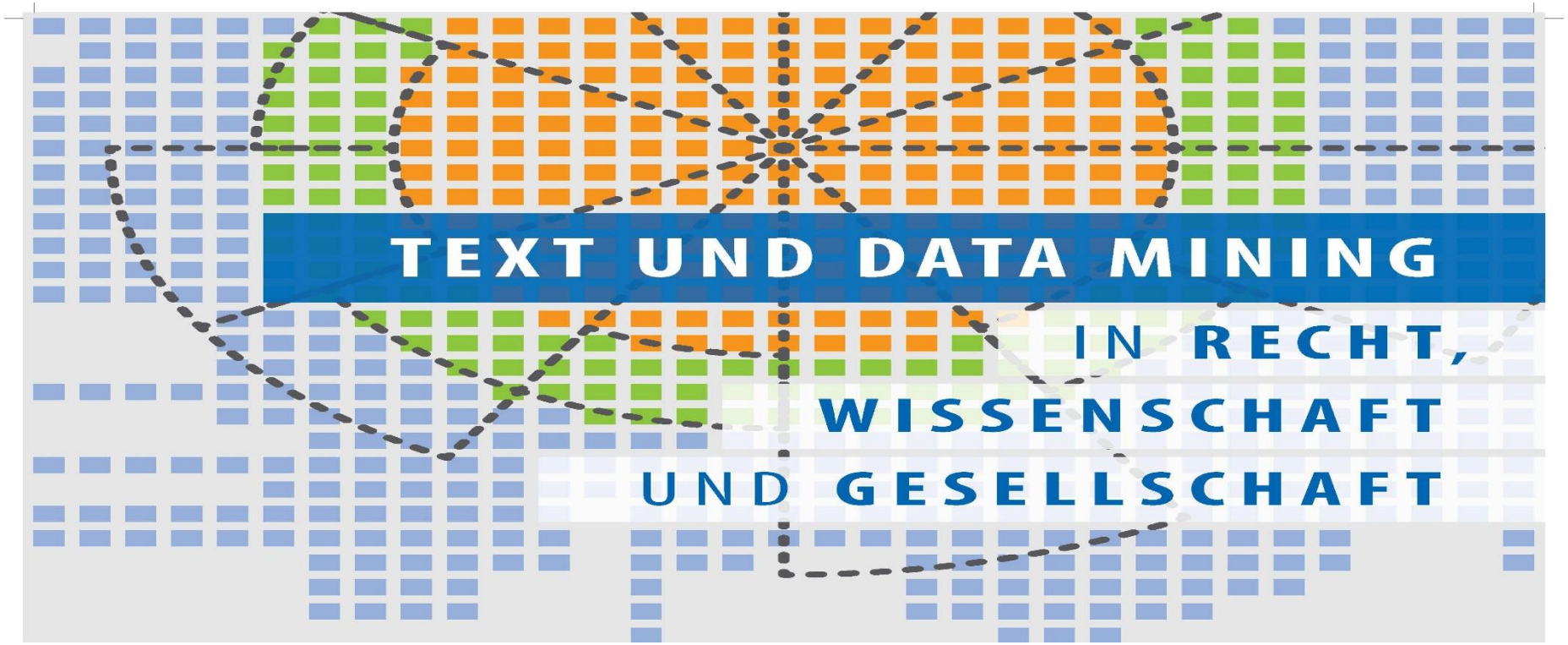
Universität Trier

#tdm18trier



Tagungswebseite:
http://www.text-und-data-mining.de/

Einführung in das Text und Data Mining

Prof. Dr. Benjamin Raue | Prof. Dr. Christof Schöch

Annäherung: Text und Data Mining

Text and Data Mining (TDM)

"the discovery […] of new, previously unknown information, by automatically extracting and relating information from diferent […] resources, to reveal otherwise hidden meanings" (Hearst, 1999)

Interdisziplinäres Methodeninventar

- Machine Learning
- Statistik
- Information Retrieval / Knowledge Discovery
- Computerlinguistik / Natural Language Processing

Gegenstände

- unstrukturierte oder schwach strukturierte Daten (bspw. Texte, Bilder, Audio/Video)
- strukturierte Daten (bspw.
- Messdaten, digitale Spurdaten, Befragungsdaten)

Anwendungshorizonte

Psychologie, Soziologie, Informatik, Linguistik Literaturwissenschaften, Medienwissenschaften, Geschichte, Kunstgeschichte uvm.

