mdt.

Ex Datis, Claritas

Vas a oír mucho sobre datos sintéticos, ya verás

La industria de la IA ve en los datos sintéticos una posible solución al problema de la escasez de textos con los que entrenar sus nuevos modelos.

Manuel Delgado Tenorio · 29 de marzo de 2025

Publicado originalmente en www.manueldelgado.com/blog/vas-a-oir-mucho-sobre-datos-sinteticos



La industria de la IA ve en los datos sintéticos una posible solución al problema de la escasez de textos con los que entrenar sus nuevos modelos.

Nvidia ha comprado una startup, Gretel, dedicada a la generación de datos sintéticos. Veo esta compra como otro paso más por parte de Nvidia en dejar de ser un fabricante de hardware e integrarse en todas las fases del ciclo de vida de los modelos de IA.

También la veo como una llamada de atención sobre **la enorme relevancia que están adquiriendo los datos sintéticos en este mercado.**

Si no tienes claro qué es esto de los datos sintéticos, échale un vistazo al artículo mío que enlazo al final del todo. **Lo escribí hace ya cinco años, antes de que la IA fuera sexy,** y tiene un sesgo claro hacia lo que en ese momento ocupaba nuestras cabezas, el *Machine Learning*, pero el concepto es exactamente igual: poder usar datos creados artificialmente, asegurándonos de que poseen la misma estructura estadística interna que los datos reales en los que se basan.

(Por cierto: minipunto para ti si entiendes por qué elijo esa imagen para ilustrar este artículo. Probablemente, significa que tienes ya más años que un bosque.)

Lo que cambia respecto a 2020 es, principalmente, el uso que se le va a dar a los datos sintéticos. En aquel momento, resolvían problemas relacionados con, por ejemplo, cómo compartir datos sensibles sobre el comportamiento de tus clientes con una consultora que te va a generar modelos de Machine Learning asegurando que:

1. los datos reales no salen de tu organización para, así, evitar multitud de riesgos;
2. y, a la vez, los datos que compartimos con ese tercero contienen la misma información sobre el comportamiento de nuestros clientes que los datos reales, por lo que los modelos creados con ellos seguirán siendo de utilidad.

Hoy, sin embargo, la prioridad está en evitar uno de los principales obstáculos de la industria de los grandes modelos de lenguaje (LLM): **se están agotando los datos reales con los que se puede entrenar estos modelos.** Y no lo digo solo yo o los (verdaderos) expertos de esta industria, también te lo dice el propio Elon Musk, nada sospechoso de ser un crítico de estas cosas.

La clave, entonces tanto como hoy, está en generar confianza en que los modelos entrenados con datos sintéticos son igual de fiables que los entrenados con datos reales.

Ahora mismo, parece haber consenso en que, sí, es posible complementar los datos reales con datos sintéticos para entrenar LLMs, sin que el producto final se vea afectado negativamente. **De hecho, no solo es posible, sino que es muy habitual.**

My two cents: si bien hablamos mucho de datos sintéticos para el entrenamiento "desde cero" de modelos de IA generativa, **pronto veremos una auténtica explosión en el uso de datos sintéticos para actividades de personalización de LLM,** ya sea por la vía del *fine-tuning* o, sobre todo, del RAG, principalmente en aquellos entornos en los que no nos fiamos por completo de que los datos con los que personalizamos el modelo vayan a seguir siendo confidenciales al usarlos de esta forma.

Y, ahora, como te prometí unos párrafos más arriba, el enlace al artículo que escribí hace un tiempo: ¿Qué son los datos sintéticos?



Sobre el autor

Manuel Delgado Tenorio

Manuel Delgado Tenorio es asesor independiente de estrategia de datos e inteligencia artificial, profesor e investigador. Cofundó y vendió una *startup* de inteligencia artificial, fue socio de la práctica de Data & Analytics para servicios financieros en EY y responsable de Data Analytics para el sur de Europa en AWS. Es profesor asociado en IE University, donde imparte la primera asignatura de grado sobre Customer Lifetime Value en España, investiga sobre el CLV en su doctorado en la Universidad Complutense de Madrid y es autor del libro "Customer Lifetime Value (CLV): No todos tus clientes valen lo mismo".

manueldelgado.com · **LinkedIn** · **X**

Recibe cada artículo en tu correo con la newsletter Ex Datis, Claritas:
manueldelgado.com/newsletter/suscribete