TECHNISCHE UNIVERSITÄT CHEMNITZ

Kurze Einführung „Text- und Datamining“

Agenda

1. Was ist Text- und Data-Mining?
   - Definition
   - Beispiele
2. Rechtlicher Rahmen – was ist gestattet?
   - Urheberrecht
   - Policies einzelner Anbieter
   - Open Access und TDM
3. Rolle der Bibliothek(en)
4. Ausblick
5. Weiterführende Informationen

1. Was ist Text- und Data-Mining?

Definition:

Unter Text- und Data-Mining (TDM) fasst man verschiedene Forschungsmethoden zusammen, bei der sehr große Mengen meist unstrukturierter Daten zunächst systematisch und maschinenlesbar aufbereitet und schließlich mithilfe von computergestützten Analysen automatisiert auf Muster, Korrelationen und andere forschungsrelevante Zusammenhänge hin untersucht werden.

1. Was ist Text- und Data-Mining?

Beispiele

- „CORD-19“ → Korpus zu Covid-19-Forschung: https://www.semanticscholar.org/cord19/

  → TDM-Anwendung, z.B. „SciFact“: https://scifact.apps.allenai.org

- „GDELT“ → Projekt zur Analyse weltweiter Nachrichten-Berichterstattung (1979-heute)

  → TDM-Anwendung auf Grundlage des Korpus „Television News Archive“: https://api.gdeltproject.org/api/v2/summary/summary?d=iatv

2. Rechtlicher Rahmen – Was ist gestattet?

Urheberrecht – Theoretische Grundlagen

- Bis 2018 → Recht auf Einsatz von TDM-Methoden musste vom Rechteinhaber gestattet werden (z.B. Verlag)
- Seit 2018 → „Gesetz zur Angleichung d. Urh.-Rechts an aktuelle Erfordernisse der Wissensgesellschaft“ (UrhWissG)

→ Neuregelung: Welche Nutzungen sind in Bildung + Wissenschaft ohne Zustimmung d. Urheber + sonst. Rechteinhaber gesetzlich erlaubt

- Recht auf TDM ist nun mit §60 d UrhG gesetzlich verbrieft
- Zweck der Forschung darf dafür ausschließlich nicht-kommerzieller Natur sein
- UrhG schlägt Vertragsrecht

→ Recht auf TDM setzt sich gegen vertragliche Ausschlussklauseln durch, sofern diese nach dem 1.3.2018 abgeschlossen

2. Rechtlicher Rahmen – Was ist gestattet?

Urheberrecht – Praktische Bedeutung

- Zur Erstellung von Korpora dürfen urheberrechtlich geschützte Daten kopiert und bearbeitet werden
- Zugang darf *„einem […] abgegrenzten Kreis von Personen für die gemeinsame […] Forschung sowie einzelnen Dritten zur Überprüfung […] öffentlich zugänglich“* gewährt werden (§ 60d UrhG Abs 1 Satz 2)
- Korpus muss nach Beendigung der Forschungsarbeiten gelöscht werden
- Darf der Bibliothek zur Archivierung übermittelt werden

2. Rechtlicher Rahmen – Was ist gestattet?

Urheberrecht - Perspektive

- UrhWissG zunächst befristet gültig bis 01.03.2023 → Evaluation
- Europäische Ebene → 2 Gesetzesentwürfe zur „Anpassung des Urheberrechts an die Erfordernisse des digitalen Binnenmarkts“
- Neu: „Rechtsinhaber können Nutzungen […] untersagen.“ → unter Einhaltung formaler Kriterien
- Gesetzesentwürfe müssen bis Sommer 2021 in nationales Recht überführt sein
- Kann die Wissenschaftsschranke des UrhWissG beeinträchtigen
- Ggfs. muss TDM dann wieder grundsätzlich verhandelt werden

