Pseudonymisierung

Eder, Elisabeth; elisabeth.eder@aau.at

Pseudonymisierung stellt ein mögliches Verfahren zum Schutz persönlicher Daten in Korpora und Datensammlungen dar. Dabei wird Information, die eine Identifikation natürlicher Personen ermöglicht, durch realistische Bezeichnungsalternativen ersetzt, um die Integrität der Daten zu wahren. 'Irene Adler' könnte etwa durch 'Hermine Granger' ersetzt werden. Dazu sind zwei Schritte notwendig: die Erkennung der individuenidentifizierenden Information und die Substitution der entsprechenden Textteile. Erstere Aufgabe weist Ähnlichkeiten zur *Named Entity Recognition* auf. Die möglichen Entitätstypen können je nach Texttyp jedoch abweichen und differenzieren beispielsweise Personennamen in Nachnamen und weibliche und männliche Vornamen aus, erfassen verschiedene Unterkategorien für Orte oder enthalten IDs sowie Kontaktdaten (Telefonnummer, URIs, E-Mail-Adressen) etc. (siehe z. B. Stubbs/Uzuner 2015b oder Eder et al. 2019) Aufgrund dessen müssen vorhandene Tools für NER und *Sequence*-Tagging potentiell auf dementsprechend annotierten Daten trainiert werden. Explizit für die automatische Erkennung von personenidentifizierender Information entwickelte Programme existieren vor allem für den medizinischen Bereich. (Überblick z. B. in Stubbs et al. 2015c, 2017) In einem zweiten Schritt werden die gefundenen Entitäten durch realistische Alternativen substituiert. Mit dem *Surrogate Generation-Tool* lassen sich Ersetzungen fürs Deutsche automatisch generieren.

Literatur:

- Eder, Elisabeth; Krieg-Holz, Ulrike; Hahn, Udo: De-identification of emails: pseudonymizing privacy-sensitive data in a German email corpus. In: Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP. Varna, Bulgaria: 2019, S. 259–269.
- Medlock, Ben: An introduction to NLP-based textual anonymization. In: Proceedings of the Fifth International Conference on Language Resources and Evaluation LREC. Genoa, Italy: 2006.
- Stubbs, Amber; Uzuner, Özlem; Kotfila, Christopher; Goldstein, Ira; Szolovits, Peter: Challenges in synthesizing surrogate PHI in narrative EMRs. In: Medical Data Privacy Handbook. Cham, Heidelberg, New York, Dordrecht, London: 2015, S. 717–735.
- Stubbs, Amber; Uzuner, Özlem: Annotating longitudinal clinical narratives for de-identification: the 2014 i2b2/UTHealth corpus. In: Journal of Biomedical Informatics 58: 2015, S. 20–29.
- Stubbs, Amber; Kotfila, Christopher; Uzuner, Özlem: Automated systems for the de-identification of longitudinal clinical narratives: overview of 2014

i2b2/UTHealth Shared Task Track 1. In: Journal of Biomedical Informatics 58: 2015, S. 11–19.

- Stubbs, Amber; Filannino, Michele; Uzuner, Özlem: De-identification of psychiatric intake records. Overview of 2016 CEGS NGRID Shared Task Track 1. In: Journal of Biomedical Informatics 75: 2017, S. 4-18.
- Yeniterzi, Reyyan; Aberdeen, John S.; Bayer, Samuel; Wellner, Benjamin; Hirschman, Lynette; Malin, Bradley A.: Effects of personal identifier resynthesis on clinical text de-identification. In: Journal of the American Medical Informatics Association 17: 2010, S. 159–168.

Software:

spacy , flair, German NER, GermaNER, Surrogate Generation, NeuroNER

Verweise:

Named Entity Recognition, Digitalisierung: Rechtliches, NLP

Themen:

Natural Language Processing

Zitiervorschlag:

Eder, Elisabeth. 2021. Pseudonymisierung. In: KONDE Weißbuch. Hrsg. v. Helmut W. Klug unter Mitarbeit von Selina Galka und Elisabeth Steiner im HRSM Projekt "Kompetenznetzwerk Digitale Edition". URL: https://gams.uni-graz.at/o:konde.159