

Möglichkeiten und Grenzen von Textanalytics im eCommerce

Laura Hoyden¹ Frank Rosenthal² Jonas Hausrucking³

Abstract

Die Synchronisation von Produktstammdaten ist eine zentrale Aufgabe im eCommerce, denn schließlich muss sichergestellt werden, dass wirklich die Variante des Artikels ausgeliefert wird, die ein Kunde bestellt hat. Schlüsselssysteme wie EAN/GTIN und UPC leisten dabei einen guten Beitrag, lösen das Problem in der Praxis aber bei weitem nicht vollständig, da viele Handelsgüter erst gar nicht mit herstellerunabhängigen Schlüsseln gekennzeichnet sind oder z.B. Packungsgrößen und Varianten zu häufig verwechselt werden. Die Abbildung von Produktstammdaten anhand der Artikelbezeichnungen ist daher eine Aufgabe für moderne Textanalyseverfahren an der Grenze von strukturierten und natürlichsprachlichen Ansätzen. Als Marktforschungsunternehmen setzt GfK solche Textanalyseverfahren ein, um unterschiedliche Artikelbezeichnungen für eine Produktvariante zu erkennen und für Marktberichte zu normieren.

Im ersten Teil des Vortrags wird erklärt, wie ein proprietäres Machine Learning Verfahren genutzt werden kann, um den Suchraum für die Zuordnung einer unbekanntem Artikelvariante so einzuschränken, dass ein Bearbeiter mit einem Blick die passendste von nur wenigen wahrscheinlichen Varianten wählen und bestätigen kann. Zentral dafür sind ein mehrstufiger Prozess sowie komplementäre Verfahren (Ensemble) in einzelnen Schritten, da Textquellen von unterschiedlicher Güte verarbeitet werden müssen, aber auch die Größe des Suchraums von Warengruppe zu Warengruppe stark differiert.

Im zweiten Teil der Präsentation werden die Voraussetzungen und Herausforderungen untersucht, die für eine vollautomatische Zuordnung zu schaffen bzw. zu lösen sind. Im Kern steht dabei die Frage, wie man die qualifizierten Vorschläge mit einem intervall-normierten Konfidenzmaß versehen kann. Über einen Schwellwert soll dann eine automatische Zuordnung möglich sein, wobei entscheidend ist, dass eine sehr hohe Zuordnungsgüte erreicht wird.

Der Vortrag arbeitet mit vielen konkreten Beispielen aus der Warenwelt von Gebrauchsgütern und gibt einen Einblick in die Praxis eines datengetriebenen Unternehmens, welches Business Intelligence als Produkt anbietet.

Zielgruppe der Veranstaltung sind an Textanalytics interessierte Zuhörer, die idealerweise entsprechende Grundkenntnisse mitbringen.

¹ GfK SE, Nordwestring 101, 90419 Nürnberg, Laura.Hoyden@gfk.com

² GfK SE, Nordwestring 101, 90419 Nürnberg, Frank.Rosenthal@gfk.com

³ GfK SE, Nordwestring 101, 90419 Nürnberg, Jonas.Hausrucking@gfk.com