Karlsruher Institut für Technologie

Bachelorarbeit: Bestimmung der semantischen Funktion von Quelltextabschnitten

Timo Januschke, betreut von Tobias Hey

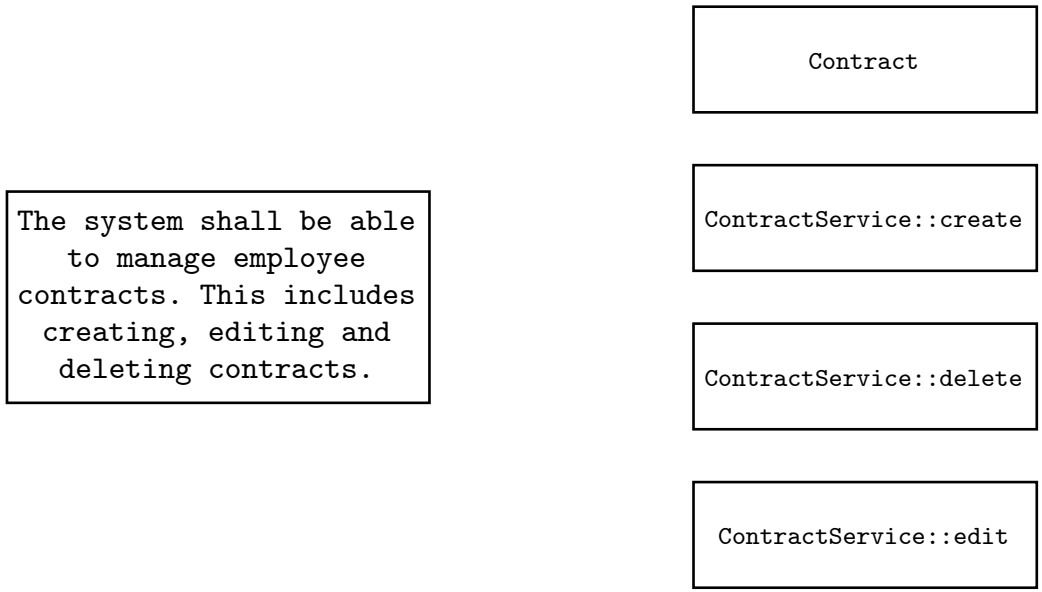
IPD Tichy, Fakultät für Informatik

Motivation

```
The system shall be able
   to manage employee
contracts. This includes
  creating, editing and
   deleting contracts.
```

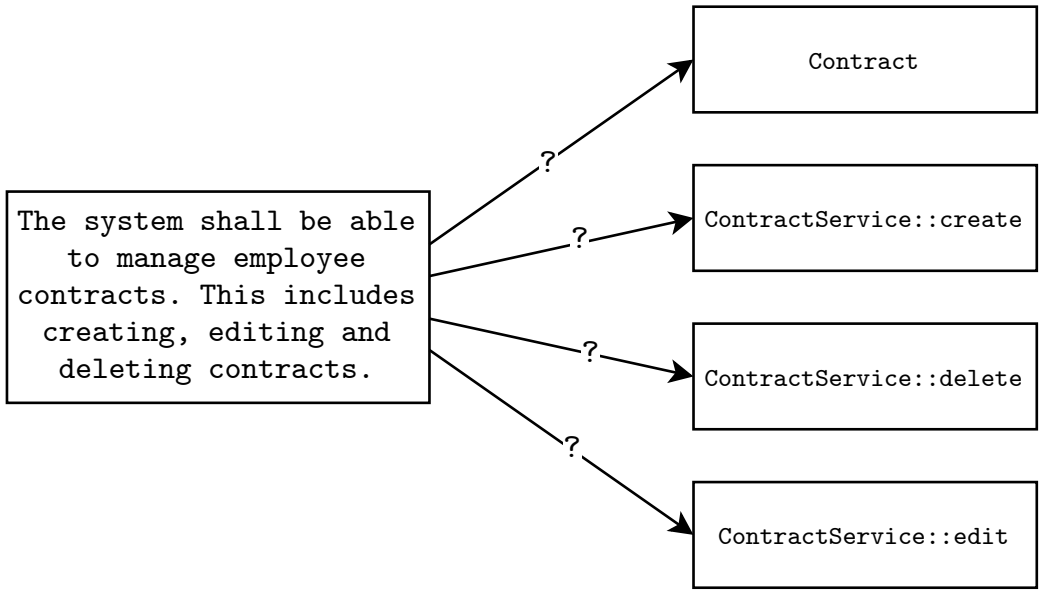
- Anforderungen spezifizieren Verhalten...

