Análisis de obtención de datos para el reto

Grupo 501

Conjuntos de datos

Primer conjunto:

Enlace:
https://data.world/vizwiz/car-sales-mock-data/workspace/file?filename=Cars+Mock+Data.csv

El archivo tiene 20 campos con información de los cuales solamente 4 son relevantes para el modelo de datos que se busca utilizar en el reto, mientras que el resto de los campos no son relevantes ya que tienen datos mucho más administrativos que descriptivos de los autos.

Los campos relevantes del conjunto son, marca, modelo, color, precio, y un tipo de rendimiento midiendo el tiempo que hace el auto a partir de 0 a 60 millas por hora. Se consideran útiles porque son campos descriptivos del automóvil, los cuales van a ser el pilar de los datos que guardemos de un automóvil.

Mientras que los campos que no encontramos relevantes se refieren a elementos administrativos que maneja una agencia, tales como depreciación de precio, descuentos y valor de reventa. Este tipo de datos no son de nuestro interés, por lo que en su mayoría no podría ser usado.

En cuanto a la implementación de los datos encontrados al modelo no relacional actual, nos ayuda confirmar que los campos descriptivos son correctos en cuanto a estándares básicos, mientras que los que no son relevantes al momento podrían incluirse en un futuro si la funcionalidad de la plataforma creciera hacia acatar las tareas administrativas que realiza una agencia.

Concluyendo con este primer conjunto de datos, el archivo encontrado nos da la oportunidad de explorar qué otro tipo de información sería relevante para las agencias, es este tipo de perspectiva que nos va a permitir generar gráficas y análisis relevantes para nuestros clientes de grupos automotrices. Puede que no utilicemos los mismos campos, pero si podemos obtener y presentar posteriormente un contexto que brinde valor a la plataforma y al cliente.

Segundo conjunto:

Enlace:
https://deepvisualmarketing.github.io/

Este conjunto a diferencia del primero es mucho más extensivo y se asemeja mucho más a los campos que se encuentran al momento en el modelo no relacional que se está manejando para el desarrollo de la plataforma.

Para empezar, el conjunto está dividido en secciones, cada sección tiene sus propios campos, estas secciones contienen información descriptiva de autos, al igual que detalles de ventas y precios. Se puede notar que las distintas secciones tienen un propósito particular que apela a una infraestructura de datos más grande que la que se está buscando manejar con el modelo no relacional de desarrollo.

Ahondando un poco más en los campos específicos que se encuentran en el conjunto y en el modelo no relacional actual, algunos de los campos descriptivos para autos son marca, modelo, un id, año del auto, color, tipo de vehículo, kilometraje, motor, y velocidad máxima. Existen registros adicionales que no están siendo contemplados a lo largo del desarrollo y que si se encuentran en el conjunto.

Ciertamente este segundo archivo tiene un gran potencial para comparar y mejorar nuestro modelo actual, comenzando con campos descriptivos que podrían servir para autos usados, o información adicional que puede no estar incluida en descripciones. En cuanto a lo anterior, estamos hablando de cantidad de asientos y puertas, número que estamos esperando vengan de una descripción para entonces ser explotada por el sistema de búsqueda de nuestra base de datos, pero en caso de que no la incluya, sería importante incluirla como contingencia o mitigación.

En cuanto a los datos administrativos, la existencia de ciertos campos y la repetición constante de algunos datos nos permite asumir o aproximar el tipo de sistema que maneja el responsable del conjunto de datos. Y aunque bastante amplia, es una mirada a lo que podría crecer y ser la plataforma que estamos desarrollando. Nuestro modelo no relacional va a ser la entrada a la repetición constante y redundancia de información.

Hay una seccion especifica que concierne a enlaces de fotografías de los autos, esta muestra del conjunto llega a pasar el millón de registros o entradas, lo cual podría parecer excesivo al inicio, sin embargo, si se pone en contexto que alguna imágenes o funcionalidad de visualización del auto a 360 grados involucran muchas fotos, entonces podemos contextualizar la gran cantidad de enlaces guardados en el conjunto.

