

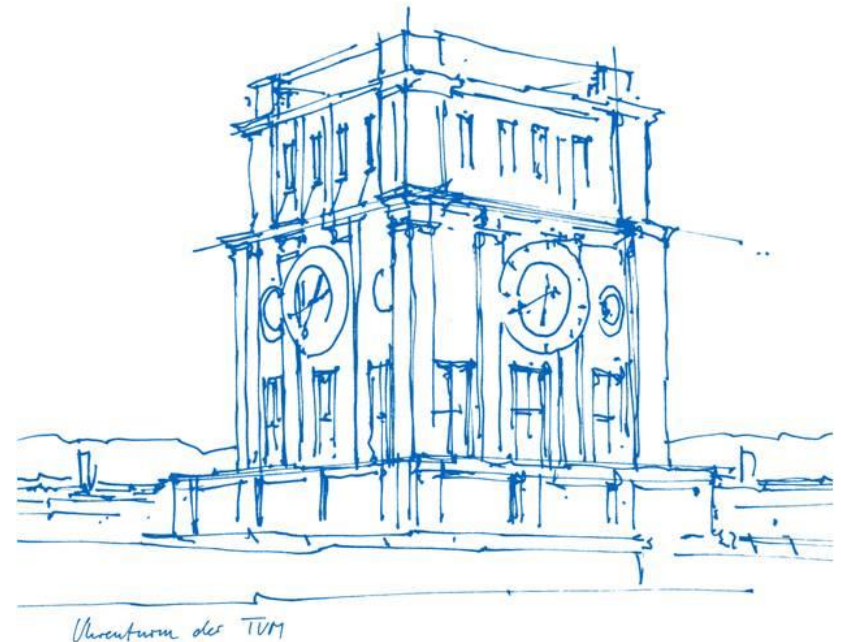
# Bachelor's Thesis in Information Systems

