Text Mining als Methode der qualitativen Sozialforschung

VO Qualitative Methoden der empirischen Sozialforschung, 3.6.2015

Nikolaus Pöchhacker, Institut für Höhere Studien

Was ist Text Mining?

- Analyse von unstrukturiertem Text
- Basiert auf Algorithmen
- Suche nach (Sinn-)Strukturen im Text
- Oft ein „uninformiertes“ Verfahren

Warum machen wir das?

- Empirische Sozialforschung basiert sehr stark auf Textanalyse
- Oft große Datenmengen
- Generiert neue Perspektiven auf die Texte
- Geschwindigkeit!

Strukturierte vs. Unstrukturierte Daten

- Strukturierte Daten
  - Haben eine vordefinierte Form
  - Lassen sich formal beschreiben
  - Beispiele: Excel Spreadsheets, Graphen, etc.

- Unstrukturierte Daten
  - Zeigen keine offensichtlichen Muster
  - Formale Beschreibung schwer
  - Beispiele: Diese Folien, ihre Bachelorarbeit, etc.

Arten des Text Minings

- Search and Information Retrieval (IR): Abspeichern und wiederauffinden von Dokumenten anhand von Schlüsselwörtern

Elder IV, J., & Hill, T. (2012). Practical text mining and statistical analysis for non-structured text data applications. Academic Press.

Arten des Text Minings

- Search and Information Retrieval (IR): Abspeichern und wiederauffinden von Dokumenten anhand von Schlüsselwörtern
- Document Clustering: Gruppieren von Dokumenten (oder Teile davon) anhand von Ähnlichkeiten

Elder IV, J., & Hill, T. (2012). Practical text mining and statistical analysis for non-structured text data applications. Academic Press.

Arten des Text Minings

- Search and Information Retrieval (IR): Abspeichern und wiederauffinden von Dokumenten anhand von Schlüsselwörtern
- Document Clustering: Gruppieren von Dokumenten (oder Teile davon) anhand von Ähnlichkeiten
- Document Classification: Tagging von Dokumenten aufgrund von vorher definierten Modellen

Elder IV, J., & Hill, T. (2012). Practical text mining and statistical analysis for non-structured text data applications. Academic Press.

Arten des Text Minings

- Search and Information Retrieval (IR): Abspeichern und wiederauffinden von Dokumenten anhand von Schlüsselwörtern
- Document Clustering: Gruppieren von Dokumenten (oder Teile davon) anhand von Ähnlichkeiten
- Document Classification: Tagging von Dokumenten aufgrund von vorher definierten Modellen
- Web Mining: Text Mining von Internettexten (Twitter, Blogs, Websites, etc.). Aufgrund von Links können Netzwerke erstellt werden.

Elder IV, J., & Hill, T. (2012). Practical text mining and statistical analysis for non-structured text data applications. Academic Press.

Arten des Text Minings

- Search and Information Retrieval (IR): Abspeichern und wiederauffinden von Dokumenten anhand von Schlüsselwörtern
- Document Clustering: Gruppieren von Dokumenten (oder Teile davon) anhand von Ähnlichkeiten
- Document Classification: Tagging von Dokumenten aufgrund von vorher definierten Modellen
- Web Mining: Text Mining von Internettexten (Twitter, Blogs, Websites, etc.). Aufgrund von Links können Netzwerke erstellt werden.
- Information Extraction: Verwandeln von unstrukturiertem Text in strukturierte Daten. Muster und Beziehungen werden hier aufgezeigt.

Elder IV, J., & Hill, T. (2012). Practical text mining and statistical analysis for non-structured text data applications. Academic Press.

Arten des Text Minings

- Search and Information Retrieval (IR): Abspeichern und wiederauffinden von Dokumenten anhand von Schlüsselwörtern
- Document Clustering: Gruppieren von Dokumenten (oder Teile davon) anhand von Ähnlichkeiten
- Document Classification: Tagging von Dokumenten aufgrund von vorher definierten Modellen
- Web Mining: Text Mining von Internettexten (Twitter, Blogs, Websites, etc.). Aufgrund von Links können Netzwerke erstellt werden.
- Information Extraction: Verwandeln von unstrukturiertem Text in strukturierte Daten. Muster und Beziehungen werden hier aufgezeigt.
- Natural Language Processing (NLP): „Verstehen“ von Texten. Erkennen von Phrasen und einfachen Aussagen.

Elder IV, J., & Hill, T. (2012). Practical text mining and statistical analysis for non-structured text data applications. Academic Press.