2. Rechtlicher Rahmen – Was ist gestattet?

Policies und Tools von Verlagen + Informations-Providern

- Einige wiss. Verlage haben – unabhängig von nationalen Urheberrechtsregelungen – generelle Policies erlassen, die regeln, was bzgl. TDM mit ihren Daten gestattet ist
- Einige Anbieter stellen zusätzlich eigene Korpora, Werkzeuge und Zugänge bereit, die TDM erleichtern und regeln sollen

2. Rechtlicher Rahmen – Was ist gestattet?

Policies und Tools von Informations-Providern (Beispiele)

- CrossRef = Registrierungsagentur für DOI:
  - eigene API
  - anbieterübergreifendes Mining
  - Deduplizierung von Ressourcen, die auf mehreren Plattformen existieren
  - Konsistenz der Daten im Korpus bleibt gewahrt, weil DOIs permanente Identifier

- Springer Nature:
  - eigene Policy –regelt TDM aus Verlagssicht
    - (z.T. restriktiver als das Urheberrecht)
  - eigene API für TDM: https://dev.springernature.com/
  - gesonderte Funktionen für OA-Content
  - Kostenpflichtige API für lizenzpflichtige Inhalte
  - Eigenes Whitepaper zu TDM

2. Rechtlicher Rahmen – Was ist gestattet?

Open Access und TDM

- Offener Zugang zu Inhalten im Sinne von Open Science erleichtert die Durchführung von TDM
- standardisierte, maschinenlesbare und Open-Content-gerechte Lizenzen tragen zur rechtssicheren Anwendung von TDM-Methoden auf Daten- und Textkorpora bei
- Inhalte müssen nicht mehr lizenziert oder Erlaubnis eingeholt werden
- Einige Verlage bieten bereits gesonderte Zugänge für offenen Content auf ihren Plattformen, z.B. Springer
- Open-Access-Transformation wird das TDM in Zukunft erleichtern und neue Möglichkeiten schaffen

3. Rolle und Aufgaben der Bibliothek(en)

- Wissenschaftsschranke im UrhG → ABER: welche Ressourcen sind genau lizenziert?
- Wir helfen bei Anfragen an Lizenzinhaber zur technischen Umsetzung von TDM!
  - Kann der Zugang zu Schnittstellen ausgehandelt/eingerichtet werden?
  - Ankündigung von Abfragen beim Anbieter → Verhinderung der Sperrung
- Aufbewahrung und Langzeitarchivierung von Text- und Datenkorpora → Realisierung noch zu klären
- Retrodigitalisierung von Ressourcen unter freien Lizenzen
- 2014 → erfolgreiche Initiative von LIBER gegen restriktive Elsevier-TDM-Lizenz

4. Ausblick

- Derzeit: **nicht** notwendig, Recht auf TDM in Lizenzverträgen mitzuverhandeln
  - Kann sich ab Sommer 2021 ändern
  - Sinnvoll: weiterführende Rechte mit Lizenzgebern verhandeln, z.B. Nutzung komfortable technische Umsetzung (APIs, Abfragehäufigkeit, etc.)
- OA-Transformation wird Erleichterung bringen
- Dass viele (Text-)Quellen noch als PDF statt als XML veröffentlicht werden, ist ein Manko
- z.T. unklare Lizenzen → z. Bsp. Elsevier + "freier Zugang" zu Corona-Forschung
- Mining von bestimmten Inhalten nicht aktiv unterstützt
  - z.B. Image-Mining bei Springer
- Nachnutzung muss künftig besser und weiter geregelt sein:
  - FDM, Anschlussforschung und Reproduzierbarkeitsprüfung
    → Es muss im Urheberrecht mehr erlaubt sein, als nur die Archivierung der Korpora

5. Weiterführende Informationen

- Dynamische Quellen-Sammlung zu urheberrechtlichen Aspekten rund um Korpusbildung, Methoden des Text- und Data-Mining und Anschlussverwendung:
  https://www.zotero.org/groups/2547430/text-_und_data-mining_tdm/library
- Informationen der UB Chemnitz:
  https://www.tu-chemnitz.de/ub/suchen-und-finden/recherche/datamining.html