésimo componente del vector es el número de veces en que la palabra w ocurre en el documento (Vajjala *et al.*, 2020, p. 87).

1.4.3. One-Hot Encoding

Los vectores One-Hot son representaciones de palabras cuya dimensión es el tamaño del vocabulario, se representan mediante 0s y una única entrada 1. Se caracteriza por contener información sobre cada palabra del documento sin tener que considerar el orden de este, por lo tanto se considera un *bag-of-single-word* (Goldberg, 2017, p. 23).

Esta representación depende de la posición en el cual se encuentra la palabra. Por ejemplo, en el vector 1.3, si el vocabulario es la frase “el material es largo”, entonces el vector asociado para representar la palabra “material” tendrá un 1 y 0s para el resto.

$$\begin{bmatrix} el \\ material \\ es \\ largo \end{bmatrix} \rightarrow \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} \tag{1.3}$$

Su notación matemática está data por la fórmula 1.4, donde un vector de entrada x contiene un conteo de *bigramas*, es decir, pares de letras consecutivas que aparecen en un documento D. Los bigramas están representados por D_i donde i es la posición de la palabra y cada vector $x^{D_{[i]}}$ es un vector one-hot

$$x = \frac{1}{|D|} \sum_{i=1}^{D} x^{D_{[i]}} \tag{1.4}$$

Cuando se aumenta el vocabulario, su dimensionalidad también se incrementa.

1.4.4. Word Embeddings

Los “Word Embeddings” o vectores de palabras son representaciones particulares de palabras a vectores de números reales para representar el significado semántico y sintáctico de estas. Se distribuyen mediante la hipótesis de distribución lingüística, donde las palabras que ocurren en contextos similares tienen significados similares (Harris, 1954). Cuando las representaciones de palabras son aprendidas en un corpus grande de texto, capturan relaciones semánticas y sintácticas de palabras y por lo tanto vectores de palabras similares tienen vectores similares.

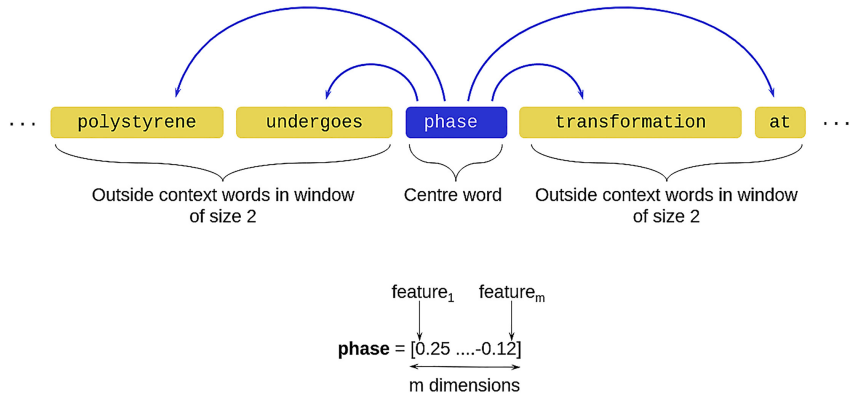

Figura 1.3: Ejemplo de representación de palabra dado su contexto. Fuente: Shetty y Ramprasad (2021)

En la imagen 1.3 se muestra un ejemplo de una representación de word embedding de la palabra *phase* utilizando la arquitectura Skip-gram. La frase es “polystyrene undergoes transformation at” y tiene como palabras contexto *polystyrene*, *undergoes*, *transformation*, *at*.

En concreto, “las regularidades se observan como compensaciones vectoriales constantes entre pares de palabras que comparten una relación particular” (Mikolov, Yih *et al.*, 2013, p. 746), por ejemplo, las palabras (“piña”) y (“fruta”) están más juntas en el espacio vectorial que (“piña”) y (“auto”). Estos vectores son el componente principal para utilizar redes neuronales debido a la capa de embedding.

Para capturar las propiedades distributivas de las palabras y medir su similitud, se utilizan las matrices de contexto, donde cada fila i representa una palabra y cada columna j representa un contexto lingüístico en que las palabras pueden ocurrir (Bullinaria & Levy, 2007). La matriz $M_{[i,j]}$ cuantifica la fuerza f de asociación entre una palabra y un contexto en un corpus. La definición formal de la matriz de contexto se representa en 1.5 donde V_W es el conjunto de palabras en el vocabulario y V_C un conjunto de posibles contextos en los que ocurre.

$$M^f \in \mathbb{R}^{|V_W| \times |V_C|} \tag{1.5}$$

Existen diferentes definiciones de contextos y formas de medir la asociación entre una palabra y un contexto, los cuales obtienen diferentes representaciones de palabras.

Los contextos son las palabras que rodean a una palabra *target* los cuales dan una

característica a la frase, ya que las palabras que están más cerca del target a menudo son más informativas sobre ella que las palabras que están más separadas. Los contextos se denominan como *windows*, por ejemplo, si la siguiente oración es "*the brown fox jumped over the lazy dog*" y la palabra target es *jumped* (Goldberg, 2017, p. 70). Si su windows es igual 2 entonces el producto será un conjunto de las siguientes características:

$$\{\text{palabra = brown, palabra = fox, palabra = over, palabra = the}\}$$

Similitud de Word Embeddings

Para calcular la similitud de los vectores de palabras, se calcula que tan cerca están dos vectores (dos palabras). Las formas que se utilizan para calcular la similitud entre vectores es utilizando la similitud de coseno y la distancia euclidiana. Sin embargo, esta última no siempre es precisa debido a que al comparar corpus de diferentes tamaños, tienen distancias mayores, por lo tanto su similitud cambia (Vajjala *et al.*, 2020). Para resolver ese problema, comúnmente se utiliza la similitud de coseno.

La similitud del coseno es el ángulo entre los vectores correspondientes. Sean dos vectores, A y B respectivamente, cada uno con n componentes, la similitud entre ellos se calcula como:

$$\text{similitud} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{|\mathbf{A}||\mathbf{B}|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \tag{1.6}$$

donde A_i y B_i son los i-ésimo componentes de los vectores A y B respectivamente.

1.5. Word2Vec

Word2Vec es un algoritmo creado por Google que implementa dos arquitecturas para entrenar word embeddings: Skip-Gram y Continuos Bag of Word (CBOW). Estas arquitecturas implementan dos modelos de optimización para entrenar los parámetros: Hierarchical Softmax y Negative Sampling. Los modelos Skip-gram y CBOW son típicamente entrenados usando descenso de gradiente estocástico. El gradiente se calcula utilizando la regla de backpropagation (Rumelhart *et al.*, 1986). Estos son modelos de

redes neuronales para distinguir grupos de palabras que realmente coexisten de palabras agrupadas al azar.

La capa de entrada toma una representación dispersa de una palabra de destino junto con una o más palabras de contexto (Ma & Zhang, 2015). Word2Vec ha demostrado tener significados semánticos útiles para desarrollar tareas en PLN, se considera muy eficaz y altamente escalables permitiendo entrenar word embeddings con vocabularios muy amplios sobre miles de millones de palabras de texto en cuestión de horas (Mikolov, Chen *et al.*, 2013b). Las representaciones de palabras que resultan de las redes neuronales codifican explícitamente muchas regularidades y patrones lingüísticos. Por ejemplo, el resultado de un cálculo vectorial vec("Madrid") - vec("España") + vec("Francia") está más cerca de vec("París") que de cualquier otro vector de palabras por la distancia del coseno (Mikolov, Chen *et al.*, 2013a).

1.5.1. Modelo Skip-Gram

El modelo Skip-gram es una red neuronal de aprendizaje no supervisado con una sola capa oculta. El objetivo de esta red es entrenar para encontrar representaciones de palabras, la cual predice las palabras que rodean a una palabra en específico. Recibe como input una palabra *target* $w(t)$ y el output son las palabras que se encuentran en su contexto *windows* de tamaño k como se muestra en la figura 1.4. Al usar contexto posicional junto con windows más pequeñas, tiende a producir similitudes que son más sintácticas, con una fuerte tendencia a agrupar palabras que comparten una parte de la oración. Además si se consideran windows de mayor tamaño, entonces el resultado son vectores de palabras que capturan el mismo tema (Goldberg, 2017, p. 113).

Las palabras iniciales son representadas como vectores one-hot y se define una capa oculta de dimensión d, la cual determina el tamaño de la matriz de embedding, esta debe ser menor al tamaño del vocabulario, es decir, $d < |V|$ debido a que logra un buen rendimiento en el conjunto de datos de la frase (Mikolov, Chen *et al.*, 2013a).

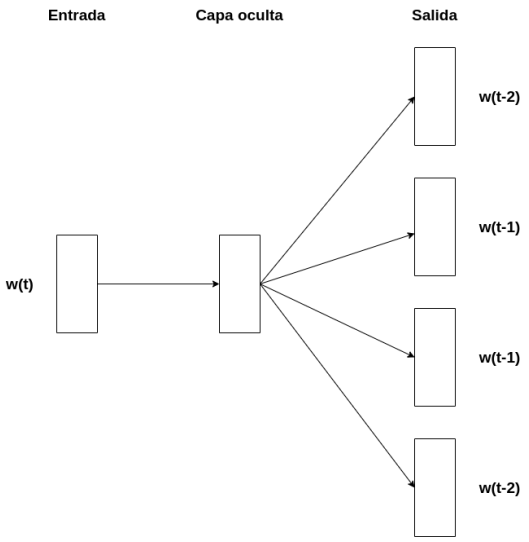

Figura 1.4: Representación de arquitectura Skip-gram. Fuente: Adaptado de Mikolov, Chen *et al.* (2013a)

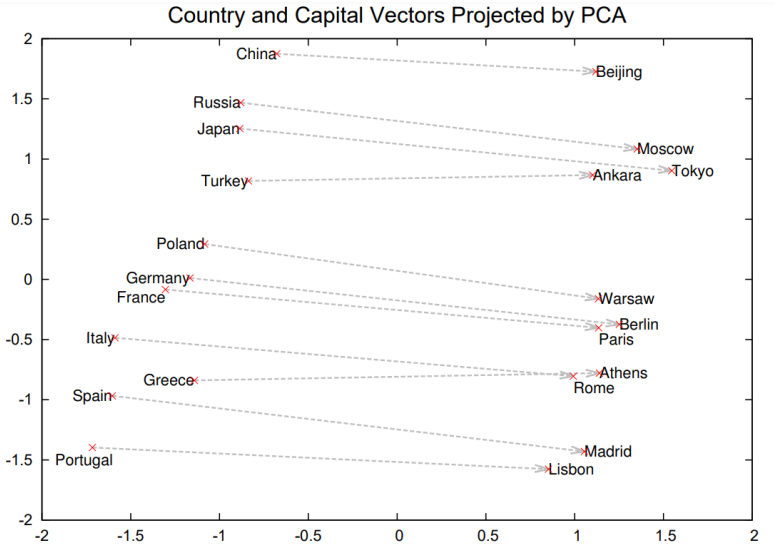

Figura 1.5: Proyección PCA bidimensional de los vectores Skip-gram de 1000 dimensiones de países y sus capitales. Fuente: Mikolov, Chen *et al.* (2013a)

En la figura 1.5 se muestra el modelo Skip-gram, donde sin la necesidad de proporcionar alguna información supervisada, este puede organizar conceptos automáticamente y aprender implícitamente las relaciones entre ellos. Generalmente se utilizan visualizaciones como PCA y t-SNE para reducir un conjunto de datos complejos a un espacio dimensional más bajo. Lo cual ayuda a obtener una base más significativa para volver a

expresar cualquier conjunto de datos ruidoso.

Al definir un windows de tamaño k, se calcularán las palabras contexto c la cual será una secuencia de c_{1-k} de una palabra target w_t. Formalmente dada una secuencia de palabras de entrenamiento $w_1, w_2, w_3, ..., w_T$, el objetivo del modelo Skip-gram es maximizar la probabilidad promedio las palabras contexto c dada la palabra target w_t, esto queda representado en 1.7.

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{c\in c_{1:k}}\log P(c|w_t) \tag{1.7}$$

Un tamaño mayor de c da lugar a más ejemplos de entrenamiento y, por lo tanto, puede conducir a una mayor precisión a expensas del tiempo de entrenamiento. La probabilidad condicional se modela con la función softmax 1.8.

$$P(c|w) = \frac{e^{\vec{c}\cdot\vec{w}}}{\sum_{c'\in C} e^{\vec{c'}\cdot\vec{w}}} \tag{1.8}$$

Donde C es el conjunto de todas las palabras contexto que suele ser el mismo que el vocabulario. La red neuronal aprenderá las estadísticas cuando la palabra target y una palabra contexto ocurre más frecuentemente juntas en el texto, entonces su probabilidad será mayor. En la figura 1.6 se muestra una representación del entrenamiento con la frase ¨The quick brown fox jumps over the lazy dog.” con windows de tamaño 2. La palabra target (azul) y el contexto (todo el vocabulario) son los pares de palabras que utilizará la arquitectura Skip-gram para calcular la softmax.

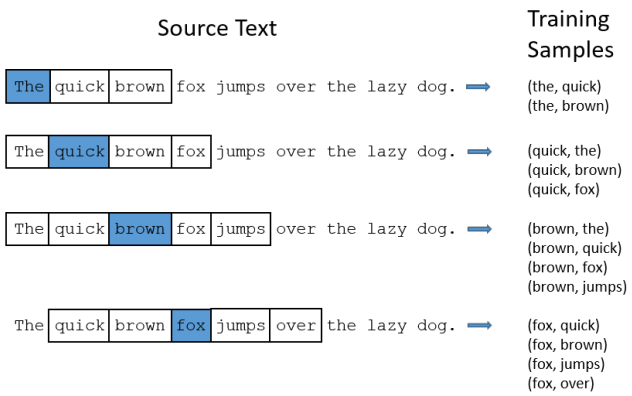

Figura 1.6: Entrenamiento de Skip-gram a partir de la cantidad de veces que aparece cada palabra-contexto. Fuente: McCormick (2016)

Al maximizar 1.7 se obtiene la siguiente ecuación:

$$\arg\max_{\vec{c},\vec{w}} \sum_{(w,c)\in D} \log P(c|w) = \sum_{(w,c)\in D} (\log e^{\vec{c}\cdot\vec{w}} - \log \sum_{c'\in C} e^{\vec{c'}\cdot\vec{w}}) \tag{1.9}$$

La ecuación 1.9 dará como resultado buenas representaciones de palabras $\vec{w}\ \forall w\ \in V$, donde palabras similares tendrán vectores similares (Goldberg *et al.*, 2014). Sin embargo, 1.9 es costoso computacionalmente, P(c|w) es muy caro de calcular debido a la suma $\sum_{c'\in C} e^{\vec{c'}\cdot\vec{w}}$ para todos los contextos c'. Para solucionar el costo, existe un modelo más eficiente para calcular la representación por Skip-gram llamado *Negative Sampling* (Mikolov, Chen *et al.*, 2013b), el cual maximiza la probabilidad de un par palabra-contexto que proceden del conjunto del corpus original D mediante una función sigmoide, este no itera sobre todo el vocabulario por lo tanto hace más veloz el entrenamiento.

Negative Sampling de Word2Vec funciona entrenando la red para distinguir pares de palabra-contexto "buenos" (i.e, palabras que ocurren en el corpus) de los "malos". Word2Vec reemplaza el objetivo de clasificación basado en el margen por uno probabilístico. Se considera el conjunto D los pares de palabra-contexto y $\bar{D}$ como los pares de palabra-contexto incorrectos. El objetivo del algoritmo es estimar la probabilidad $P(D = 1|w, c_i)$ que el par palabra-contexto proviene del conjunto correcto D (Goldberg, 2017, p. 124). Este debe ser alto (1) para pares de D y bajo (0) para pares de $\bar{D}$. La restricción de probabilidad dicta que $P(D = 1|w, c_i) = 1 - P(D = 0|w, c_i)$. La función de probabilidad se modela como un sigmoide:

$$P(D = 1|w, c_i) = \frac{1}{1 + e^{-\vec{w}\cdot\vec{c_i}}} \tag{1.10}$$

El objetivo de todo el corpus del algoritmo es maximizar la probabilidad logarítmica de los datos $D \cup \bar{D}$:

$$\arg\max_{\vec{c},\vec{w}} \sum_{(w,c)\in D} \log P(D = 1|w, c_i) + \sum_{(w,c)\in \bar{D}} \log P(D = 0|w, c_i) \tag{1.11}$$

En un gran corpus de texto existen muchas palabras que se repiten frecuentemente y que no entregan mayor información, por ejemplo, "el" , "los", "se", "un". Es por ello que existe una técnica de submuestreo para contrarrestar el desequilibrio entre las palabras y las palabras frecuentes vacías: cada palabra w_i del conjunto de entrenamiento se descarta con una probabilidad calculada mediante la fórmula 1.12

$$P(w_i) = 1 - \sqrt{\frac{t}{f(w_i)}} \tag{1.12}$$

Donde $f(w_i)$ es la frecuencia de una palabra w_i y t es el umbral elegido, normalmente tiene un valor de 10^{-5}. Mediante a este submuestreo de las palabras frecuentes se obtiene un aumento de la velocidad debido a que descarta palabras sin significado. Sin embargo, puede provocar la eliminación de palabras que son importantes. Por lo tanto, esta técnica dependerá del caso de uso específico y de la naturaleza del corpus de entrenamiento.

1.6. Aplicación de Word2Vec

Las publicaciones científicas crecen exponencialmente. En algunos campos específicos se publican decenas de miles de artículos científicos al año (Larsen & von Ins, 2010). Actualmente existen herramientas de PLN para analizar cualquier literatura científica para acelerar el ritmo del conocimiento científico, permitiendo llegar a una comprensión más completa analizando características del texto (Spangler *et al.*, 2014). A continuación, se expondrá un trabajo sobre la extracción de conocimiento y relaciones que existen en la ciencia de materiales mediante la literatura científica utilizando Word2Vec.

El trabajo publicado por Tshitoyan *et al.*, 2019 presenta un estudio sobre el conocimiento de la ciencia de los materiales presente en la literatura, la cual se codifica eficazmente como incrustaciones de palabras con aprendizaje no supervisado. Estas representaciones de palabras se comportan de forma coherente con la intuición química cuando se combinan utilizando varias operaciones vectoriales (proyección, adición, sustracción) sin la inserción explícita de conocimiento químico. Estas incrustaciones capturan conceptos complejos de la ciencia de los materiales, como la estructura subyacente de la tabla periódica y las relaciones estructura-propiedad de los materiales.