Motivation



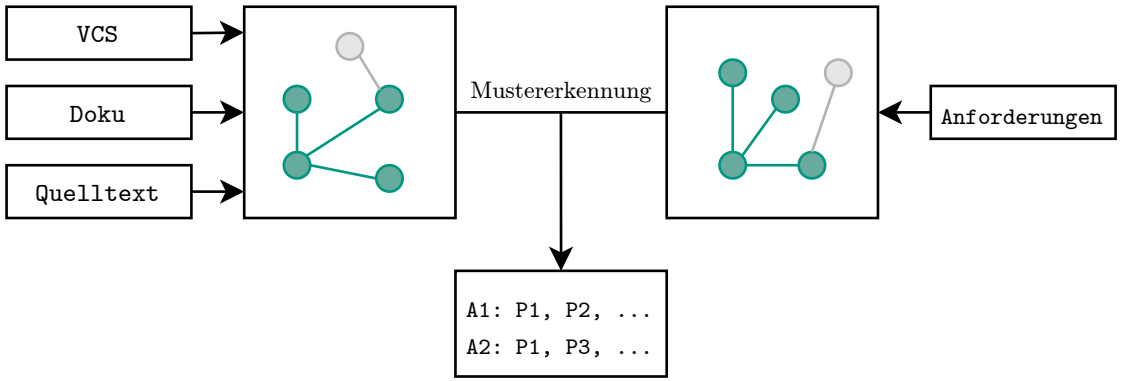
- ...Quelltext implementiert es

Motivation



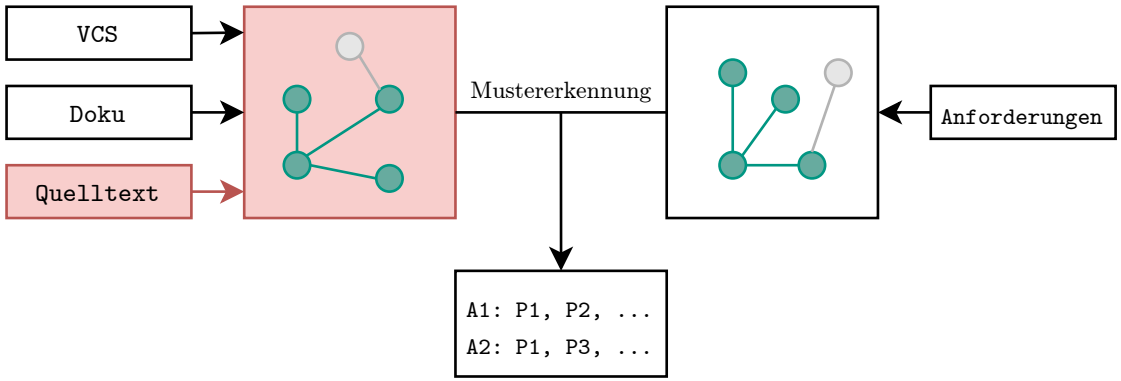
- Lücke in der Dokumentation

INDIRECT



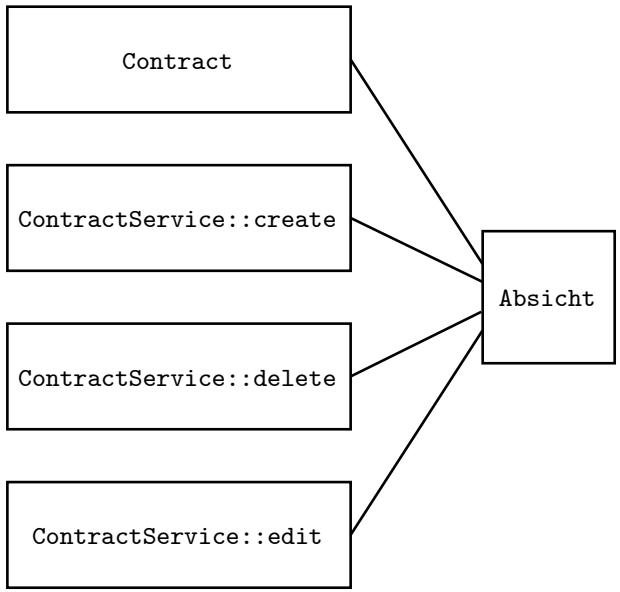
- Graphen als Absichtsmodelle

INDIRECT



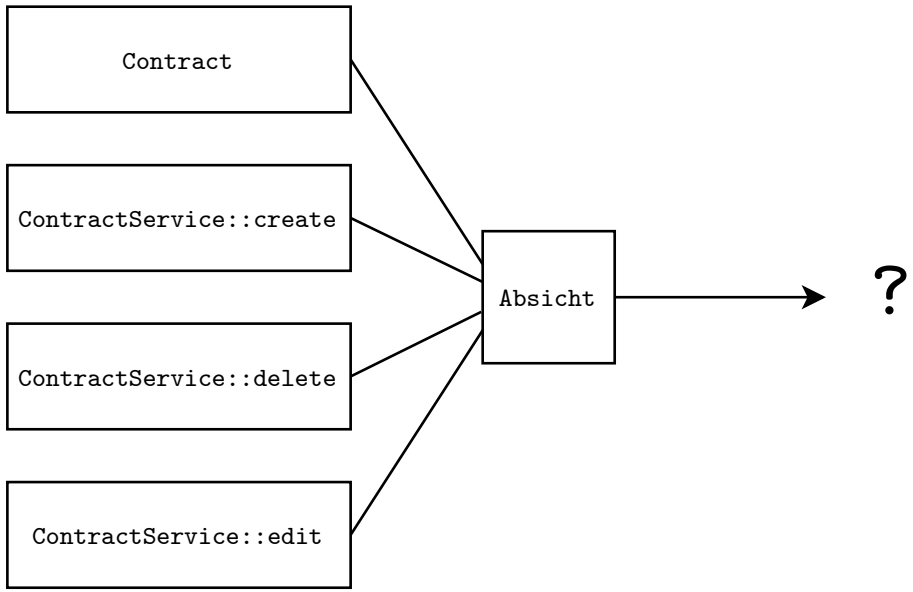
- Graphen als Absichtsmodelle

Quelltextgruppierungen [Eur]



- Geteilte Semantik

Quelltextgruppierungen



- Wie Semantik beschreiben?

Zielsetzung

- Ziel: Beschreibungen der geteilten Semantik von Quelltextgruppierungen
  - Vergleich verschiedener Repräsentationen
  - Finden eines Ausgangsverfahrens

Zielsetzung

- Ziel: Beschreibungen der geteilten Semantik von Quelltextgruppierungen
  - Vergleich verschiedener Repräsentationen
  - Finden eines Ausgangsverfahrens

Zielsetzung

- Ziel: Beschreibungen der geteilten Semantik von Quelltextgruppierungen
  - Vergleich verschiedener Repräsentationen
  - Finden eines Ausgangsverfahrens

Verwandte Arbeiten

- Beschreibung einzelner Quelltextelemente:
  - code2seq [Alo+19]
  - NeuralCodeSum [Ahm+20]
  - [APS16]
- Übertragbarkeit auf Quelltextgruppierungen?
- Kein Goldstandard →Überwachte Verfahren problematisch

