Taller de Tesis I – Entrega I

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.