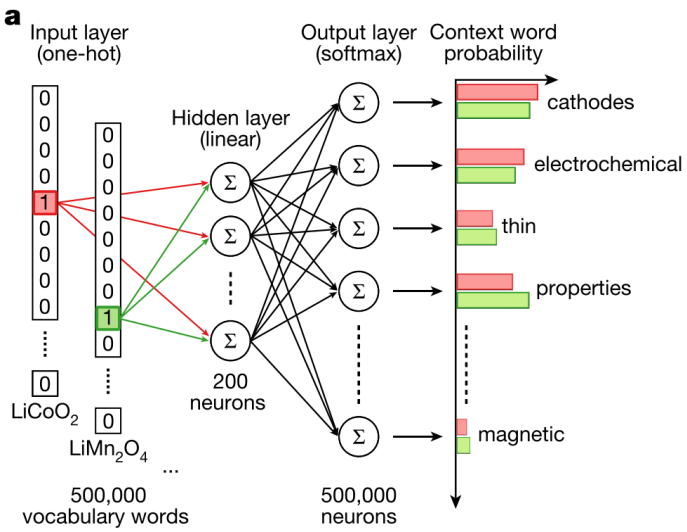

Figura 1.7: Modelo Skip-gram y analogías. Fuente: Tshitoyan *et al.* (2019)

En la figura 1.7 se muestra el modelo skip-gram de Word2vec para las palabras target $LiCoO_2$ y $LiMn_2O_4$. Los vectores one-hot se utilizan como entradas para una red neuronal con una sola capa oculta lineal, en este caso 200 neuronas, la cual está entrenada para predecir todas las palabras mencionadas dentro de una cierta distancia (palabras de contexto) de la palabra target dada. Las palabras $LiCoO_2$ y $LiMn_2O_4$ resultaron tener propiedades similares, lo cual es correcto debido a que son materiales de cátodo de batería similares. Finalmente, la función softmax se usa en la capa de salida para normalizar las probabilidades.

1.6.1. Características del corpus

Para entrenar las incrustaciones, procesaron aproximadamente 3,3 millones de abstracts científicos publicados entre 1922 y 2018 en más de 1.000 revistas que contienen investigaciones relacionadas con ciencia de materiales. Lo que dio como resultado un vocabulario de aproximadamente 500.000 palabras. Además, estos deben tener temas relacionados a la rama de estudio para eliminar el ruido innecesario, ya que no necesariamente tener mas datos es más eficiente, sino que la especificidad del dominio de los corpus determinan la utilidad de las incrustaciones de palabras.

Para no perder una fórmula química o un símbolo de elemento importante, se utilizaron técnicas de tokenización y extracción de datos utilizando pymatgen (Ong *et al.*, 2013) y ChemDataExtractor (Swain & Cole, 2016).

1.7. Caracterización del material bioplástico

Los plásticos se han convertido en un contaminante importante en los sistemas ecológicos debido a la falta de biodegradabilidad, las bajas tasas de reciclaje y las malas prácticas de gestión de residuos. Entre el 80 % y el 85 % de toda la contaminación marina se atribuye al plástico (Jambeck *et al.*, 2015). Los envases de un solo uso, incluidas las bolsas de plástico ligeras y las films de plástico, constituyen la mitad de toda la contaminación plástica marina (Foschi & Bonoli, 2019). Los plásticos de un solo uso a menudo se descomponen en partículas microplásticas; estos son de creciente preocupación como un contaminante mundial en los ecosistemas oceánicos y de agua dulce, donde pueden bioacumularse y transportar toxinas (Zhao *et al.*, 2017).

Los bioplásticos o biopolímeros son materiales con propiedades similares a los plásticos derivados del petróleo, pero que se producen a partir de fuentes de carbono renovables (azúcares, ácidos, lípidos, etc.) y son generalmente biodegradables. Se clasifican de acuerdo a su origen como biopolímeros de origen animal, marino, agrícola, y microbianos (Iles & Martin, 2013). Los materiales biodegradables son las alternativas más atractivas para la sustitución de polímeros de origen fósil o plásticos tradicionales (Bartolo *et al.*, 2021). Además, estos pueden reemplazar el segmento de empaques, donde la vida útil del empaque es corta y la cantidad de desechos posconsumo es grande y genera problemas significativos con su uso (Izdebska-Podsiadły, 2019).

Los bioplásticos son plásticos de base biológica, biodegradables o son de materiales compostables. El creciente interés por salvaguardar el mundo ha llevado a la comunidad científica desarrollar plásticos hechos a base biológica y totalmente biodegradables, como el ácido poliláctico (PLA), el policaprolactona (PCL), poli(adipato-co-tereftalato de butileno) (PBAT), polihidroxiácidos (PHA), así como bio-polietileno (bio-PE), bio-polipropileno (bio-PP) y el tereftalato de bio-polietileno (bio-PET). Aunque la biodegradabilidad se asocia típicamente con materiales de base biológica, no depende del origen del polímero, sino solo de su composición química (Brizga *et al.*, 2020).

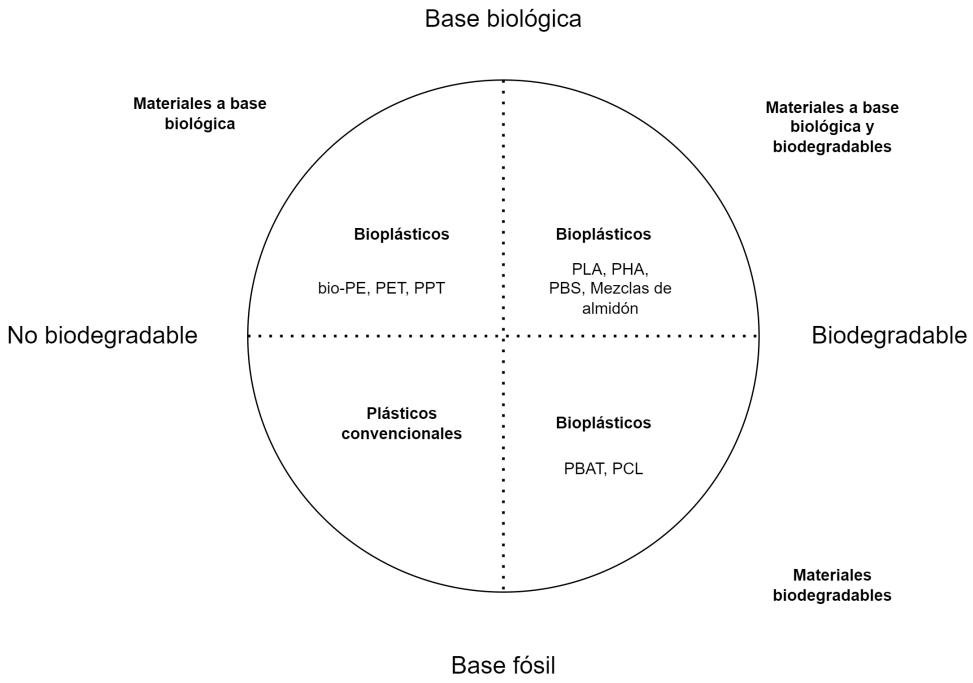

Figura 1.8: Subdivisión de polímeros de base biológica y fósil. Fuente: Adaptado de Visco *et al.* (2022)

1.7.1. Materiales y métodos de producción de bioplásticos

Las algas marinas han sido reconocidas como materia prima sostenible para la producción de bioplásticos dado que no requieren agua dulce, tierra cultivable o fertilizantes para crecer, absorben el exceso de nutrientes del agua de mar y actúan como sumideros de carbono, lo que tiene un efecto mitigador sobre el cambio climático (Radulovich *et al.*, 2015). Los ficocoloides son polisacáridos que se derivan de las algas marinas. Existen tres ficocoloides principales, estos son el alginato, el agar y la carragenina. Los alginatos se extraen principalmente de las algas pardas, en cambio el agar y la carragenina se extraen de las algas rojas. En los últimos años, se han agregado más usos de los ficocoloides como agentes gelificantes en medicina y tecnología (Häder, 2021).

Las diferentes algas utilizadas como materia prima en la producción de agar han dado lugar a productos con comportamientos diferentes, aunque todas pueden incluirse en la definición general de agar. Los tratamientos con algas previos a la extracción son muy importantes ya que condicionarán en gran medida las características del agar obtenido (Ramisen & Galatas, 1987). La química de los procesos utilizados para fabricar

alginato de sodio a partir de algas pardas es relativamente simple. Las dificultades de los procesos surgen de las separaciones físicas que se requieren, como la necesidad de filtrar residuos viscosos de soluciones viscosas o separar precipitados gelatinosos que contienen grandes cantidades de líquido dentro de su estructura y que resisten tanto la filtración como la centrifugación (McHugh, 1987). La funcionalidad de las carrageninas en diversas aplicaciones depende en gran medida de sus propiedades reológicas. Como polímeros lineales solubles en agua, típicamente forman soluciones acuosas altamente viscosas. Para aplicaciones de gelificación, una baja viscosidad en solución caliente es generalmente deseable para facilitar el manejo y, afortunadamente, la carragenina de alta fuerza de gel (sales mixtas de calcio y potasio de kappa o iota) cumplen este requisito debido a su menor hidrofilicidad (Stanley, 1987).

Las films de biopolímeros de agar-glicerina tienen propiedades físicas y mecánicas fundamentales para ser aplicadas a empaques de un solo uso, esto es debido a su resistencia, elasticidad y ductabilidad. Además al cambiar las concentraciones de agar-glicerina, se pueden lograr ciertas propiedades de tracción comparables al almidón termoplástico (TPS), el acrilonitrilo butadieno estireno (ABS) y el polipropileno (PP) (Hernández *et al.*, 2022).

Existen varios materiales en base biológica que han sido utilizados para la creación de finos films para el cubrimiento de alimentos frescos o procesados para prolongar su vida útil. Uno de ellos es el almidón natural, que combinado con las propiedades espesantes y gelificantes de la carragenina dan origen a films bioplásticos con propiedades tanto mecánicas y comestibles (Abdou & Sorour, 2014). Las ventajas de tener un recubrimiento comestible es la biocompatibilidad del material, apariencia estética, propiedades de barrera a los gases, no toxicidad, no contaminación y su bajo costo (Malhotra *et al.*, 2015).

Las propiedades del alginato varían de una especie a otra, por lo que la elección de qué algas cosechar se basa tanto en la disponibilidad de especies particulares como en las propiedades del alginato que contienen. Esto indica la importancia de explorar una amplia gama de concentraciones de ingredientes al evaluar y cómo influyen en una propiedad material de interés. Actualmente, la necesidad de tener productos amigables al medio ambiente ha producido varias investigaciones y exploraciones de nuevos materiales bioplásticos. Generalmente se centran en varios frentes que incluyen mejorar el rendimiento de los polímeros biodegradables naturales mediante tratamientos físicos, químicos y

enzimáticos, sintetizar nuevos polímeros biodegradables, mejorar las características de los polímeros y escalar los procesos, mejorar la producción de polímeros convencionales de base biológica, y buscar nuevas fuentes renovables (Shlush & Davidovich-Pinhas, 2022).

1.8. Propiedades de bioplásticos

Cada tipo de plástico tiene diferentes combinaciones de propiedades funcionales, lo que los hace adecuados para una amplia gama de aplicaciones. Una de estas propiedades funcionales es la propiedad antimicrobiana, esta agrega al bioplástico la capacidad de resistir el crecimiento de microorganismos, como bacterias, hongos y virus en su superficie. Esta propiedad ayuda a reducir la propagación de enfermedades y mejorar la higiene en diferentes entornos. Los plásticos antimicrobianos son especialmente útiles en aplicaciones que requieren una alta higiene, como en la fabricación de envases de alimentos, equipo médico y dispositivos electrónicos (Huang *et al.*, 2019).

Los bioplásticos de un solo uso se utilizan en productos altamente desechables. Para estos productos, es importante que tengan ciertas propiedades como la transparencia, permitir que el contenido del envase sea visible y atractivo para los consumidores. Por lo tanto, es deseable que los bioplásticos tengan un alto índice de refracción para lograr la transparencia necesaria. Otra propiedad funcional que pueden agregar valor a los plásticos es la permeabilidad al vapor. Esta propiedad permite que el vapor de agua o de otros gases pase a través de su estructura. Esta propiedad es importante en aplicaciones como el envasado de alimentos y productos farmacéuticos, donde es necesario controlar la cantidad de vapor que se mueve dentro y fuera del paquete para mantener la calidad y frescura del producto.

Algunas de las propiedades funcionales que permiten que un plástico pueda resistir fuerzas y deformaciones sin romperse o agrietarse, son las propiedades mecánicas. A continuación, se describirán algunas de las propiedades mecánicas más importantes:

- *Tensión de rotura (Tensile Strength)*: Es la máxima tensión que un material puede soportar bajo tensión antes de que su sección transversal se contraiga de manera significativa. Es una propiedad intensiva; por lo tanto su valor no depende del tamaño de la muestra, sino de factores, tales como la preparación, la presencia o no de defectos superficiales, y la temperatura del medioambiente y del material.(Smith &

Hashemi, 2006)

- *Alargamiento a la rotura (Elongation at Break)*: Es la medida de la ductilidad de un material, es decir, indica la capacidad de un material para sufrir una deformación significativa antes de fallar. Una alta ductilidad indica que es más probable que un material se deforme y no se rompa, mientras que una baja ductilidad indica que un material es frágil y se fracturará antes de deformarse mucho bajo una carga de tracción.

- *Young's Modulus (E)*: Se define como la relación entre la tensión aplicada al material a lo largo del eje longitudinal de la muestra ensayada y la deformación medida en ese mismo eje. Para ello es necesario saber cuanta *stress* (σ) definida como fuerza en el area.

Teniendo en cuenta estas propiedades, es posible identificar formulaciones que cumplan con los requisitos de rendimiento específicos. Sin embargo, para encontrar formulaciones que sean relevantes para aplicaciones específicas, es necesario comprender las tensiones mecánicas a las que estará expuesto el material en la aplicación prevista. Por ejemplo, los films de embalaje poliméricas a menudo se seleccionan por su alta ductilidad y su resistencia relativamente alta a la rotura (Sangroniz *et al.*, 2019).

Capítulo 2

Marco Metodológico

2.1. Definición del problema

La escalabilidad de los films de polímeros naturales derivados de algas marinas sigue siendo limitada debido a sus propiedades físicas, mecánicas y de barrera escasas en comparación con los materiales de envasado tradicionales. Por lo tanto, es importante abordar este problema mediante la investigación de diversos aditivos que afecten las propiedades funcionales de los compuestos a base de algas. A medida que se publican más artículos cada año en todos los subdominios de la ciencia e ingeniería de materiales, incluida la ciencia e ingeniería de biopolímeros y bioplásticos, se vuelve cada vez más difícil para cualquier persona dominar toda la información de manera exhaustiva. Por lo tanto, se requiere un esfuerzo continuo de aprendizaje y actualización para mantenerse al tanto de los avances y descubrimientos en este campo, lo que limita el razonamiento inductivo que aprovecha plenamente el conocimiento pasado y la evolución de nuevas hipótesis.

Por otro lado, la gran mayoría del conocimiento científico se publica como texto, lo que dificulta su análisis tanto mediante métodos estadísticos tradicionales como mediante métodos modernos de aprendizaje automático. Por el contrario, las bases de datos estructuradas de propiedades son la principal fuente de datos interpretables por máquinas para la comunidad de investigación de materiales, pero solo abarcan una pequeña fracción del conocimiento presente en la literatura de investigación (Butler *et al.*, 2018). Además de los valores de las propiedades, las publicaciones contienen conocimientos valiosos sobre las conexiones y relaciones entre los elementos de datos, tal como los interpretan los autores.

Para mejorar la identificación y el uso de este conocimiento, se han llevado a cabo varios estudios enfocados en la recuperación de información de la literatura científica mediante el procesamiento del lenguaje natural supervisado, lo cual requiere grandes conjuntos de datos etiquetados a mano para el entrenamiento (Spangler *et al.*, 2014).

2.1.1. Propuesta de la solución

La propuesta de solución consiste en utilizar técnicas de procesamiento de lenguaje natural, como los word embeddings, para extraer atributos relevantes de dos mil abstracts relacionados con polímeros naturales derivados de algas marinas. Con base en los atributos obtenidos, se aplicarán modelos clásicos de minería de texto para preprocesar los datos y utilizar word embeddings para predecir de manera probabilística las palabras relacionadas en el corpus sin supervisión. Para lograr esto, se utilizará t-SNE para reducir la dimensionalidad y preservar la mayor cantidad posible de la estructura significativa de los datos de alta dimensionalidad en el mapa de baja dimensión.

El resultado de los métodos aplicados en esta investigación pretende dar una mejor comprensión de la relación entre ingredientes que den origen a bioplásticos con propiedades y aplicaciones similares a plásticos convencionales. Como se trata de un tema relacionado con ciencia de materiales, se necesita la opinión de expertos relacionados a esta área para discernir si estos son o no relevantes, por lo que el trabajo multidisciplinario será esencial en este trabajo.

2.1.2. Metodología propuesta

La Metodología propuesta para este trabajo de tesis se divide en cuatro pasos, los cuales se describirán a continuación:

- **Análisis de tendencias**: El propósito de este análisis es identificar las tendencias más frecuentes en los abstracts de artículos científicos, incluyendo qué ingredientes y propiedades se mencionan con mayor frecuencia. También se realizará un seguimiento de cómo ha variado la popularidad de diversas aplicaciones de bioplásticos a lo largo del tiempo, señalando las disminuciones o aumentos en su uso.
- **Preprocesar datos**: Para obtener resultados que entreguen mayor información, es necesario procesar y eliminar símbolos que no entregan mayor significados. Además,

es necesario obtener datos como ingredientes y propiedades que al procesarlos no pierdan sus características, tanto de abreviaturas como duplicidad de tokens.

- **Entrenar modelo de Word2Vec**: Obtención de representación vectorial del contexto en que se encuentran de todas las palabras del corpus procesado.
- **Análisis de similitud de word embeddings**: Una vez que se hayan obtenido los word embeddings, se procederá a analizar los vectores de palabras para obtener información sobre cómo se agrupan los ingredientes en función de su aplicación.