## Topic Model Visualization for Opinion Mining

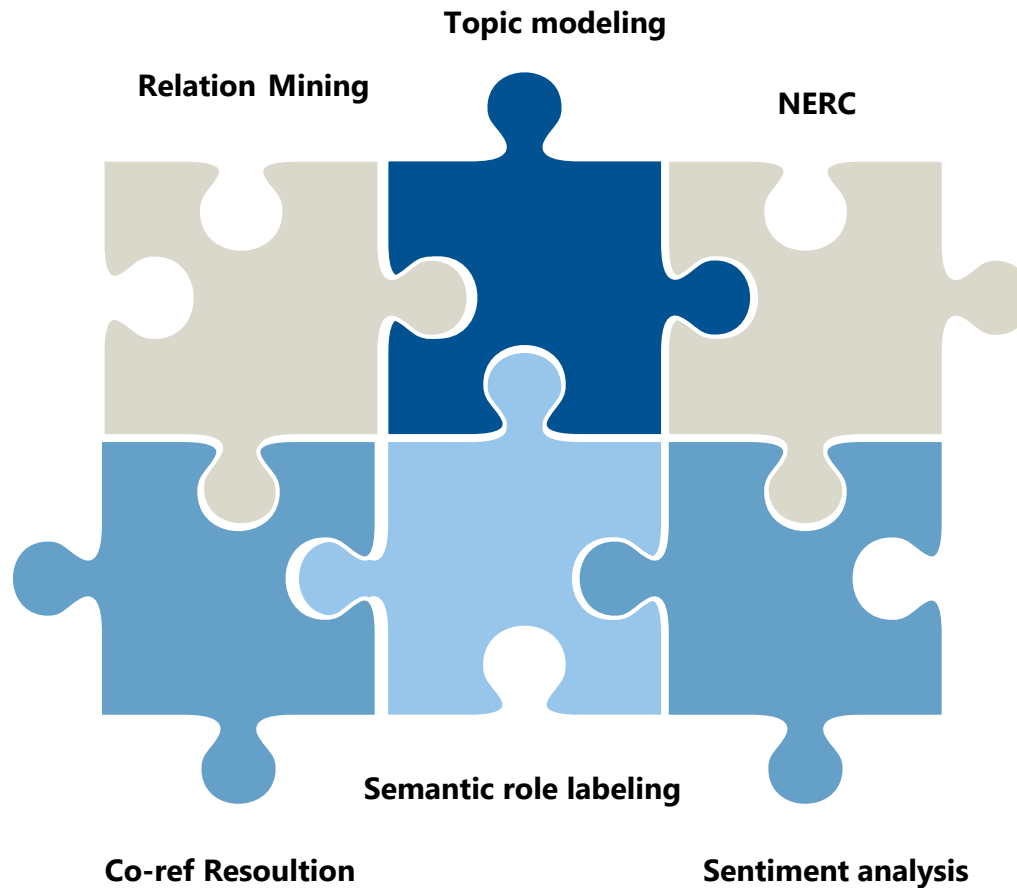
Maria Potzner

Advisor: PD Dr. Georg Groh

17.01.2019



# Social ROM



- Analyse opinion of large population towards organic food
- Based on online user generated data

# Topic Modelling

2017

## What was done before

- Scraped online user generated data
  - Topic Generation
- Themes from topics correspond qualitative research
- Trends evolving over time (e.g. fipronil scandal)

2018

## What have I done

- Automatic Topic Labeling
  - find meaningful labels
  - reduce cognitive overhead of interpreting
- Internal Consistency
  - provide overview how topic models change

# Automatic Topic Labeling


## Intrinsic approach

- Working only on our texts and datasets
- No ontologies or embeddings
- We used the approach of Mei et al. 2007

## Extrinsic approach

- Working with ontologies and embeddings
- We used WordNet

# Intrinsic approach



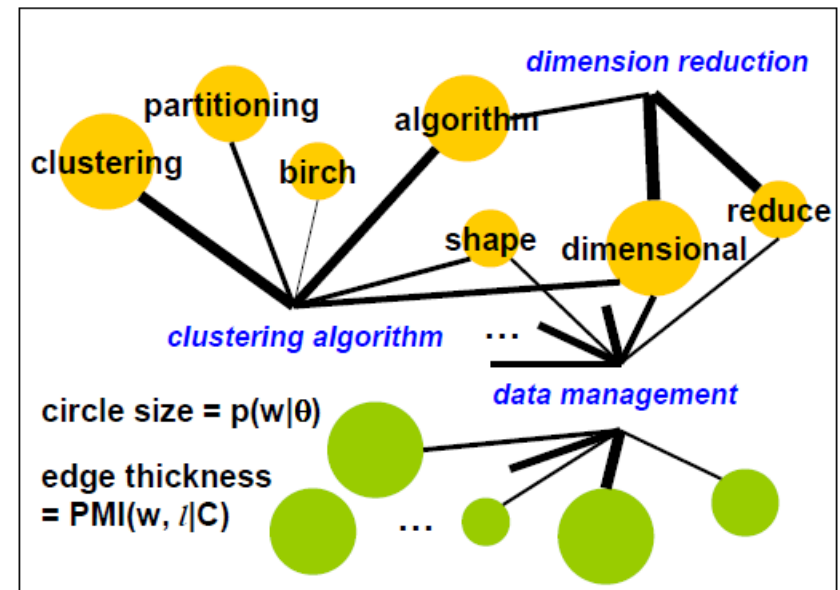
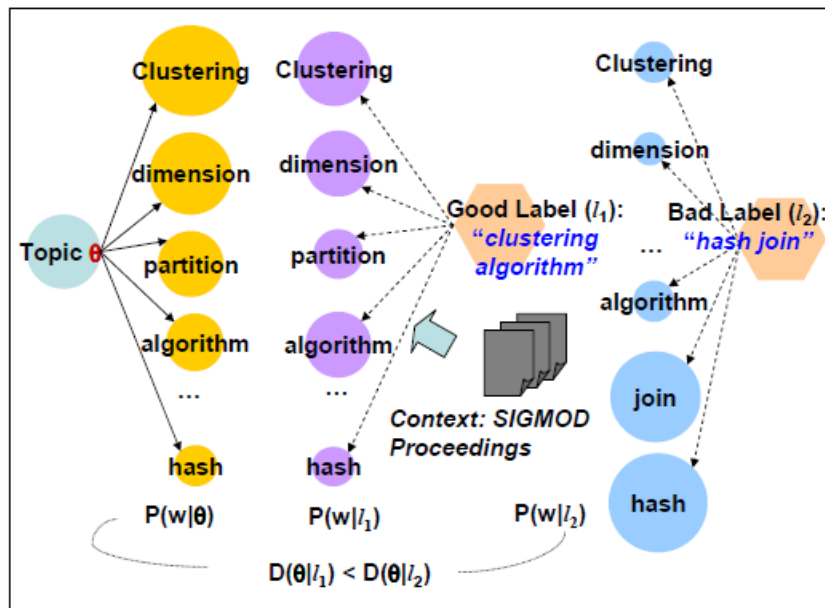
```
graph LR; A["Noun phrases and n-grams as labels"] --> B["Label extraction with POS-tags"]; B --> C["Rank labels By semantic similarity"]
```

