VolkswagenStiftung

Data Extraction and Interactive Visualization of Unexplored Textual Datasets for Investigative Data-Driven Journalism

Initiative: Wissenschaft und Datenjournalismus

Bewilligung: 20.10.2015

Laufzeit: 9 Monate

Projekt-Website: http://www.newsleak.io/

This project combines the latest scientific findings in natural language processing and information visualization in a powerful exploration tool that provides accelerated access to text document collections for investigative data-driven journalism. Whereas today's data-driven journalism is mostly based on structured data, i.e. data available in spreadsheets or databases, this project applies the principles of data journalism to unstructured text documents. Using statistical methods of language processing, such as named entity recognition and keyphrase extraction, important elements from these text documents can be identified and visualized in a network. A journalist can browse this network to quickly grasp the content of the text document collection, view the developments of groups of entities over time, annotate the elements of the network with labels and comments to attach a layer of interpretation to the data, and eventually publish part of the network along with an (online) article that is developed in a data-driven fashion from the findings in the document collection. The project consortium combines complementary expertise from "Der Spiegel" and the TU Darmstadt: investigative journalism, information visualization and text mining.

Projektbeteiligte

Prof. Dr. Chris Biemann
Technische Universität Darmstadt
Fachbereich Informatik
Sprachtechnologie
Darmstadt

Kathrin Ballweg
Technische Universität Darmstadt
Fachbereich Informatik
Grafisch Interaktive Systeme
Darmstadt

Dr. Alexander Panchenko
Technische Universität Darmstadt
Fachbereich Informatik
FG Language Technology
Darmstadt

Marcel Rosenbach
Der SPIEGEL
Berlin

Dr. Michaela Regneri
Der SPIEGEL
Hamburg

Dr.-Ing. Tatiana von Landesberger
Technische Universität Darmstadt
Fachbereich Informatik
Interactive Graphics Systems Group
Darmstadt

Open Access-Publikationen

new/s/leak Information Extraction and Visualization for an Investigative Data Journalists
new/s/leak A Tool for Visual Exploration of Large Text Document Collections in the Journalistic Domain
new/s/leak Anforderungsanalyse einer interaktiven Visualisierung für Data-Driven Journalism.
Guidance for Multi-Type Entity Graphs from Text Collections