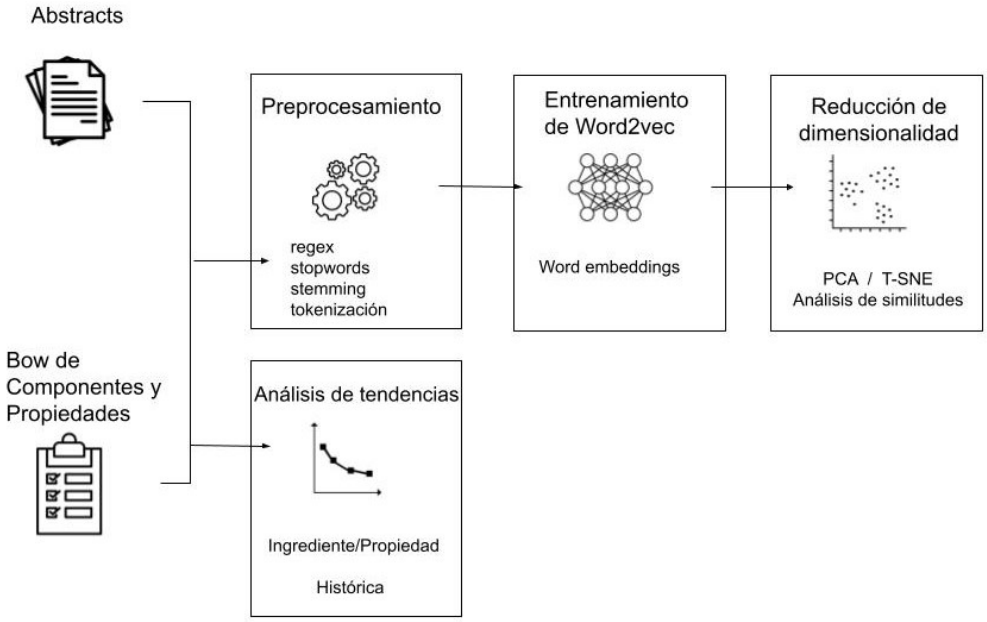

Figura 2.1: Diagrama resumen del marco metodológico. Fuente: Elaboración propia

2.2. Datos utilizados

En esta tesis se emplearon los abstracts de artículos científicos obtenidos del sitio web de la biblioteca digital de la Universidad de Santiago de Chile, suministrados por el Proyecto FONDEF IT-USACH “Bioplástico Territorial”. Se aplicó un filtro de datos para capturar términos relevantes relacionados con bioplásticos a partir de biopolímeros de algas marinas, como “Alginate”, “Agar”, “Carrageenan” o “Packaging”. Los datos resultantes fueron los siguientes:

Tabla 2.1: *Base de datos ocupados en el trabajo de tesis.*

Total de abstracts	2.000
Rango de años	1958 - 2022
Artículos de acceso abierto	404

En la tabla 2.1 se muestra que se obtuvieron un total de dos mil abstracts, con un rango de años comprendido entre 1958 y 2022. Sin embargo, la mayor cantidad de texto recolectado se encuentra en el periodo que va del año 2000 al 2022.

Los abstracts seleccionados cumplen con un requisito especial para el análisis de texto, ya que presentan sistemáticamente la información más relevante de los estudios científicos. Para un mejor análisis de tendencias, se utilizará una bolsa de palabras o "bag of words" también proporcionada por el proyecto FONDEF IT-USACH "Bioplástico Territorial", la cual contiene una clasificación de 255 tipos de materiales/compuestos y 111 propiedades.

2.3. Pre-procesar datos

En esta sección se mostrarán los métodos de minería de texto utilizados para el preprocesamiento de los datos, cuyo fin es estandarizar la entrada al modelo word2vec para poder entrenar un conjunto de datos de manera limpia y bien formada. Esto garantizará que el modelo pueda mejorar la calidad de los word embedding resultantes.

2.3.1. Expresiones regulares

En las primeras tareas a realizar para la metodología de este trabajo de tesis, es la reducción de vocabulario eliminando símbolos que no entreguen significado, es por ello el uso de expresiones regulares en el corpus, ya que estas pueden ayudar a identificar y eliminar patrones no deseados mejorando el rendimiento del modelo al reducir el ruido en el corpus de texto y al proporcionar una representación más limpia y coherente de las palabras individuales.

El preprocesamiento con expresiones regulares implica la creación de patrones de búsqueda que se utilizarán para identificar y eliminar patrones específicos en el corpus de

texto. Por ejemplo, se utilizarán expresiones regulares para eliminar unidades de medición de propiedades y de ingredientes, eliminar números o letras específicas que no aportan información relevante al modelo. Además, no se eliminaron las abreviaciones debido a que en los textos de abstracts, comúnmente se utilizan abreviaturas para referirse a términos específicos.

2.3.2. Stop words

El paso siguiente de esta etapa es la eliminación de palabras, para ello, se aplicará la Ley de Zipf, lo que permitirá saber qué palabras con más frecuencia aparecen en el corpus para poder realizar una lista de stop words. Además, los abstracts originales tendrán una eliminación de copyrights debido a que los datos finales deben tener relevancia sólo con los ingredientes y propiedades de materiales bioplásticos, como también generar tokens sin autores.

2.3.3. Tokenización

La tokenización se utiliza en el procesamiento del lenguaje natural para dividir párrafos y oraciones en unidades más pequeñas a las que se les puede asignar un significado más fácilmente. El primer paso del proceso de PLN es recopilar los datos (una oración) y dividirlos en partes comprensibles (palabras). Este trabajo de tesis captura términos específicos de la ciencia de los materiales, como propiedades (ej: tensile strength) e ingredientes (ej: sodium alginate). Para ello, se captura tokens por medio de la herramienta *tokenize* de la librería *nltk*. Esta herramientas extrae automáticamente información de grandes volúmenes de datos no estructurados como la literatura científica (Swain & Cole, 2016).

Tabla 2.2: *Diferencia entre corpus original con procesamiento de texto.*

Abstract Original	Abstract tokenizado
main property index follow liquid viscosity 434 mPa·s; capsule thickness 0.11-0.12 mm; water content 10.09 % disintegration time 8-15 min simulate intestinal fluid disintegrated 2 h simulate gastric fluid light transmittance capsule film 85.5 % tensile strength 27.97 MPa elongation break 2.0 % conclusion: study lay theoretical foundation replace gelatin capsule vegetable enteric hollow capsule	['main', 'property', 'index', 'follow', 'liquid', 'viscosity', '434', 'mPa·s', 'capsule', 'thickness', '0.11', '-', '0.12', 'mm', 'water', 'content', '10.09', '%', 'disintegration', 'time', '8-15', 'min', 'simulate', 'intestinal', 'fluid', 'disintegrated', '2', 'h', 'simulate', 'gastric', 'fluid', 'light', 'transmittance', 'capsule', 'film', '85.5', '%', 'tensile', 'strength', '27.97', 'MPa', 'elongation', 'break', '2.0', '%', 'conclusion', 'study', 'lay', 'theoretical', 'foundation', 'replace', 'gelatin', 'capsule', 'vegetable', 'enteric', 'hollow', 'capsule']

Esta técnica es usada como input para entrenar un modelo de vectores de palabras para cada token en el corpus.

2.3.4. Stemming

Al utilizar el proceso de stemming en los datos de texto antes de entrenar el modelo word2vec puede ayudar a reducir la dimensionalidad de los datos, ya que las diferentes variaciones de una palabra se reducirán a la misma raíz, que luego el algoritmo word2vec trata como una sola palabra. Esto puede generar incrustaciones de palabras más significativas y un mejor rendimiento en algunos casos.

2.3.5. Bigramas

Después de realizar el preprocesamiento de los datos para reducir el ruido en el corpus de texto, el uso de bigramas puede mejorar significativamente el rendimiento del modelo Word2vec al proporcionar más información contextual para cada palabra. Esta técnica es especialmente útil para capturar las relaciones semánticas entre palabras adyacentes, como se muestra en ejemplos como “food_packaging”, “biodegradable_packaging”

y “polyvinyl_alcohol”.

Una vez que se ha creado la lista de bigramas, se utiliza el conjunto de palabras para entrenar el modelo Word2vec. Esto se logra con el mismo algoritmo de aprendizaje que se utiliza para las palabras individuales, pero esta vez se considerando tanto las palabras como los bigramas como entradas al modelo. Sin embargo, esta técnica puede aumentar el tamaño del corpus de texto, lo que puede requerir más recursos computacionales para entrenar el modelo.

2.4. Word Embeddings

En esta etapa de la metodología se unifican todos los tokens ya preprocesados para ser generados en vectores numéricos extrayendo la forma semántica y sintáctica del texto dependiendo del contexto en el cual se encuentra la palabra. El modelo a utilizar para la generación de vectores es un algoritmo de Machine Learning llamado Word2Vec, este obtiene las representaciones de palabras y logra la extracción de información de todos los abstracts. Cabe señalar que los resultados de embedding dependerán directamente del orden y la cantidad de texto que entrene la red neuronal. Para trabajar con Word2Vec, se utilizará la librería *Gensim* de *Python* (Řehůřek, 2010), en un entorno *Jupyter Notebook*. Los parámetros de entrada son los siguientes:

- **size**: Tamaño del embedding de la palabra, su valor por defecto es de doscientos.
- **windows**: Distancia máxima entre la palabra actual y la predicha.
- **min_count**: Ignora todas las palabras con una frecuencia inferior a esta.
- **sg**: Parámetro booleano para utilizar Skip-gram o CBOW.
- **min_alpha**: Especifica el valor mínimo que puede tomar la tasa de aprendizaje durante el proceso de entrenamiento del modelo.
- **negative**: Número de muestras negativas a utilizar en cada iteración del proceso de entrenamiento.
- **iter**: Número de iteraciones de entrenamiento sobre todo el conjunto de datos.

2.5. Relaciones entre ingredientes y propiedades

Antes de realizar el análisis de word embeddings, se llevará a cabo un análisis de palabras clave que incluye los ingredientes y propiedades relevantes para el cálculo de tendencias. En este análisis se considerarán únicamente las frecuencias con las que aparecen en el corpus los elementos más relevantes, tales como las propiedades mecánicas fundamentales para las aplicaciones de empaque y los ingredientes con mayor ocurrencia. El objetivo es evidenciar los datos sin generar un sesgo en el análisis. Para lograr esto, se utilizarán los siguientes métodos:

- **Análisis de tendencias**: Para analizar las tendencias significativas en los abstracts, se llevará a cabo un cálculo de frecuencia de ciertos ingredientes y propiedades en los corpus de texto. Con base en estos cálculos, se obtendrá un gráfico que muestre las frecuencias de los ingredientes y propiedades más comunes. Este análisis permitirá visualizar las tendencias más relevantes en los abstracts de los artículos científicos y ayudará a identificar los patrones y temas que se presentan con mayor frecuencia.

- **Análisis histórico**: El objetivo de este análisis es evaluar la popularidad de ficocoloides con respecto al estudio de propiedades a lo largo del tiempo, tomando en cuenta los artículos científicos publicados en cada período de cinco años. Se identificarán las tendencias y cambios significativos en su uso y popularidad. Con esta información, se podrá tener una mejor comprensión de la evolución y el desarrollo de ciertos ingredientes en la investigación científica para bioplásticos a base de algas.

2.6. Análisis de analogías de ingredientes

Al capturar cada palabra con vectores de salida, la semántica del texto se puede reproducir utilizando aritmética vectorial. Los word embeddings caracterizan de manera única cada palabra de una manera en que se pueden comparar. Por ejemplo, si dos palabras tienden a coexistir con una colección común de otras palabras, es probable que sean similares; si una palabra tiende a coexistir con palabras de un tema en particular, entonces es probable que se relacione con ese tema. Como tal, al reflejar las estadísticas

de co-ocurrencia, los word embeddings capturan algo del significado de las palabras (Allen & Hospedales, 2019).

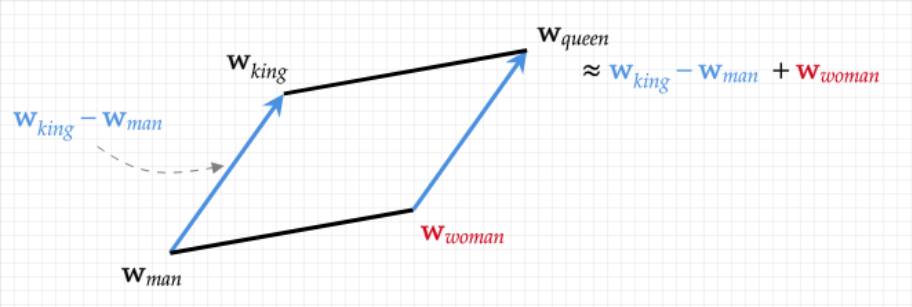

Figura 2.2: king, man y woman forman una relación geométrica específica. Fuente: Allen y Hospedales (2019)

Estos exhiben un comportamiento aparentemente lineal donde las relaciones de palabras de la forma "w_a is to $w_{a\prime}$ as w_b is to $w_{b\prime}$".

Para reproducir patrones entre ingredientes-propiedad e ingrediente-aplicación, se utilizarán analogías de palabras. Se seleccionará un conjunto de ingredientes conocidos en la literatura científica y se realizará la suma vectorial para generar vectores de palabras con cada propiedad y aplicación correspondiente. Por ejemplo, los vectores más similares a la operación vec("ingrediente") + ("propiedad") serán visualizados.

2.7. Representación de puntos en el espacio

Para visualizar las proyecciones de los vectores de palabra de alta dimensionalidad generados por el algoritmo Word2Vec, se utilizarán dos técnicas de reducción de dimensionalidad: PCA y t-SNE. El algoritmo PCA es un enfoque no supervisado que permite reducir la dimensionalidad de los datos y visualizar las similitudes entre los embeddings de palabras de una manera más clara. Esto puede proporcionar información más precisa sobre cómo se agrupan los ingredientes en función de sus aplicaciones y propiedades.

Otro algoritmo para la visualización de los embeddings de palabras es el t-SNE (t-Distributed Stochastic Neighbor Embedding). Este algoritmo también reduce la dimensionalidad de los datos, pero se enfoca en preservar las relaciones de distancia relativa entre los puntos en el espacio original. Esto permite una representación más clara de las similitudes debido a la agrupación de palabras similares, ayudando a identificar patrones

en la distribución de los embeddings (Van der Maaten & Hinton, 2008). Matemáticamente, t-SNE construye una distribución de probabilidad de pares de puntos en alta dimensión y una distribución de probabilidad de pares de puntos en baja dimensión. La distribución de probabilidad de pares de puntos en alta dimensión se modela mediante una distribución gaussiana que tiene como centro un punto de datos y una desviación estándar que se ajusta según la distancia de ese punto de datos a los demás puntos (Maaten & Hinton, 2008).

Al visualizar los vectores más cercanos tanto a propiedades como aplicaciones, se buscará abarcar diferentes aspectos importantes relacionados con la creación de bioplásticos. Se considerarán los materiales y sustancias utilizadas, así como las características funcionales deseadas.

2.8. Limitaciones

En este trabajo de tesis se capturarán las relaciones entre las palabras contexto utilizando métodos de PLN, por lo tanto, no se espera entregar los valores en cada propiedad mecánica por cada ingrediente. Además, hay palabras que presentan duplicidad de token debido a nombres comunes en la literatura, por ejemplo, “agar”, “agar-agar”, “agar-based” y “agarose”. Los resultados serán evaluados de forma cualitativa. Una advertencia a tener en cuenta es que las predicciones apuntan a ingredientes con probabilidad de ocurrencia en el contexto de una propiedad y se pueden usar en particular de varias maneras diferentes.

Capítulo 3

Resultados

En este capítulo se presentarán los resultados obtenidos al aplicar la metodología propuesta. Se mostrarán las tendencias identificadas a partir de la información textual del corpus de datos, incluyendo los ingredientes y propiedades más comúnmente mencionados en la literatura sobre plásticos a base de algas. Luego se visualizarán y analizarán las similitudes de palabras al entrenar los datos preprocesados con el algoritmo Word2vec.

3.1. Preprocesamiento de corpus

En la primera fase de la metodología 2.3, luego de eliminar stop words y procesar expresiones regulares para eliminar las unidades de medida asociadas a las propiedades, aún quedaron algunos símbolos irrelevantes en el corpus, aunque en menor cantidad. Además, el uso de stemming generó la aparición de varias palabras no identificadas, como “simul” y “vesicl”. Dado este inconveniente, se tomó la decisión de no utilizar stemming en el análisis del corpus. Finalmente, una vez preprocesado el corpus completo, se utilizaron bigramas para generar los textos de entrada para el modelo Word2vec.

3.2. Análisis de tendencias

En esta sección se mostrarán las tendencias de ingredientes y propiedades relevantes para la creación de plásticos basados en algas. Los resultados de la primera etapa se representan mediante un gráfico que mostrará los diez ingredientes más frecuentes en la literatura científica.

3.2.1. Análisis de tendencia de ingrediente y propiedades

Al utilizar bag of words de ingredientes/compuestos, se identificaron que los datos del corpus existe una tendencia significativa para ingredientes. En el gráfico 3.1, es posible observar que el Alginato es el componentes más popular seguido por el Alginato de sodio, Agar y la Carragenina. También se pueden apreciar polisacáridos como el Chitosan o quitina y la Celulosa. Estos ingredientes se caracterizan por una amplia gama de usos, donde el más relevante es el uso de films de quitosano biodegradables, los cuales tienen potencial para conservar varios productos alimenticios, también conservar su firmeza y restringir la pérdida de peso debido a la deshidratación (Al-Tayyar *et al.*, 2020).

Otro aspecto importante a considerar en la composición de los bioplásticos de algas son los plastificantes naturales que se utilizan para aumentar su flexibilidad y suavidad. Glycerol es el tipo de plastificante más común y el Starch o almidón el cual es el polímero más utilizado en el mercado de los bioplásticos (Abe *et al.*, 2021).

También se encuentra la presencia de la Gelatina. Esta es utilizada como agente gelificante, lo que ayuda a mejorar la resistencia y elasticidad del material en un bioplástico (Hanani *et al.*, 2014).

Otro ingrediente presente en la literatura es el Pectin o pectina, este también puede utilizarse como un agente espesante y gelificante, mejorando la textura y la elasticidad del material.

Finalmente, Clay o arcilla, este se utiliza a menudo como un agente de refuerzo o como un material de carga para mejorar las propiedades mecánicas del material. La adición de arcilla en bioplásticos de algas puede mejorar su resistencia a la tracción, la dureza y la estabilidad dimensional. También puede aumentar la resistencia a la penetración de gases y líquidos, lo que mejora la barrera del material y su capacidad de proteger los productos empaquetados (Boey *et al.*, 2022).

Como se muestra en el gráfico 3.2 se puede apreciar diversas propiedades. Tensile Strength es la propiedad que ocurre con más frecuencia en los datos recolectados. Esto indica que la propiedad más común para los bioplásticos a base de algas en la literatura científica es la evaluación de la calidad de ruptura para aplicaciones como de envasado, esto debido a todas las tensiones que puede tener el material durante su manipulación. Luego se visualiza que las propiedades funcionales se distribuyen en cantidades similares. Con esto, se puede observar otras que son igualmente importantes para el envasado

de productos. Entre ellas, se destacan la permeabilidad al vapor de agua y la actividad antimicrobial. Estas son propiedades críticas debido a que miden la capacidad de que el interior del envase se encuentre seco y libres del crecimiento de microorganismos.

Después de analizar ambos gráficos, se puede destacar que Alginate y Tensile strength aparecen con una frecuencia casi el doble que los demás componentes y propiedades, lo que sugiere que los textos proporcionan información sobre la evaluación de la calidad y resistencia de los materiales. Además, se observa que varios ingredientes se describen como aditivos para mejorar las propiedades de los bioplásticos. Este resultado puede indicar que los datos están principalmente relacionados con bioplásticos fabricados a partir de algas.