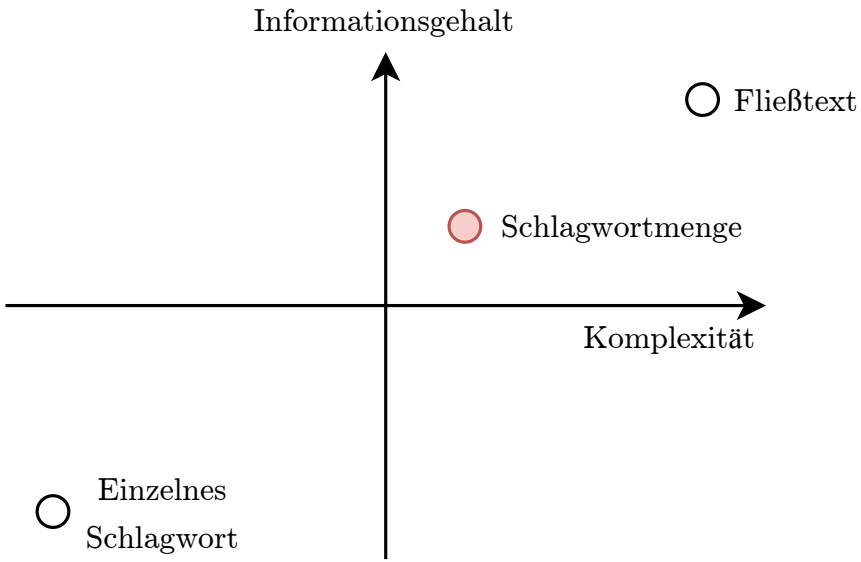
Verwandte Arbeiten

- Beschreibung einzelner Quelltextelemente:
  - code2seq [Alo+19]
  - NeuralCodeSum [Ahm+20]
  - [APS16]
- Übertragbarkeit auf Quelltextgruppierungen?
- Kein Goldstandard $\rightarrow$Überwachte Verfahren problematisch

Verwandte Arbeiten

- Beschreibung einzelner Quelltextelemente:
  - code2seq [Alo+19]
  - NeuralCodeSum [Ahm+20]
  - [APS16]
- Übertragbarkeit auf Quelltextgruppierungen?
- Kein Goldstandard $\rightarrow$Überwachte Verfahren problematisch

Repräsentation von Beschreibungen



- Kompromiss aus Komplexität und Informationsgehalt

Ansatz

```java
/**
*  This class deals with logging
*/
class Logger {
  LoggerFactory factory;

  /**
  *  Gets a Logger instance
  */
  Logger getLogger();
}
```

```java
/**
*  An employee contract
*/
class Contract {
  Logger getLogger();

  void create();

  void delete();

  void edit();
}
```

- Idee: Nutze natürliche Sprache

Vorverarbeitung

- Kleinbuchstaben: `contract`, `Contract`, `cOnTrAcT` $\rightarrow$ `contract`
- Subworttokenisierung: `getLogger` $\rightarrow$ `get`, `Logger`
- Stopwortentfernung: `an`, `employee`, `contract` $\rightarrow$ `employee`, `contract`
- Stammformreduktion: `logger`, `logging` $\rightarrow$ `log`

Vorverarbeitung

- Kleinbuchstaben: `contract`, `Contract`, `cOnTrAcT` $\rightarrow$ `contract`
- Subworttokenisierung: `getLogger` $\rightarrow$ `get`, `Logger`
- Stopwortentfernung: `an`, `employee`, `contract` $\rightarrow$ `employee`, `contract`
- Stammformreduktion: `logger`, `logging` $\rightarrow$ `log`

Vorverarbeitung

- Kleinbuchstaben: `contract`, `Contract`, `cOnTrAcT` $\rightarrow$ `contract`
- Subworttokenisierung: `getLogger` $\rightarrow$ `get`, `Logger`
- Stopwortentfernung: `an`, `employee`, `contract` $\rightarrow$ `employee`, `contract`
- Stammformreduktion: `logger`, `logging` $\rightarrow$ `log`

Vorverarbeitung

- Kleinbuchstaben: `contract`, `Contract`, `cOnTrAcT` → `contract`
- Subworttokenisierung: `getLogger` → `get`, `Logger`
- Stopwortentfernung: `an`, `employee`, `contract` → `employee`, `contract`
- Stammformreduktion: `logger`, `logging` → `log`

Vorstudie

- Vergleich verschiedener Verfahren:
  - LDA [BNJ03]
  - HDP [Teh+05]
  - Worteinbettungen [Zha+16]
- Vergleiche Konfigurationen:
  - Trainingsdokumente
  - Parameter
- Sieger: **LDA**

