**INSTITUTO TECNOLÓGICO Y DE ESTUDIOS SUPERIORES DE MONTERREY**

**A blue and white logo

Description automatically generated**

# 

# Clasificación automática de descripciones de carga contenerizada en códigos del Sistema Armonizado (HS) usando modelos de Procesamiento de Lenguaje Natural

**Prof. Titular: Dra. Grettel Barceló Alonso**

**Juan Pablo Acosta López (A01794035)**

**Oscar Villa Cárdenas (A01794052)**

**Genaro Rodriguez Vázquez (A01150931)**

**29 septiembre 2024**

**Índice**

[Clasificación automática de descripciones de carga contenerizada en códigos del Sistema Armonizado (HS) usando modelos de Procesamiento de Lenguaje Natural](#_heading=h.crzpksqdu7ux)

[Análisis descriptivo de manifiestos validados (Input)](#_heading=h.r0z25814vwo2)

[Análisis descriptivo del diccionario (Objetivo)](#_heading=h.wjnl5x7ghx8v)

[1. Naturaleza de los datos:](#_heading=h.pj3cdn56v6pq)

[2. Transformaciones necesarias:](#_heading=h.9zt0iti36ede)

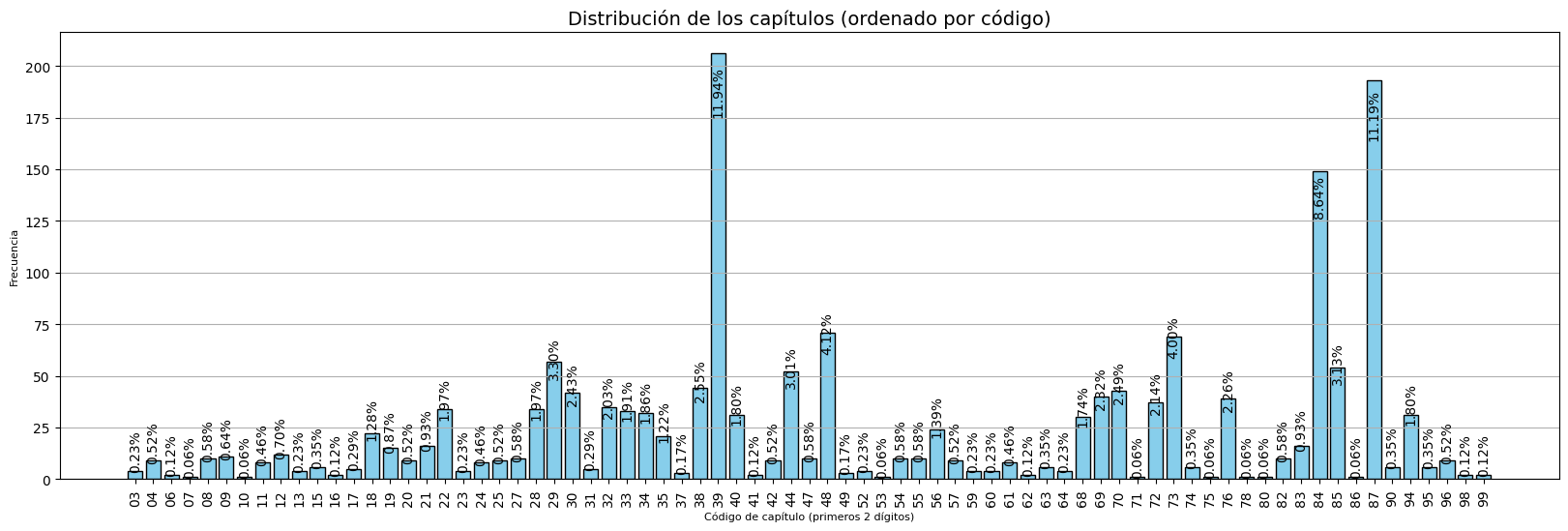
[3. Ambigüedad y variabilidad:](#_heading=h.beyrh5dh0ibw)

[4. Exploración de patrones lingüísticos:](#_heading=h.40dvk6tt9235)

[4. Bibliografía](#_heading=h.21bo6563hf1s)

## Análisis descriptivo de manifiestos validados (Input)

**¿Hay valores faltantes en el conjunto de datos? ¿Se pueden identificar patrones de ausencia?**



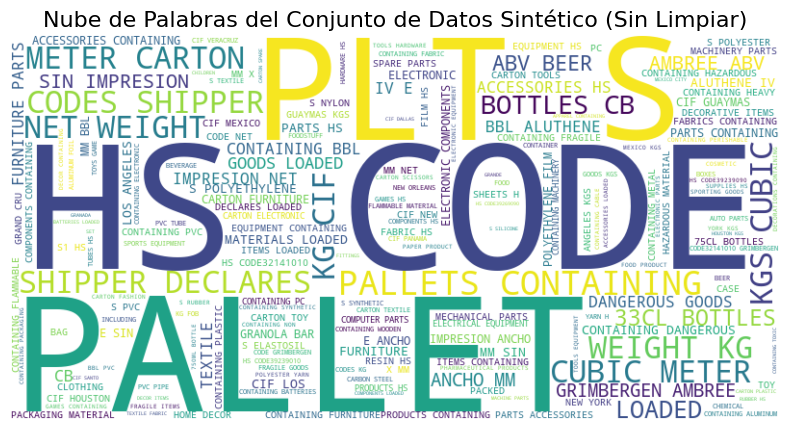
Al analizar el **dataset de validación**, notamos que hay algunos capítulos (primeros 2 dígitos) que no se encuentran, sin embargo esto no supone ningún tipo de anomalía. Para esta aplicación, se pudiera considerar como “dato faltante” alguna descripción muy corta, al analizar el conjunto de datos masivo que necesitaremos predecir, éste es el TOP 5 de descripciones más cortas encontradas:

| **ID** | **Descripción** | **Cantidad de caracteres** |
| --- | --- | --- |
| 3361 | 12,000 CARTON PACKING PEANUTS. | 31 |
| 5321 | IN 30 PALLETS OF TOYS AND GAMES. | 33 |
| 6015 | IN 25 PALLETS OF TOYS AND GAMES. | 33 |
| 6374 | IN 20 PALLETS OF FERTILIZERS. | 30 |
| 11376 | GRIPON H20 HS-CODE38189010. | 27 |

**¿Cuáles son las estadísticas resumidas del conjunto de datos?**

Debido a la naturaleza de los datos de entrada (códigos HS), no es posible aplicar estadísticas descriptivas tradicionales. Sin embargo, se pueden realizar análisis alternativos que permitan resaltar características clave de los datos. A continuación, se describen algunos de los enfoques utilizados para este análisis:

1. Nube de Palabras del Conjunto de Datos Sintético: Se generó una nube de palabras a partir de los datos sintéticos del manifiesto, creados por el equipo. Esta visualización permite identificar patrones y posibles sesgos en los datos originales. Al observar las palabras más frecuentes, es posible detectar términos dominantes que podrían influir en la calidad de los modelos de clasificación.



Se puede observar que debido a que no se hace un pre-procesamiento, las palabras más comunes son relacionadas al contexto de la descripción de la carga y no a la descripción como tal, podemos observar palabras como “HS Code”, “Pallets”, “Weight”, etc.

1. Nube de Palabras Tras Limpieza de Datos: Posteriormente, se realizó una limpieza de datos eliminando palabras de relleno (stop words) y otros términos irrelevantes como por ejemplo “'hs', 'code', 'pallets', 'plts', 'shipper', 'declares', 'containing',etc La nueva nube de palabras muestra una representación más precisa de los términos significativos, ayudando a mejorar la interpretación y uso de los datos para el modelo.



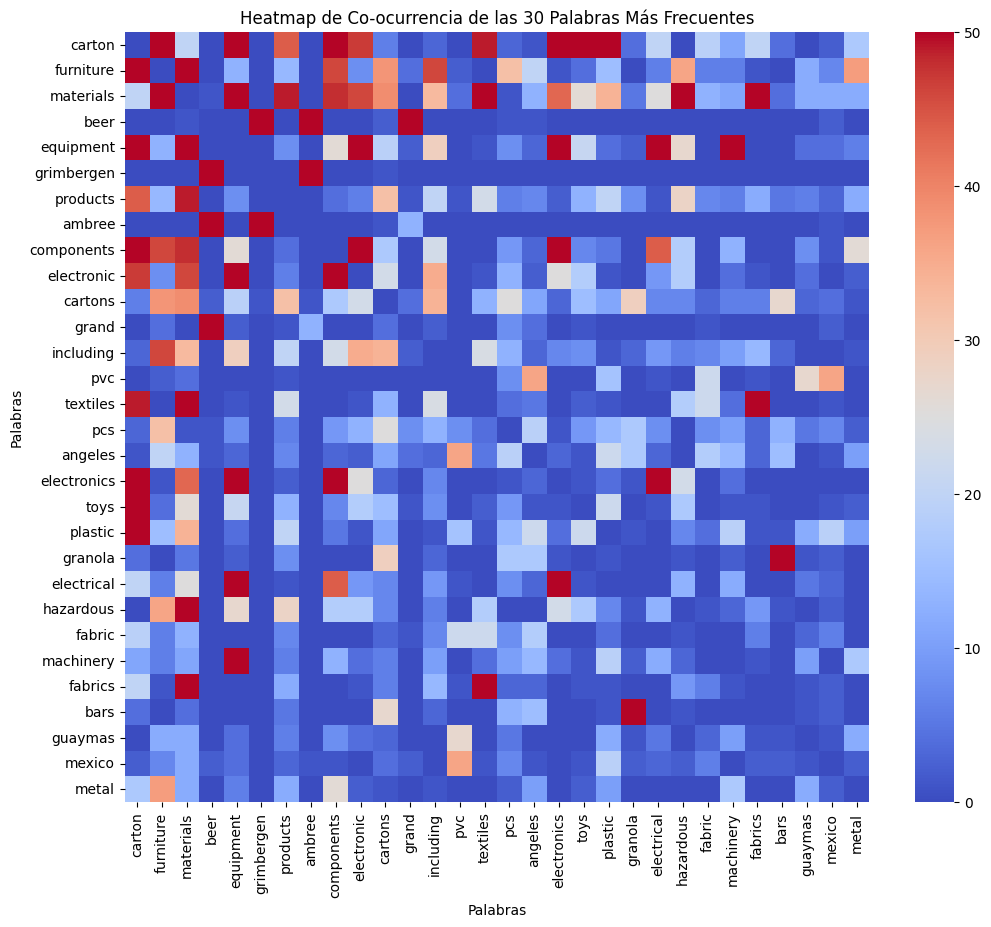
La nube de palabras generada tras la limpieza de datos revela las principales características y patrones presentes en los manifiestos de carga. A diferencia de la versión sin limpiar, en esta iteración se eliminan términos irrelevantes y repetitivos como "HS code," "declaraciones," y otros relacionados con el formato, permitiendo un enfoque más claro en los contenidos significativos.