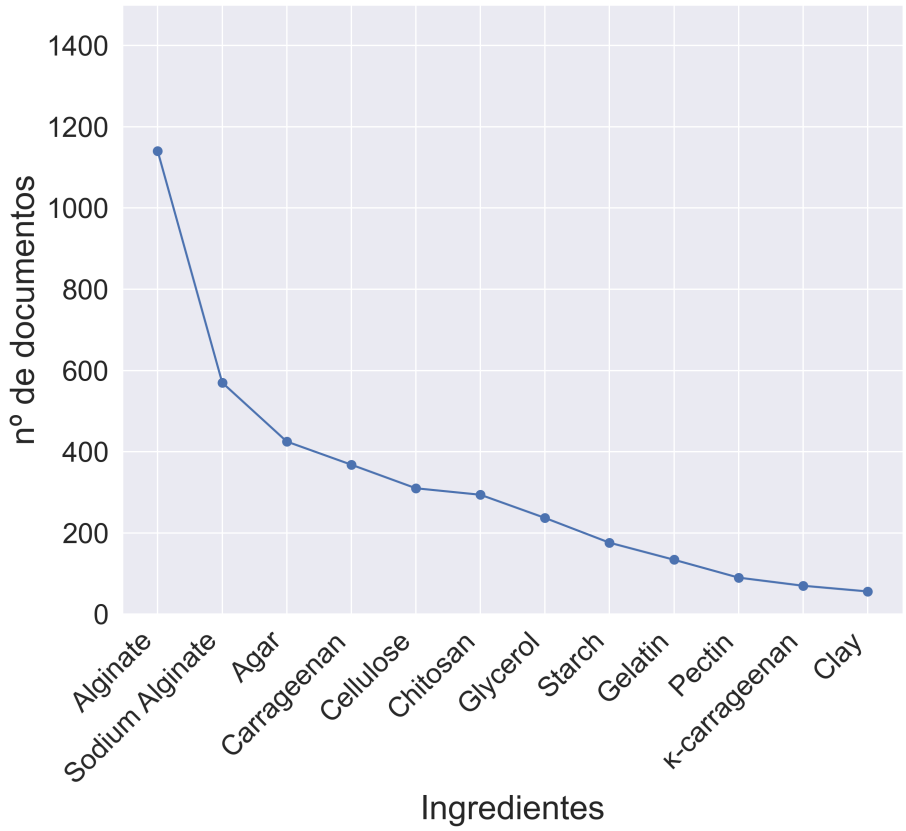

Figura 3.1: Gráfico de línea que representa los ingredientes con más frecuencia en el corpus

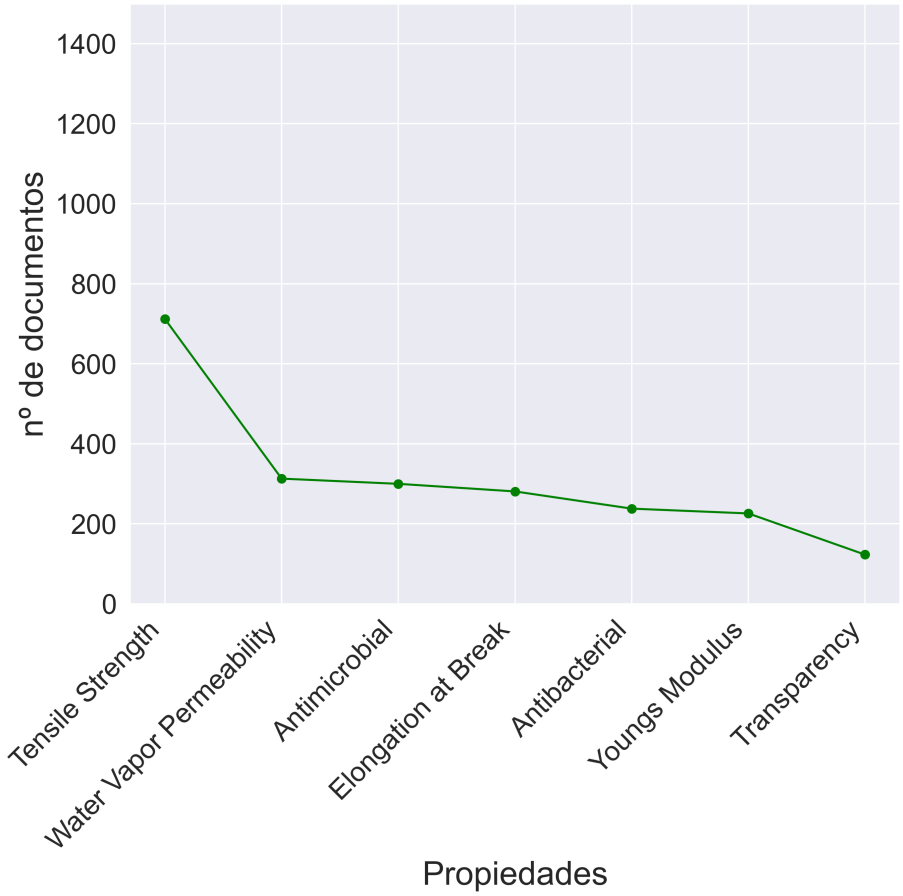

Figura 3.2: Gráfico de línea que representa las propiedades más frecuentes en el corpus.

3.2.2. Análisis histórico de ingredientes

El Agar, el Alginato y la Carragenina son tres ingredientes que se han utilizado históricamente en la creación de bioplásticos. Estos tres ingredientes naturales han sido utilizados debido a sus propiedades únicas y su disponibilidad. Al estudiar la tendencia de estas palabras a través de los años en la literatura científica, en el gráfico 3.3 es posible observar cómo ha evolucionado el interés en estos compuestos. Desde el año 1958 al 2022 se logran identificar diversas tendencias, los cuales se muestran en rangos de cinco años.

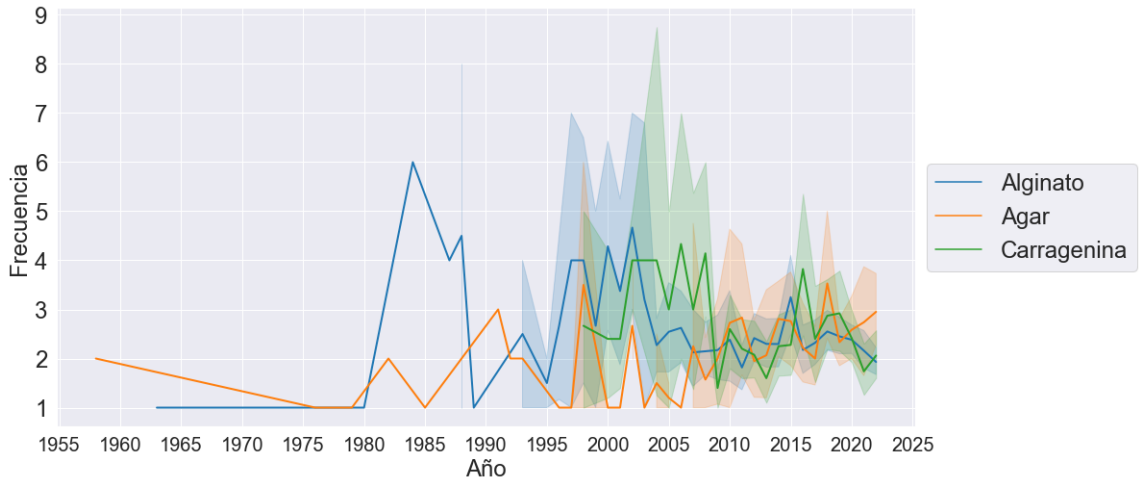

Figura 3.3: Gráfico de ingredientes durante 1958 a 2022

El agar fue el primer ingrediente en aparecer en la literatura analizada, y los primeros abstracts sugieren que fue utilizado como un agente de recubrimiento en análisis de bacterias (Whitt, 1958). Además, la literatura científica muestra que el alginato empezó a ser estudiado utilizando el método de casting en los años 80, lo que permitió mostrar opciones de recubrimiento de alimentos y tratamientos gelificantes. Esto pudo haber contribuido a su popularidad en la época como uno de los biopolímeros más prometedores y versátiles.

A partir de principios de los años 90, se comenzó a notar un estimador de frecuencia, debido a que en algunos casos, los ingredientes fueron mencionados levemente en la literatura, mientras que en otros casos su presencia fue reportada con mayor frecuencia. Posteriormente, a finales de los años 90, se empezó a observar la presencia de la carragenina, lo cual marcó el comienzo de una tendencia similar para el alginato desde

entonces. Por otro lado, el agar perdió popularidad con la aparición de la carragenina, lo que podría deberse a que ésta última es más fácil de extraer y tiene mayor capacidad para formar geles en comparación con el agar (Usov, 2011).

Finalmente, cabe destacar que durante los años 2010 y 2022, la tendencia de los ingredientes se ha mantenido en rangos similares de frecuencia.

3.3. Análisis de similitudes entre ingredientes

Una vez obtenidos todos los word embeddings mediante Word2vec utilizando Skip-gram, los datos mostraron diversos ingredientes en función de la similitud coseno entre polímeros a base de algas. A continuación, se muestran palabras más similares relacionados entre alginatos, agar y carrageninas. Cada palabra en la tabla está acompañada de un valor de similitud que indica qué tan cercana es la palabra.

Tabla 3.1: *Ranking de similitudes para alginate y sodium alginate*

alginate	similitud	sodium alginate	similitud
calcium_alginate	0.5112	polyvinyl_alcohol	0.5708
hydrophilic	0.5060	calcium_chloride	0.5615
gelatin	0.4995	methyl_cellulose	0.5457
aerogels	0.4959	carboxymethyl_cellulose	0.5415
alginate_dialdehyde	0.4897	k-carrageenan	0.5308
pectin	0.4823	konjac_glucomannan	0.5300

En la tabla 3.1 se presentan resultados para dos palabras de consulta relacionadas con alginato y alginato de sodio. Se empleó el alginato de sodio debido a su amplia presencia en la literatura como uno de los ingredientes más populares. Este alginato ha sido procesado con hidróxido de sodio. Los resultados revelan palabras similares asociadas a las características de fabricación de hidrogeles de alginatos, que incluyen propiedades gelificantes y espesantes, como konjac glucomannan, así como ingredientes utilizados para mejorar las propiedades del hidrogel, como el polyvinyl alcohol. Además, es posible observar materiales biodegradables como methyl cellulose y carboxymethyl cellulose, los cuales presentan propiedades físicas similares al alginato de sodio y son

utilizados en la preparación de bioplásticos (Strnad *et al.*, 2019) (Lee *et al.*, 2022) (R. Zhang *et al.*, 2022).

Tabla 3.2: *Ranking de similitudes para agar*

agar	similitud
k-carrageenan	0.4957
lignin	0.4775
Naalg	0.4513
agar_maltodextrin	0.4420
soy_protein	0.4323
zein	0.4302

En la tabla 3.2, se presentan similitudes de menor valor en comparación con los alginatos y carrageninas. Se destacan diversos materiales naturales derivados principalmente de plantas que pueden ser utilizados en la fabricación de bioplásticos, como la proteína de soja, la cual al mezclarse con agar, mejora las propiedades mecánicas del film resultante (Tian *et al.*, 2011). Además, se observa la presencia de polisacáridos, polímeros y proteínas entre estos materiales.

Tabla 3.3: *Ranking de similitudes para carrageenan y k-carrageenan*

carrageenan	similitud	k-carrageenan	similitud
rice_starch	0.6611	gelatine	0.7654
lignin	0.6497	pullulan	0.7636
polyethylene_glycol	0.6473	xanthan	0.7567
pullulan	0.6299	curdlan	0.7543
gum	0.6270	gellan_gum	0.7487
sorbitol	0.6202	bean_gum	0.7482

Finalmente en la tabla 3.3, las palabras más similares de carrageenan tienen diferentes estructuras y propiedades. La palabra rice starch tiene una relación directa con la carragenina, ya que ambos son materiales espesantes. También se muestra la

presencia de lignina, la cual se utiliza como agente adhesivo y relleno debido a su alta viscosidad y resistencia. Lo sigue el polyethylene glycol, generalmente usado como agente humectante y lubricante debido a su capacidad para retener la humedad y reducir la fricción. Por otro lado, se encuentran ingredientes como pullulan, gum y sorbitol, comúnmente utilizados en la industria alimentaria como estabilizantes y emulsionantes. Para la palabra k-carrageenan, al ser un material con mayor capacidad de formar geles, la similitudes de palabras son hidrocoloides y polisacáridos producidos por bacterias y hongos. Todos ellos utilizados en la industria alimentaria para diferentes funciones.

3.4. Visualización de relaciones entre ingrediente-propiedad e ingrediente-aplicación mediante t-SNE

La visualización de similitudes de palabras permite agrupar relaciones funcionales entre ellas en función de los patrones en los que se emplean las palabras en el texto original. Esta herramienta es útil para obtener una comprensión más profunda de los temas y conceptos que se abordan mediante un determinado corpus.

Para llevar a cabo este análisis, se entrenará el modelo Word2vec con el corpus preprocesado y se utilizará el algoritmo t-SNE, el cual proyectará la predicción con mayor probabilidad de vectores de palabras de cada grupo en un espacio de dos dimensiones, lo que permitirá analizar y visualizar las similitudes entre ellos. Los términos serán representados gráficamente en grupos de diferentes colores, lo que facilitará la identificación y comprensión de relaciones entre ellos. Cabe destacar que se utilizará la totalidad de los abstract para poder capturar los patrones sintácticos y semánticos. Se mostrarán las cien palabras que tengan mayor probabilidad de aparecer tanto para propiedades funcionales de bioplásticos como para aplicaciones de uso.

Para las propiedades funcionales de bioplásticos, se han seleccionado las siguientes palabras clave: “Antimicrobial”, “Permeability”, “Water vapor”, “Termoplastic”, “Antibacterial”, “Transparent” y “Tensile strength”.

Por otro lado, para las aplicaciones de uso, se han seleccionado las siguientes palabras clave: “Food packaging”, “Biodegradable packaging”, “Pharmaceutical industry” y “Edible coating”.

Se analizarán acercamientos en distintos puntos del espacio, identificados como A,

B, C y D. Se explorarán los puntos solapados como puntos de conjunto de datos agrupados para identificar patrones y extraer información valiosa de esta visualización.

3.4.1. Identificación de las relaciones entre ingrediente-propiedad

Para garantizar que el análisis de similitudes se recopile información sobre los ingredientes que han sido evaluados con ciertas características de propiedad funcional, utilizándose la técnica de aritmética vectorial en Word2vec. En este caso, se seleccionarán palabras claves basadas en analogías de conjuntos de datos conocidos en los abstracts más recientes (Roy *et al.*, 2022) (Yang *et al.*, 2022) (Hernández *et al.*, 2022) (Bora *et al.*, 2022). Para esto son escogidos términos que aparecen cercanos a los abstracts, estos son:

```
vectors = [["antimicrobial","gelatin/agar"],
["water_vapor","sodium_alginate"],
["thermoplastic","agar"],
["transparent","boehmite_alumina"],
["uv","alvarezii"],
["tensile_strength","agar/chitosan/halloysite"]]
```

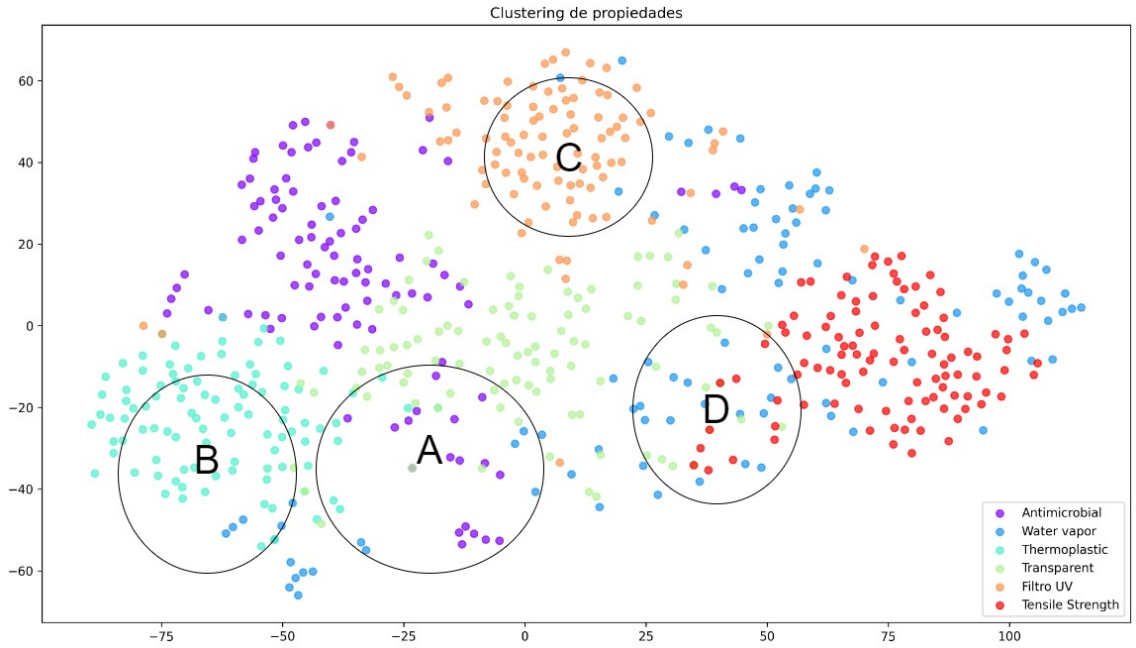

Figura 3.4: Top 100 predicciones de palabras de contexto a propiedades funcionales de bioplásticos

En la figura 3.4, se aprecia que los grupos de 6 propiedades de bioplásticos están bien definidos, lo que indica una clara relación entre los ingredientes y las propiedades funcionales de los bioplásticos.

La elección de los parámetros en el modelo t-SNE se basó en varios factores importantes. En primer lugar, se estableció una perplejidad de 15, lo que se considera un valor adecuado para equilibrar la agrupación y la dispersión de los puntos en el espacio de dos dimensiones. Además, se estableció que la visualización se realizaría en dos dimensiones, por lo que la cantidad de componentes se fijó en 2.

Otro factor importante en la elección de los parámetros fue el método de inicialización, que se configuró como PCA para reducir la complejidad del modelo y aumentar su eficiencia. También puede ayudar a eliminar parte del ruido y la información redundante en los datos de alta dimensión.

En la imagen de la figura 3.5, se pueden observar varios grupos de ingredientes en el clúster Antimicrobial, así como otros que están más cercanos a las propiedades Termoplásticas y Transparentes. Al analizar algunos ingredientes como los activos antimicrobianos, se encontraron ingredientes como *SRIC* (un film basado en carragenina de iota

semirrefinada), *triclosán*, *NLC* (lípidos nanoestructurados), *alginate roselle* y *cinnamon eo* (aceite de canela). Hay una excepción para aceite esencial de Dracocephalum moldavica (DEO), ya que en los abstracts no mencionan específicamente la propiedad antimicrobiana, sin embargo, estudios previos (Aćimović *et al.*, 2022) han demostrado su eficacia contra varias cepas de bacterias y su actividad antioxidante y antimicrobiana. Por lo que el modelo puede predecir su comportamiento a esta propiedad.