Noun phrases  
and n-grams  
as labels

Label extraction  
with POS-tags

Rank labels  
By semantic  
similarity

# Visual interpretation of the intrinsic approach



Adapted from Mei et al., 2007

# Evaluation and Results

	Topic 107	Topic 23
method	waste, compost, use, scrap, material, landfill, ton, environmental, throw, gas	grow, garden, plant, farm, vegetable, seed, year, tomato, produce, farming
intrinsic manual	rahm emanuel waste	hairy vetch homegrown food

- Meaningful and specific labels  
BUT: not really fitting to the topics
- often names occure as labels e.g Rahm Emanuel, Safran Foer

# Evaluation and Results

	Topic 6	Topic 10
with POS-tags	restaurant, fast, chain, meal, say, menu, ingredient, burger, chipotle, mcdonald	child, eat, kid, parent, family, healthy, school, who, health,can
(NN, NN)	music festival	anorexia nervosa
(JJ, NN)	hot fudge	premature aging
-	dunkin donuts	anorexia nervosa

→ seem minimal better than labels without POS-tags

→ names of persons decrease



# Extrinsic Topic Labeling

- Used english online database WordNet
- Organizes words in synsets

## Noun

- **S: (n) farming, agriculture, husbandry** (the practice of cultivating the land or raising stock)
- **S: (n) farming, land** (agriculture considered as an occupation or way of life) *"farming is a strenuous life"; "there's no work on the land any more"*

## Verb

- **S: (v) farm** (be a farmer; work as a farmer) *"My son is farming in California"*
- **S: (v) farm** (collect fees or profits)
- **S: (v) grow, raise, farm, produce** (cultivate by growing, often involving improvements by means of agricultural techniques) *"The Bordeaux region produces great red wines"; "They produce good ham in Parma"; "We grow wheat here"; "We raise hogs here"*

