Beratung Text-and-Data-Mining („TDM-Support“)

15.07.2024

Die Verfügbarkeit großer digitaler Datenmengen und die rasante Entwicklung computerbasierter Analysemethoden erlauben den Wissenschaften, auf neuen Wegen Informationen zu extrahieren und Wissen zu generieren. Am Anfang dieses Weges steht häufig das sog. Text-and-Data-Mining (TDM). Mit Hilfe von Methoden aus der Statistik, dem maschinellen Lernen oder der künstlichen Intelligenz wird hierbei eine möglichst autonome und effiziente Identifizierung von Mustern innerhalb großer Datensätze vorgenommen. Das Text-Mining stellt hierbei ein Teilgebiet des Data-Mining dar und „verwendet regelbasierte, statistische sowie neuronale Verfahren und erlaubt somit innovative Anwendungen, bei denen sehr große Mengen an Text sehr schnell und sehr umfangreich ausgewertet werden“.[1]

Um die Fachinformationsdienste (FID) in der Entwicklung und Durchführung eigener TDM-Vorhaben zu unterstützen, bietet das Kompetenzzentrum für Lizenzierung (KfL) nun mit dem TDM-Support einen Beratungsservice zu Fragen des Text-and-Data-Mining an. Die Beratungsleistungen des KfL, das von der Deutschen Forschungsgemeinschaft (DFG) gefördert wird, erfolgen entgeltlos.[2]

Zielgruppe des TDM-Support

Der TDM-Supportservice richtet sich ausschließlich an Fachinformationsdienste, die TDM erproben oder einsetzen möchten. Die Kommunikation mit der Fachcommunity findet ausschließlich über die FID statt.

Angebotene Leistungen des TDM-Supports

- Beratung: Der TDM-Support des KfL bietet individuelle Beratungen für FID an, um ihnen bei der Planung und Durchführung von TDM-Projekten zu helfen. Dabei werden verschiedene Aspekte des TDM wie die Auswahl geeigneter Datenquellen – darunter gegebenenfalls auch FID-Lizenzen –, die Anwendung von Analysewerkzeugen und die Interpretation von Ergebnissen behandelt.

- Technische Unterstützung: Der TDM-Support berät bei der Einrichtung und Nutzung von TDM-Tools und -Plattformen. Dies umfasst die Installation von Software, die Konfiguration von Analyseumgebungen und die Behebung technischer Probleme während des Forschungsprozesses. Die Expertise unseres Teams liegt schwerpunktmäßig auf Tools in der Programmiersprache Python, es kann aber auch bei Vorhaben, die auf der Basis von Java oder R durchgeführt werden, unterstützt werden. Es ist allerdings nicht vorgesehen, dass der TDM-Support Programmierarbeiten für die FID übernimmt. Es handelt sich um einen Beratungsservice auf konzeptioneller Ebene. Für Programmierarbeiten ist jedoch eine Finanzierung über Drittmittel denkbar.[3]

- Methodenentwicklung und -optimierung: Der TDM-Support berät die FID bei der Entwicklung und Optimierung von TDM-Methoden, um effektivere und präzisere Ergebnisse zu erzielen. Dies beinhaltet die Anpassung von Analysestrategien an spezifische Forschungsfragen sowie die Bewertung und Verbesserung der Methodeneffizienz und Skalierbarkeit.

- Rechtliche Fragen: Da TDM-Projekte häufig mit Unsicherheiten auf rechtlichem Gebiet wie dem Urheber- oder Datenschutzrecht verbunden sind, bietet der TDM-Support auch hier den FID Unterstützung, indem die KfL-RechtsexpertInnen TDM-spezifische-Rechtsfragen klären.[4]

---

[1] Biemann, C., Heyer, G. & Quasthoff, U. (2022). *Wissensrohstoff Text: Eine Einführung in das Text Mining* (2., wesentlich überarbeitete Auflage). Springer. S. 1. Online verfügbar unter https://doi.org/10.1007/978-3-658-35969-0. Zum Einstieg in das Thema bietet sich daneben auch an *forTEXT: Literatur digital erforschen.* https://fortext.net/.

[2] Sie dienen dem Erhalt der öffentlichen Informationsinfrastruktur und damit der Wahrnehmung der öffentlichen Aufgabe „Wissenschaft“ und finden ihre gesetzliche Grundlage in § 3 NHG, der die Zuweisung des Projektinhalts zum staatlichen Handeln normiert.

[3] Weitere Informationen hierzu sind zu finden unter https://www.sub.uni-goettingen.de/digitale-bibliothek/service-research-software-engineering/.

[4] Als einführende Lektüre zu diesem Thema siehe Rack, F. (2024). Rechtsfragen zur generativen KI. *ABI Technik*, *44*(1), 39–47.