Aparecen también pequeños grupos de ingredientes que aparecen más frecuentemente con la propiedad antimicrobial, como el *Ti2* (dióxido de titanio), este tiene la capacidad antifúngica y se han realizado pruebas para evaluar sus propiedades antibacterianas. A pesar de que la palabra *alginate* no parece estar relacionada con propiedades antimicrobianas y transparentes, estudios han demostrado que este material adquiere dichas propiedades cuando se combina con otros aditivos para formar nanocompuestos (Mohamadinia *et al.*, 2021). Al analizar la presencia de *zeína prolamina* en dos clústers (transparent, thermoplastic), esta proteína está relacionada con la caracterización de propiedades transparentes en el conjunto de datos (Sanchez-Garcia *et al.*, 2010). Sin embargo, el algoritmo también la relaciona con los termoplásticos, lo cual puede ser debido a que, al incorporarse con otros biopolímeros, actúa como un plastificante de refuerzo, lo que justifica su clasificación. No obstante, su asociación como termoplástico no indica necesariamente que posea buenas propiedades termoplásticas, sino que simplemente está relacionada con ellas.

En la figura 3.6, se observan términos relacionados con la propiedad Thermoplastic, tales como *hydrocolloid matrices*, *gelling agents* y *reinforced compounds*. Al analizar los ingredientes, se encontró que los compuestos basados en agar y GA-k-carragenina (una forma modificada de la carragenina) están más cercanos entre sí. Además, la literatura (Adam *et al.*, 2020) indica que GA-k-carragenina se ha utilizado como agente gelificante en la producción de cápsulas duras de HPMC, lo que demuestra una asociación con materiales termoplásticos. Sin embargo, /agar-based puede estar asociado con starch/agar-based, cellulose/agar-based y gelatin/agar-based. Los compuestos basados en starch/agar-based, cellulose/agar-based son polisacáridos naturales que se utilizan a menudo como materiales de carga o espesantes en productos farmacéuticos, cosméticos y alimentarios (Roy & Rhim, 2022). Por otro lado, gelatin/agar-based es un compuesto que contiene gelatina, un polímero termoplástico comúnmente utilizado en la fabricación

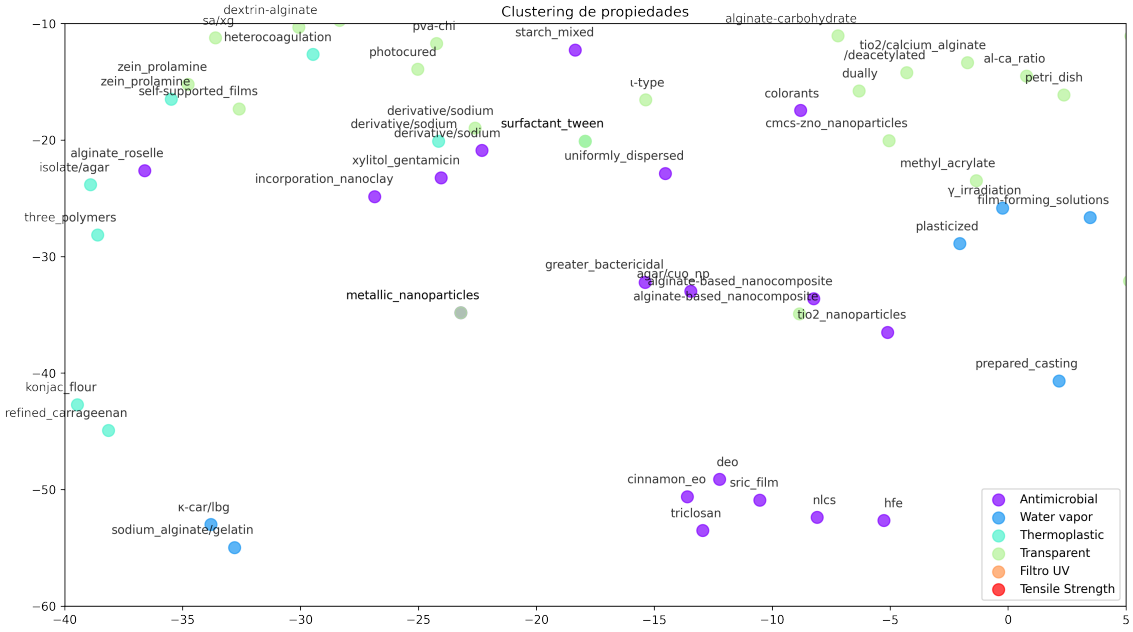

Figura 3.5: Región correspondiente a elementos A

de cápsulas blandas (Kim *et al.*, 2022). Es importante destacar que estos compuestos no son termoplásticos, sino que son termoestables, lo cual es probable que el modelo asocia la estabilidad térmica con la propiedad clasificada. Además se encuentran palabras como *multifunctional composite*, *reinforced*, *highly suitable*, estas palabras describen un material compuesto altamente especializado y adecuado para una determinada aplicación o propósito debido al fortalecimiento de materiales.

En el corpus de abstracts, se ha encontrado que la utilización de otros ingredientes como la *k-carragenina* y *Prunus maackii* presentes en la figura, pueden mejorar la funcionalidad de los films termoplásticos, haciéndolos más flexibles y resistentes (Sun *et al.*, 2019). Por lo tanto, es correcto asociar estos ingredientes como parte de los componentes que contribuyen a la propiedad termoplástica de los biocompuestos duros.

Al analizar la región de Filtro UV 3.7, se encuentran palabras claves relacionadas con la propiedad, como *absorptivity*, *stadistical_significance*, *uv/visible*, *trabsmittance* y *orange-brown*. En la figura observada, se pueden ver varios ingredientes que han demostrado poseer la propiedad de barrera contra la luz ultravioleta, como los films de k-carragenina, que mejoraron significativamente su capacidad al incorporar PFE o PPE, tal como se indica en (Xiao *et al.*, 2020). Los términos PFE o PPE y carragenina se encuentran cercanos en el agrupamiento, junto con palabras como *relating*, *benchmark* y

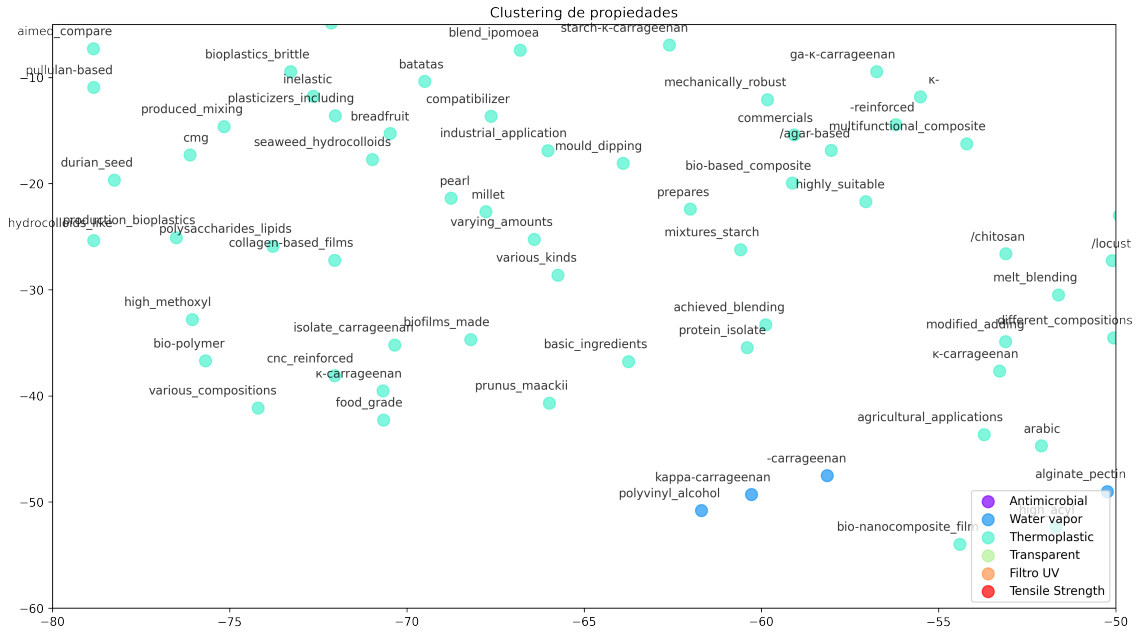

Figura 3.6: Región correspondiente a elementos B

outperforming.

Sin embargo, también aparecen otros ingredientes como el *seaweed-neem* (hojas de neem en la matriz del biopolímero de algas), que si bien no menciona directamente la propiedad de filtro UV, se asocia en esta categoría debido a la presencia de experimentos con radiación, según se indica en Uthaya Kumar *et al.* (2020). Lo mismo ocurre con los términos *CMC/chitosan*, *alginato/glicerol* y *CMC/sodio*, que aparecen cercanos entre sí y han sido caracterizados por su transmitancia de luz, transparencia (transmisión de luz UV y visible) y absorción UV.

Aunque se han obtenido buenos resultados, también se encontraron valores relacionados con el vector vec(“alvarezii”), como las nanopartículas de lignina (lnps), que no están directamente relacionados con la propiedad de filtro UV, como se señala en Rizal *et al.* (2021).

Finalmente, en la figura 3.8 se destaca la aparición de ingredientes como *pectin/agar*, *sa/gg* (sodium alginate/ gum ghatti), *sal/pal_nanocomposite*, *cas* (casein) y *chitin_nanofibers*. En todos estos compuestos e ingredientes, algunos adheridos a otras nanopartículas, no se menciona directamente “Tensil Strength”, sin embargo, se hace referencia a buenas propiedades físicas, refuerzo del film o propiedades mecánicas satisfactorias (Akman *et al.*, 2021) (Anter *et al.*, 2018) (Saito *et al.*, 2007) (Roy & Rhim,

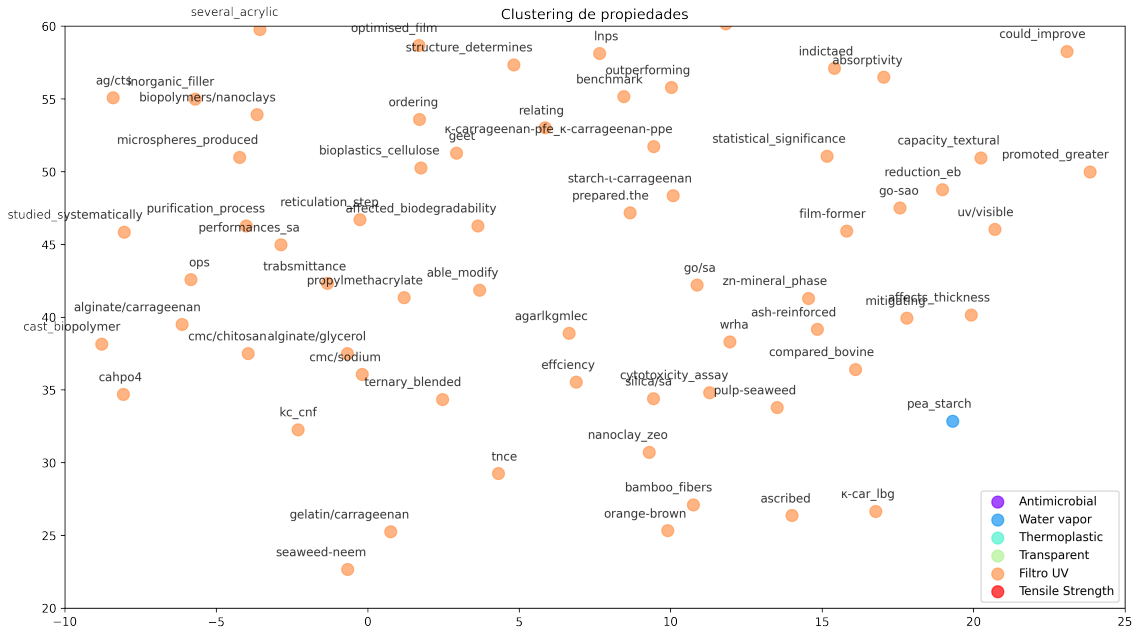

Figura 3.7: Región correspondiente a elementos C

2021).

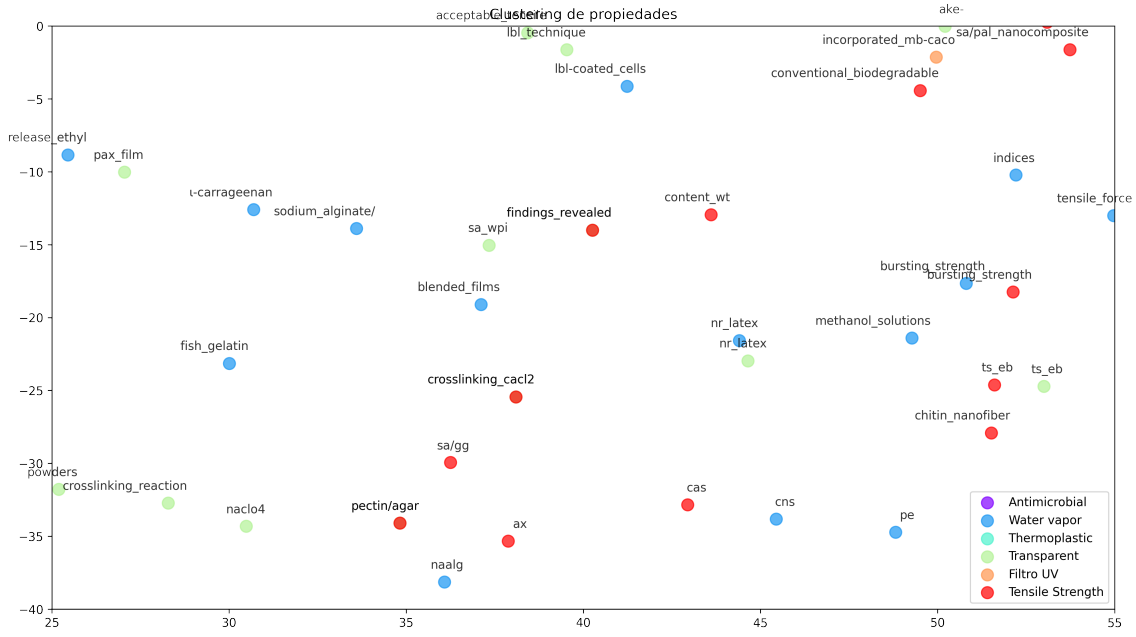

Figura 3.8: Región correspondiente a elementos D

Analizando los resultados obtenidos mediante el análisis de la similitud con suma vectorial entre propiedades e ingredientes, se identifican patrones de clustering entre

ingredientes y técnicas, lo que sugiere que existen ciertas tendencias en cuanto a las propiedades que pueden tener los ingredientes. Además, se puede observar que algunos ingredientes están presentes en varios clústeres, lo que indica que podrían tener múltiples propiedades funcionales. No obstante, es importante destacar que aún hay ruido en los datos, lo que significa que el modelo no es capaz de distinguir palabras ambiguas.

De los resultados, es posible resaltar que existen palabras que aparecen en distintas clasificaciones sin la presencia de la palabra “propiedad” en el corpus. Esto podría indicar que el modelo Word2vec está aprendiendo representaciones vectoriales dependiendo del contexto, entregando ingredientes con mayor probabilidad de ocurrencia debido a la literatura científica, mostrando ingredientes con tendencia a tener ciertas propiedades. Esta capacidad para identificar y agrupar palabras relacionadas, incluso sin una instrucción explícita, ofrece una valiosa herramienta para el análisis de grandes cantidades de datos y la identificación de patrones.

3.4.2. Identificación de las relaciones entre ingrediente-aplicación

Al igual que las similitudes entre ingrediente y propiedad, nuevamente se empleará la técnica de aritmética vectorial. En este caso, se han elegido términos basados en asociaciones entre aplicaciones mostrados en la literatura (Amariei *et al.*, 2022) (Giordano *et al.*, 2022) (Bibi *et al.*, 2022) (Mesgari *et al.*, 2022), estos son:

```
vectors = [["food_packaging","agar/alginate/glycerol"],
          ["biodegradable_packaging","agar"],
          ["pharmaceutical_industry","sodium_alginate"],
          ["edible_coating","carrageenan"]]
```

En la figura 3.9 se muestra una visualización mediante t-SNE de predicciones de palabras para aplicaciones. En ella es posible observar que hay presencia de agrupamientos, sin embargo también hay elementos que se encuentras en diferentes regiones.

Se observaron solapamientos en los datos de similitud por Word2vec, lo que podría sugerir que existen algunas aplicaciones bioplásticas que comparten características similares.

Al revisar los datos de la figura 3.10, se pueden observar que hay ingredientes repetidos. En particular, al analizar el ingrediente *cottonii* (films de k-carragenina extraí-

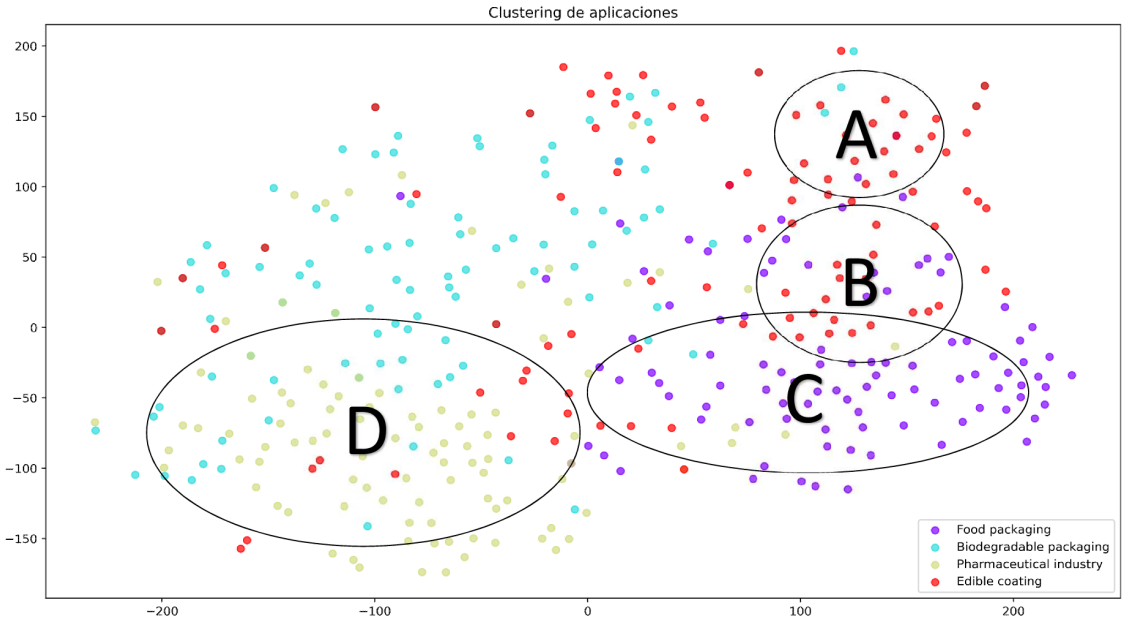

Figura 3.9: Top 100 predicciones de palabras de contexto de aplicaciones

das de Eucheuma cottonii) en la literatura científica, se puede comprobar que existen relaciones entre este ingrediente y las aplicaciones de empaque y comestible. Además, estudios recientes sugieren que la mezcla de este ingrediente podría ser una alternativa de empaque activo degradable que reduce la necesidad de agregar antioxidantes directamente a los productos alimenticios, lo que hace relevante su cercanía a la palabra *make antimicrobial* (Fransiska *et al.*, 2020) (Hamid *et al.*, 2018).

