CIENCIA DE DATOS - TIPOS DE DATOS

Tanto en Data Science como en Big Data, se encontrará con muchos tipos diferentes de datos, y cada uno de ellos tiende a requerir diferentes herramientas y técnicas. Las principales categorías de tipos de datos son las siguientes:

Datos estructurados

Datos no estructurados

Lenguaje natural

Generado por máquina

Basado en gráficos

Audio, video e imágenes

Ahora vamos a explorar todos estos tipos de datos interesantes.

Datos Estructurados

Los datos estructurados son datos que dependen de un modelo de datos y residen en un campo fijo dentro de un registro. Como tal, a menudo es fácil almacenar datos estructurados en tablas dentro de bases de datos o archivos de Excel como en la siguiente imagen. SQL es la forma preferida de administrar y consultar datos que residen en bases de datos. También puede encontrar datos estructurados que podrían dificultar su almacenamiento en una base de datos relacional tradicional. Los datos jerárquicos, como puede ser un árbol genealógico.

Datos no estructurados

Los datos no estructurados son datos que no son fáciles de encajar en un modelo de datos porque el contenido es específico del contexto o varía. Un ejemplo de datos no estructurados puede ser un correo electrónico. Aunque el correo electrónico contiene elementos estructurados como el remitente, el título y el cuerpo del texto, es un desafío encontrar el número de personas que han escrito un mensaje de correo electrónico de queja sobre un empleado específico porque existen muchas maneras de referirse a una persona, por ejemplo. Los miles de idiomas y dialectos diferentes que hay por ahí complican aún más esto. Un correo electrónico escrito por un humano es también un ejemplo perfecto de datos en lenguaje natural.

Lenguaje natural

El lenguaje natural es un tipo especial de datos no estructurados; es difícil de procesar porque requiere el conocimiento de técnicas específicas de ciencia de datos y lingüística.

La comunidad de procesadores de lenguaje natural ha tenido éxito en el reconocimiento de entidades, el reconocimiento de temas, el resumen, la finalización de textos y el análisis de sentimientos, pero los modelos formados en un dominio no se generalizan bien a otros dominios. Ni siquiera las técnicas más avanzadas son capaces de descifrar el significado de cada trozo de texto. Sin embargo, esto no debería ser una sorpresa: los humanos también luchan con el lenguaje natural. Es ambiguo por naturaleza. El concepto de significado en sí mismo es cuestionable aquí. Que dos personas escuchen la misma conversación. ¿Conseguirán el mismo significado? El significado de las mismas palabras puede variar cuando vienen de alguien molesto o alegre.

Datos generados por máquinas

Los datos generados por máquinas es información que se crea automáticamente por un ordenador, proceso, aplicación u otra máquina sin intervención humana. Los datos generados por máquinas se están convirtiendo en un importante recurso de datos y lo seguirán siendo. La IDC (International Data Corporation) ha estimado que habrá 50 veces más máquinas conectadas que personas en 2024. Esta red es comúnmente conocida como la Internet de las cosas. El análisis de los datos de las máquinas se basa en herramientas altamente escalables, debido a su gran volumen y velocidad. Ejemplos de datos de máquinas (por ejemplo en la siguiente imagen) son los registros del servidor web, los registros de detalles de llamadas o los registros de eventos de red.

Los datos de la máquina mostrados en la figura de arriba encajarían muy bien en una base de datos clásica estructurada en forma de tabla con la peculiaridad que han sido generados por una máquina

Datos gráficos

"Gráfico de datos" puede ser un término confuso porque cualquier dato puede ser mostrado en un gráfico. "Gráfico" en este caso apunta a la teoría de gráficos matemáticos. En la teoría de grafos, un gráfico es una estructura matemática para modelar las relaciones entre los objetos. Los datos de los gráficos o de las redes son, en resumen, datos que se centran en la relación de los objetos. Las estructuras de los gráficos utilizan nodos, bordes y propiedades para representar y almacenar los datos de los gráficos. Los datos basados en gráficos son una forma natural de representar las redes sociales, y su estructura permite calcular métricas específicas como la influencia de una persona y el camino más corto entre dos personas. Se pueden encontrar ejemplos de datos basados en gráficos en muchos sitios web como redes sociales. Su lista de seguidores en Twitter es un ejemplo de datos basados en gráficos. El poder y la sofisticación proviene de múltiples gráficos superpuestos de los mismos nodos. Por ejemplo, imagina los bordes de conexión aquí para mostrar "amigos" en Facebook. Imagina otro gráfico con la misma gente que conecta a los colegas de negocios a través de LinkedIn. Imagina un tercer gráfico basado en los intereses de las películas en Netflix. Superponiendo los tres gráficos podemos encontrar unas relaciones interesantes.

Audio, imagen y vídeo

Audio, imagen y vídeo son tipos de datos que plantean desafíos específicos a un Data Scientist. Tareas que son triviales para los seres humanos, como reconocer objetos en imágenes, resultan ser un desafío para las computadoras. MLBAM (Major League Baseball Advanced Media) anunció en 2018 que aumentará la captura de vídeo a aproximadamente 14 TB por juego con el fin de realizar análisis en directo dentro del juego. Las cámaras de alta velocidad en los estadios capturarán los movimientos de la pelota y del atleta para calcular en tiempo real, por ejemplo, el camino que toma un defensor en relación con dos líneas de base. Recientemente una compañía llamada DeepMind tuvo éxito en la creación de un algoritmo que es incapaz de aprender a jugar a los videojuegos. Este algoritmo toma la pantalla de vídeo como entrada y aprende a interpretar todo a través de un complejo proceso de Deep Learning. Es una hazaña notable que llevó a Google a comprar la compañía para sus propios planes de desarrollo de Inteligencia Artificial (IA).