En conclusión, es bueno confirmar con este conjunto que los campos que estamos guardando en cuanto a la descripción del auto son los adecuados, además nos permitió considerar utilizar más campos para todo tipo de autos. Y de la misma manera que el conjunto anterior, nos da una perspectiva hacia lo que podría ser el manejo de datos de la plataforma en un futuro, específicamente con los enlaces a fotografías, que podrían extenderse a ser cientos de miles algún día.

Desarrollo de conjunto de datos dummy/prueba:

En cuanto a crear nuestro propio conjunto de datos de prueba, exploramos varios métodos, entre ellos la creación de un programa en python que generará listas de datos aleatorios dentro de ciertos parámetros. De la misma manera se exploró el uso de herramientas externas como Chat GPT, del cual se pedía una muestra pequeña de 10 o 100 entradas de algún campo que estuviéramos considerando en el modelo de datos.

Después de considerar y probar los diferentes métodos mencionados anteriormente, decidimos utilizar Chat GPT para generar listas cortas que podríamos expandir usando programas como Excel. Fue entonces que utilizando esta herramienta generamos listas de los campos que se estaban considerando en el modelo no relacional incluso antes de la inspección de los conjuntos de datos.

La lista de campos generados es la siguiente: marca, modelo, año, precio, color, combustible, rendimiento, transmision, cantidad, motos, agencia_id, estado_agencia, municipio_agencia, tipo_vehiculo, descripcion, url_fotografia, grupo_automotriz_id, grupo_automotriz, y gerente_id.

Se puede apreciar que incluimos campos no descriptivos del auto en este conjunto, y esto es para poder en un futuro cercano poblar esa colección de la base de datos con el conjunto generado. Estos campos incluyen id 's de cuentas de usuarios gerentes y de agencias registradas, esto para mantener consistencia en el formato de dummy data que manejamos, usando todos los campos.

Aunque si se buscara la funcionalidad de las imágenes, este conjunto debería ser modificado ya que los enlaces guardados no funcionan y son de ejemplo. Incluso si se quisiera completar las otras colecciones con información de prueba, sería recomendable aprovechar los datos de agencias y mantener coherencia con los distintos id ’s usados a lo largo de las colecciones de la base de datos.

Este conjunto podría ser utilizado a lo largo de las distintas etapas de prueba, poblando así la base de datos y permitiendo un desarrollo más estandarizado en cuanto al manejo de datos.
Es importante partir de esta base para después extender este conjunto a las pruebas de la integración de la herramienta de Elasticsearch, porque aunque puede servir para filtrar un simple catálogo, se debe definir y crear coherencia entre campos si esperamos aprovechar al máximo la capacidad de procesamiento de lenguaje natural.

En conclusión, esta etapa sirvió mucho para tener una base de datos de prueba y un fundamento para continuar el desarrollo de la plataforma. El conteo final de entradas o registros creados es de 10,000, alcanzando los criterios descritos de la actividad.

Conjunto obtenido por scraping:

Para este caso, se decidió hacer el web scraping usando la herramienta de Scrapy con la integración de scrapy_playwright, gracias a la flexibilidad y robustez que nos ofrece. La manera principal en la que lo aprovechamos fue por medio de la interacción con headless browsers, pues las tarjetas de los autos no cargaban de manera completa, también interactuando con el sitio, pues ciertos elementos solamente cargaban al dar click en un “acordeón”.

Los campos que se seleccionaron están basados en la disponibilidad de datos que tiene Kavak, en contraste con los datos que nosotros deseamos implementar en nuestro sitio, pues había algunos que no existían en Kavak o aparecían raramente en el sitio. Un comentario que puede valer la pena mencionar es que este hecho de que hay inconsistencia de campos, nos dice que Kavak puede estar usando algún tipo de base de datos no relacional, dada la flexibilidad que se da para registrar los autos.

Este conjunto de datos puede ser expandido para poder extraer las imágenes si es que necesitamos hacer distintos tipos de pruebas. y posiblemente integrarlo con los demás conjuntos como el de dummy data, por sus campos como el de descripción, para tener un conjunto que cada vez más se acerque al que usaría nuestra plataforma. La limitación aquí se encuentra en los datos que se encuentren disponibles en los sitios web, así como en la calidad del código del scraping.