Observaciones clave:

1. **Productos Destacados:** Palabras como "furniture," "textile," "beer," y "machinery" son prominentes, lo que sugiere que estos son algunos de los productos más comunes o relevantes en la carga. Esto indica una concentración de mercancías relacionadas con muebles, textiles, bebidas alcohólicas, y maquinaria, lo cual puede influir en las dinámicas de comercio en términos de aduanas y regulaciones.
2. **Materiales y Componentes:** Términos como "material," "plastic," "fabric," y "electronic" sugieren una alta presencia de bienes manufacturados y componentes industriales. Esto refleja la diversidad y complejidad de las descripciones de carga, abarcando desde artículos de consumo hasta insumos para la producción.
3. **Marcas y Nombres Específicos:** Se observan términos como "Grimbergen" y "Ambree," que parecen ser nombres de marcas o tipos de productos específicos (posiblemente cervezas). Esto puede reflejar la inclusión de bienes de consumo altamente identificables que son parte significativa de las exportaciones/importaciones.
4. **Diversidad Geográfica y Usos:** La presencia de términos como "machinery," "chemical," y "polyethylene" indica una amplia gama de productos, desde químicos y plásticos hasta equipos industriales. Esto podría sugerir que el conjunto de datos incluye bienes destinados a diferentes industrias y propósitos, lo cual es relevante para entender la composición y la complejidad de la carga.

En resumen, esta limpieza de datos nos ha permitido destacar los elementos realmente valiosos y representativos del conjunto de datos, evidenciando la diversidad en los tipos de productos y materiales que se transportan. Estos hallazgos pueden ayudar a enfocar el análisis en áreas clave y mejorar la precisión de modelos predictivos en el contexto de clasificación de carga.

1. Mapa de Calor de Distribución de Palabras: Se llevó a cabo un análisis detallado de co-ocurrencias de palabras mediante la creación de un heatmap, con el objetivo de visualizar las relaciones entre términos en las descripciones de carga. Este análisis permitió identificar patrones y agrupaciones de palabras que son relevantes para la clasificación de los códigos del Sistema Armonizado (HS).



Se hacen las siguientes observaciones clave:

Alta correlación entre términos relacionados:

* "Carton" y "furniture" muestran una fuerte co-ocurrencia, sugiriendo que podría haber muchos muebles de cartón o embalajes de cartón para muebles.
* "Electronic" y "components" también presentan una alta correlación, lo que indica la presencia frecuente de componentes electrónicos.

Clusters temáticos:

* Se observa un cluster alrededor de "materials", "equipment", y "products", sugiriendo que estos términos a menudo aparecen juntos en el dataset.
* Otro cluster notable incluye "plastic" y "electrical", posiblemente relacionado con productos o materiales eléctricos de plástico.

Patrones diagonales:

* La diagonal principal muestra valores altos, lo cual es esperado ya que representa la co-ocurrencia de cada palabra consigo misma, es decir, difícilmente se encuentran con palabras repetidas en el dataset.

Asimetrías interesantes:

* Algunas palabras muestran patrones de co-ocurrencia asimétricos. Por ejemplo, "hazardous" parece co-ocurrir más fuertemente con "materials" que viceversa.

Términos centrales:

* Palabras como "materials", "equipment", y "products" parecen tener una co-ocurrencia moderada a alta con muchas otras palabras, sugiriendo que podrían ser términos centrales o generales en el contexto del dataset.

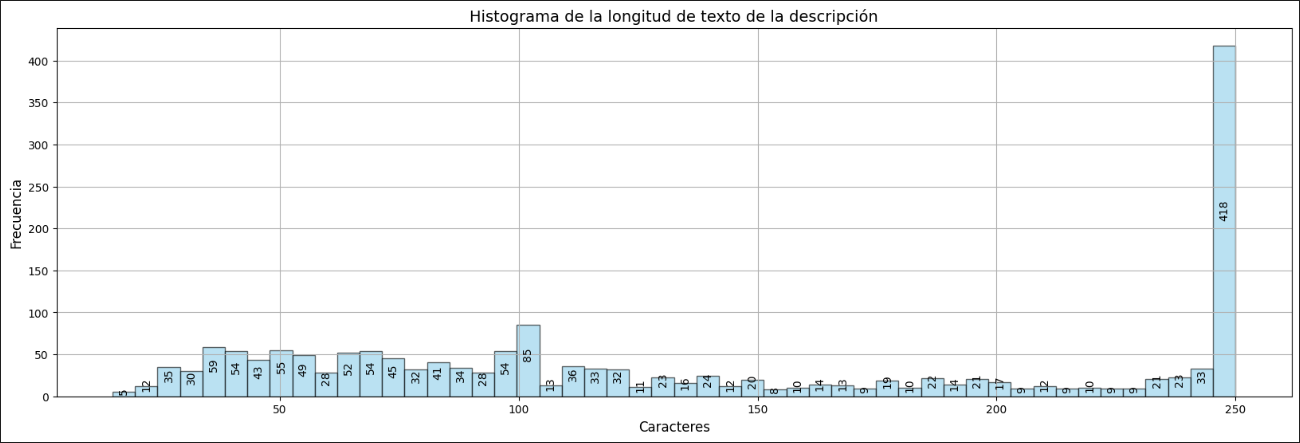
Posibles categorías de productos:

* Se pueden inferir posibles categorías de productos como muebles, electrónicos, textiles y plásticos basándose en los patrones de co-ocurrencia.

Términos geográficos:

* "México" y "guaymas" aparecen en el dataset, pero con baja co-ocurrencia general, lo que podría indicar una presencia limitada de referencias geográficas específicas.

**¿Hay valores atípicos en el conjunto de datos (Dataset de validacion)?**



**Forma de la distribución:**

La distribución presenta una asimetría moderada. Observamos una tendencia hacia valores más bajos en cuanto al conteo de caracteres, con una disminución gradual en la frecuencia a medida que los conteos de caracteres aumentan.

**Concentración de frecuencias:**

La mayoría de los datos están concentrados en un rango más bajo de conteo de caracteres. Hay pocos casos con conteos de caracteres más altos, lo que genera una "cola" que se extiende hacia la derecha, típica de una distribución sesgada positivamente.

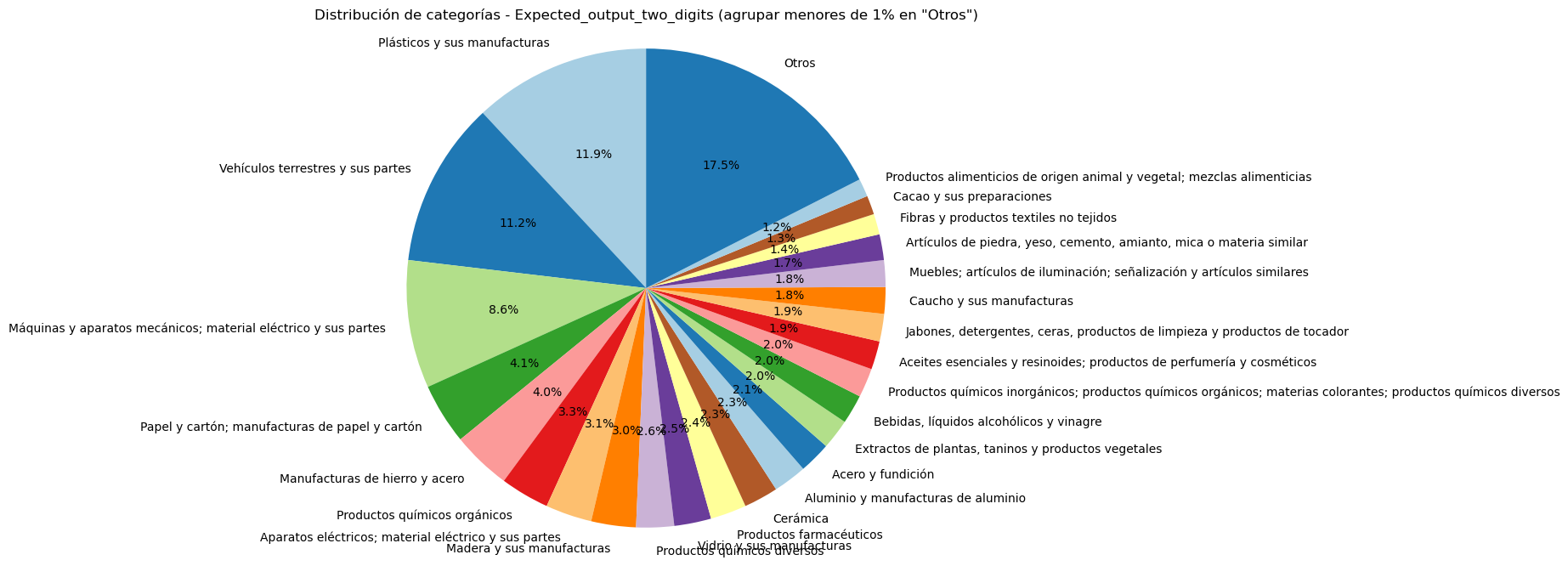
**Posibles valores atípicos:**

**Aunque no hay picos dramáticos que sugieran outliers extremos**, el histograma tiene algunas barras más pequeñas en los rangos de conteos más altos. Esto podría indicar una distribución larga, con algunos casos que tienen más caracteres de lo habitual.

