Lista de comprobación de limpieza de datos

Le damos la bienvenida a la era de la inteligencia artificial (IA), donde los negocios se ven supeditados a tecnologías de uso intensivo de datos, como el aprendizaje automático y el aprendizaje profundo. Para aprovechar las ventajas de estas nuevas herramientas de IA, debe asegurarse de que el "hogar" en el que almacena los datos de su organización está ordenado.

A continuación dispone de una lista de comprobación para comenzar a limpiar los datos almacenados, que se desglosa en dos fases clave del proceso de limpieza: formación e inferencia.

Siga estos pasos para poder convertirse en un experto en IA. Para más información sobre cómo aplicar la IA a gran escala y en la producción, y no solo como prueba de concepto, consulte este informe de IDC titulado *"Accelerate and Operationalize AI Deployments Using AI-Optimized Infrastructure" (Acelerar y poner en marcha las implementaciones de IA mediante infraestructura optimizada para la IA)*.

Formación

Durante la fase de formación del proceso de preparación para el uso de la IA, desarrollará algoritmos con el fin de comprender un conjunto de datos. Su principal objetivo será el de agrupar los datos existentes y utilizar la IA para aprender una nueva capacidad.

- [ ] Averigüe cuál es el problema de negocio específico que desea resolver con el uso de la IA (comience con proyectos más pequeños para que pueda aprender)
- [ ] Localice los datos que puedan resolver ese problema en las fuentes pertinentes (lo más probable es que no se encuentren todos en un mismo lugar)
- [ ] Prepare sus datos con etiquetas de metadatos con el fin de reducir drásticamente el tiempo que dedica a buscar los datos pertinentes
- [ ] Asegúrese de que sus datos se han sincronizado y vinculado adecuadamente en todos los conjuntos de datos que usará (también temporalmente)
- [ ] Marque cualquier dato confidencial de sus clientes, así como otros datos personales, para garantizar la seguridad completa de su almacenamiento y el cumplimiento de todas las normas y gobernanzas (el proceso de etiquetado de metadatos puede ser útil en este paso)
- [ ] Escoja el entorno de desarrollo apropiado para el tipo de datos que vaya a usar y la forma en que tendrán formato (es decir, las imágenes, los vídeos, los textos en formato libre y los audios suelen tener un tipo de entorno determinado)
- [ ] Extraiga los conjuntos de datos de su repositorio e introdúzcalos en su entorno de desarrollo
- [ ] Divida sus datos en dos grupos con el objetivo de facilitar la mejora de su proceso de desarrollo de modelos (guarde un conjunto en una carpeta denominada "train" [formación], y otro en una carpeta con el nombre "test" [prueba])
- [ ] Mantenga un seguimiento de sus datos registrando el origen de estos o la fuente de la que provienen (considere utilizar herramientas que puedan ayudarle a automatizar el proceso)
- [ ] Lleve a cabo tareas de higiene de datos básicas para prepararlos para la elaboración de un modelo (por ejemplo, rellene entradas de datos vacías y elimine las entradas de datos no válidas)
- [ ] Utilice una muestra de un subconjunto de aquellos datos cuya respuesta a la actividad de predicción ya conozca (lo que se conoce como "conjunto de formación") e identifique todos los pasos previos al procesamiento que sean necesarios para prepararlos para la predicción
- [ ] Aproveche lo que conozca de este conjunto de formación para calcular puntuaciones de precisión que le puedan aportar la seguridad necesaria como para aplicar el mismo modelo a los nuevos datos con los que nunca se ha entrenado al modelo de manera explícita

Inferencia

Una vez que haya desarrollado un modelo que sea capaz de solucionar su problema de negocio, pasará de la fase de formación a la de inferencia. En esta fase, aprovechará ese modelo que funciona y lo aplicará a los nueavos datos que reciba, lo que también requiere que realice algo de limpieza de datos de manera continua.

- [ ] Sitúe su modelo de IA cerca de sus datos con el objetivo de reducir la latencia y los requerimientos del ancho de banda, además de para mejorar el rendimiento general del modelo
- [ ] Desarrolle un proceso de gestión de los datos eficiente y aplique un etiquetado de metadatos a medida que estos lleguen, de modo que pueda agrupar los nuevos datos y utilizarlos para mejorar el modelo en el futuro
- [ ] Etiquete los datos de manera que estén vinculados y sincronizados (por ejemplo, si los datos se encuentran secuenciados en el tiempo, puede sincronizarlos en varios conjuntos de datos o vincularlos escogiendo un campo determinado para todos los datos nuevos que reciba; como el nombre de un cliente)
- [ ] Desarrolle un plan de almacenamiento a largo plazo de los datos durante su ciclo de vida para gestionar el volumen y la velocidad de estos a medida que los recibe y archiva
- [ ] Analice la posibilidad de contratar a un director de datos que se encargue de gestionar los datos de su organización en futuros proyectos de IA, aprendizaje profundo y otros proyectos basados en datos.

IBM