También, es posible observar la presencia de *myrtle berries* en dos clusteres. Al revisar la literatura, se encontró que este ingrediente ha sido estudiado como sustituto de embalaje sintético y como una nueva estrategia para mejorar la seguridad microbiana y la vida útil de los alimentos (Cheikh *et al.*, 2020).

La *tapioca* se ha utilizado tanto como film bioindicador comestible como fuente de almidón para mejorar las propiedades de uso de empaques debido a sus propiedades plastificantes y gelificantes (Wardana & Widyaningsih, 2017) (de Lima Barizão *et al.*, 2020). Por lo tanto, los tres ingredientes mencionados muestran una similitud de patrones en el texto con diferentes clasificación de aplicaciones.

En la figura 3.11, se pueden observar las similitudes en el espacio vectorial de palabras relacionadas “Edible coating”. Entre los ingredientes utilizados en este tipo de

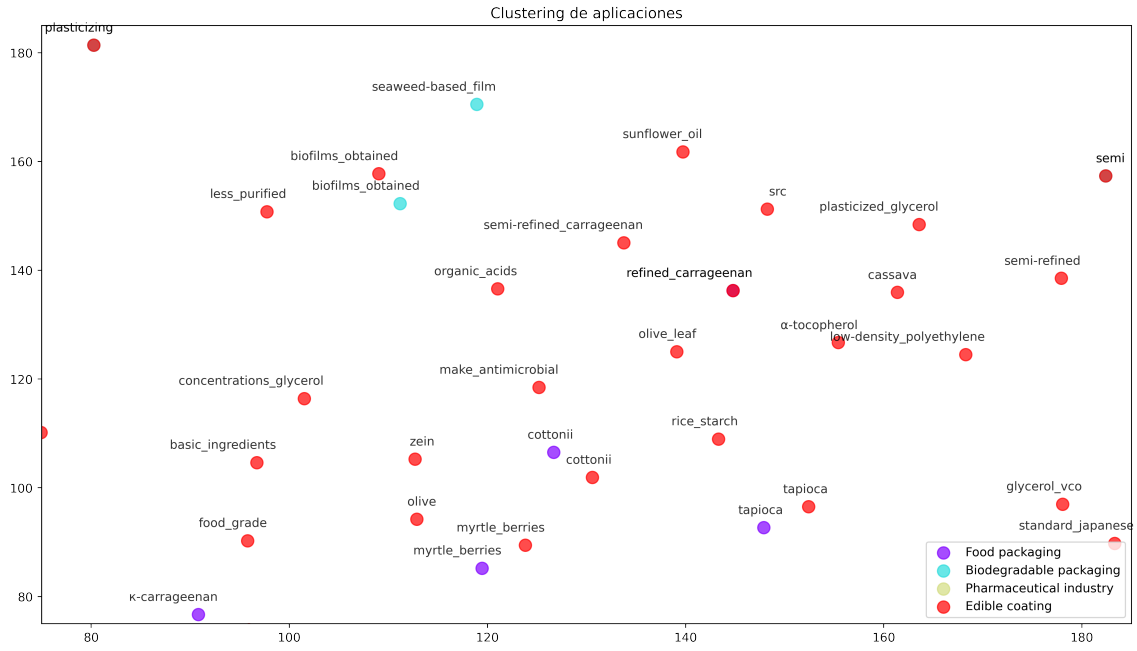

Figura 3.10: Región correspondiente a elementos solapados A

recubrimiento destacan la *spirulina* y la *tilapia*, así como las frutas *strawberries*, *papaya cubes* y los *groundnut kernels*. Estos ingredientes han sido utilizados en diversos estudios para evaluar la calidad y la vida útil de los alimentos tratados con este tipo de recubrimiento, utilizando polímeros como *alginate/carrageenan*, *chitosan*, *alginate/gum arabic*, *agar* y *starch carragenina* (Tabassum & Khan, 2020) (Wang *et al.*, 2022).

Por otro lado, *garlic* es un ingrediente natural que cuenta con propiedades antibacterianas, lo que lo hace adecuado como agente antimicrobiano en diversas aplicaciones. Varios estudios han demostrado su eficacia como agente antipatógeno, lo que resalta la importancia de considerar su cercanía en el espacio vectorial con la palabra *anti-pathogenic* (Campa-Siqueiros *et al.*, 2020) (Halimah *et al.*, 2022). En general, la cercanía en el espacio vectorial de palabras relacionadas con ingredientes comestibles y polímeros utilizados para el envasado de alimentos, sugiere que existe una relación entre ellos.

Además, se ha encontrado la presencia de ingredientes abreviados como *Carr/CuS*, que corresponden a la carragenina natural (Carr) y el sulfuro de cobre (CuS NP). Al revisar la literatura (Li *et al.*, 2020) muestra que los films de Carr/CuS presentan una alta transparencia, propiedades mecánicas mejoradas, actividad microbiana y una mayor estabilidad térmica en comparación con un film de solo Carr. Aunque la combinación de

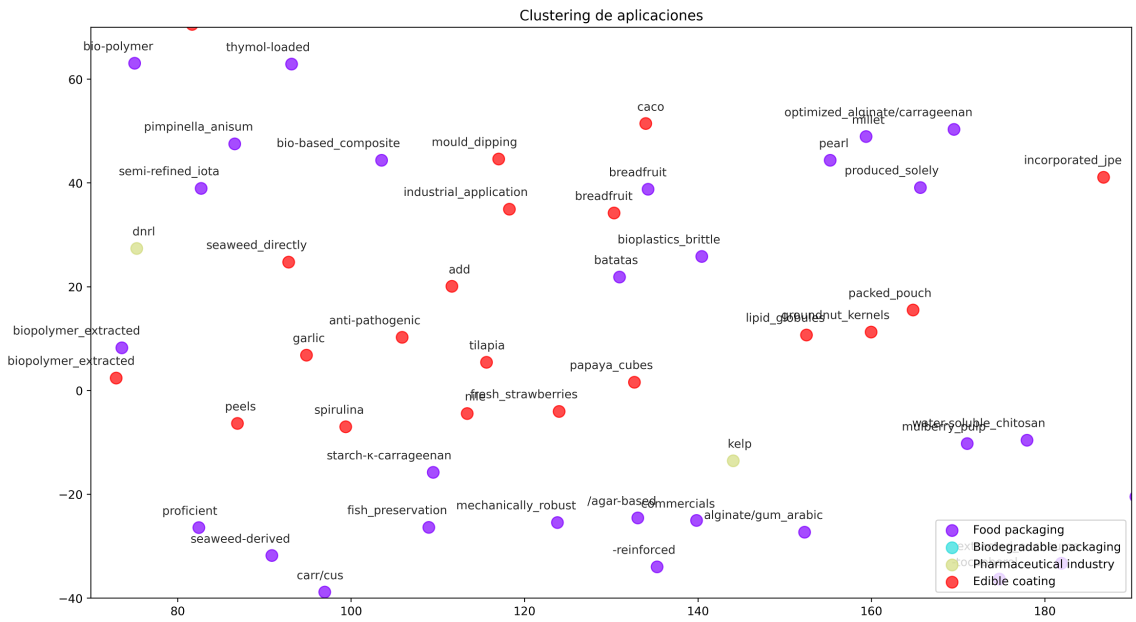

Figura 3.11: Región correspondiente a elementos B

palabras “food packaging” no aparece explícitamente en el texto, Word2vec ha identificado patrones en la literatura que sugieren que este films puede ser un buen indicador para la industria de empaquetado de alimentos.

Para el caso de la figura 3.12 se ha encontrado una variedad de compuestos biodegradables y aditivos que se utilizan en la producción de envases de alimentos a base de bioplásticos. Al ampliar el espectro de la dimensión, se han identificado compuestos biodegradables como *sps/spncc* (almidón de palma de azúcar/celulosa nanocristalina de palma de azúcar) y aditivos como *tocopherol*, *sepiolita*, *nanoclay*, *xilitol*, *Prunin laurate*, *satureja* y *Khuzestanica*.

Tocopherol es un antioxidante natural que mejora la resistencia a la oxidación de los bioplásticos (Al-Malaika, 2000), mientras que sepiolita y nanoclay se utilizan como carga para mejorar la resistencia mecánica y reducir la fragilidad de los bioplásticos (Khan *et al.*, 2022). *Xilitol* y *Prunin laurate* actúan como plastificantes (Chen *et al.*, 2019) y *satureja* y *Khuzestanica* son agentes antimicrobianos utilizados en los bioplásticos (Sallak *et al.*, 2021).

La presencia de estos aditivos en Food packaging sugiere que son comúnmente utilizados en la producción de envases de alimentos a base de bioplásticos y que probablemente tienen características o propiedades similares en la literatura.

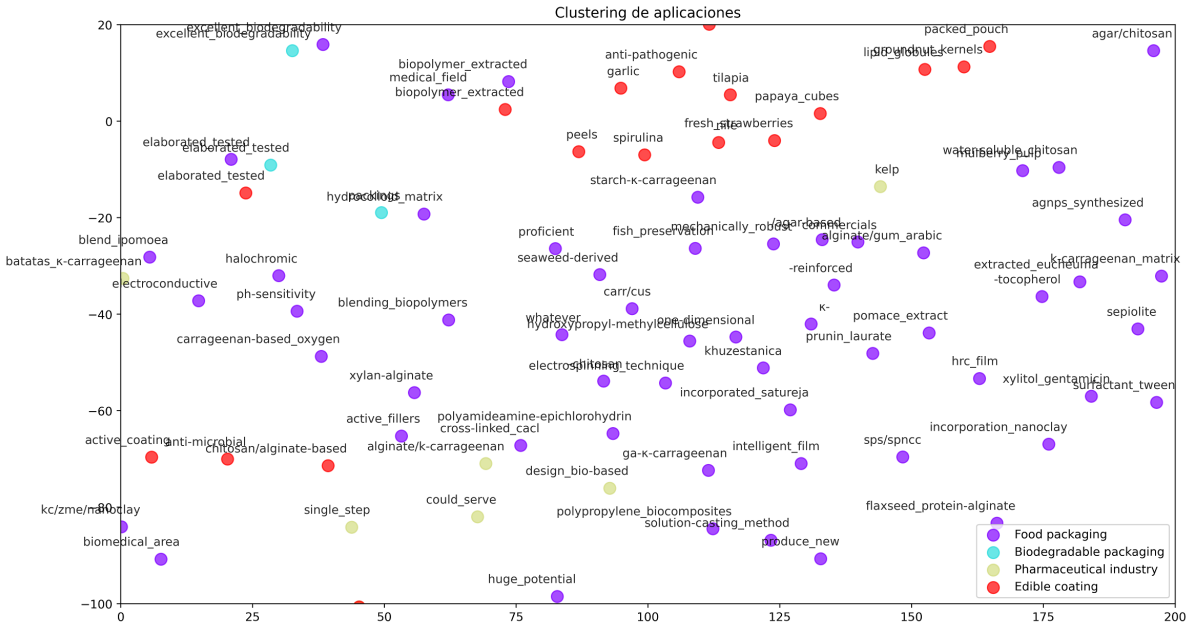

Figura 3.12: Región correspondiente a elementos C

Después de analizar la figura 3.13, se encontraron sinónimos y técnicas relacionadas con “Biodegradable packaging” y “Pharmaceutical industry”, como por ejemplo, *active pharmaceutical*, *pharmaceutical capsules*, *new biodegradable* y *development sustainable*. Una técnica destacada es la *Inactivación Fotodinámica de Microorganismos* (PDIM), esta es una técnica de desinfección utilizada en ambos campos debido a que su combinación con recubrimientos es una alternativa para extender la vida útil de alimentos como el queso ricotta (Miazaki *et al.*, 2022). Además, se encontraron ingredientes cercanos con esas aplicaciones, como la *papaya*, la *jaboticaba* (Plinia cauliflora) y la *banana*, que se identificaron frutas cercanas a palabras como *low cost biodegradable* y *biodegradable nature*. La literatura científica sugiere que algunos de estos ingredientes tienen potencial como aditivos naturales en los materiales de empaque (Avila *et al.*, 2020).

3.4.3. Comparación de resultados con diferencia de vectores

A continuación, se compararán los resultados de similitudes asociados a aplicaciones de ingredientes, por lo tanto se utilizarán suma de vectores de diferentes ingredientes que han sido señalados en la literatura mostrando una aplicación prometedora (Y. Zhang *et al.*, 2022) (Islamiyah *et al.*, 2022) (Liu *et al.*, 2022) (Aziz & Salama, 2022). Cabe destacar que no se compararán los clústeres de propiedades, ya que no se han encontrado

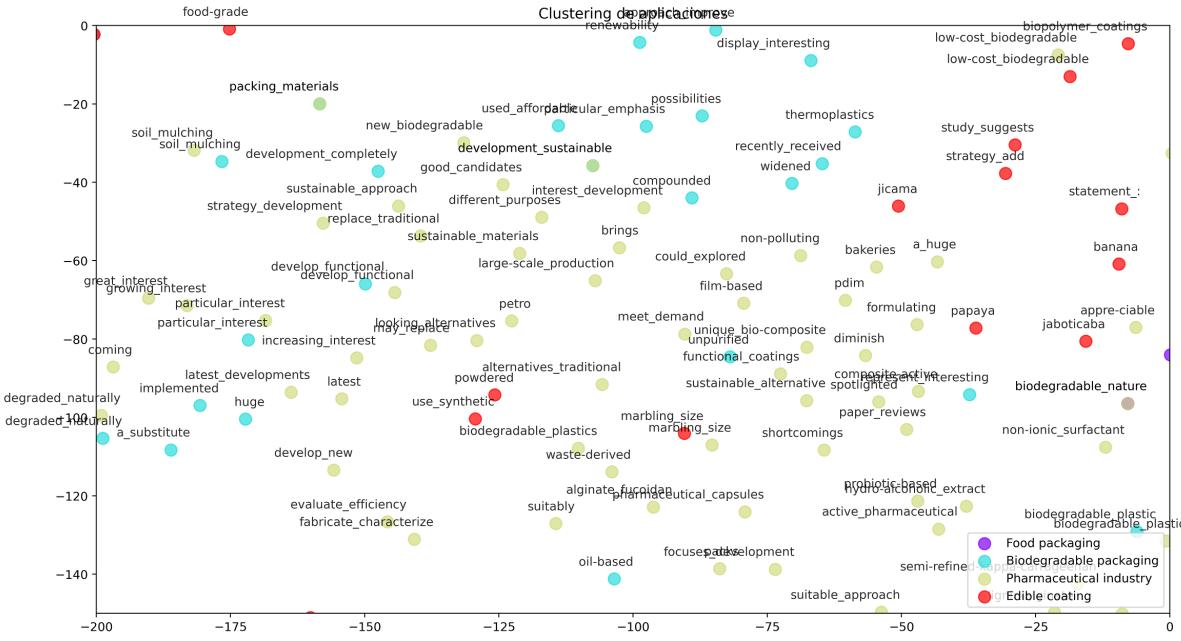

Figura 3.13: Región correspondiente a elementos D

asociaciones significativas debido a que la mayoría de los ingredientes son caracterizados por múltiples propiedades.

Los vectores a utilizar son:

```
vectors_2 = [["food_packaging","alginate/carrageenan"],
             ["biodegradable_packaging","carrageenan-based"],
             ["pharmaceutical_industry","ssps/cg/agnp"],
             ["edible_coating","sodium_alginate"]]
```

En particular, la figura A 3.14 muestra solapamiento en los datos, a pesar de esto, aún es posible agrupar ingredientes en la aplicación correspondiente al cambiar el vector de ingrediente. Por ejemplo, se observa que Edible coating tiene resultados similares a *spirulina*, *garlic*, *iota-carrageenan* y *tapioca*. Además, se destaca que algunas palabras genéricas aparecen en distintos clustering, como *elaborated_tested* y *biodegradable_nature*. Por lo tanto, las similitudes de palabras generadas por Word2vec pueden mantener la estructura general del conjunto de datos y preservar las relaciones semánticas entre las palabras en el espacio reducido de baja dimensión.

En la figura 3.15 se puede observar el impacto de las similitudes de palabras en el ajuste de la perplejidad en el algoritmo t-SNE. A pesar de cambiar las perplejidades,

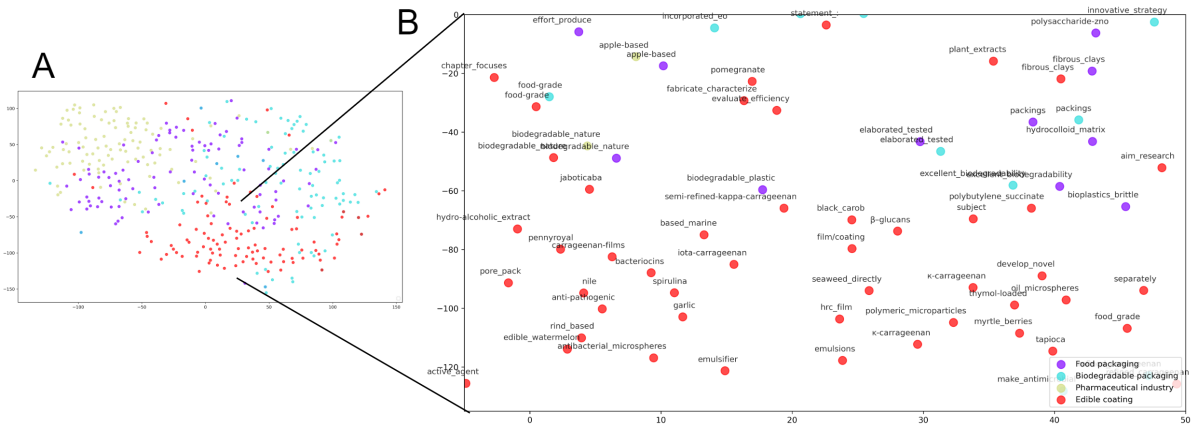

Figura 3.14: (A) Vectores de palabras asociados con seis aplicaciones comunes proyectadas en 2 dimensiones usando t-SNE, (B) región correspondiente a edible coating.

se puede apreciar que la información no afecta significativamente el clustering de las aplicaciones. Los clusters como Pharmaceutical industry tiene puntos cercanos en el espacio a Food packaging y Edible coating con biodegradable packaging. Esto sugiere que, aunque se cambie la perplejidad, los vectores se basan en una comprensión del contexto y los patrones en los datos distintos, lo cual se evidencia que las relaciones semánticas entre las palabras son relativamente consistentes con diferentes ajustes de t-SNE.

