Tarea 1 - Análisis de Nombres de Productos

Nelson Alfonso Beas Ham
Matrícula: 1942687
Universidad Autónoma de Nuevo León
Facultad de Ciencias Físico Matemáticas

28 de enero de 2025

1. Introducción

En este reporte se presenta un análisis detallado de una lista de nombres de productos. El objetivo principal es realizar un preprocesamiento de los datos, incluyendo la limpieza del texto, tokenización, eliminación de palabras vacías y análisis estadístico. Además, se generan visualizaciones para comprender mejor la distribución de las palabras más comunes en los nombres de los productos.

2. Metodología

El análisis se realizó siguiendo los siguientes pasos:

1. **Preprocesamiento:** Limpieza del texto para eliminar caracteres especiales, convertir todo a minúsculas y normalizar los datos.
2. **Tokenización:** Segmentación del texto en palabras individuales.
3. **Eliminación de palabras vacías:** Filtrado de palabras comunes que no aportan significado contextual.
4. **Análisis estadístico:** Cálculo de frecuencias de palabras, bigramas y trigramas.
5. **Visualización:** Generación de una nube de palabras y tablas resumen.

3. Preprocesamiento

En esta etapa, se realizó la limpieza del texto de los nombres de los productos. Se eliminaron caracteres especiales y se convirtió todo el texto a minúsculas. Además, se conservó la columna original de nombres de productos y se agregó una nueva columna con el texto preprocesado.

4. Tokenización

El texto preprocesado se dividió en palabras individuales (tokens) para facilitar su análisis. Cada nombre de producto se convirtió en una lista de palabras.

5. Eliminación de palabras vacías

Se utilizó una lista personalizada de palabras vacías (conectores, artículos, preposiciones, etc.) para filtrar palabras comunes que no aportan significado contextual. Esto permitió enfocar el análisis en términos más relevantes.

6. Análisis estadístico

Se calcularon las frecuencias de las palabras, bigramas (pares de palabras consecutivas) y trigramas (tres palabras consecutivas). A continuación, se presentan los resultados más relevantes:

6.1. Frecuencia de palabras

Las palabras más comunes en los nombres de los productos son:

- **chocolate**: 15 apariciones
- **cookies**: 10 apariciones
- **salt**: 8 apariciones

6.2. Bigramas más comunes

Los bigramas más frecuentes son:

- **chocolate cookies**: 5 apariciones
- **sandwich cookies**: 4 apariciones

6.3. Trigramas más comunes

Los trigramas más frecuentes son:

- **chocolate sandwich cookies**: 3 apariciones

7. Visualización

Para visualizar los resultados, se generaron varias gráficas que muestran las palabras más frecuentes, los trigramas más comunes y una nube de palabras.

7.1. Top 20 palabras más frecuentes

A continuación se presenta un histograma con las 20 palabras más frecuentes en los nombres de los productos:

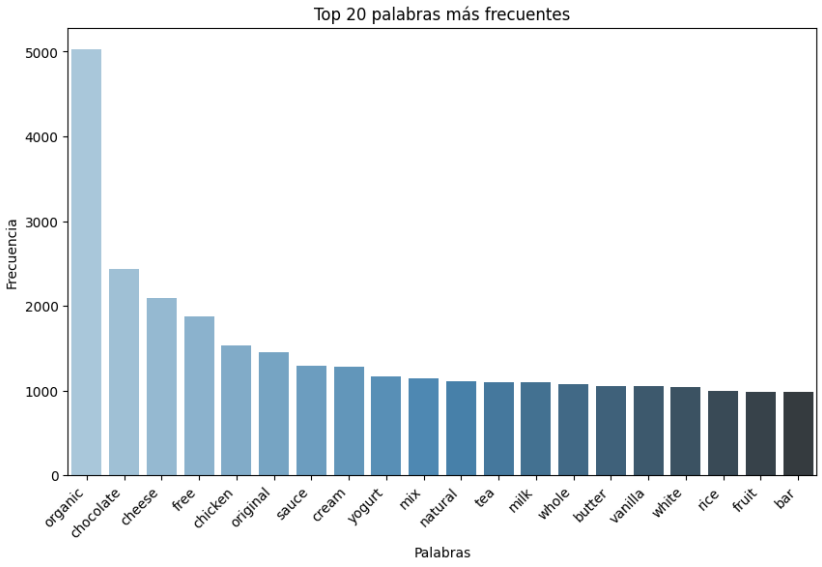

Figura 1: Top 20 palabras más frecuentes en los nombres de productos

7.2. Top 10 trigramas más comunes

A continuación se presenta un histograma con los trigramas más comunes:

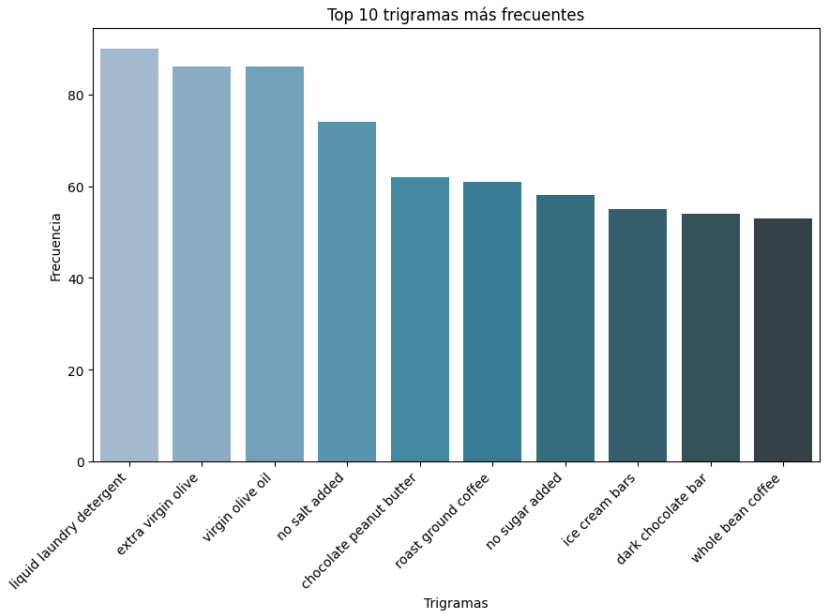

Figura 2: Top 10 trigramas más comunes en los nombres de productos

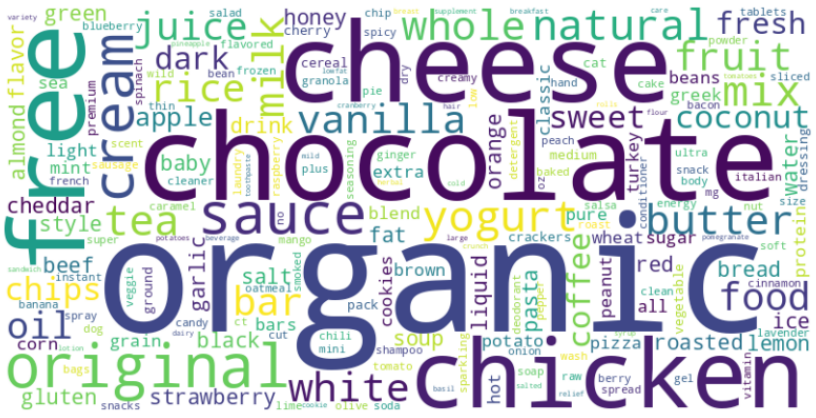

Figura 3: Nube de palabras de los nombres de productos

8. Conclusiones

El análisis de los nombres de productos permitió identificar patrones y términos comunes. La nube de palabras, el histograma de las palabras más frecuentes y los trigramas más comunes proporcionan una visión clara de las palabras más relevantes en los nombres de los productos. Este tipo de

análisis puede ser útil para mejorar la estrategia de naming de productos y comprender mejor las tendencias del mercado.