Arten des Text Minings

- Search and Information Retrieval (IR): Abspeichern und wiederauffinden von Dokumenten anhand von Schlüsselwörtern
- Document Clustering: Gruppieren von Dokumenten (oder Teile davon) anhand von Ähnlichkeiten
- Document Classification: Tagging von Dokumenten aufgrund von vorher definierten Modellen
- Web Mining: Text Mining von Internettexten (Twitter, Blogs, Websites, etc.). Aufgrund von Links können Netzwerke erstellt werden.
- Information Extraction: Verwandeln von unstrukturiertem Text in strukturierte Daten. Muster und Beziehungen werden hier aufgezeigt.
- Natural Language Processing (NLP): „Verstehen“ von Texten. Erkennen von Phrasen und einfachen Aussagen.
- Concept Extraction: Gruppieren von Wörter und Phrasen in semantisch ähnliche Gruppen.

Elder IV, J., & Hill, T. (2012). Practical text mining and statistical analysis for non-structured text data applications. Academic Press.

Arten des Text Minings

- Search and Information Retrieval (IR): Abspeichern und wiederauffinden von Dokumenten anhand von Schlüsselwörtern
- Document Clustering: Gruppieren von Dokumenten (oder Teile davon) anhand von Ähnlichkeiten
- Document Classification: Tagging von Dokumenten aufgrund von vorher definierten Modellen
- Web Mining: Text Mining von Internettexten (Twitter, Blogs, Websites, etc.). Aufgrund von Links können Netzwerke erstellt werden.
- **Information Extraction: Verwandeln von unstrukturiertem Text in strukturierte Daten. Muster und Beziehungen werden hier aufgezeigt.**
- Natural Language Processing (NLP): „Verstehen“ von Texten. Erkennen von Phrasen und einfachen Aussagen.
- **Concept Extraction: Gruppieren von Wörter und Phrasen in semantisch ähnliche Gruppen.**

Elder IV, J., & Hill, T. (2012). Practical text mining and statistical analysis for non-structured text data applications. Academic Press.

Beispiel: Das Projekt

Evaluation des Förderinstruments „Diskursprojekte zu ethischen, rechtlichen und sozialen Fragen in den modernen Lebenswissenschaften“ des BMBF

Mein Arbeitspaket: Analyse der Webseiten

Beispiel: Inhaltsanalyse

- Mittels Skript die Textinhalte herunterladen
- Stopp-Words entfernen
  - Häufig genutzte Wörter wie: *und, oder, mit*, …
  - Wörterbücher online verfügbar
- Word Cloud erzeugen
- Concept Extraction: Gibt Aufschluss über die thematische Ausrichtung

Beispiel: Ergebnisse

Probleme?

- Nur ca. 10% aller Wörter ausgewertet

Probleme?

- Nur ca. 10% aller Wörter ausgewertet
- Kategorisierung muss trainiert werden
  - Zeitaufwendig
  - Kategorien je nach Analyse unterschiedlich

Probleme?

- Nur ca. 10% aller Wörter ausgewertet
- Kategorisierung muss trainiert werden
  - Zeitaufwendig
  - Kategorien je nach Analyse unterschiedlich
- Mehrdeutigkeit der Wörter
  - „Ungenauigkeit“ der Sprache

