**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)

[Descripción de algoritmo utilizado como baseline](#_heading=h.xhep8orgpmfj)

[¿Cuál es la métrica adecuada para este problema de negocio?](#_heading=h.tx7oyewvzx15)

[¿Cuál debería ser el desempeño mínimo a obtener?](#_heading=h.84zuticc6o1i)

[Resultados obtenidos con baseline](#_heading=h.i3z0nteionzm)

[Conclusión](#_heading=h.9btp7rs39z2q)

[Referencias](#_heading=h.ud6c1uqe1vfv)

[(N.d.). Retrieved from https://www.dripcapital.com/hts-code/section/all](#_heading=h.yxuj5vdccito)

Nota: Para este proyecto las siguientes preguntas no aplican:

* ¿Se puede determinar la importancia de las características para el modelo generado?

*En los modelos de deep learning basados en embeddings, como Word2Vec, BERT o RoBERTa, la relación entre las características y la importancia es más compleja porque el modelo no utiliza características explícitas como palabras aisladas, sino que transforma las palabras en vectores de alta dimensión. Estos vectores encapsulan la semántica y el contexto de las palabras, lo que dificulta la interpretación directa de la importancia de características.*

* ¿El modelo está sub/sobreajustando los datos de entrenamiento?

***No es un modelo supervisado****, por lo que no tiene un objetivo explícito de generalización.*

***No hay una salida directa o una predicción*** *que pueda evaluarse en términos de ajuste en los datos de entrenamiento o prueba.*

***La evaluación se hace de forma indirecta*** *a través de tareas downstream, no en términos de sobreajuste en el corpus original.*

## Descripción de algoritmo utilizado como baseline

El algoritmo utilizado como baseline se centra en la limpieza de datos, la búsqueda de tokens y el conteo de coincidencias para determinar la asignación correcta de códigos HS. Este proceso implica varios pasos clave:

* Limpieza de Datos: Se inicia con la limpieza de las descripciones de mercancías, asegurando que los datos sean consistentes y estén libres de errores que puedan afectar la clasificación.
* Búsqueda de Tokens: A continuación, el algoritmo realiza una búsqueda de tokens en las descripciones, extrayendo términos relevantes que se utilizarán para la clasificación. Esto permite identificar palabras clave que son indicativas del código HS correspondiente.
* Conteo de Coincidencias: El algoritmo cuenta las coincidencias de tokens entre las descripciones de mercancías y los códigos HS en el catálogo. Se busca la coincidencia con la mayor cantidad de tokens, lo que aumenta la precisión de la clasificación.
* Evaluación de Desempeño: Para obtener métricas de desempeño confiables, se utiliza el enfoque de Monte Carlo. Este método permite ejecutar múltiples iteraciones para evaluar la robustez del algoritmo y su precisión en la asignación de códigos HS. Al final de estas iteraciones, se calculan métricas de precisión que sirven como indicadores del rendimiento del modelo.

## ¿Cuál es la métrica adecuada para este problema de negocio?

Para los agentes aduanales, la métrica más relevante es la precisión a 4 dígitos del código HS. Este nivel de detalle es fundamental, ya que permite clasificar las mercancías de manera más específica, facilitando así el cumplimiento de regulaciones aduaneras y optimizando los procesos de despacho. Conocer el código a 4 dígitos permite a los agentes determinar de manera precisa los aranceles aplicables, así como cumplir con las normativas internacionales de comercio.

Sin embargo, en ciertas circunstancias, la precisión a 2 dígitos puede ser suficiente para fines arancelarios generales. Por ejemplo, cuando se trata de mercancías de categorías amplias donde los aranceles son homogéneos, un código de 2 dígitos puede ser adecuado para la clasificación y el cálculo de impuestos. Esto puede simplificar el proceso aduanal, reduciendo el tiempo y los costos asociados con la verificación de mercancías.

Dado esto, es esencial redefinir las métricas de precisión en función de estos dos niveles de detalle. Se deben establecer criterios de éxito que abarquen tanto la precisión a 2 dígitos como la precisión a 4 dígitos, permitiendo a los agentes evaluar la eficacia del modelo en contextos diferentes. La medición de ambas precisiones ayudará a determinar la capacidad del algoritmo para satisfacer las necesidades específicas de los agentes aduanales, garantizando así que la clasificación sea precisa y eficiente.

## ¿Cuál debería ser el desempeño mínimo a obtener?

El desempeño mínimo a obtener se establece comparando el modelo con una asignación aleatoria de códigos HS. En 2024, existen 97 categorías a 2 dígitos y 1,244 categorías a 4 dígitos.

Si el modelo realiza asignaciones de manera aleatoria, la precisión esperada sería de aproximadamente 1.03% para la clasificación a 2 dígitos y de aproximadamente 0.08% para la clasificación a 4 dígitos. Por lo tanto, se sugiere que el modelo propuesto busque superar estas cifras de precisión aleatoria. Se recomienda un objetivo de al menos un 5% de precisión a 2 dígitos y un 1% a 4 dígitos, lo que demostraría el valor y la efectividad del modelo en comparación con un enfoque aleatorio.