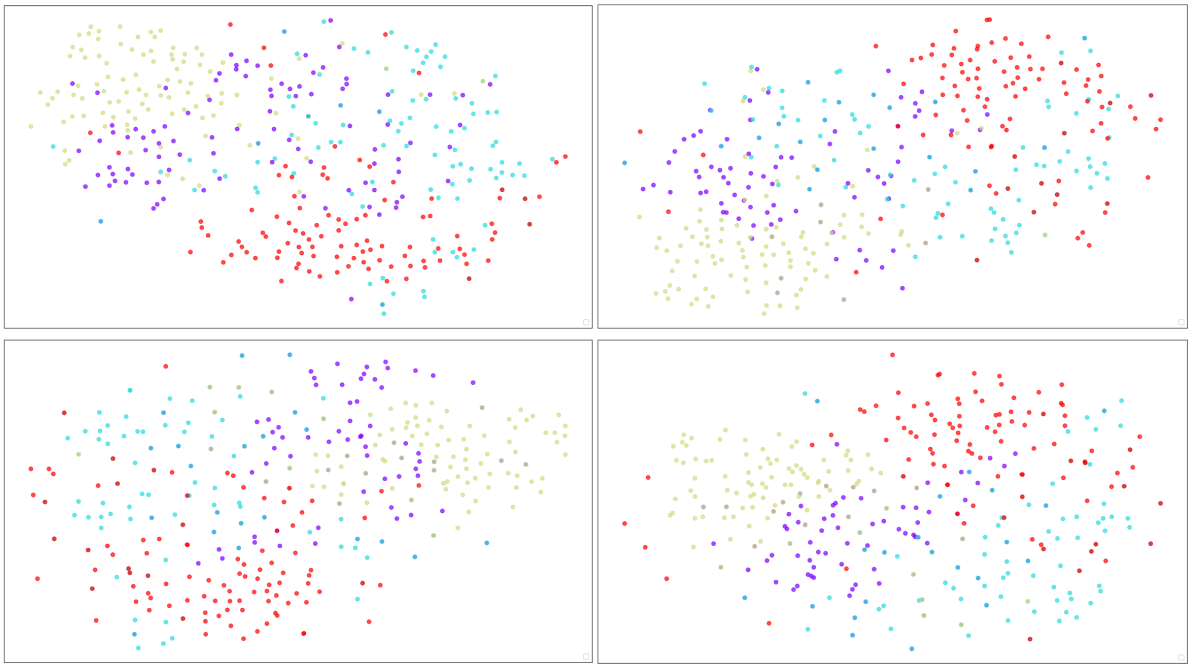

Figura 3.15: Visualización utilizando t-SNE con distintas perplejidades (perp). Izquierda superior perp 15, izquierda inferior perp 50, derecha superior perp 30 y derecha inferior perp 100

Conclusiones

En este trabajo de tesis se estudiaron las principales características de técnicas de procesamiento de lenguaje natural y aprendizaje no supervisado capaces de aprender texto humano sin una etiqueta previa, se mostraron como se representa el texto de gran volumen mediante Word embeddings.

Al analizar las tendencias de ingredientes y propiedades, se observa que los datos están mayormente relacionados con polímeros naturales derivados de algas marinas, y con aditivos naturales. En cuanto a las propiedades, se destaca que las propiedades mecánicas son las más mencionadas en la literatura, seguidas por las propiedades antimicrobianas. Al analizar la tendencia histórica de los tres principales ficocoloides, se observa que estos han generado un gran interés a partir de la década de los 2000, lo cual muestra una popularidad relativamente reciente en estos tres polímeros.

Las palabras obtenidas relacionadas con la similitud coseno entre polímeros a base de algas, destacan la capacidad de retención de agua en los alginatos y la fabricación de hidrogeles en el caso del alginato de sodio. En el caso del agar, se encontraron probabilidades menores en comparación con los alginatos y carrageninas, pero se destacaron varios materiales naturales derivados mayormente de plantas para la fabricación de bioplásticos. En cuanto a la carragenina, las palabras más similares tienen diferentes estructuras y propiedades, pero se encontraron ingredientes comúnmente utilizados en la industria alimentaria como estabilizantes y emulsionantes.

El resultado del entrenamiento por Word2vec, resultaron vectores de dimensionalidad de 200, lo cual para una mejor visualización y aproximación en el espacio vectorial de las palabras predichas, se utilizó el algortimo de t-SNE, el cual al reducir el vocabulario en dos dimensiones, es posible encontrar ciertos patrones y palabras significativas y relevantes a propiedades funcionales que dan la característica a un bioplástico. Dichos

resultados fueron analizados de forma cualitativa basado en técnicas de análisis textual utilizando el corpus de datos con abstracts.

Al realizar la visualización de datos, se observaron varias relaciones entre propiedades conocidas y la literatura. Sin embargo, también se identificó que, aunque no se mencionen directamente las palabras específicas de una propiedad, el modelo puede asociar palabras genéricas a dicha propiedad, lo cual no siempre proporciona información relevante. Por otro lado, el modelo logra captar varios conjuntos de ingredientes debido a la estructura que poseen los abstracts, lo que representa una posible ventaja en comparación con el uso de un listado de materiales individuales

En cuanto a los ingredientes más relevantes con vectores de aplicaciones, se observó que gran parte de ellos mostró coherencia en su relación con las aplicaciones correspondientes. Esto podría ser debido a que la literatura científica presenta una visión más concreta acerca de cómo utilizar distintos films y formulaciones para aplicaciones específicas. Sin embargo, también se detectaron varias presencias de geles como alginato, agar y carragenina en diferentes contextos analizados, lo que resultó en la presencia de varios tokens duplicados y ambigüedad en los resultados obtenidos.

Los resultados brindan una visión del progreso y avance en la investigación de los bioplásticos a base de algas, y cómo la tecnología del procesamiento de lenguaje natural y aprendizaje no supervisado pueden ayudar en la identificación de patrones y relaciones claves de la literatura científica. Destacando que cada cambio en el preprocesamiento y en los vectores, pueden predecir distintas palabras con distintas relaciones.

Referencias bibliográficas

Abdou, E. S. & Sorour, M. A. (2014). Preparation and characterization of starch/carrageenan edible films. *International Food Research Journal*, *21*, 189-193.

Abe, M. M., Martins, J. R., Sanvezzo, P. B., Macedo, J. V., Branciforti, M. C., Halley, P., Botaro, V. R. & Brienzo, M. (2021). Advantages and Disadvantages of Bioplastics Production from Starch and Lignocellulosic Components. *Polymers*, *13*. https://doi.org/10.3390/POLYM13152484

Aćimović, M., Šovljanski, O., Šeregelj, V., Pezo, L., Zheljazkov, V. D., Ljujić, J., Tomić, A., Ćetković, G., Čanadanović-Brunet, J., Miljković, A. & Vujisić, L. (2022). Chemical Composition, Antioxidant, and Antimicrobial Activity of Dracocephalum moldavica L. Essential Oil and Hydrolate. *Plants*, *11*, 941. https://doi.org/10.3390/PLANTS11070941/S1

Adam, F., Jamaludin, J., Bakar, S. H. A., Rasid, R. A. & Hassan, Z. (2020). Evaluation of hard capsule application from seaweed: Gum Arabic-Kappa carrageenan biocomposite films. *Cogent Engineering*, *7*. https://doi.org/10.1080/23311916.2020.1765682/SUPPL_FILE/OAEN_A_1765682_SM4859.DOCX

Akman, P. K., Bozkurt, F., Dogan, K., Tornuk, F. & Tamturk, F. (2021). Fabrication and characterization of probiotic Lactobacillus plantarum loaded sodium alginate edible films. *Journal of Food Measurement and Characterization*, *15*, 84-92. https://doi.org/10.1007/S11694-020-00619-6/METRICS

Allen, C. & Hospedales, T. (2019). Analogies Explained: Towards Understanding Word Embeddings, 1-11.

Al-Malaika, S. (2000). Vitamin E: An effective biological antioxidant for polymer stabilisation. *Polymers and Polymer Composites*, 537-542. https://www.researchgate.net/

publication/279702825_Vitamin_E_An_effective_biological_antioxidant_for_polymer_stabilisation

Alpaydin, E. & Bach, F. (2014). *Introduction to Machine Learning* (3.ª ed.). MIT Press. https://ebookcentral-proquest-com.ezproxy.usach.cl/lib/usach-ebooks/detail.action?docID=3339851

Al-Tayyar, N. A., Youssef, A. M. & Al-hindi, R. (2020). Antimicrobial food packaging based on sustainable Bio-based materials for reducing foodborne Pathogens: A review. *Food Chemistry*, *310*, 125915. https://doi.org/10.1016/J.FOODCHEM.2019.125915

Amariei, S., Ursachi, F. & Petraru, A. (2022). Development of New Biodegradable Agar-Alginate Membranes for Food Packaging. *Membranes 2022, Vol. 12, Page 576*, *12*, 576. https://doi.org/10.3390/MEMBRANES12060576

Anter, H. M., Hashim, I. I. A., Awadin, W. & Meshali, M. M. (2018). Novel anti-inflammatory film as a delivery system for the external medication with bioactive phytochemical .ᴬpocynin". *Drug design, development and therapy*, *12*, 2981-3001. https://doi.org/10.2147/DDDT.S176850

Arumugam, R. & Shanmugamani, R. (2018). *Hands-On Natural Language Processing with Python : A Practical Guide to Applying Deep Learning Architectures to Your NLP Applications*. Packt Publishing, Limited. https://ebookcentral-proquest-com.ezproxy.usach.cl/lib/usach-ebooks/detail.action?pq-origsite=primo&docID=5456142

Avila, L. B., Barreto, E. R. C., de Souza, P. K., Silva, B. D. Z., Martiny, T. R., Moraes, C. C., Morais, M. M., Raghavan, V. & da Rosa, G. S. (2020). Carrageenan-Based Films Incorporated with Jaboticaba Peel Extract: An Innovative Material for Active Food Packaging. *Molecules 2020, Vol. 25, Page 5563*, *25*, 5563. https://doi.org/10.3390/MOLECULES25235563

Aziz, M. S. A. & Salama, H. E. (2022). Development of alginate-based edible coatings of optimized UV-barrier properties by response surface methodology for food packaging applications. *International Journal of Biological Macromolecules*, *212*, 294-302. https://doi.org/10.1016/J.IJBIOMAC.2022.05.107

Bahdanau, D., Cho, K. & Bengio, Y. (2016). NEURAL MACHINE TRANSLATION BY JOINTLY LEARNING TO ALIGN AND TRANSLATE, 1-15. https://arxiv.org/pdf/1409.0473.pdf

Bartolo, A. D., Infurna, G. & Dintcheva, N. T. (2021). A Review of Bioplastics and Their Adoption in the Circular Economy. *Polymers 2021, Vol. 13, Page 1229*, *13*, 1229. https://doi.org/10.3390/POLYM13081229

Bibi, S., Mir, S., Rehman, W., Menaa, F., Gul, A., Alaryani, F. S. S., Alqahtani, A. M., Haq, S. & Abdellatif, M. H. (2022). Synthesis and In Vitro/Ex Vivo Characterizations of Ceftriaxone-Loaded Sodium Alginate/poly(vinyl alcohol) Clay Reinforced Nanocomposites: Possible Applications in Wound Healing. *Materials 2022, Vol. 15, Page 3885*, *15*, 3885. https://doi.org/10.3390/MA15113885

Blanco-Pascual, N., Montero, M. P. & Gómez-Guillén, M. C. (2014). Antioxidant film development from unrefined extracts of brown seaweeds Laminaria digitata and Ascophyllum nodosum. *Food Hydrocolloids*, *37*, 100-110. https://doi.org/10.1016/J.FOODHYD.2013.10.021

Boey, J. Y., Lee, C. K. & Tay, G. S. (2022). Factors Affecting Mechanical Properties of Reinforced Bioplastics: A Review. *Polymers*, *14*. https://doi.org/10.3390/POLYM14183737

Bora, D., Jayaramudu, J., Saikia, P., Bohra, R. C., Phukan, L., Selvam, S. P., Ray, S. S. & Sadiku, E. R. (2022). Effect of boehmite alumina nanoparticles on the physical and chemical characteristics of eco-friendly sodium alginate/polyvinyl alcohol bionanocomposite film. *International Journal of Polymer Analysis and Characterization*, *27*, 236-251. https://doi.org/10.1080/1023666X.2022.2061749

Brizga, J., Hubacek, K. & Feng, K. (2020). The Unintended Side Effects of Bioplastics: Carbon, Land, and Water Footprints. *One Earth*, *3*(1), 45-53. https://doi.org/https://doi.org/10.1016/j.oneear.2020.06.016

Bullinaria, J. A. & Levy, J. P. (2007). Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior research methods*, *39*(3), 510-526.

Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O. & Walsh, A. (2018). Machine learning for molecular and materials science. *Nature 2018 559:7715*, *559*, 547-555. https://doi.org/10.1038/s41586-018-0337-2

Campa-Siqueiros, P. I., Vargas-Arispuro, I., Quintana-Owen, P., Freile-Pelegrín, Y., Azamar-Barrios, J. A. & Madera-Santana, T. J. (2020). Physicochemical and transport properties of biodegradable agar films impregnated with natural semiochemical

based-on hydroalcoholic garlic extract. *International Journal of Biological Macromolecules*, *151*, 27-35. https://doi.org/10.1016/J.IJBIOMAC.2020.02.158

Cheikh, D., Martín-Sampedro, R., Majdoub, H. & Darder, M. (2020). Alginate bionanocomposite films containing sepiolite modified with polyphenols from myrtle berries extract. *International Journal of Biological Macromolecules*, *165*, 2079-2088. https://doi.org/10.1016/J.IJBIOMAC.2020.10.052

Chen, H., Wang, J., Cheng, Y., Wang, C., Liu, H., Bian, H., Pan, Y., Sun, J. & Han, W. (2019). Application of Protein-Based Films and Coatings for Food Packaging: A Review. *Polymers*, *11*. https://doi.org/10.3390/POLYM11122039

de Lima Barizão, C., Crepaldi, M. I., de Oliveira S. Junior, O., de Oliveira, A. C., Martins, A. F., Garcia, P. S. & Bonafé, E. G. (2020). Biodegradable films based on commercial k-carrageenan and cassava starch to achieve low production costs. *International Journal of Biological Macromolecules*, *165*, 582-590. https://doi.org/10.1016/J.IJBIOMAC.2020.09.150

Foschi, E. & Bonoli, A. (2019). The Commitment of Packaging Industry in the Framework of the European Strategy for Plastics in a Circular Economy. *Administrative Sciences 2019, Vol. 9, Page 18*, *9*, 18. https://doi.org/10.3390/ADMSCI9010018

Fransiska, D., Giyatmi, Basmal, J. & Susanti, E. (2020). The effect of organic powdered cottonii concentration and types of plasticizers on the characteristics of edible film. *IOP Conference Series: Earth and Environmental Science*, *483*, 012008. https://doi.org/10.1088/1755-1315/483/1/012008

Friedman, C., Kra, P., Yu, H., Krauthammer, M. & Rzhetsky, A. (2001). GO BACK GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles. *17*, 74-82. http://www.bioinformatics.oupjournals.org

Giordano, A., Caruso, M. R. & Lazzara, G. (2022). New tool for sustainable treatments: agar spray—research and practice. *Heritage Science*, *10*, 1-16. https://doi.org/10.1186/S40494-022-00756-9/FIGURES/26

Glorot, X., Bordes, A. & Bengio, Y. (2011). Deep sparse rectifier neural networks. *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, 315-323.

Goldberg, Y. (2016). A primer on neural network models for natural language processing. *Journal of Artificial Intelligence Research*, *57*, 345-420.

Goldberg, Y. (2017). Neural Network Methods for Natural Language Processing. *Neural Network Methods for Natural Language Processing*. https://doi.org/10.1007/978-3-031-02165-7

Goldberg, Y., Levy, O., Mikolov, T., Sutskever, I., Chen, K., Corrado, G. & Dean, J. (2014). word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method. https://doi.org/10.48550/arxiv.1402.3722

Gomaa, M., Fawzy, M. A., Hifney, A. F. & Abdel-Gawad, K. M. (2018). Use of the brown seaweed Sargassum latifolium in the design of alginate-fucoidan based films with natural antioxidant properties and kinetic modeling of moisture sorption and polyphenolic release. *Food Hydrocolloids*, *82*, 64-72. https://doi.org/10.1016/J.FOODHYD.2018.03.053

Häder, D.-P. (2021). Phycocolloids from macroalgae. *Natural Bioactive Compounds*, 187-201. https://doi.org/10.1016/B978-0-12-820655-3.00009-4

Halimah, L. S., -, K. H., Tanti, A., Retnani, Y., -, I. R. H. S., Iin, H. & Fahma, F. (2022). Production of zeolite-cellulose nanocomposites with garlic essential oil for antimicrobial tablets. *IOP Conference Series: Earth and Environmental Science*, *1034*, 012016. https://doi.org/10.1088/1755-1315/1034/1/012016

Hamid, K. H., Saupy, N. A., Zain, N. M., Mudalip, S. K., Shaarani, S. M. & Azman, N. A. (2018). Development and characterization of semi-refined carrageenan (SRC) films from Eucheuma cottonii incorporated with glycerol and tocopherol for active food packaging application. *IOP Conference Series: Materials Science and Engineering*, *458*, 012022. https://doi.org/10.1088/1757-899X/458/1/012022

Hanani, Z. A. N., Roos, Y. H. & Kerry, J. P. (2014). Use and application of gelatin as potential biodegradable packaging materials for food products. *International Journal of Biological Macromolecules*, *71*, 94-102. https://doi.org/10.1016/J.IJBIOMAC.2014.04.027

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

Hernández, V., Ibarra, D., Triana, J. F., Martínez-Soto, B., Faúndez, M., Vasco, D. A., Gordillo, L., Herrera, F., García-Herrera, C. & Garmulewicz, A. (2022). Agar Biopolymer Films for Biodegradable Packaging: A Reference Dataset for Exploring the Limits of Mechanical Performance. *Materials*, *15*, 3954. https://doi.org/10.3390/MA15113954/S1

Hill, J., Mulholland, G., Persson, K., Seshadri, R., Wolverton, C. & Meredig, B. (2016). Materials science with large-scale data and informatics: Unlocking new opportunities. *MRS Bulletin*, *41*, 399-409. https://doi.org/10.1557/MRS.2016.93

Huang, T., Qian, Y., Wei, J. & Zhou, C. (2019). Polymeric Antimicrobial Food Packaging and Its Applications. *polymers*, 1-18. https://doi.org/10.3390/polym11030560

Iles, A. & Martin, A. N. (2013). Expanding bioplastics production: sustainable business innovation in the chemical industry. *Journal of Cleaner Production*, *45*, 38-49. https://doi.org/10.1016/J.JCLEPRO.2012.05.008

Islamiyah, H. S., Alamsjah, M. A. & Pujiastuti, D. Y. (2022). Application of Modified Starch in the Carragenan-Based Biodegradable Packaging from Eucheuma cottonii on Biodegradablility and Mechanical Properties. *IOP Conference Series: Earth and Environmental Science*, *1036*, 012072. https://doi.org/10.1088/1755-1315/1036/1/012072

Izdebska-Podsiadły, J. (2019). Application of Plasma in Printed Surfaces and Print Quality. *Non-Thermal Plasma Technology for Polymeric Materials: Applications in Composites, Nanostructured Materials, and Biomedical Fields*, 159-191. https://doi.org/10.1016/B978-0-12-813152-7.00006-8

Jambeck, J. R., Geyer, R., Wilcox, C., Siegler, T. R., Perryman, M., Andrady, A., Narayan, R. & Law, K. L. (2015). Plastic waste inputs from land into the ocean. *Science*, *347*, 768-771. https://doi.org/10.1126/SCIENCE.1260352/SUPPL_FILE/JAMBECK.SM.PDF

Khalil, H. P. A., Saurabh, C. K., Tye, Y. Y., Lai, T. K., Easa, A. M., Rosamah, E., Fazita, M. R., Syakir, M. I., Adnan, A. S., Fizree, H. M., Aprilia, N. A. & Banerjee, A. (2017). Seaweed based sustainable films and composites for food and pharmaceutical applications: A review. *Renewable and Sustainable Energy Reviews*, *77*, 353-362. https://doi.org/10.1016/J.RSER.2017.04.025

Khan, Z. I., Habib, U., Mohamad, Z. B., Rahmat, A. R. B. & Abdullah, N. A. S. B. (2022). Mechanical and thermal properties of sepiolite strengthened thermoplastic polymer nanocomposites: A comprehensive review. *Alexandria Engineering Journal*, *61*, 975-990. https://doi.org/10.1016/J.AEJ.2021.06.015

Kim, H.-J., Roy, S. & Rhim, J.-W. (2022). Gelatin/agar-based color-indicator film integrated with Clitoria ternatea flower anthocyanin and zinc oxide nanoparticles for monitoring

freshness of shrimp. *Food Hydrocolloids*, *124*, 107294. https://doi.org/10.1016/J.FOODHYD.2021.107294

Kocijan, V., Lukasiewicz, T., Davis, E., Marcus, G. & Morgenstern, L. (2020). A Review of Winograd Schema Challenge Datasets and Approaches. https://doi.org/10.48550/arxiv.2004.13831

Kyriakides, G. & Margaritis, K. G. (2019). *Hands-On Ensemble Learning with Python: Build highly optimized ensemble machine learning models using scikit-learn and Keras*. Packt Publishing Ltd.