Mehrdeutigkeit der Sprache

- Kontext ist wichtig
  - Ich gehe zur Prüfung

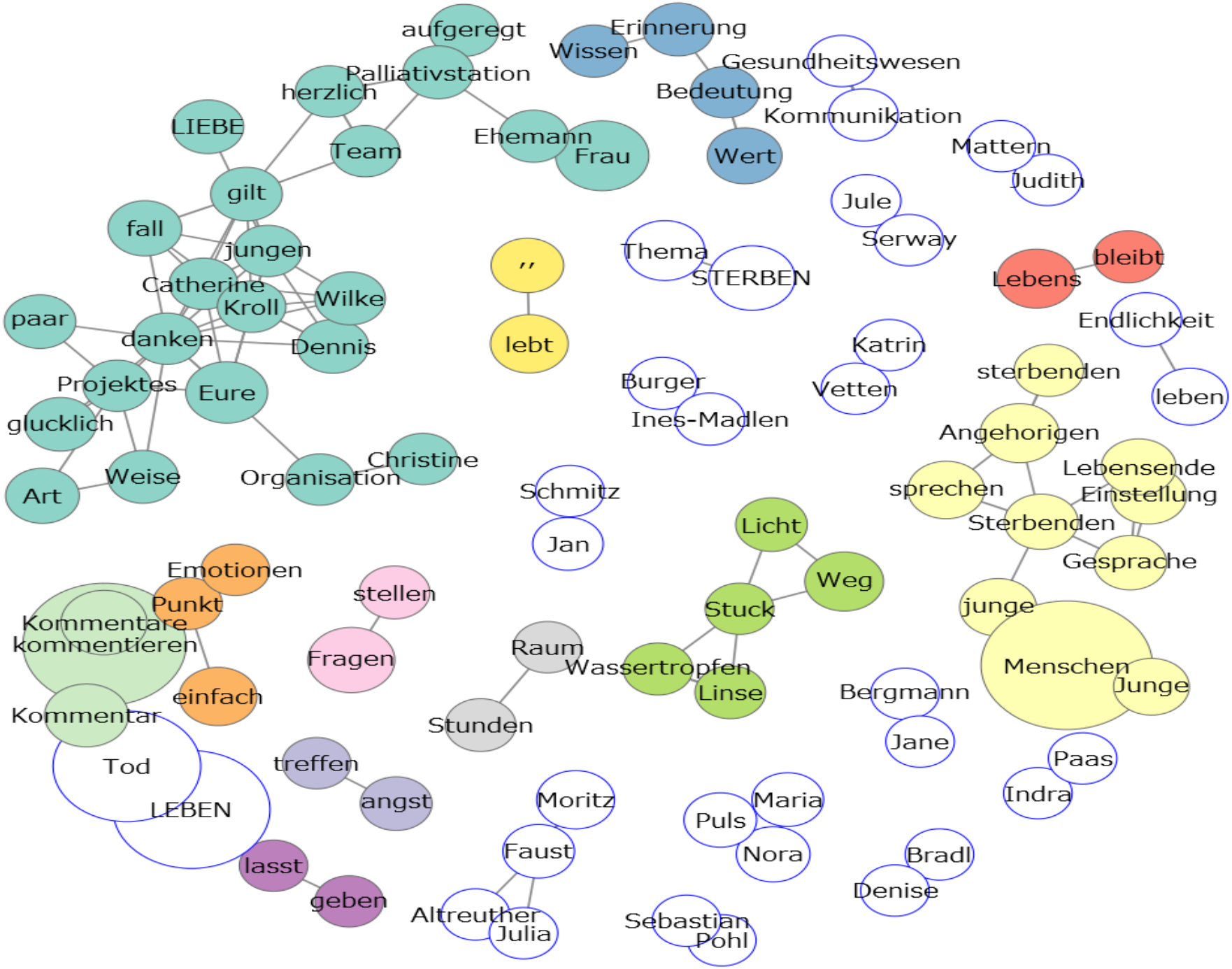
    vs.
  - Ich gehe nur dann zur Prüfung, **wenn die Hölle zufriert**

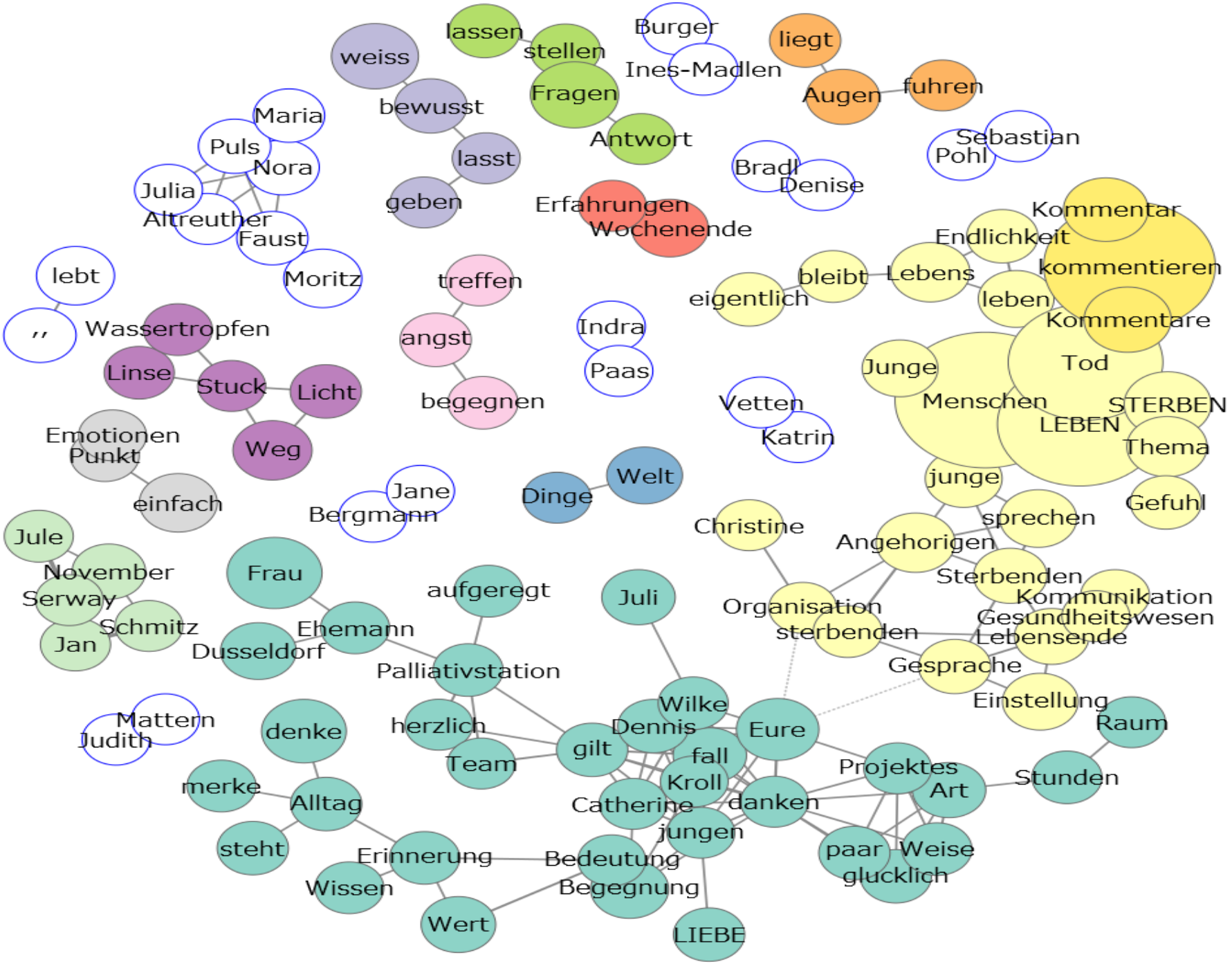
- Wörter sind mehrdeutig
  - Ich sitze auf der **Bank**

