expertenReport

Scharf auf AI: Warum plötzlich selbst verstaubte Antiquariatsbücher zum begehrten Rohstoff werden

Dirk Pappelbaum

Wer heute ein Antiquariat betritt, erwartet selten Überraschungen. Zwischen vergriffenen Fachbüchern, jahrzehntealten Monografien und wissenschaftlichen Ladenhütern herrscht meist das Gegenteil von Hektik. Doch wie eine Recherche der Washington Post anhand freigegebener Gerichtsunterlagen zeigt, geraten genau diese Bücher zunehmend in den Fokus der Künstlichen Intelligenz. Nicht, weil Menschen sie lesen wollen – sondern weil Sprachmodelle sie verschlingen.

Das Internet als Datenquelle stößt an Grenzen

Die rasante Entwicklung generativer KI basiert auf einem einfachen Prinzip: Je mehr hochwertige Texte einem Sprachmodell zur Verfügung stehen, desto besser kann es Sprache verstehen und erzeugen.

Die ersten Generationen moderner KI-Systeme wurden deshalb überwiegend mit öffentlich zugänglichen Inhalten trainiert. Webseiten, Zeitungsarchive, Diskussionsforen, wissenschaftliche Veröffentlichungen und frei verfügbare Dokumente lieferten Milliarden von Wörtern.

Mit jeder neuen Modellgeneration wächst jedoch der Bedarf an zusätzlichen Trainingsdaten. Gleichzeitig wird deutlich, dass das frei verfügbare Internet keine unerschöpfliche Quelle ist. Viele Inhalte wiederholen sich, manche sind qualitativ schwach, andere rechtlich nicht nutzbar.

Für Unternehmen wie OpenAI, Anthropic oder Google stellt sich deshalb eine neue strategische Frage: Wo finden sich hochwertige Texte, die bislang kaum digital erschlossen wurden?

Die Antwort führt überraschend häufig in Antiquariate.

Gedruckte Bücher als neue Datenquelle

Besondere Aufmerksamkeit erhielt dieses Thema durch ein Urheberrechtsverfahren gegen Anthropic.

Freigegebene Gerichtsunterlagen zeigen, dass das Unternehmen unter dem Projektnamen Project Panama plante, große Mengen gedruckter Bücher anzukaufen, um sie für das Training seiner KI-Modelle zu digitalisieren.

In internen Dokumenten findet sich sogar die Formulierung eines Vorhabens, „destructively scan all the books in the world“. Gemeint war damit keine tatsächliche Digitalisierung sämtlicher Bücher weltweit, sondern eine Vision, die den enormen Bedarf an Trainingsdaten verdeutlicht.

Der technische Ablauf ist vergleichsweise unkompliziert. Die Bücher werden gekauft, der Buchrücken entfernt und sämtliche Seiten automatisiert eingescannt. Anschließend wandelt eine Texterkennung die Inhalte in maschinenlesbare Daten um. Für das Training des Sprachmodells wird nur noch der digitale Text benötigt.

Das gedruckte Exemplar hat seine Aufgabe erfüllt.

Warum gerade alte Fachbücher wertvoll werden

Bemerkenswert ist, welche Bücher besonders gefragt sind.

Es handelt sich häufig nicht um aktuelle Bestseller oder bekannte Romane. Viel interessanter sind wissenschaftliche Monografien, technische Handbücher, medizinische Fachliteratur, historische Nachschlagewerke oder längst vergriffene Standardwerke.

Diese Bücher besitzen oft einen geringen Wiederverkaufswert. Für Sprachmodelle dagegen liefern sie genau jene Vielfalt, die moderne KI benötigt.

Sie enthalten Fachbegriffe, unterschiedliche Sprachregister, seltene Formulierungen und Wissensgebiete, die online nur lückenhaft vertreten sind. Gerade diese sprachliche Diversität verbessert die Qualität großer Sprachmodelle erheblich.

Damit verschiebt sich der ökonomische Wert eines Buches. Entscheidend ist nicht mehr allein, wie viele Menschen es lesen möchten. Relevant wird zunehmend auch sein Potenzial als Datenträger.

