Taller de Tesis I – Entrega II Año 2025

Data Mining - UBA

Alumno: Santiago Tedoldi

Contexto y Objetivo

El tema elegido para cumplir con las consignas del Taller de Tesis I se relaciona con la clasificación arancelaria/tarifario de mercaderías, partiendo de descripciones comerciales y según la nomenclatura del Sistema Armonizado (SA) internacional. La finalidad es entrenar y evaluar un modelo de clasificación de texto, considerando los desafíos propios de un dataset presuntamente acotado -para la cantidad de clases objetivo- y desbalanceado.

Descripción del Conjunto de Datos

Contenido y Origen:
El dataset está formado por tuplas con información de mercaderías sometidas al comercio internacional (descripción_comercial, código_tarifario) en idioma inglés, en un total de 500 mil ítems.

El origen de estos datos no es preciso. Fue material de un curso suministrado en una capacitación internacional en Corea del Sur, en el marco del Programa BACUDA de la Organización Mundial de Aduanas.

Características:

- **Muestra Variada:** El conjunto de datos presenta una diversidad considerable en cuanto a tipos de productos y mercaderías, clasificadas en ~1200 códigos a 4 dígitos y ~5600 códigos a 6 dígitos.
- **Desbalanceo:** Predominan las descripciones de productos relacionados con máquinas y material eléctrico/electrónico, como también de vehículos automotores.

Preguntas y Retos del Proyecto

Con estos datos y la problemática planteada, surgen preguntas que podrías abordarse en el siguiente orden:

> **¿Es posible entrenar un modelo de clasificación arancelaria basado en descripciones comerciales, incluso con una muestra acotada?**

Técnicas y Enfoques Propuestos

Para enfrentar los desafíos planteados, las potenciales técnicas identificadas son:

- **Encoders de texto y Clasificación:** Se propone el uso de redes neuronales usando modelos “tradicionales” -Doc2Vec, FastText- y con arquitectura transformer pre-entrenados -BERT o similar-.
- **PCA y Clústering:** Cómo método analítico y exploratorio, para reducir la

dimensionalidad del conjunto de datos, facilitar el manejo de las características extraídas e identificar patrones de los textos.

Metodología

EDA

En la problemática a trabajar sobre clasificación arancelaria/tarifaria de mercaderías del comercio internacional, el dataset disponible es particularmente sencillo, partiendo de solo dos variables:

- HS06_code: código tarifario armonizado a 6 dígitos (Categórica).
- goods_desc: texto libre para describir comercialmente la mercadería (Texto).

Por esta razón, para la etapa exploratoria, además de evaluar temas de calidad (nulos, duplicados, outliers) y distribución/frecuencias, corresponde trabajar con técnicas de **ingeniería de variables**.

En términos generales, el análisis exploratorio busca entender cómo se compone el dataset, que mercaderías abarca, que información contiene y que limitaciones tenemos que considerar.

Ingenería de variables

Para mejorar el análisis del dataset, y contemplar toda la información "escondida" en el mismo, se procede a:

- Desagregar la variable HS06 en otros códigos subyacentes.
- Buscar fuentes externas que complementen a los datos disponibles.
- Aplicar técnicas de encoding de texto, buscando representaciones latentes.
- Reducir dimensiones (PCA, UMAP y/o t-SNE).

Técnicas no-supervisadas

Las técnicas que no requieren el uso del resultado o variable target pueden presentar una solución simplificada a problemáticas de clasificación. Sin embargo, en este caso parece difícil que esto se cumpla.

La variadad de mercaderías, con aprox. 1200 códigos posibles, y la baja calidad que suelen acarrear campos de texto libre, parece que la solución requerirá entrenamiento y ajuste de los encoders mediante el uso de técnicas supervisadas.

De todas formas, se implementa una técnica no-supervisada, haciendo uso de un encoder de texto (BERT, distiltBERT, o similar) pre-entrenado, luego aplicando reducción de dimensiones, k-means esférico (con distancia del coseno) y clustering jerárquico.

Definición del target/ variable objetivo

La desagregación del código tarifario HS06 tiene fundamento de negocio y permite considerar cambios en la variable objetivo, de cara a técnicas supervisadas. El sistema armonizo de

clasificación de mercaderías, a nivel internacional, se componde de:

- Secciones: No se reflejan en el código tarifario, pero tienen un papel normativo relevante. Ejemplo: Sección XVI refiere a las máquinas, aparatos y equipamiento eléctrico/electrónico y comprende los capítulos 84 y 85.