    vs.
  - Ich habe auf der **Bank** ein Konto

Bespiel: Beziehungen

- Welche Eigenschaften sind wichtig?
- Welchen Eindruck erhalten sie davon?
- Wo liegen ihrer Ansicht nach die Probleme?


aufgeregt
Palliativstation
herzlich
LIEBE
Team
Ehemann
Frau
gilt
fall
jungen
Catherine
Kroll
Wilke
paar
danken
Dennis
Projektes
Eure
glucklich
Art
Weise
Organisation
Christine
Erinnerung
Wissen
Bedeutung
Gesundheitswesen
Kommunikation
Wert
Mattern
Judith
Jule
Serway
Thema
STERBEN
''
lebt
bleibt
Lebens
Endlichkeit
leben
Katrin
Vetten
Burger
Ines-Madlen
sterbenden
Angehorigen
Lebensende
sprechen
Einstellung
Sterbenden
Gesprache
junge
Menschen
Junge
Bergmann
Jane
Paas
Indra
Schmitz
Jan
Licht
Weg
Stuck
Wassertropfen
Linse
Raum
Stunden
Emotionen
Punkt
Kommentare
kommentieren
einfach
Kommentar
stellen
Fragen
Tod
LEBEN
treffen
angst
lasst
geben
Moritz
Faust
Altreuther
Julia
Maria
Puls
Nora
Bradl
Denise
Sebastian
Pohl


lassen
stellen
Burger
weiss
Ines-Madlen
liegt
Fragen
Augen
fuhren
bewusst
Maria
Antwort
Sebastian
Puls
lasst
Pohl
Nora
Bradl
Julia
Denise
geben
Erfahrungen
Kommentar
Altreuther
Wochenende
Endlichkeit
Faust
kommentieren
lebt
Moritz
treffen
bleibt
Lebens
eigentlich
leben
,,
Wassertropfen
Kommentare
angst
Indra
Junge
Tod
Linse
Stuck
Licht
Paas
Menschen
STERBEN
Vetten
begegnen
LEBEN
Emotionen
Katrin
Thema
Punkt
Weg
junge
Welt
Jane
Dinge
Gefuhl
einfach
Bergmann
Christine
sprechen
Jule
Angehorigen
November
Frau
Sterbenden
Serway
aufgeregt
Juli
Kommunikation
Organisation
Gesundheitswesen
Schmitz
Ehemann
sterbenden
Lebensende
Jan
Dusseldorf
Gesprache
Palliativstation
Einstellung
Wilke
Mattern
Raum
herzlich
Dennis
Eure
Judith
denke
gilt
fall
Projektes
Stunden
Team
Kroll
merke
Art
Catherine
danken
Alltag
jungen
steht
paar
Weise
Erinnerung
Bedeutung
glucklich
Wissen
Begegnung
Wert
LIEBE

Conclusio

- **Aufbrechen von Texten**
  - Überraschen lassen
  - Beziehungen finden, die man sonst übersehen hätte
- **Überblick über große Datenmengen erhalten**
  - Stimmungslage
  - Grobe Einschätzungen möglich
- **Fundiertes Sampling für genauere Analysen**
  - Abseits vom Mainstream analysieren
- **Was es nicht ist:**
  - Ersatz für tiefergehende Analysen
  - Ausgelagerte Intuition