**Simetría:**

La distribución no es perfectamente simétrica; tiene una inclinación hacia la derecha, lo que sugiere que un pequeño porcentaje de las entradas tiene un mayor número de caracteres en comparación con la mayoría.

**¿Cómo se distribuyen los datos en función de diferentes categorías?**r

* + Pie chart genérico a nivel capítulo - de la muestra del manifiesto
  + 
* ¿Se deberían normalizar el texto para visualizarlo mejor? (cambiamos la pregunta texto)
  + Convertimos el texto a minúsculas, eliminamos puntuación y números, y eliminamos palabras vacías (stopwords).

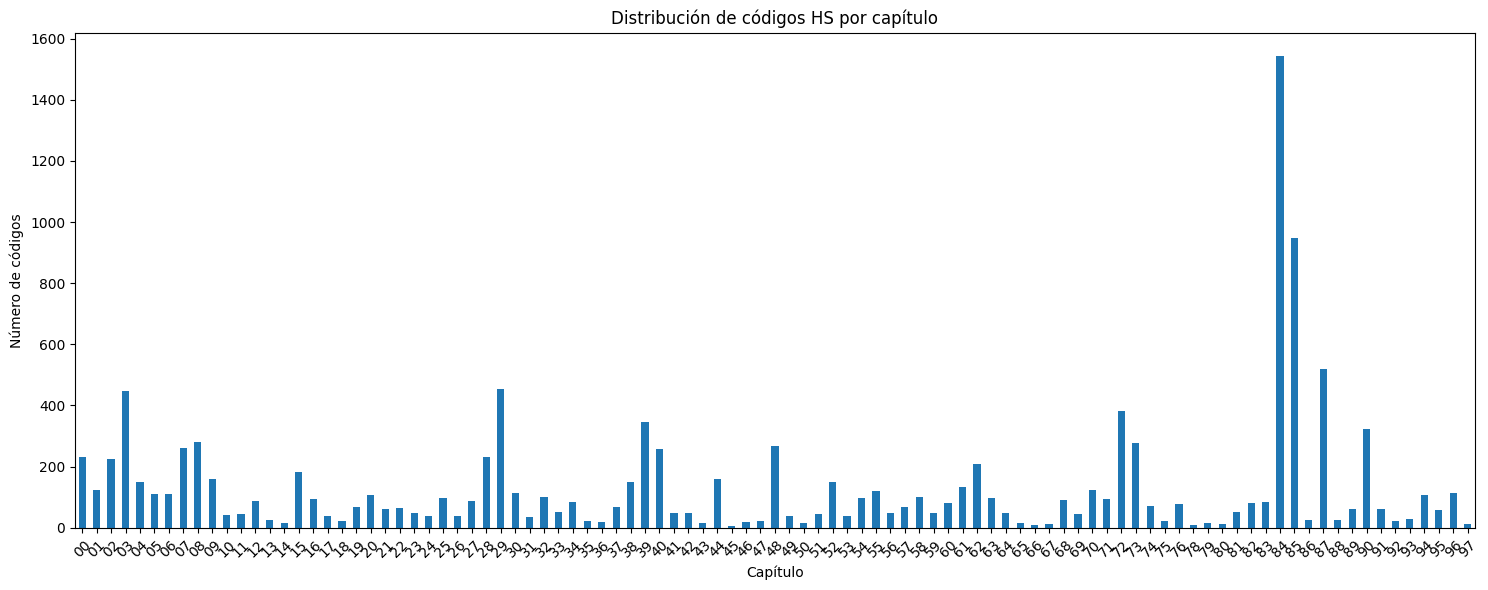
## Análisis descriptivo del diccionario (Objetivo)

**¿Hay desequilibrio en las clases de la variable objetivo?**

Para poder responder esta pregunta se plantearon 2 análisis:

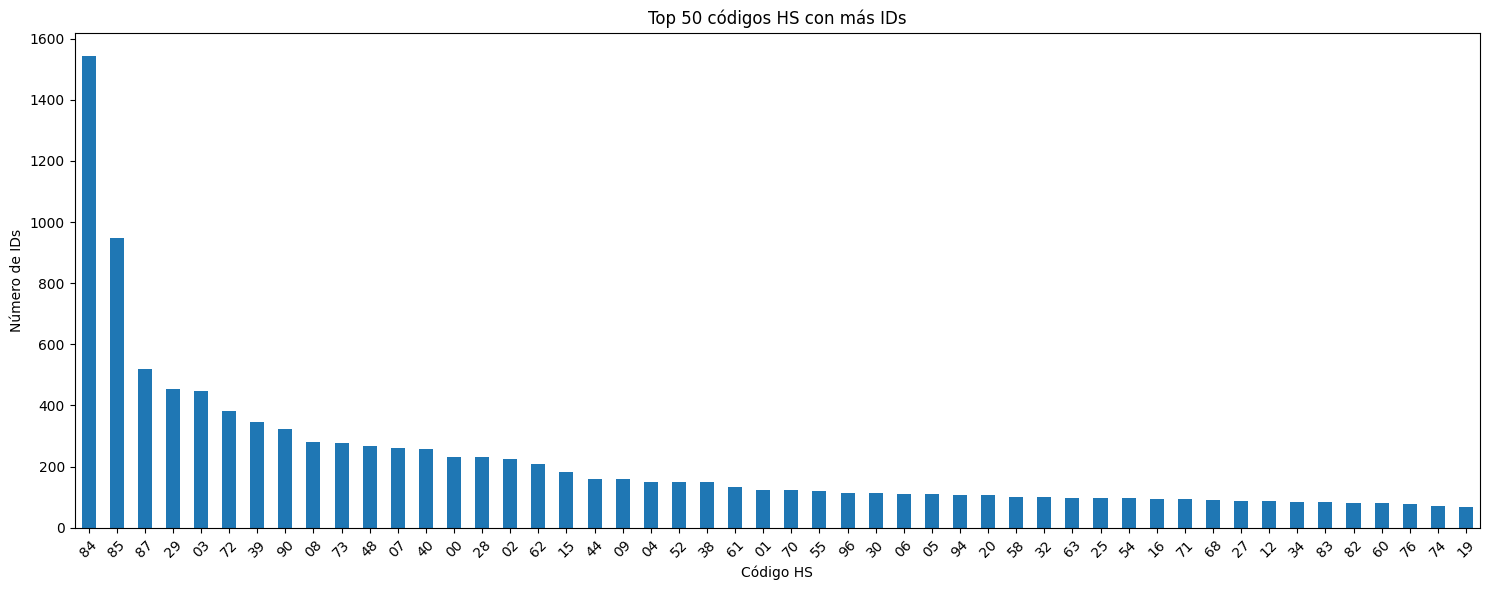
**Distribución de Códigos armonizados por Capítulo**

Se realizó un análisis de la distribución de códigos HS según los primeros dos dígitos, representando el capítulo al que pertenecen. Este gráfico de barras muestra la cantidad de códigos que corresponden a cada capítulo, permitiendo identificar cuáles son los capítulos más comunes en el catálogo. Esta información es crucial para comprender cómo se distribuyen los productos en diferentes categorías y para facilitar el análisis de tendencias en el comercio internacional.



**Top 50 Capítulos con Más Códigos**

Además, se identificaron los 50 capítulos con mayor cantidad de códigos HS. Este análisis proporciona información valiosa sobre cuáles son las categorías más representadas en el catálogo. Al visualizar la cantidad de códigos en cada capítulo, es posible entender mejor la estructura del comercio y la variedad de productos que se manejan en el sistema de clasificación HS. Este gráfico permite a los analistas y a los tomadores de decisiones identificar áreas de interés y potenciales enfoques para optimizar la gestión de la clasificación de mercancías.



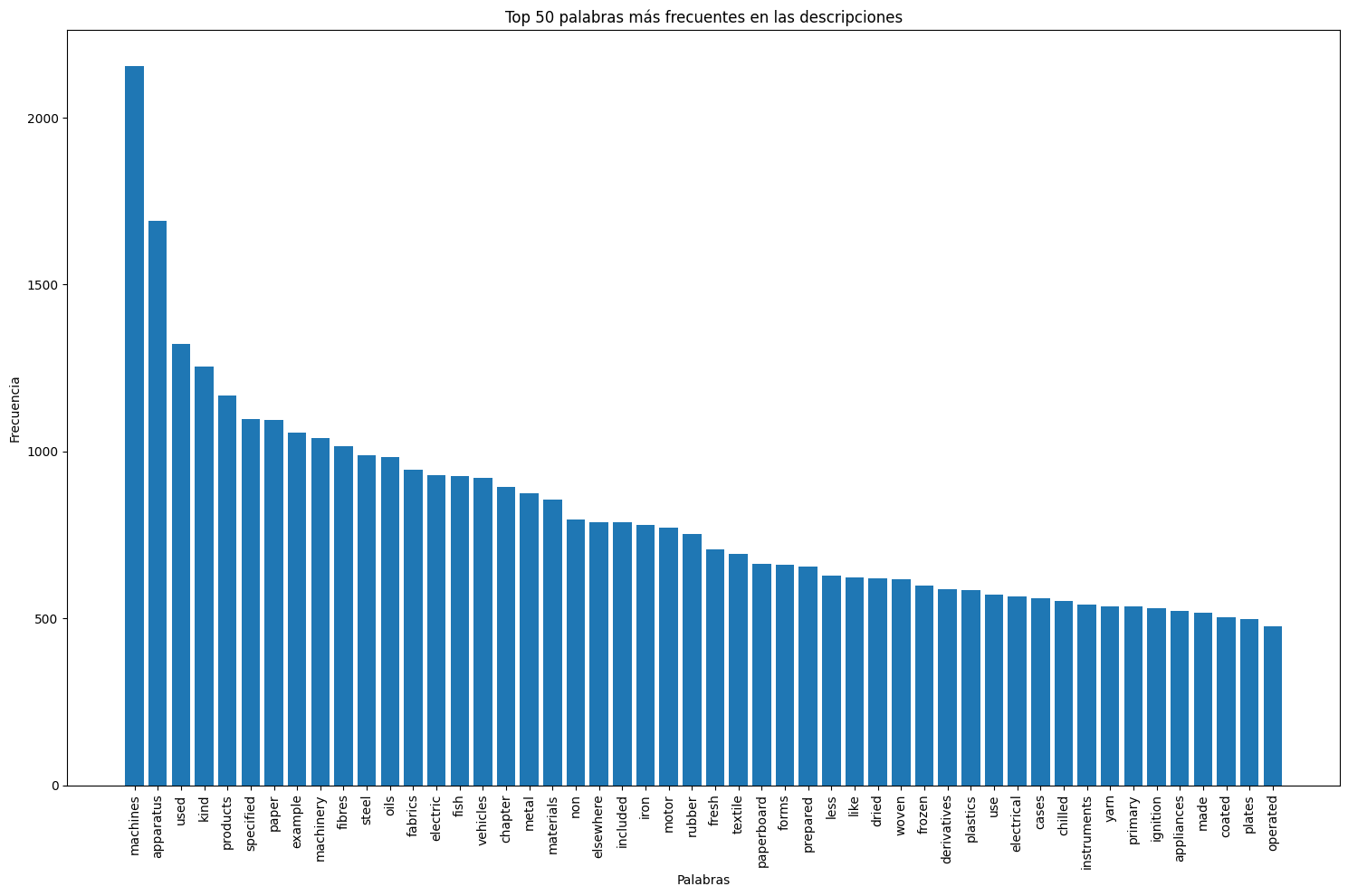
Se puede observar que sí existe un ligero desbalance con ciertos capítulos, El top 10 de capítulos son:

* 84 Máquinas y aparatos mecánicos
* 85 Máquinas, aparatos y materiales eléctricos
* 87 Vehículos automóviles y sus partes y accesorios
* 29 Productos químicos orgánicos
* 03 Pescados y crustáceos
* 72 Acero y fundición
* 39 Plásticos y sus manufacturas
* 90 Instrumentos y aparatos ópticos; instrumentos y aparatos de medida; instrumentos médicos
* 08 Frutas y frutos comestibles, cáscaras de cítricos y melones
* 73 Manufacturas de hierro y acero

Además del análisis anterior, se analizará si hay desbalance en el catálogo de códigos armonizados en cuestión de las palabras que aparecen con más frecuencia en el catálogo:

**Top 50 Palabras Más Frecuentes en las Descripciones**

Para el análisis de catálogo de HS Code, se llevó a cabo un examen de las descripciones de los códigos, donde se identificaron las 50 palabras más frecuentes. Este gráfico de barras ilustra cómo ciertas palabras aparecen con mayor frecuencia en las descripciones, proporcionando una visión sobre los términos y conceptos predominantes utilizados en la categorización de los productos. La limpieza de texto y la eliminación de palabras irrelevantes (stopwords) permitieron resaltar las palabras que realmente aportan significado a las descripciones de los códigos.

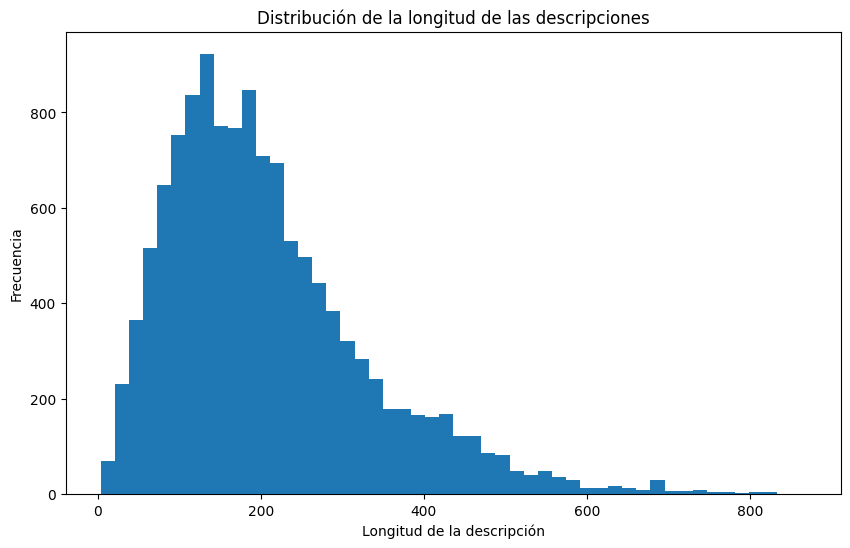


Se observa que palabras relacionadas a los capítulos con mayor número de subpartidas tienden a aparecer con más frecuencia, como los son máquinas, materiales, aparatos, etc.

Por último se hizo un análisis respecto a la longitud de las descripciones en el diccionario.

**Distribución de la Longitud de las Descripciones**

Se llevó a cabo un análisis de la longitud de las descripciones de los códigos HS para entender la variabilidad en la extensión de la información proporcionada. El histograma resultante revela cómo se distribuyen las longitudes de las descripciones, lo que puede influir en la claridad y precisión de la identificación de productos. Adicionalmente, se calcularon estadísticas descriptivas como la longitud promedio, mínima y máxima de las descripciones, lo que proporciona un contexto adicional sobre la extensión del contenido en el catálogo.



Se observan las siguientes estadísticas:

Longitud promedio de descripción: 208.34

Longitud mínima de descripción: 4

Longitud máxima de descripción: 868

Podemos observar una distribución relativamente normal cuyo promedio es una descripción de 208 caracteres.