- Capítulos (HS02): Primeros dos dígitos del código y agrupa mercaderías según su tipo o naturaleza. Ejemplo: Capítulo 84 contiene los reactores nucleares, calderas, maquinaria y aparatos mecánicos.

- Partidas (HS04): Primeros cuatro dígitos del código y define a mercaderías relevantes para el comercio internacional. Ejemplo: Partida 8471 define a las máquinas automáticas para tratamiento o procesamiento de datos (computadoras).

- Subpartidas (HS06): Códigos tarifarios a 6 dígitos, con una "apertura" de la partida, permitiendo definir mercaderías más específicas. Ejemplo: 847130 define a Máquinas automáticas para tratamiento o procesamiento de datos, portátiles, de peso inferior o igual a 10 kg, que estén constituidas, al menos, por una unidad central de proceso, un teclado y un visualizador.

Lo descripto, permitiría ejercitar la clasificación a 2, 4 o 6 dígitos. A menor cantidad de dígitos, menos cantidad de clases, lo que también simplifica la tarea.

Sin embargo, en el comercio internacional, una clasificación debe tener tantos dígitos como sea necesario. En el MERCOSUR, las mercaderías se comercializan con 8 dígitos (se agregan códigos regionales) y en Argentina la nomenclatura moderna usa 11 dígitos tarifarios, que definen aranceles y tributos, certificaciones, prohibiciones y cupos, entre otras cuestiones de relevancia para el comercio legítimo.

Técnicas supervisadas y transferencia del aprendizaje

Para abordar la pregunta, sobre si es posible entrenar un modelo para clasificación arancelaria, cubriendo las técnicas de explotación de datos y aprendizaje automático disponibles, debemos cubrir técnicas de aprendizaje supervisado.

Algunas técnicas "tradicionales" de explotación de texto (Doc2Vec o FastText) utilizan redes neuronales para generar representaciones latentes y luego clasificar usando medidas de similitud. Estos modelos deben entrenarse desde cero, con el corpus disponible.

Además, es posible partir de modelos con arquitectura Transformer (BERT, distiltBERT, o similar), pre-entrenados con corpus de texto extensos. Esto puede usarse como encoder (codificador o cuerpo) del modelo, mientras que se desarrolla y entrena un classifier (clasificador o cabeza) que se encarga de la clasificación final.

La arquitectura encoder-classifier se utiliza mediante:

- Transfer learning (transferencia del aprendizaje), ajustando durante el entrenamiento solo al classifier.

- Fine-tuning (ajuste fino) total o parcial, ajustando el conjunto cabeza y encoder (en todos sus capas -total- o en algunas de sus capas de salidad -parcial-.

Desarrollo

Quick EDA

En una revisión rápida del dataset crudo, se analizan nulos, duplicados y frecuencias respecto a códigos de capítulo (HS02), partida (HS04) y subpartida (HS06):

Valores nulos por columna:

Columna	Valores nulos
HS06	0
GOODS_DESCRIPTION	0
HS04	0
HS02	0

Frecuencias por HS02, TOP10:

HS02	Cantidad	Frecuencia relativa	Frecuencia acumulada
84	54.901	20,5%	20,5%
85	33.571	12,54%	33,04%
87	28.476	10,63%	43,67%
73	16.173	6,04%	49,71%
39	12.218	4,56%	54,28%
90	11.611	4,34%	58,61%
82	7.972	2,98%	61,59%
94	7.921	2,96%	64,55%
40	7.526	2,81%	67,36%
83	4.285	1,60%	68,96%

Frecuencias por HS02, BOTTOM10:

HS02	Cantidad	Frecuencia relativa	Frecuencia acumulada
41	22	0,01%	99,96%
81	19	0,01%	99,97%
45	19	0,01%	99,97%
05	15	0,01%	99,98%
80	15	0,01%	99,98%
50	11	0,00%	99,99%
14	11	0,00%	99,99%
78	10	0,00%	100,0%

51	6	0,00%	100,0%
43	5	0,00%	100,0%

Frecuencias por HS04, TOP10:

HS04	Cantidad	Frecuencia relativa	Frecuencia acumulada
8708	8.524	3,18%	3,18%
8703	7.341	2,74%	5,92%
7318	5.700	2,13%	8,05%
8536	5.487	2,05%	10,10%
8482	4.895	1,83%	11,93%
8421	4.593	1,72%	13,65%
8431	3.910	1,46%	15,11%
8483	3.819	1,43%	16,53%
8481	3.562	1,33%	17,86%
8714	3.485	1,30%	19,16%

Frecuencias por HS06, TOP10:

HS06	Cantidad	Frecuencia relativa	Frecuencia acumulada
870323	3.869	1,44%	1,44%
848280	2.924	1,09%	2,54%
871120	2.779	1,04%	3,57%
731815	2.632	0,98%	4,56%
870322	2.247	0,84%	5,40%
870899	2.074	0,77%	6,17%
950300	1.988	0,74%	6,91%
940540	1.874	0,70%	7,61%
848180	1.857	0,69%	8,31%
901890	1.836	0,69%	8,99%

Respecto a las frecuencias por HS04 y HS06, BOTTOM10, los mínimos están en 1 única muestra por código.

Nomenclatura HS06 en ingles

Como desarrollo del dataset original, se trabajó con la nomenclatura de los códigos de HS06 que consiste en descripciones genéricas para cada código y que, además puede separarse en capítulos,

partidas y subpartidas.

Primeros 5 códigos HS06:

HS06	full_eng	HS04	HS02
010120	Live horses, asses, mules and hinnies. && - Horses :	0101	01
010121	Live horses, asses, mules and hinnies. && - Horses : && -- Pure-bred breeding animals	0101	01
010129	Live horses, asses, mules and hinnies. && - Horses : && -- Other	0101	01
010130	Live horses, asses, mules and hinnies. && - Asses	0101	01
010190	Live horses, asses, mules and hinnies. && - Other	0101	01

Últimos 5 códigos HS06:

HS06	full_eng	HS04	HS02
961590	Combs, hair-slides and the like; hairpins, curling pins, curling grips, hair-curlers and the like, other than those of heading 85.16, and parts thereof. && - Other	9615	96
961610	Scent sprays and similar toilet sprays, and mounts and heads therefor; powder-puffs and pads for the application of cosmetics or toilet preparations. && - Scent sprays and similar toilet sprays, and mounts and heads therefor	9616	96
961620	Scent sprays and similar toilet sprays, and mounts and heads therefor; powder-puffs and pads for the application of cosmetics or toilet preparations. && - Powder-puffs and pads for the application of cosmetics or toilet preparations	9616	96
970110	Paintings, drawings and pastels, executed entirely by hand, other than drawings of heading 49.06 and other than hand-painted or hand-decorated manufactured articles; collages and similar decorative plaques. && - Paintings, drawings and pastels	9701	97
970190	Paintings, drawings and pastels, executed entirely by hand, other than drawings of heading 49.06 and other than hand-painted or hand-decorated manufactured articles; collages and similar decorative plaques. && - Other	9701	97

IMPORTANTE: Encontramos que 4.5 % de los datos no encuentran su descripción de nomenclatura, lo que puede deberse a que se están utilizando una nomenclatura diferente a la usada en el registro de los datos.

EDA detallado

Para analizar la composición de las descripciones, desde parámetros estadísticos, se procedió a generar las siguientes variables:

['GOODS_DESCRIPTION_len_words',

'GOODS_DESCRIPTION_len_chars',

'subtokenization_indicator']

Para cada descripción, se evalúan el largo en cantidad de palabras y de caracteres y un indicador de tokenización, que mide como se tokeniza la descripción en relación a la cantidad de palabras.

El tokenizer utilizado es un algoritmo de Hugging Face Transformers, llamado "distilbert-base-uncased". Este algoritmo corresponde al modelo pre-entrenado que se utiliza luego para procesar los textos en estudio:

https://huggingface.co/distilbert/distilbert-base-uncased

Luego, se procesaron las variables en agregaciones estadísticas, para agrupaciones en HS06, HS04 y HS02, según el siguiente diccionario:

{'HS06': ['count'],

'GOODS_DESCRIPTION_len_words': ['sum', 'min', 'mean', 'median', 'max', 'std'],

'GOODS_DESCRIPTION_len_chars': ['sum', 'min', 'mean', 'median', 'max', 'std'],

'subtokenization_indicator': ['sum', 'min', 'mean', 'median', 'max', 'std']}

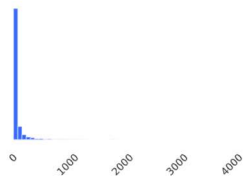
A continuación, se muestran algunas distribuciones, con datos básicos de las variables:

HS06_count

Real number (ℝ)

High correlation

Distinct	430	**Minimum**	1
Distinct (%)	10.3%	**Maximum**	3869
Missing	0	**Zeros**	0
Missing (%)	0.0%	**Zeros (%)**	0.0%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	63.909308	**Memory size**	65.5 KiB



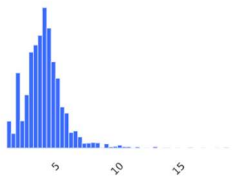
El desbalance en las clases/códigos se cuantifica razonablemente en variables como: HS06_count, GOODS_DESCRIPTION_len_words_sum y GOODS_DESCRIPTION_len_chars_sum, con valores medios lejos de la mediana y distribuciones de colas largas.

GOODS_DESCRIPTION_len_words_mean

Real number (ℝ)

High correlation

Distinct	1692	**Minimum**	1
Distinct (%)	40.4%	**Maximum**	19
Missing	0	**Zeros**	0
Missing (%)	0.0%	**Zeros (%)**	0.0%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	4.0534472	**Memory size**	65.5 KiB



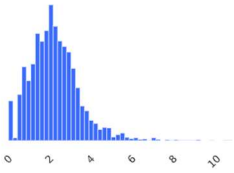
A su vez, surgen distribuciones más acampanadas para variables como el largo promedio (tanto para palabras como para caracteres), cuando se agrupa en los distintos códigos HS06, HS04 y HS02.

GOODS_DESCRIPTION_len_words_std

Real number (ℝ)

High correlation Missing Zeros

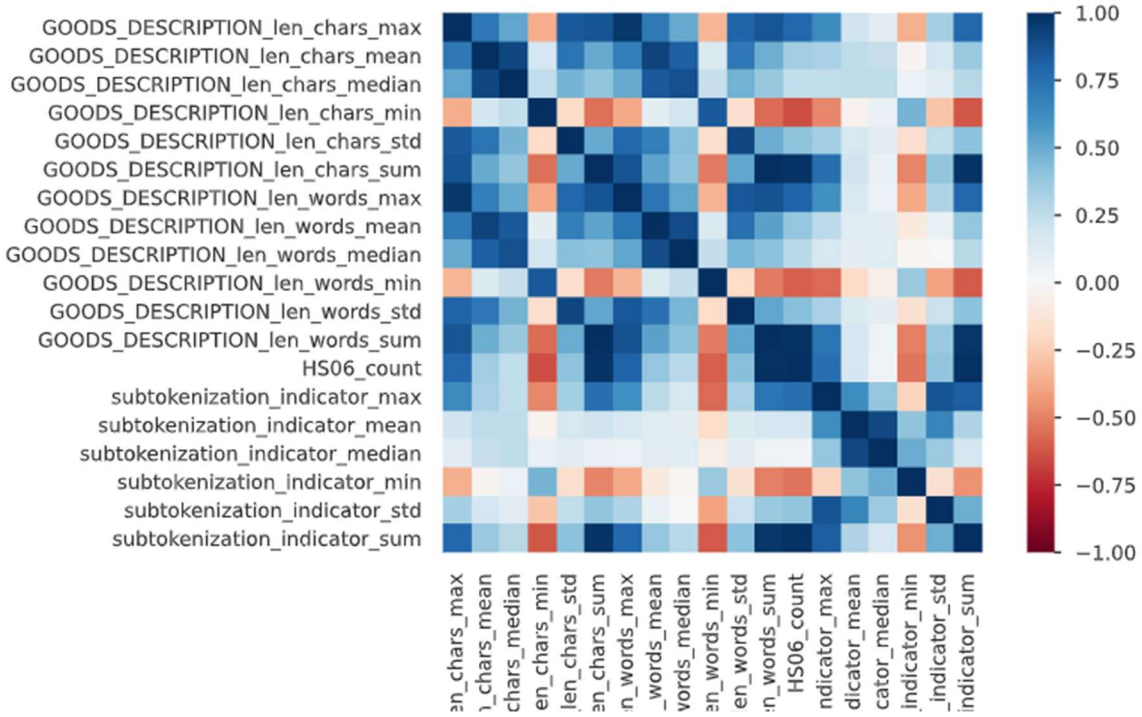
Distinct	2713	**Minimum**	0
Distinct (%)	74.3%	**Maximum**	10.843585
Missing	540	**Zeros**	100
Missing (%)	12.9%	**Zeros (%)**	2.4%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	2.2447975	**Memory size**	65.5 KiB



