NEWS EYE

A Digital Investigator for Historical Newspapers

In diesem Projekt arbeiten gemeinsam
Geisteswissenschaftler*innen,
Informatiker*innen und Bibliothekar*innen
an historischen Zeitungen
aus drei Nationalbibliotheken
(Österreich, Finnland, Frankreich)

https://www.univie.ac.at/newseye/

www.newseye.eu @newseyeeu

NEWS EYE

Semantische Textan-reicherung

von OCR-Inhalten

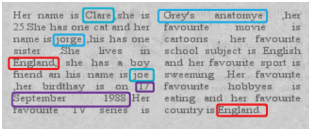
Kontakt:
Antoine Doucet
antoine.doucet@univ-lr.fr

In der semantischen Textanreicherung werden Dokumente analysiert und anschließend semantische Metadaten zu deren Inhalten hinzugefügt. Bei der Vielzahl an semantischen Metadaten legen wir den Fokus auf benannte Entitäten (named entities). Genauer gesagt, wollen wir diese Entitäten in Dokumenten erfassen und zu einer Wissensdatenbank zusammenführen. Zusätzlich werden diese Entitäten mit der Haltung des Textes, in dem sie erwähnt werden, in Verbindung gebracht.

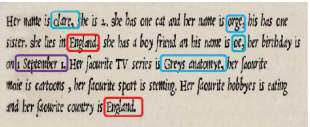
Je nach Zustand der Dokumente, bedingt durch Alterung, schlechte Lagerbedingungen und/oder die schlechte Qualität der verwendeten Druckermaterialien, kann die OCR jedoch zahlreiche Fehler aufweisen. Diese Fehler reduzieren die Leistung aller nachfolgenden Verfahren der natürlichen Sprachverarbeitung (z. B. NER, NEL und Stance Detection).

Als Lösung dieser Probleme werden in NewsEye Ansätze auf der Grundlage neuester Techniken (Neuronale Netze und Deep Learning) entworfen, die robust gegenüber OCR-Problemen und sprachunabhängig sind. Die Ansätze von NewsEye erreichten bei der CLEF HIPE 2020 Competition zu NER- und NEL-Aufgaben in Englisch, Französisch und Deutsch von insgesamt 13 Teilnehmern in 50 von 52 Bestenlisten den ersten Platz.

Angemessene Qualität



Mittlere Qualität



Schlechte Qualität