Larsen, P. O. & von Ins, M. (2010). The rate of growth in scientific publication and the decline in coverage provided by science citation index. *Scientometrics*, *84*, 575-603. https://doi.org/10.1007/S11192-010-0202-Z/TABLES/8

Lee, C., Volpi, F., Fiocco, G., Weththimuni, M. L., Licchelli, M. & Malagodi, M. (2022). Preliminary Cleaning Approach with Alginate and Konjac Glucomannan Polysaccharide Gel for the Surfaces of East Asian and Western String Musical Instruments. *Materials*, *15*, 1100. https://doi.org/10.3390/MA15031100/S1

Li, F., Liu, Y., Cao, Y., Zhang, Y., Zhe, T., Guo, Z., Sun, X., Wang, Q. & Wang, L. (2020). Copper sulfide nanoparticle-carrageenan films for packaging application. *Food Hydrocolloids*, *109*, 106094. https://doi.org/10.1016/J.FOODHYD.2020.106094

Liu, J., Dong, Y., Ma, Z., Rao, Z., Zheng, X. & Tang, K. (2022). Soluble Soybean Polysaccharide/Carrageenan Antibacterial Nanocomposite Films Containing Green Synthesized Silver Nanoparticles. *ACS Applied Polymer Materials*, *4*, 5608-5618. https://doi.org/10.1021/ACSAPM.2C00635/SUPPL_FILE/AP2C00635_SI_001.PDF

Ma, L. & Zhang, Y. (2015). Using Word2Vec to process big text data. *Proceedings - 2015 IEEE International Conference on Big Data, IEEE Big Data 2015*, 2895-2897. https://doi.org/10.1109/BIGDATA.2015.7364114

Maaten, L. V. D. & Hinton, G. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research*, *9*, 2579-2605.

Malhotra, B., Keshwani, A. & Kharkwal, H. (2015). Antimicrobial food packaging: Potential and pitfalls. *Frontiers in Microbiology*, *6*, 611. https://doi.org/10.3389/FMICB.2015.00611/BIBTEX

McCormick, C. (2016). *Word2Vec Tutorial - The Skip-Gram Model*. http://mccormickml.com/2016/04/19/word2vec-tutorial-the-skip-gram-model/

McHugh, D. J. (1987). *CHAPTER 2 - PRODUCTION, PROPERTIES AND USES OF ALGINATES*. https://www.fao.org/3/x5822e/x5822e04.htm#chapter%5C%202%5C%20%5C%20%5C%20production,%5C%20properties%5C%20and%5C%20uses%5C%20of%C2%BA%5C%20alginates

Mesgari, M., Aalami, A. H., Sathyapalan, T. & Sahebkar, A. (2022). A Comprehensive Review of the Development of Carbohydrate Macromolecules and Copper Oxide Nanocomposite Films in Food Nanopackaging. *Bioinorganic Chemistry and Applications*, *2022*. https://doi.org/10.1155/2022/7557825

Miazaki, J. B., dos Santos, A. R., de Freitas, C. F., Stafussa, A. P., Mikcha, J. M. G., de Cássia Bergamasco, R., Tonon, L. A. C., Madrona, G. S., Caetano, W., da Silva, L. H. & da Silva Scapim, M. R. (2022). Edible coatings and application of photodynamics in ricotta cheese preservation. *LWT*, *165*, 113697. https://doi.org/10.1016/J.LWT.2022.113697

Mikolov, T. (2016). *Machine Learning Prague 2016 – conference on machine learning in practice*. http://2016.mlprague.com/

Mikolov, T., Chen, K., Corrado, G. & Dean, J. (2013a). Distributed Representations of Words and Phrases and their Compositionality. https://arxiv.org/pdf/1310.4546.pdf

Mikolov, T., Chen, K., Corrado, G. & Dean, J. (2013b). Efficient Estimation of Word Representations in Vector Space. https://arxiv.org/pdf/1301.3781.pdf

Mikolov, T., Yih, W.-T. & Zweig, G. (2013). Linguistic Regularities in Continuous Space Word Representations, 9-14. http://research.microsoft.com/en-

Mohamadinia, P., Anarjan, N. & Jafarizadeh-Malmiri, H. (2021). *Green Processing and Synthesis*, *10*(1), 860-873. https://doi.org/doi:10.1515/gps-2021-0081

Moreira, D., Cruz, I., Gonzalez, K., Quirumbay, A., Magallan, C., Guarda, T., Andrade, A. & Castillo, C. (s.f.). Análisis del Estado Actual de Procesamiento de Lenguaje Natural Analysis of the Current State of Natural Language Processing.

Ong, S. P., Richards, W. D., Jain, A., Hautier, G., Kocher, M., Cholia, S., Gunter, D., Chevrier, V. L., Persson, K. A. & Ceder, G. (2013). Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis. *Computational Materials Science*, *68*, 314-319. https://doi.org/10.1016/J.COMMATSCI.2012.10.028

Perkins, J., Chopra, D. & Hardeniya, N. (2016). Natural Language Processing : Python and NLTK Table of Contents, 1-7. https://www.oreilly.com/library/view/natural-language-processing/9781787285101/

Prieto, A., Prieto, B., Ortigosa, E. M., Ros, E., Pelayo, F., Ortega, J. & Rojas, I. (2016). Neural networks: An overview of early research, current frameworks and new challenges. *Neurocomputing*, *214*, 242-268. https://doi.org/10.1016/J.NEUCOM.2016.06.014

Radulovich, R., Neori, A., Valderrama, D., Reddy, C. R., Cronin, H. & Forster, J. (2015). Farming of seaweeds. *Seaweed Sustainability: Food and Non-Food Applications*, 27-59. https://doi.org/10.1016/B978-0-12-418697-2.00003-9

Ramisen, R. & Galatas, F. (1987). *CHAPTER 1 - PRODUCTION, PROPERTIES AND USES OF AGAR*. Department of Chemistry, University College. https://www.fao.org/3/x5822e/x5822e03.htm#TopOfPage

Ramsundar, B., Kearnes, S., Riley, P., Webster, D., Konerding, D., Pande, V. & Edu, P. (2015). Massively Multitask Networks for Drug Discovery. https://tripod.

Řehůřek, R. (2010). Gensim: Python Framework for Vector Space Modelling. https://radimrehurek.com/gensim/index.html

Rizal, S., Alfatah, T., H. P. S., A. K., Mistar, E. M., Abdullah, C. K., Olaiya, F. G., Sabaruddin, F. A., Ikramullah & Muksin, U. (2021). Properties and Characterization of Lignin Nanoparticles Functionalized in Macroalgae Biopolymer Films. *Nanomaterials*, *11*(3). https://doi.org/10.3390/nano11030637

Rong, X. (2016). word2vec Parameter Learning Explained, 1-21. https://arxiv.org/pdf/1411.2738.pdf

Roy, S., Priyadarshi, R. & Rhim, J. W. (2022). Gelatin/agar-based multifunctional film integrated with copper-doped zinc oxide nanoparticles and clove essential oil Pickering emulsion for enhancing the shelf life of pork meat. *Food Research International*, *160*, 111690. https://doi.org/10.1016/J.FOODRES.2022.111690

Roy, S. & Rhim, J. W. (2021). Fabrication of pectin/agar blended functional film: Effect of reinforcement of melanin nanoparticles and grapefruit seed extract. *Food Hydrocolloids*, *118*, 106823. https://doi.org/10.1016/J.FOODHYD.2021.106823

Roy, S. & Rhim, J. W. (2022). Starch/agar-based functional films integrated with enoki mushroom-mediated silver nanoparticles for active packaging applications. *Food Bioscience*, *49*, 101867. https://doi.org/10.1016/J.FBIO.2022.101867

Rumelhart, D. E., Hinton, G. E. & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature 1986 323:6088*, *323*, 533-536. https://doi.org/10.1038/323533a0

Russell Stuart, J. & Norvig, P. (2009). *Artificial intelligence: a modern approach*. Prentice Hall.

Saito, T., Kimura, S., Nishiyama, Y. & Isogai, A. (2007). Cellulose nanofibers prepared by TEMPO-mediated oxidation of native cellulose. *Biomacromolecules*, *8*, 2485-2491. https://doi.org/10.1021/BM0703970

Sallak, N., Moghanjoughi, A. M., Ataee, M., Anvar, A. & Golestan, L. (2021). Antimicrobial biodegradable film based on corn starch/Satureja khuzestanica essential oil/Ag-TiO2nanocomposites. *Nanotechnology*, *32*. https://doi.org/10.1088/1361-6528/AC0A15

Sanchez-Garcia, M. D., Hilliou, L. & Lagaron, J. M. (2010). Nanobiocomposites of Carrageenan, Zein, and Mica of Interest in Food Packaging and Coating Applications. *Journal of Agricultural and Food Chemistry*, *58*, 6884-6894. https://doi.org/10.1021/JF1007659

Sangroniz, A., Zhu, J.-B., Tang, X., Etxeberria, A., Chen, E. Y.-X. & Sardon, H. (2019). Packaging materials with desired mechanical and barrier properties and full chemical recyclability, 1-7. https://doi.org/10.1038/s41467-019-11525-x

Shetty, P. & Ramprasad, R. (2021). Automated knowledge extraction from polymer literature using natural language processing. *iScience*, *24*, 101922. https://doi.org/10.1016/J.ISCI.2020.101922

Shlush, E. & Davidovich-Pinhas, M. (2022). Bioplastics for food packaging. *Trends in Food Science & Technology*, *125*, 66-80. https://doi.org/10.1016/J.TIFS.2022.04.026

Smith, W. F. & Hashemi, J. (2006). *Fundamentos de La Ciencia e Ingenieria de Materiales*. McGRAW-HILL Interamericana Editores S.A. de C.V. https://es.scribd.com/document/539044075/William-F-Smith-and-Javad-Hashemi-Fundamentos-de-La-Ciencia-e-Ingenieria-de-Materiales-McGraw-Hill-2006

Spangler, S., Wilkins, A. D., Bachman, B. J., Nagarajan, M., Dayaram, T., Haas, P., Regenbogen, S., Pickering, C. R., Comer, A., Myers, J. N., Stanoi, I., Kato, L., Lelescu, A., Labrie, J. J., Parikh, N., Lisewski, A. M., Donehower, L., Chen, Y. & Lichtarge, O. (2014). Automated Hypothesis Generation Based on Mining Scientific Literature. https://doi.org/10.1145/2623330.2623667

Stanley, N. (1987). *CHAPTER 3 - PRODUCTION, PROPERTIES AND USES OF CARRAGEENAN*. https://www.fao.org/3/x5822e/x5822e05.htm#TopOfPage

Strnad, S., Oberhollenzer, Z., Šauperl, O., Kreže, T. & Zemljič, L. F. (2019). MODIFYING PROPERTIES OF FEATHER KERATIN BIOPLASTIC FILMS USING KONJAC GLUCOMANNAN. *CELLULOSE CHEMISTRY AND TECHNOLOGY Cellulose Chem. Technol*, *53*, 1017-1027.

Sun, G., Chi, W., Zhang, C., Xu, S., Li, J. & Wang, L. (2019). Developing a green film with pH-sensitivity and antioxidant activity based on k-carrageenan and hydroxypropyl methylcellulose incorporating Prunus maackii juice. *Food Hydrocolloids*, *94*, 345-353. https://doi.org/10.1016/J.FOODHYD.2019.03.039

Swain, M. C. & Cole, J. M. (2016). ChemDataExtractor: A Toolkit for Automated Extraction of Chemical Information from the Scientific Literature. *Journal of Chemical Information and Modeling*, *56*, 1894-1904. https://doi.org/10.1021/ACS.JCIM.6B00207/ASSET/IMAGES/LARGE/CI-2016-00207J_0010.JPEG

Tabassum, N. & Khan, M. A. (2020). Modified atmosphere packaging of fresh-cut papaya using alginate based edible coating: Quality evaluation and shelf life study. *Scientia Horticulturae*, *259*, 108853. https://doi.org/10.1016/J.SCIENTA.2019.108853

Tian, H., Xu, G., Yang, B. & Guo, G. (2011). Microstructure and mechanical properties of soy protein/agar blend films: Effect of composition and processing methods. *Journal of Food Engineering*, *107*, 21-26. https://doi.org/10.1016/J.JFOODENG.2011.06.008

Tolle, K. M., Tansley, D. S. W. & Hey, A. J. (2011). The fourth Paradigm: Data-intensive scientific discovery. *Proceedings of the IEEE*, *99*, 1334-1337. https://doi.org/10.1109/JPROC.2011.2155130

Tshitoyan, V., Dagdelen, J., Weston, L., Dunn, A., Rong, Z., Kononova, O., Persson, K. A., Ceder, G. & Jain, A. (2019). Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature 2019 571:7763*, *571*, 95-98. https://doi.org/10.1038/s41586-019-1335-8

Usov, A. I. (2011). Polysaccharides of the red algae. *Advances in Carbohydrate Chemistry and Biochemistry*, *65*, 115-217. https://doi.org/10.1016/B978-0-12-385520-6.00004-2

Uthaya Kumar, U. S., Abdulmadjid, S. N., Olaiya, N. G., Amirul, A. A., Rizal, S., Rahman, A. A., Alfatah, T., Mistar, E. M. & Abdul Khalil, H. P. S. (2020). Extracted Compounds from Neem Leaves as Antimicrobial Agent on the Physico-Chemical Properties of Seaweed-Based Biopolymer Films. *Polymers*, *12*(5). https://doi.org/10.3390/polym12051119

Vajjala, S., Majumder, B., Gupta, A. & Surana, H. (2020). *Practical Natural Language Processing: A Comprehensive Guide to Building Real-World NLP Systems*. O'Reilly Media.

Van der Maaten, L. & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of machine learning research*, *9*(11).

Visco, A., Scolaro, C., Facchin, M., Brahimi, S., Belhamdi, H., Gatto, V. & Beghetto, V. (2022). Agri-Food Wastes for Bioplastics: European Prospective on Possible Applications in Their Second Life for a Circular Economy. *Polymers 2022, Vol. 14, Page 2752*, *14*, 2752. https://doi.org/10.3390/POLYM14132752

Wang, X. H., Song, X. J., Zhang, D. J., Li, Z. J. & Wang, H. J. (2022). Preparation and characterization of natamycin-incorporated agar film and its application on preservation of strawberries. *Food Packaging and Shelf Life*, *32*, 100863. https://doi.org/10.1016/J.FPSL.2022.100863

Wardana, A. A. & Widyaningsih, T. D. (2017). Development of edible films from tapioca starch and agar, enriched with red cabbage (Brassica oleracea) as a sausage deterioration bio-indicator. *IOP Conference Series: Earth and Environmental Science*, *109*, 012031. https://doi.org/10.1088/1755-1315/109/1/012031

Whitt, E. W. (1958). A surface film method of evaluating quaternary ammonium compounds. *Journal of Applied Bacteriology*, *21*, 272-277. https://doi.org/10.1111/J.1365-2672.1958.TB00143.X

Xiao, L., Liu, Y., Zhang, X., Li, C., Qin, Y. & Liu, J. (2020). Comparison of the structural, physical and functional properties of k-carrageenan films incorporated with pomegranate flesh and peel extracts. *International Journal of Biological Macromolecules*, *147*, 1076-1088. https://doi.org/10.1016/J.IJBIOMAC.2019.10.075

Yang, Y., Yu, X., Zhu, Y., Zeng, Y., Fang, C., Liu, Y., Hu, S., Ge, Y. & Jiang, W. (2022). Preparation and application of a colorimetric film based on sodium alginate/sodium carboxymethyl cellulose incorporated with rose anthocyanins. *Food Chemistry*, *393*, 133342. https://doi.org/10.1016/J.FOODCHEM.2022.133342

Zhang, R., Zhao, W., Ning, F., Zhen, J., Qiang, H., Zhang, Y., Liu, F. & Jia, Z. (2022). Alginate Fiber-Enhanced Poly(vinyl alcohol) Hydrogels with Superior Lubricating Property and Biocompatibility. *Polymers 2022, Vol. 14, Page 4063*, *14*, 4063. https://doi.org/10.3390/POLYM14194063

Zhang, Y., Man, J., Li, J., Xing, Z., Zhao, B., Ji, M., Xia, H. & Li, J. (2022). Preparation of the alginate/carrageenan/shellac films reinforced with cellulose nanocrystals obtained from enteromorpha for food packaging. *International Journal of Biological Macromolecules*, *218*, 519-532. https://doi.org/10.1016/J.IJBIOMAC.2022.07.145

Zhao, T., Liu, Y., Ju, W. & Shi, S. (2017). Materials discovery and design using machine learning. *Journal of Materiomics*, *3*, 159-177. https://doi.org/10.1016/J.JMAT.2017.08.002