Finalmente, la desviación estándar del largo de las descripciones tiene una distribución parecida

a un Xi cuadrado, tanto para el largo en palabras como en caracteres. Se destaca que hay códigos con una única muestra, resultado en NULLs al calcular la desviación.

En términos de correlaciones, no hay grandes sorpresas, ya que se correlacionan variables que agregan cantidades como: HS06_count, GOODS_DESCRIPTION_len_words_sum, GOODS_DESCRIPTION_len_chars_sum, GOODS_DESCRIPTION_len_words_max, GOODS_DESCRIPTION_len_chars_max.



Embeddings con DistilBERT y PCA

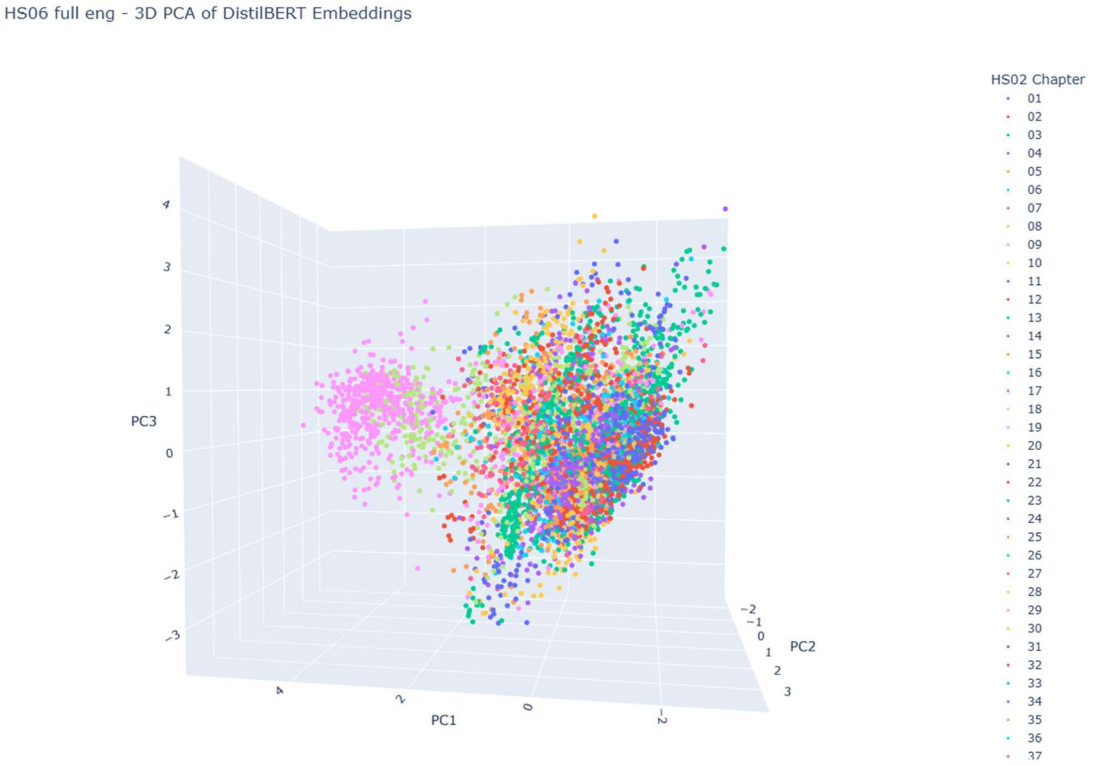
Para evaluar las descripciones se procede a utilizar un modelo pre-entrenado, disponible en Hugging Face, llamado "distilbert-base-uncased": https://huggingface.co/distilbert/distilbert-base-uncased

Los tokens generados se procesan y se obtienen embeddings de 764 dimensiones, uno para cada caso. A su vez, a modo de referencia, se procesan los textos de la nomenclatura de los códigos HS06.