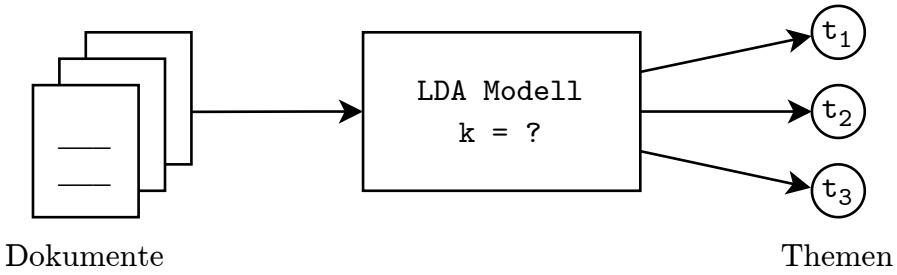
Vorstudie

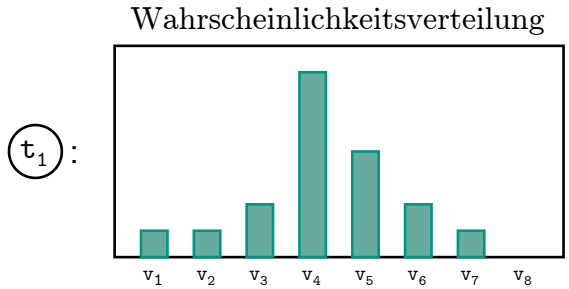
- Vergleich verschiedener Verfahren:
  - LDA [BNJ03]
  - HDP [Teh+05]
  - Worteinbettungen [Zha+16]
- Vergleiche Konfigurationen:
  - Trainingsdokumente
  - Parameter
- Sieger: **LDA**

Vorstudie

- Vergleich verschiedener Verfahren:
  - LDA [BNJ03]
  - HDP [Teh+05]
  - Worteinbettungen [Zha+16]
- Vergleiche Konfigurationen:
  - Trainingsdokumente
  - Parameter
- Sieger: **LDA**

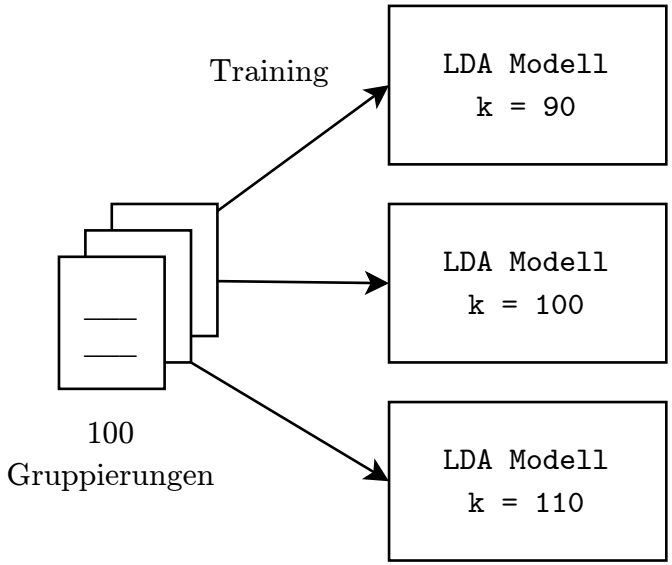
LDA [BNJ03]





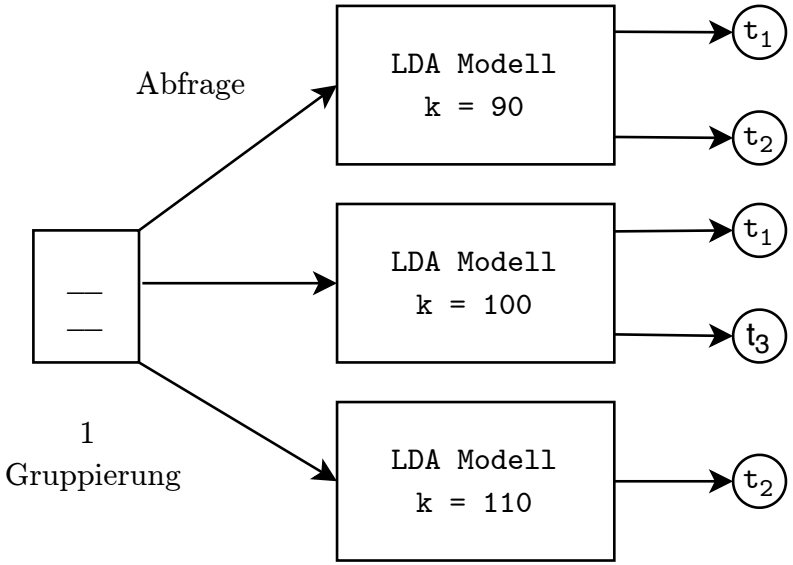
- Ein Dokument pro Gruppierung
- Themen sind Wahrscheinlichkeitsverteilungen