Warum Kaufen günstiger sein kann als Lizenzieren

Aus wirtschaftlicher Sicht ist die Strategie nachvollziehbar.

Ein gebrauchtes Fachbuch kostet im Antiquariat häufig nur wenige Euro. Demgegenüber wäre es deutlich aufwendiger, für Millionen einzelner Werke individuelle Lizenzvereinbarungen mit Verlagen auszuhandeln.

Der Erwerb eines physischen Exemplars ermöglicht zunächst den legalen Besitz des Buches. Die juristische Debatte beginnt allerdings bei der Frage, ob dessen Digitalisierung und Nutzung zum KI-Training urheberrechtlich zulässig sind.

In den USA entschied ein Gericht im konkreten Verfahren, dass das Einscannen rechtmäßig erworbener Bücher unter bestimmten Voraussetzungen als Fair Use bewertet werden könne. Diese Entscheidung betrifft jedoch den Einzelfall und beendet die internationale Diskussion keineswegs. Insbesondere das europäische Urheberrecht setzt teilweise andere Maßstäbe.

Eine neue Datenökonomie entsteht

Interessanter als die juristische Auseinandersetzung ist die wirtschaftliche Entwicklung, die dahinter sichtbar wird.

Im Wettlauf um leistungsfähigere KI-Systeme werden hochwertige Sprachdaten zu einer strategischen Ressource. Neben Halbleitern, Energieversorgung und Rechenleistung entwickelt sich Wissen selbst zu einem entscheidenden Wettbewerbsfaktor.

Bibliotheken, Archive und Antiquariate erhalten dadurch eine neue Bedeutung. Sie bewahren nicht nur kulturelles Erbe, sondern verfügen über Textbestände, die für die nächste Generation künstlicher Intelligenz von erheblichem wirtschaftlichem Interesse sein können.

Die KI-Industrie konkurriert damit zunehmend um Inhalte, die jahrzehntelang kaum Beachtung fanden.

Antiquariate vor einer neuen Rolle

Ob sich daraus dauerhaft ein neuer Markt entwickelt, ist noch offen. Dennoch könnte sich die Preisbildung verändern.

Während bislang vor allem Sammlerwert, Seltenheit oder Nachfrage den Preis eines Buches bestimmten, könnte künftig ein weiterer Faktor hinzukommen: sein Wert als Trainingsmaterial.

Gerade unscheinbare Fachbücher könnten dadurch wirtschaftlich interessanter werden, obwohl sie im klassischen Buchhandel kaum noch nachgefragt werden.

Für Antiquariate eröffnet diese Entwicklung neue Perspektiven. Gleichzeitig wirft sie Fragen nach Urheberrecht, kulturellem Erbe und der Zukunft des Buches auf.

Der eigentliche Wandel beginnt erst

Der eigentliche Wandel besteht nicht darin, dass Bücher digitalisiert werden. Das geschieht seit Jahrzehnten.

Neu ist ihre Funktion.

Gedruckte Werke werden zunehmend als Rohstoff einer datengetriebenen Industrie betrachtet. Sie sind nicht mehr ausschließlich Wissensspeicher für Menschen, sondern Trainingsmaterial für Maschinen.

Damit verändert sich die wirtschaftliche Bedeutung des Buches grundlegend. Aus einem kulturellen Gut wird zugleich eine strategische Ressource der globalen KI-Ökonomie.

KI hat das Internet fast leer gelesen

Die Entwicklung großer Sprachmodelle begann mit dem naheliegendsten Rohstoff: dem Internet. Webseiten, Foren, Nachrichten und frei zugängliche Dokumente lieferten Milliarden Wörter für das Training von Systemen wie Claude, ChatGPT oder Gemini.

Doch dieser Vorrat ist endlich. Hochwertige, sprachlich vielfältige Texte werden zunehmend knapp. Für KI-Unternehmen stellt sich deshalb eine neue Frage: Woher kommen die nächsten Daten?

Die Antwort führt überraschend häufig ins Antiquariat.

Versicherungs- und Finanznachrichten

expertenReport

https://www.experten.de/id/4950712/scharf-auf-ki-antiquariatsbuecher-trainingsdaten/