Para la visualización de estos embeddings se utiliza PCA, en 2 y 3 dimensiones. A continuación, se observan los resultados, según HS06 descripciones de mercaderías - muestra de un 5 % (13,389 casos):

En amarillo, se observa un marcado cluster que consiste en las descripciones del capítulo HS02 87 (vehículos y material automotor); y

HS06 full eng (nomenclatura):



En este caso, en rosa y verde claro se observan un cluster que corresponde a la nomenclatura de

los capítulos HS02 28 y 29 (químicos orgánicos e inorgánicos).

Medición de similitud del coseno

Cómo último ejercicio exploratorio, se procede a medir la similitud, para cada caso, entre GOODS_DESCRIPTION y full_eng (descripción de nomenclatura), para evaluar cuando distan los embeddings de estos textos.

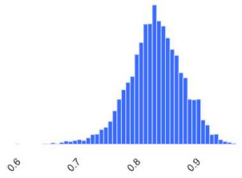
En una agregación por HS06 se observan valores medios y medianos altos y coincidentes, del orden de 0.84:

cosine_sim_gd_vs_hs_text_mean

Real number (ℝ)

High correlation Missing

Distinct	3960	**Minimum**	0.60158304
Distinct (%)	100.0%	**Maximum**	0.96980786
Missing	230	**Zeros**	0
Missing (%)	5.5%	**Zeros (%)**	0.0%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	0.83501187	**Memory size**	65.5 KiB



Quantile statistics

Minimum	0.60158304
5-th percentile	0.75920722
Q1	0.80768058
median	0.83575511
Q3	0.86522526
95-th percentile	0.9077215
Maximum	0.96980786
Range	0.36822482
Interquartile range (IQR)	0.057544687

Descriptive statistics

Standard deviation	0.045425905
Coefficient of variation (CV)	0.054401508
Kurtosis	0.57477478
Mean	0.83501187
Median Absolute Deviation (MAD)	0.028919566
Skewness	-0.31314854
Sum	3306.647
Variance	0.0020635128
Monotonicity	Not monotonic

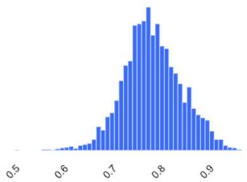
En términos de similitud mínima por código HS06, la similitud también es relativamente alta, con una media de 0.79.

cosine_sim_gd_vs_hs_text_min

Real number (ℝ)

High correlation Missing

Distinct	3957	**Minimum**	0.50259566
Distinct (%)	99.9%	**Maximum**	0.96980786
Missing	230	**Zeros**	0
Missing (%)	5.5%	**Zeros (%)**	0.0%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	0.78660514	**Memory size**	65.5 KiB



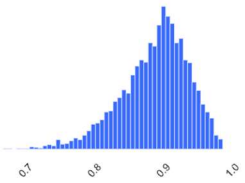
En términos de similitud máxima por código HS06, el valor medio ronda el 0.89, aunque se muestran valores muy próximos a la unidad.

cosine_sim_gd_vs_hs_text_max

Real number (ℝ)

High correlation | Missing

Distinct	3951	**Minimum**	0.66693276
Distinct (%)	99.8%	**Maximum**	0.98533964
Missing	230	**Zeros**	0
Missing (%)	5.5%	**Zeros (%)**	0.0%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	0.88693692	**Memory size**	65.5 KiB



Lo antes descripto, motiva a observar los casos de similitud media, mínima y máxima.

Casos de similitud media o regular:

HS06	Descripción de la mercadería	Descripción completa del HS06	Similitud de coseno
853950	ASSY LED Base Strobe Upgrade	Electric filament or discharge lamps, including sealed beam lamp units and ultra-violet or infra-red lamps – Light-emitting diode (LED) lamps	0.827174
220870	VODKA FRAISE JELZIN STRAWBERRY	Undenatured ethyl alcohol of an alcoholic strength by volume of less than 80 % vol; spirits, liqueurs and other spirituous beverages – Liqueurs and cordials	0.831804
620590	SHORT SLEEVE REPAIR SHIRT	Men's or boys' shirts – Of other textile materials	0.830535
732620	HANGER ROD	Other articles of iron or steel – Articles of iron or steel wire	0.830890
843143	8-3/8 SH Extension Overshot C-17208	Parts suitable for use solely or principally with the machinery of headings 84.25 to 84.30 – Parts for boring or sinking machinery	0.828724
870323	SUZUKI ESCUDO 2006	Motor cars and other motor vehicles principally designed for the transport of persons – Of a cylinder capacity exceeding 1 500 cm^3 but not exceeding 3 000 cm^3	0.817948
841899	Spare Parts for 10 TR Air Cooled Water Chiller	Refrigerators, freezers and other refrigerating or freezing equipment – Parts	0.822742
871120	USED CHANGZHOU KWANGYANG MOTORBYKE	Motorcycles (including mopeds) and cycles fitted with an auxiliary motor – With reciprocating internal combustion piston engine of a cylinder capacity exceeding 50	0.826228

		cm³ but not exceeding 250 cm³	
940540	SURFACE MOUNTED LIGHTS	Lamps and lighting fittings including searchlights and spotlights and parts thereof – Other electric lamps and lighting fittings	0.817389
842131	FILTER ASSY, AIR CLEANER	Centrifuges, including centrifugal dryers; filtering or purifying machinery and apparatus for liquids or gases – Intake air filters for internal combustion engines	0.831361

BOTTOM5 casos de similitud mínima:

HS06	Descripción de la mercadería	Descripción completa del HS06	Similitud de coseno
741820	SHOWER HEAD SQUARE BLACK 260X 190mm	Table, kitchen or other household articles and parts thereof, of copper; pot scourers and scouring or polishing pads, gloves and the like, of copper; sanitary ware and parts thereof – Sanitary ware and parts thereof	0.553591
741820	74182000000	Table, kitchen or other household articles and parts thereof, of copper; pot scourers and scouring or polishing pads, gloves and the like, of copper; sanitary ware and parts thereof – Sanitary ware and parts thereof	0.552399
741820	FREESTANDING BATH TOWER	Table, kitchen or other household articles and parts thereof, of copper; pot scourers and scouring or polishing pads, gloves and the like, of copper; sanitary ware and parts thereof – Sanitary ware and parts thereof	0.552331
741820	TRAP ASSEMBLY	Table, kitchen or other household articles and parts thereof, of copper; pot scourers and scouring or polishing pads, gloves and the like, of copper; sanitary ware and parts thereof – Sanitary ware and parts thereof	0.548731
741820	HG BATH MIXER WALL MOUNTED MYSPORT CHROME	Table, kitchen or other household articles and parts thereof, of copper; pot scourers and scouring or polishing pads,	0.545765

		gloves and the like, of copper; sanitary ware and parts thereof – Sanitary ware and parts thereof	

TOP5 Casos de similitud máxima:

HS06	Descripción de la mercadería	Descripción completa del HS06	Similitud de coseno
640299	Other:Other footwear with outer soles and uppers of rubber or pla:Other footwear	Other footwear with outer soles and uppers of rubber or plastics. – Other footwear – Other	0.98534
520939	Other fabrics:Woven fabrics of cotton, containing 85 % or more by weight of:Dyed	Woven fabrics of cotton, containing 85 % or more by weight of cotton, weighing more than 200 g/m² – Dyed – Other fabrics	0.984521
700729	Other:Safety glass, consisting of toughened (tempered) or:Laminated safety glass	Safety glass, consisting of toughened (tempered) or laminated glass – Laminated safety glass – Other	0.984296
848220	Tapered roller bearings, including cone and tapered roller assemblies	Ball or roller bearings – Tapered roller bearings, including cone and tapered roller assemblies	0.984269
740729	Other:Copper bars, rods and profiles.:Of copper alloys	Copper bars, rods and profiles – Of copper alloys – Other	0.984245

Los casos de similitud mínima muestran indicios de una mala calidad de la descripción, incluso con un caso donde la descripción es el mismo código. Por otro lado, los casos de similitud máxima muestran que la descripción comercial, en realidad, consiste en una variante descripción de nomenclatura.

En los casos de similitud dentro del entorno de la media, las descripciones parecen lógicas, relacionadas con la naturaleza de la mercadería, pero un tanto escuetas.

Bibliografía

Explainable Product Classification for Customs, EUNJI LEE, 2023

DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter, VICTOR SANH, 2020