LDA Training



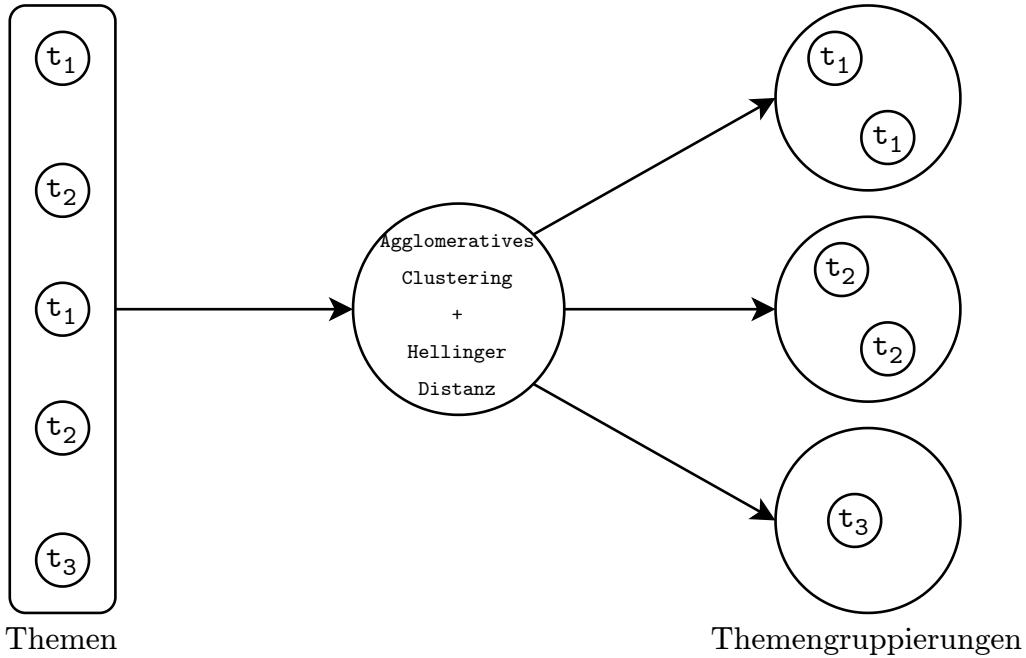
- Approximation $k = \#Gruppierungen$

LDA Abfrage



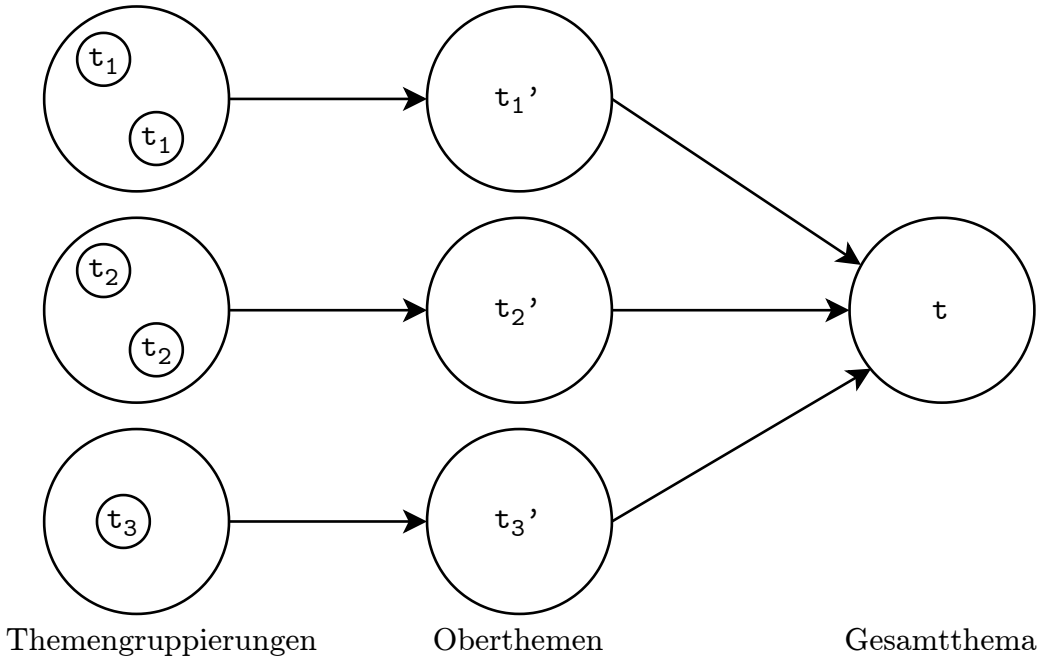
- Wahl von *k* beeinflusst Themen

Kombination mehrerer Modelle



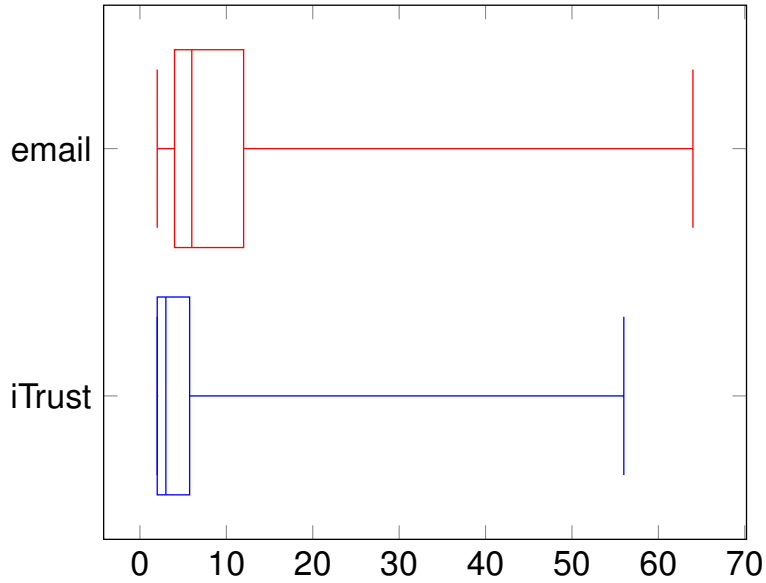
- Gruppierung von Themen

Kombination mehrerer Modelle



- Rückkombination zu Gesamtthema
- Durchschnittswerte in jeder Stufe

Evaluationskorpus



Projekt	l_{code}	l_{com}	Anz. Gruppierungen
iTrust	14573 (62%)	6338 (27%)	238
email	6776 (54%)	4177 (33%)	51

- Kein Goldstandard $\rightarrow$Nutzerstudie
- Projekte `iTrust` und `apache-commons-email`

Studiendesign

Aufgabe 1:

```
class Contract {
  void create {..}
  void edit() {..}
  void delete() {..}