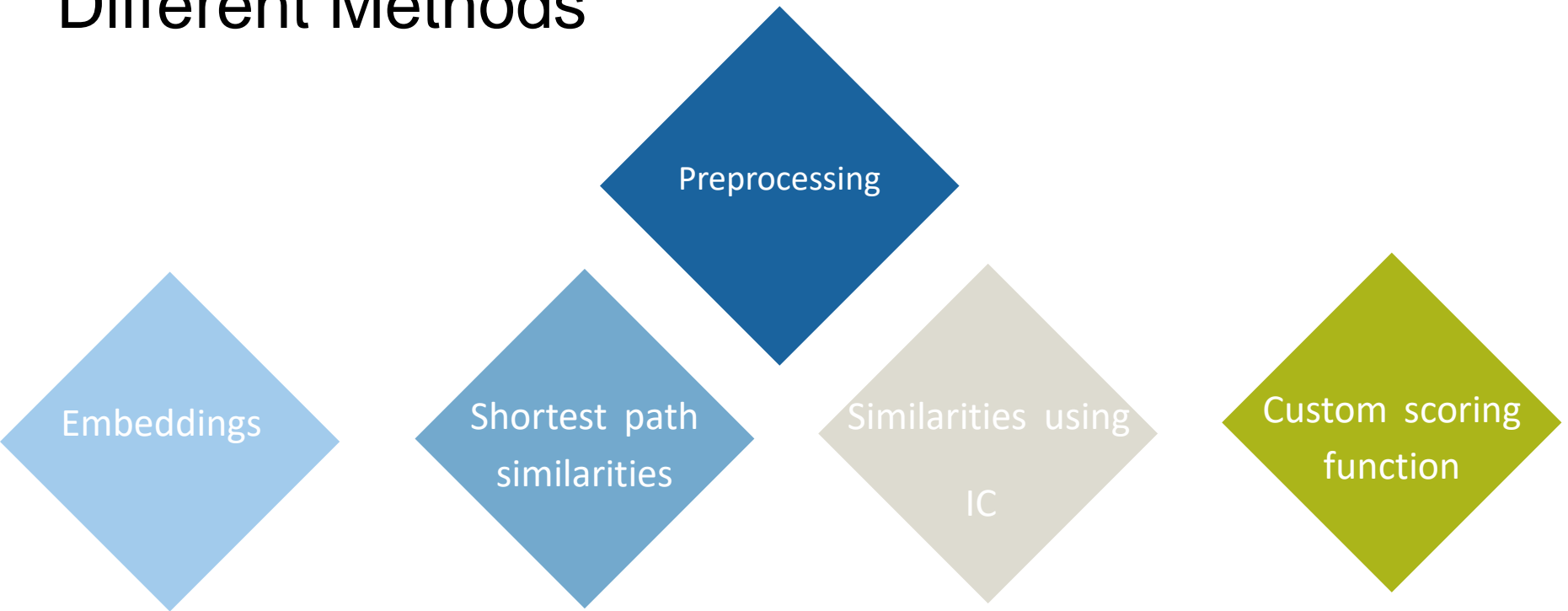
## Adjective

- **S: (adj) agrarian, agricultural, farming** (relating to farming or agriculture) *"an agrarian (or agricultural) society"; "farming communities"*

## Noun

- **S: (n) farming, agriculture, husbandry** (the practice of cultivating the land or raising stock)
  - [direct hyponym](#) / [full hyponym](#)
  - [part meronym](#)
  - [domain term category](#)
  - [direct hypernym](#) / [inherited hypernym](#) / [sister term](#)
- **S: (n) cultivation** ((agriculture) production of food by preparing the land to grow crops (especially on a large scale))
  - **S: (n) production** ((economics) manufacturing or mining or growing something (usually in large quantities) for sale) *"he introduced more efficient methods of production"*
  - **S: (n) industry, manufacture** (the organized action of making of goods and services for sale) *"American industry is making increased use of computers to control production"*
  - **S: (n) commercial enterprise, business enterprise, business** (the activity of providing goods and services involving financial and commercial and industrial aspects) *"computers are now widely used in business"*
  - **S: (n) commerce, commercialism, mercantilism** (transactions (sales and purchases) having the objective of supplying commodities (goods and services))
  - **S: (n) transaction, dealing, dealings** (the act of transacting within or between groups (as carrying on commercial activities)) *"no transactions are possible without him"; "he has always been honest in his dealings with me"*