## Resultados obtenidos con baseline

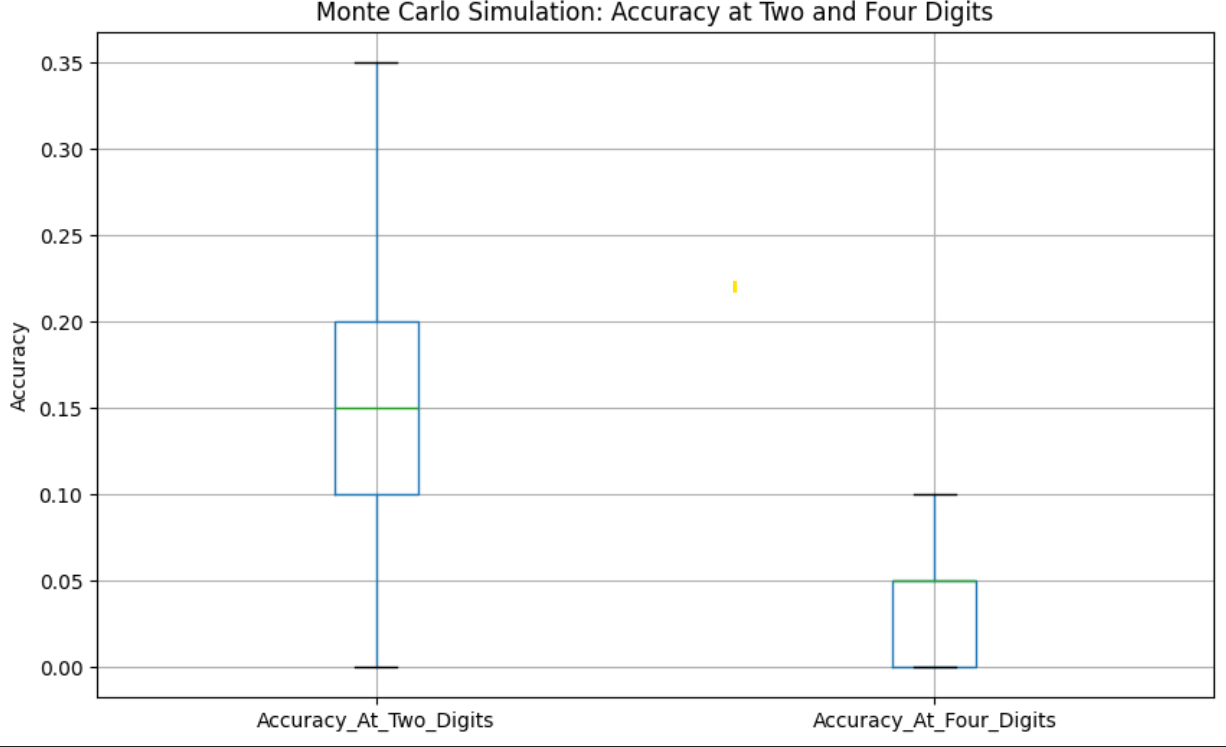


Figura 1.1: Resultados de la simulación de Montecarlo para la evaluación del algoritmo simple de búsqueda

La gráfica de la simulación de Monte Carlo presenta los resultados de precisión obtenidos al clasificar códigos HS a dos y cuatro dígitos. Se observa que el promedio de precisión para la clasificación a dos dígitos es de aproximadamente 0.15, mientras que para la clasificación a cuatro dígitos es significativamente menor, en torno a 0.05. Estos resultados indican que, si bien el algoritmo es capaz de realizar clasificaciones, la precisión es notablemente superior en el caso de los códigos de dos dígitos. La representación gráfica en forma de caja de bigotes ilustra la variabilidad de los resultados a través de las diferentes iteraciones, destacando que, aunque existen picos de rendimiento, la capacidad del modelo para mantener una precisión constante es limitada.

## Conclusión

En conclusión, los resultados obtenidos en la simulación de Monte Carlo superan las expectativas de un enfoque aleatorio, lo que sugiere que el algoritmo tiene una capacidad básica para clasificar los códigos HS. Sin embargo, la búsqueda directa de tokens tiene limitaciones significativas, ya que no logra capturar adecuadamente el contexto de términos que pueden no coincidir literalmente con las entradas del diccionario, como "chango" y "mono" a pesar de que en el diccionario se les clasifica como “primates”. Además, el algoritmo enfrenta dificultades ante errores ortográficos o datos ruidosos. En este sentido, un modelo de inteligencia artificial podría ofrecer mejoras considerables al incorporar un análisis contextual y semántico de los términos, lo que permitiría una clasificación más precisa y robusta de las descripciones de carga.

## Referencias

Inmon, B. (2017). Turning text into gold: Taxonomies and textual analytics. Bradley Beach: Technics Publications

Manning, C., Raghavan, P., & Schütze, H. (2008). Introduction to information retrieval. Cambridge: Cambridge University Press.

Wilbur, W. J., & Sirotkin, K. (1992). The automatic identification of stop words. Journal of Information Science, 18(1), 45-55

#### (N.d.). Retrieved from <https://www.dripcapital.com/hts-code/section/all>