Technology
Arts Sciences
TH Köln

Pressemitteilung

Nr. 13 vom 16. Februar 2021

Künstliche Intelligenz für den Datenjournalismus

Referat Kommunikation und Marketing
Presse- und Öffentlichkeitsarbeit
Christian Sander
0221-8275-3582
pressestelle@th-koeln.de

Technische Hochschule Köln

Postanschrift:
Gustav-Heinemann-Ufer 54
50968 Köln

Sitz des Präsidiums:
Claudiusstraße 1
50678 Köln

TH Köln und Science Media Center Germany entwickeln Lösung für die Informationsextraktion

Datenjournalisten sammeln und analysieren öffentlich zugängliche Daten und bereiten die darin verborgenen Informationen für ihre Medien auf. Dafür müssen sie häufig mit Daten arbeiten, die nur in unstrukturierter Form vorliegen. Dies macht eine automatisierte Auswertung schwierig. Im Forschungsprojekt „Journalistic Information Extraction" (JoIE) möchten die TH Köln und das Science Media Center Germany daher ein Tool entwickeln, das mit solchen Datenquellen umgehen und diese journalistisch nutzbar machen kann.

„Die Daten, in denen Journalisten nach Informationen suchen, können ungemein vielfältig sein: Es handelt sich um Texte, Tabellen oder Grafiken, Dokumente unterschiedlichen Typs wie Word, PDF oder E-Mail oder um Webseiten, die zudem noch höchst unterschiedlich formatiert sein können. All das macht es sehr schwierig, zuverlässige und konsistente Regeln zu definieren, nach denen eine automatisierte Auswertung erfolgen könnte", erläutert Prof. Dr. Philipp Schaer vom Institut für Informationswissenschaft der TH Köln die Problemstellung.

Auf Grundlage der beiden Open-Source-Werkzeuge Workbench und Fonduer soll eine Lösung entstehen, die unstrukturierte Daten in eine strukturierte und damit auswertbare Form bringt. Workbench erlaubt unter anderem die Extraktion von Webdaten. Fonduer verwendet künstliche Intelligenz, um automatisch Extraktionsmuster zum Beispiel zur Erkennung von Tabellen zu lernen.

„Kernidee unseres Projektes ist die Synthese der Nutzerfreundlichkeit von Workbench mit der hervorragenden Extraktionsleistung von Fonduer. Dabei geht es auch darum, komplexe Eingabehilfen zu entwickeln, mit denen Regeln für die Datenbearbeitung ohne Programmierkenntnisse erstellt und entwickelt werden können", sagt Björn Engelmann, der im Rahmen von JoIE seine Doktorarbeit verfassen wird.

Um die spezifischen Anforderungen von Redaktionen und Datenjournalisten zu erfahren, sind Experteninterviews und gegebenenfalls Umfragen geplant. „Mit unserem Tool möchten wir den State of the Art der Datenverarbeitung für Redakteurinnen und Redakteure verfügbar machen, damit sie Informationen aus der Wildnis des Internets schnell und zuverlässig beschaffen können. Da diese oftmals mit begrenzten Ressourcen arbeiten müssen, wird unsere Lösung kostenlos und als Open-Source-Software verfügbar sein", sagt Dr. Meik Bittkowski, Leiter Forschung und Entwicklung beim Science Media Center Germany.

Das Forschungsprojekt „Journalistic Information Extraction" (JoIE) wird über drei Jahre von der Klaus Tschira Stiftung gGmbH gefördert. In dieser Zeit soll das Grundgerüst der Anwendung entstehen. Die Überführung in ein für Externe nutzbares System ist für eine optionale Projektverlängerung von zwölf Monaten angedacht.

Masterstudiengängen eingeschrieben. Die TH Köln gestaltet Soziale Innovation – mit diesem Anspruch begegnen wir den Herausforderungen der Gesellschaft. Unser interdisziplinäres Denken und Handeln, unsere regionalen, nationalen und internationalen Aktivitäten machen uns in vielen Bereichen zur geschätzten Kooperationspartnerin und Wegbereiterin.