}
```

Schlagw.: contract, manage, edit

Aufgabe 2:

```
class Contract {
  void create {..}
  void edit() {..}
  void delete() {..}
}
```

Bewertung (gut/neutral/schlecht):

contract: gut

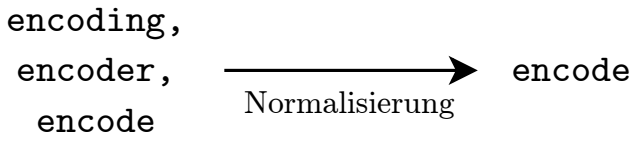
edit: gut

office: neutral

Übereinstimmung (1-5): 3

- Beschreibung und Bewertung

Studiendesign



Norm. Antworten:

contract (8/8)
edit (7/8)
delete (6/8)
create (5/8)
office
employee

Erzeugte Schlagwortmenge:

{ **contract**, **edit**, office, email }
TP TP FP FP

- Normalisierung und Konsensbildung

Auswertung

Projekt	Präzision	Ausbeute	F1	Bew.	$score_{TP}$	$score_{FP}$
iTrust	0.65	0.86	0.75	3.67	1.83	1.22
email	0.33	0.65	0.43	4.00	1.93	1.22
Gesamt	0.45	0.73	0.56	3.83	1.88	1.22