**Descripción general de la forma y los tipos de datos.**

Un análisis descriptivo previo para un proyecto de **NLP (Natural Language Processing)** no es igual que para un proyecto de **Machine Learning tradicional** debido a las diferencias fundamentales en la naturaleza de los datos y las técnicas de procesamiento. Las principales diferencias que se pueden comentar son:

### 1. Naturaleza de los datos:

* En **Machine Learning tradicional**, los datos generalmente están estructurados en forma de tablas con columnas que representan variables numéricas o categóricas. Un análisis descriptivo típico puede incluir estadísticas como media, mediana, desviación estándar, correlación, etc., junto con gráficos como histogramas o diagramas de dispersión.
* En **NLP**, los datos son principalmente texto no estructurado. El análisis descriptivo previo implica evaluar el contenido lingüístico, la frecuencia de palabras, y la estructura del lenguaje, lo cual requiere técnicas como análisis de frecuencia de términos (TF-IDF), análisis de n-gramas, distribución de longitudes de texto, y tal vez procesamiento de sentimiento o entidades nombradas.

### 2. Transformaciones necesarias:

* En **ML tradicional**, el análisis descriptivo se realiza directamente sobre las características del conjunto de datos, pero en NLP, el texto necesita ser transformado primero a una representación numérica (tokens, embeddings, bag of words) para luego aplicar algoritmos. Esto implica un proceso de **preprocesamiento lingüístico** (como tokenización, eliminación de stop words, lematización o stemming).

### 3. Ambigüedad y variabilidad:

* En los **proyectos de Machine Learning tradicionales**, las características suelen estar bien definidas y pueden ser fácilmente cuantificadas. Sin embargo, en NLP, el texto es inherentemente ambiguo y su interpretación depende de muchos factores contextuales como la sintaxis y semántica. El análisis descriptivo inicial en NLP debe lidiar con esta ambigüedad, lo que puede requerir técnicas avanzadas para capturar significado latente (como Word2Vec o BERT).

### 4. Exploración de patrones lingüísticos:

* En **NLP**, los patrones que son útiles no siempre son evidentes a través de técnicas descriptivas tradicionales. Se suelen usar métodos como el análisis de temas (LDA) para descubrir patrones ocultos o estructuras latentes en los textos. En ML tradicional, los patrones suelen estar más relacionados con la correlación entre las variables o su comportamiento ante las etiquetas de predicción.

**Conclusión**

Tras un exhaustivo análisis exploratorio de datos de manifiestos de carga y códigos armonizados, se llegó a las siguientes conclusiones:

Se descubrió que la calidad general de los datos es satisfactoria, sin valores faltantes significativos. No obstante, se identificaron algunas descripciones excepcionalmente cortas que podrían presentar desafíos en la clasificación.

En el análisis de las características textuales, se observó que las descripciones de carga contienen una mezcla de información relevante e irrelevante. Mediante la aplicación de técnicas de limpieza y normalización, se logró identificar palabras clave significativas relacionadas con productos como muebles, textiles, cerveza y maquinaria. Adicionalmente, se revelaron patrones interesantes de co-ocurrencia entre palabras, proporcionando información valiosa sobre las relaciones entre distintos tipos de productos y materiales.

Al examinar la distribución de los datos, se detectó un ligero desequilibrio en la distribución de códigos armonizados por capítulo. Se encontró que los capítulos más representados incluyen maquinaria, productos químicos y equipos eléctricos. En cuanto a la longitud de las descripciones, se determinó que siguen una distribución aproximadamente normal, con un promedio de 208 caracteres.

Se identificaron varios desafíos potenciales para el proyecto. La presencia de términos irrelevantes y la variabilidad en la longitud de las descripciones podrían complicar el proceso de modelado. Además, se notó que el ligero desequilibrio en las clases de la variable objetivo podría afectar el rendimiento del modelo si no se aborda adecuadamente en las etapas posteriores.

Por otra parte, también se reconocieron oportunidades significativas. La identificación de palabras clave y patrones de co-ocurrencia ofrece un camino prometedor para mejorar la clasificación de productos. Asimismo, se comprobó que las técnicas de normalización y limpieza del texto son efectivas para resaltar la información relevante en las descripciones.

En conclusión, el análisis exploratorio de datos ha proporcionado una comprensión profunda de la estructura y características del conjunto de datos. Estos hallazgos serán fundamentales para informar las decisiones en las etapas de preprocesamiento y modelado. Se considera que, al abordar los desafíos identificados y aprovechar las oportunidades descubiertas, será posible desarrollar un sistema de clasificación de códigos armonizados más preciso y eficiente.

## 4. Bibliografía

Guo, Y., Şener, M. F., Cairns, S., & Duval, Y. (2021). Improving the effectiveness of the harmonized system: A big data analytics approach. Journal of International Commerce and Economics, 1-22.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. Advances in neural information processing systems, 26.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Rogers, A., Kovaleva, O., & Rumshisky, A. (2020). A primer in BERTology: What we know about how BERT works. Transactions of the Association for Computational Linguistics, 8, 842-866.