AI-powered visual and interactive constrained document clustering

Studiengang: MAS Data Science

Der Begriff «data-driven», hinter dem sich die Idee verbirgt, dass Prozesse, Entscheidungen und Aktivitäten einer Unternehmung ausschliesslich auf Daten und nicht auf Intuition oder persönlicher Erfahrung basieren, ist schon seit einiger Zeit in aller Munde. Auch Machine Learning Anwendungen sind ohne Daten, deren Volumen und Verfügbarkeit in den letzten Jahren exponentiell gewachsen ist, nicht möglich.
Doch sind diese Anwendungen damit schon wirklich «datengetrieben»?

Kategorisierung durch Mensch und Maschine

Jeden Monat erhält der Swisscom-Kundensupport viele Tausend E-Mails mit Kundenanliegen, die in eine von derzeit über 1'500 Kategorien eingeordnet und einem geeigneten Bearbeiter zugewiesen werden müssen. Ein mühevoller, zeitraubender Prozess, der mittlerweile von einer Machine Learning Anwendung unterstützt wird, welche die Zuordnung wesentlich schneller durchführen kann und damit täglich viele Stunden Arbeit einsparen hilft.
Doch ähnlich wie der Mensch tut sich auch die Maschine bei der korrekten Zuordnung der Kategorien schwer, die sich teilweise überlappen, redundant oder obsolet sind, weil diese über die Jahre evolutionär gewachsen sind.
Die Qualität der Vorhersage dieser Kategorien durch einen Algorithmus, der anhand dieser widersprüchlichen Klassifizierungen «lernt» muss daher unweigerlich hinter den Möglichkeiten zurückbleiben und erreicht im besten Fall das Niveau eines Menschen.
Der Grund dafür liegt auf der Hand: er ist in der unzureichenden Kategorisierungsstruktur zu suchen.

Ein «datengetriebener» Ansatz für eine Klassifikationsstruktur

In dieser Masterarbeit wird ein neuartiger Ansatz vorgestellt, der sich auf die Daten selbst fokussiert und auf Basis deren inhärenter Muster ein rein datengetriebenes Kategorisierungsschema vorschlägt, welches interaktiv erkundet, optimiert und verfeinert werden kann.
Dies geschieht mit Hilfe der Clusteranalyse, also einem Prozess der eine Menge von Elementen, in diesem Falle die Kunden E-Mails anhand ihrer Ähnlichkeit gruppiert und damit eine inhaltliche Kategorisierung der Eingangsdaten vornimmt.
Nach der notwendigen Bereinigung dieser Dokumente wurden diese in einzelne Worte, sogenannte Tokens, zerlegt und im Anschluss daran in ein geeignetes maschinen-lesbares Format überführt.
Aus verschiedenen Vektorisierungs- und Gewichtungsparametern, der Vektordimension, der Clusterzahl und weiteren Parametern wurde experimentell eine optimale Kombination bestimmt, indem die Vorverarbeitungsschritte und der Clusteringvorgang selbst kontinuierlich erneut durchgeführt und die Metriken am Ende jedes Durchlaufs verglichen wurden.
Neben TF-IDF, einer Methode, die auf der Wortfrequenz basiert, wurden auch Vektorisierungen mit Wort-Embeddings, im Speziellen fastText Modelle verwendet, die sich letztlich als die geeignetere Wahl entpuppten.
Eine moderne Web-Applikation («aicido») ermöglicht das Clustering-Ergebnis interaktiv zu erforschen und gemäss den Bedürfnissen zu optimieren und zu verfeinern. Dies können beispielsweise Bedingungen sein, die erfüllt sein müssen, sogenannte «constraints».
Jeder Cluster erhält eine aussagekräftige Zusammenfassung, sogenannte «key phrases», die gleichzeitig als neue Kategorien verwendet werden.

aicido - das interaktive Clusteringtool

Das Endergebnis

Als Ergebnis erhält man eine optimierte, rein datengetriebene Kategorisierungsstruktur, die mit den Daten «mitwächst» und sich dynamisch dem Business anpasst.
Grösste Nutzniesser sind jedoch Machine Learning Anwendungen, deren Performance sich durch Verwendung der «neuen» Strukturierung in allen Metriken drastisch verbesserte und damit den Vorteil einer solchen datengetriebenen Struktur eindrücklich unter Beweis stellen konnte.

Georg Andreas Jaksch

georgandreas.jaksch@gmail.com