- Niedrige Präzision, hohe Ausbeute
- Schlagworte gut bewertet

Fazit

- Ziel: Beschreibung der geteilten Semantik von Quelltextgruppierungen
- Ansatz:
  - Nutzen natürlicher Sprache in Gruppierungen
  - LDA um Themen für Gruppierungen zu finden
  - Kombination mehrerer LDA-Modelle
- Ergebnis:
  - Konsensbildung →Niedrige Präzision, hohe Ausbeute
  - Durchschnittlich: 0.45 Präzision, 0.73 Ausbeute, 0.56 F1-Wert
  - Falsch positive dennoch mit 1.22/2.00 bewertet (besser als neutral)
- Ausblick:
  - Erzeugung von Fließtextsätzen
  - Abstraktion der Schlagworte durch externe Wissensquellen

Fazit

- Ziel: Beschreibung der geteilten Semantik von Quelltextgruppierungen
- Ansatz:
  - Nutzen natürlicher Sprache in Gruppierungen
  - LDA um Themen für Gruppierungen zu finden
  - Kombination mehrerer LDA-Modelle
- Ergebnis:
  - Konsensbildung →Niedrige Präzision, hohe Ausbeute
  - Durchschnittlich: 0.45 Präzision, 0.73 Ausbeute, 0.56 F1-Wert
  - Falsch positive dennoch mit 1.22/2.00 bewertet (besser als neutral)
- Ausblick:
  - Erzeugung von Fließtextsätzen
  - Abstraktion der Schlagworte durch externe Wissensquellen

Fazit

- Ziel: Beschreibung der geteilten Semantik von Quelltextgruppierungen
- Ansatz:
  - Nutzen natürlicher Sprache in Gruppierungen
  - LDA um Themen für Gruppierungen zu finden
  - Kombination mehrerer LDA-Modelle
- Ergebnis:
  - Konsensbildung →Niedrige Präzision, hohe Ausbeute
  - Durchschnittlich: 0.45 Präzision, 0.73 Ausbeute, 0.56 F1-Wert
  - Falsch positive dennoch mit 1.22/2.00 bewertet (besser als neutral)
- Ausblick:
  - Erzeugung von Fließtextsätzen
  - Abstraktion der Schlagworte durch externe Wissensquellen

Fazit

- Ziel: Beschreibung der geteilten Semantik von Quelltextgruppierungen
- Ansatz:
  - Nutzen natürlicher Sprache in Gruppierungen
  - LDA um Themen für Gruppierungen zu finden
  - Kombination mehrerer LDA-Modelle
- Ergebnis:
  - Konsensbildung $\rightarrow$Niedrige Präzision, hohe Ausbeute
  - Durchschnittlich: 0.45 Präzision, 0.73 Ausbeute, 0.56 F1-Wert
  - Falsch positive dennoch mit 1.22/2.00 bewertet (besser als neutral)
- Ausblick:
  - Erzeugung von Fließtextsätzen
  - Abstraktion der Schlagworte durch externe Wissensquellen

References I

Ahmad, Wasi Uddin u. a. (Mai 2020). „A Transformer-Based Approach for Source Code Summarization“. In: *arXiv:2005.00653 [cs, stat]*. arXiv: 2005.00653 [cs, stat].

Allamanis, Miltiadis, Hao Peng und Charles Sutton (2016). „A Convolutional Attention Network for Extreme Summarization of Source Code“. In: *ICML*.

Alon, Uri u. a. (Feb. 2019). „Code2seq: Generating Sequences from Structured Representations of Code“. In: *arXiv:1808.01400 [cs, stat]*. arXiv: 1808.01400 [cs, stat].

Blei, David M., Andrew Y. Ng und Michael I. Jordan (März 2003). *Latent Dirichlet Allocation*.

Eurich, Felix. „Entwurf Und Aufbau Einer Semantischen Repräsentation von Quelltext“. Master’s Thesis. Karlsruher Institut für Technologie (KIT) – IPD Tichy.

References II

Teh, Yee W u. a. (2005). „Sharing Clusters among Related Groups: Hierarchical Dirichlet Processes“. In: *Advances in Neural Information Processing Systems*, S. 1385–1392.

Zhang, Wei Emma u. a. (Dez. 2016). „Mining Source Code Topics Through Topic Model and Words Embedding“. In: S. 664–676. ISBN: 978-3-319-49585-9. DOI: 10.1007/978-3-319-49586-6_47.