Bereiche des TDM-Support

Das KfL-Team berät und unterstützt die FID in ihren TDM-Projekten über die verschiedenen Stationen des *life cycle* eines solchen Projekts hinweg. Ein solcher *life cycle* beginnt in der Regel damit, die Ziele und Anforderungen des Projekts zu verstehen und zu dokumentieren. Diese erste Phase, in Anlehnung an das im TDM weit verbreitete CRISP-DM-Modell (Cross-Industry Standard Process for Data-Mining) als „Problemverständnis" bezeichnet, umfasst üblicherweise die Definition der Projektziele und -fragen, die Bewertung der aktuellen Situation, die Erstellung eines Projektplans und die Festlegung von Erfolgskriterien.[5] Die nächste der insgesamt sechs Phasen in diesem Modell, die typischerweise iterativ durchlaufen werden, umfasst die Datensammlung, erste Datenexplorationen, eine Beschreibung der Daten und eine Bewertung der Datenqualität. Auf diese Phase des „Datenverständnisses" folgt die in der Regel sehr arbeits- und zeitintensive Phase der „Datenvorbereitung". Sie beinhaltet Datenbereinigung (z.B. Umgang mit fehlenden Werten oder heterogenen Formatierungen), Datenintegration und -transformation, Merkmalserstellung (Feature Engineering) und die Auswahl relevanter Daten. Hier verfügt das KfL-Team über Kenntnisse in einer Vielzahl an Tools und Methoden insbesondere im Bereich Text Mining wie z.B. Tokenisierung, Stemming, Normalisierung und Named Entity Recognition (NER) und kann in allen Schritten Beratung anbieten.

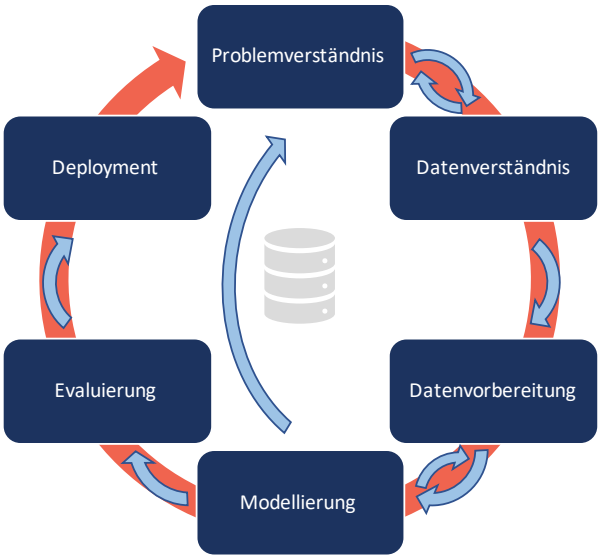

Abb. 1: Life cycle eines TDM-Projekts nach dem CRISP-DM Modell

Auch in der Modellierungsphase, die die Auswahl der Modellierungstechniken, den Aufbau und das Training der Modelle, die Modellbewertung und -optimierung umfasst, kann das KfL-Team die FID durch eine Vielzahl an Methoden und Werkzeugen zu den für ihre Zwecke am besten geeigneten leiten z.B. auf dem Gebiet der Frequenzanalyse, Kookkurrenzanalyse, Clusteranalyse oder Klassifikation etc. Das KfL-Team hilft in der Evaluierungsphase, in der die Ergebnisse überprüft und die Modelle hinsichtlich ihrer Effektivität und Genauigkeit beurteilt werden. Schließlich berät KfL auch in der letzten Phase, dem Deployment, in der die Modelle in die Praxis umgesetzt werden – sei es durch die Implementierung in ein operatives System oder die Bereitstellung von Ergebnissen und Berichten.

Im Ergebnis können unsere Experten die FID über den ganzen Prozess von unstrukturierten Rohdaten bis hin zu Visualisierungen der Ergebnisse der Analyse beraten.

Online verfügbar unter https://doi.org/10.1515/abitech-2024-0005.

[5] Chapman, P. (2000). *CRISP-DM 1.0: Step-by-step data mining guide.* Online verfügbar unter https://www.kde.cs.uni-kassel.de/wp-content/uploads/lehre/ws2012-13/kdd/files/CRISPWP-0800.pdf. Vgl. auch Nguyen, G., Dlugolinsky, S., Bobák, M., Tran, V., López García, Á., Heredia, I., Malík, P. & Hluchý, L. (2019). Machine Learning and Deep Learning frameworks and libraries for large-scale data mining: a survey. *Artificial Intelligence Review*, *52*(1), 77–124. Online verfügbar unter https://doi.org/10.1007/s10462-018-09679-z.

Künftige Entwicklungen

Auf der Basis einer fortlaufenden Evaluation prüfen wir, wie wir den TDM-Support gemäß den Bedürfnissen der FID weiterentwickeln können. Vor diesem Hintergrund führt das KfL-Team nach erfolgter Beratung eine kurzen Nachbefragung durch.

Kontakt

Weitere Informationen zum TDM-Support und den angebotenen Dienstleistungen sind jederzeit beim KfL-Team unter der E-Mail-Adresse tdm-support@fid-lizenzen.de zu erhalten. Bei der Kontaktaufnahme werden vorzugsweise bereits für die Beratung relevante Informationen bezüglich des Forschungsdesigns, des Formats, Umfangs und der Qualität der ausgewählten Daten, der Software und des konkreten TDM-Vorhabens mitgeteilt. Falls der Anfragesteller Beispiel- oder Dummy-Daten bereitstellt, kann das KfL-Team bereits eine erste Dateninspektion durchführen. Alle uns mitgeteilten Informationen werden in einem Repositorium der GWDG (OwnCloud) gespeichert, wobei der Datenschutz nach DSGVO unbedingt gewährleistet wird.

Auch bei allen anderen Fragen im Kontext TDM steht das Serviceteam des KfL den FID gerne beratend zur Verfügung!

Nicole Altmeier, Tillmann Dönicke, Mathias Göbel, Niklas Spies (SUB Göttingen)