Rohdaten, Korpus, Forschungsdaten

1. Vorbereitung des Korpus

- Rohdaten sammeln
- Artefakte entfernen
- Strukturierung, Normalisierung
- Kategorisierung (Metadaten)

⇨ *„Ursprungsmaterial“ iSv. § 60d UrhG*

⇨ *„Korpus“ iSv. § 60d UrhG*

2. Auswertung des Korpus

- Merkmale extrahieren
- Explorative Analysen
- Hypothesengeleitete Analysen
- Visualisierung / Interpretation

… eine kurze rechtliche Einführung

Ursprungsmaterial

→ urhR Schutz (Werk/LSR)

UrhR schützt **nicht** den **Inhalt**, sondern die „**Hülle**“

Ausnahme: fiktionale Werke

- Texte
- Bilder
- Videos
- Tonaufnahmen

Datenbanken

→ Vervielfältigungen iSv. § 16 UrhG

→ Verarbeitung iSv. Art. 4 Nr. 3 DSGVO

- Speicherung
- Normalisierung etc.
- Laden in Arbeitsspeicher

- Art. 6 Abs. 1 DSGVO
  - Nr. 1: Einwilligung
  - Nr. 2: Berechtigte Interessen
- Art. 89 DSGVO

Erlaubnistatbestände („Schranken“)? Session 3

- § 44a UrhG: vorübergehende Vervielfältigungen
- § 60d UrhG: für die wissenschaftliche Forschung
- …

Korpus Session 2

- entstehen *zusätzliche* Rechte durch Sammlung/Bearbeitung?

→ Löschungspflicht Session 4

- § 60d III UrhG: „nach Abschluss der Forschungsarbeiten“

§ 60d UrhG Text und Data Mining

(1) Um eine Vielzahl von Werken (**Ursprungsmaterial**) **für die wissenschaftliche Forschung** automatisiert auszuwerten, ist es **zulässig**,

1. das **Ursprungsmaterial** auch automatisiert und systematisch **zu vervielfältigen**, **um** daraus insbesondere durch Normalisierung, Strukturierung und Kategorisierung ein auszuwertendes **Korpus** **zu erstellen**, und
2. das **Korpus** einem bestimmt **abgegrenzten Kreis von Personen** für die gemeinsame wissenschaftliche Forschung sowie einzelnen Dritten zur Überprüfung der Qualität wissenschaftlicher Forschung **öffentlich zugänglich zu machen.**

Der Nutzer darf hierbei **nur nicht-kommerzielle Zwecke** verfolgen.

(2) Werden **Datenbankwerke** nach Maßgabe des Absatzes 1 genutzt, so gilt dies **als übliche Benutzung** nach § 55a Satz 1. Werden unwesentliche Teile von Datenbanken nach Maßgabe des Absatzes 1 genutzt, so gilt dies mit der normalen Auswertung der Datenbank sowie mit den berechtigten Interessen des Datenbankherstellers im Sinne von § 87b Absatz 1 Satz 2 und § 87e als vereinbar.

(3) Das **Korpus** und die **Vervielfältigungen des Ursprungsmaterials** sind **nach Abschluss der Forschungsarbeiten zu löschen**; die öffentliche Zugänglichmachung ist zu beenden. **Zulässig** ist es jedoch, das **Korpus** und Vervielfältigungen des **Ursprungsmaterials** den in **den §§ 60e und 60f genannten Institutionen** zur dauerhaften Aufbewahrung **zu übermitteln**.