# Different Methods



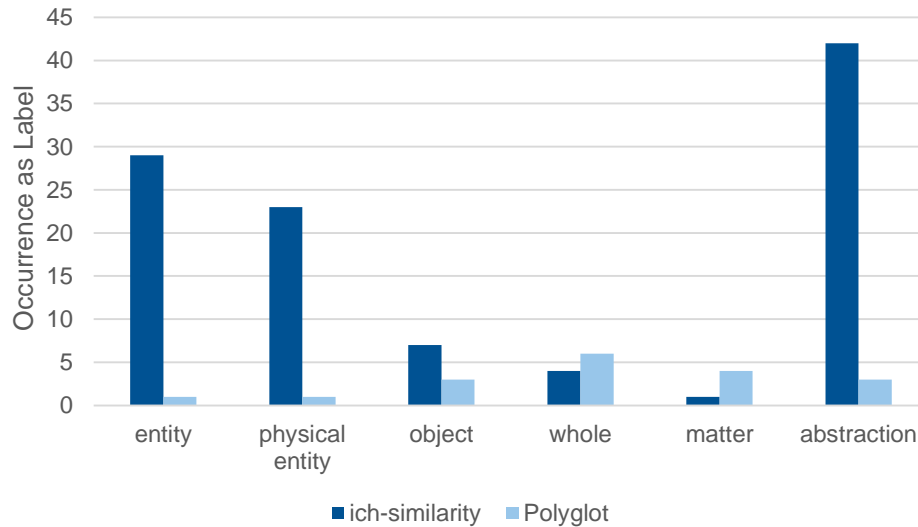
# Evaluation and Results

	Topic 74		Topic 84	
method	meat, feed, beef, grass, eat, raise, cow, buy, make, animal		company, tea, brand, product, drink, honest, new, beverage, consumer, goldman	
path	entity	meat	<b>beverage</b>	<b>beverage</b>
ich	entity	abstraction	physical entity	substance
res	matter	meat	substance	substance
jsn	food	meat	<b>beverage</b>	<b>beverage</b>
lin	matter	meat	<b>beverage</b>	<b>beverage</b>
plg	cattle	physical entity	food	food
<b>Csf</b>	cattle	be	<b>beverage</b>	<b>beverage</b>
manual	animal husbandry		<b>beverage</b>	

→ some labels are comparable with the labels from Domain experts

# Evaluation and Results

- Count of labels which are too general:  
*method, entity, physical, entity, object, whole, matter and abstraction*

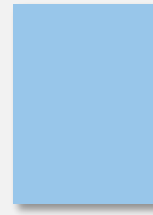


# Internal Consistency



## **Document composition**

Few and dominating topics?

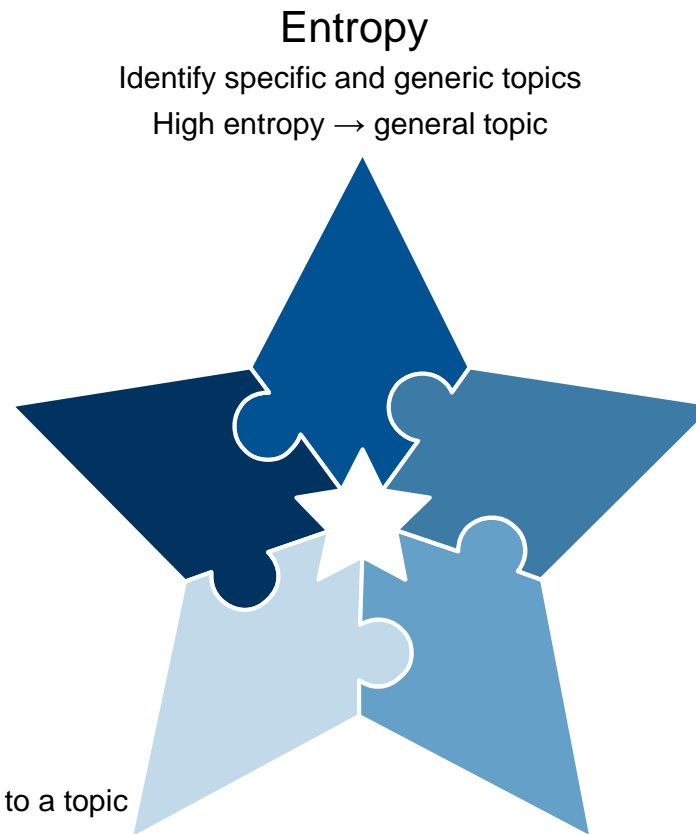


## **Topic development**

Consistent topics across topic models?

General or specific topics in and across topic models?

# Key Figures



## Entropy

Identify specific and generic topics  
High entropy → general topic

## Alpha

Measure importance of topics  
High alpha → documents are mixtures of many topics

## Coherence

Measure similarity of words in topics  
High value → high similarity/interpretability

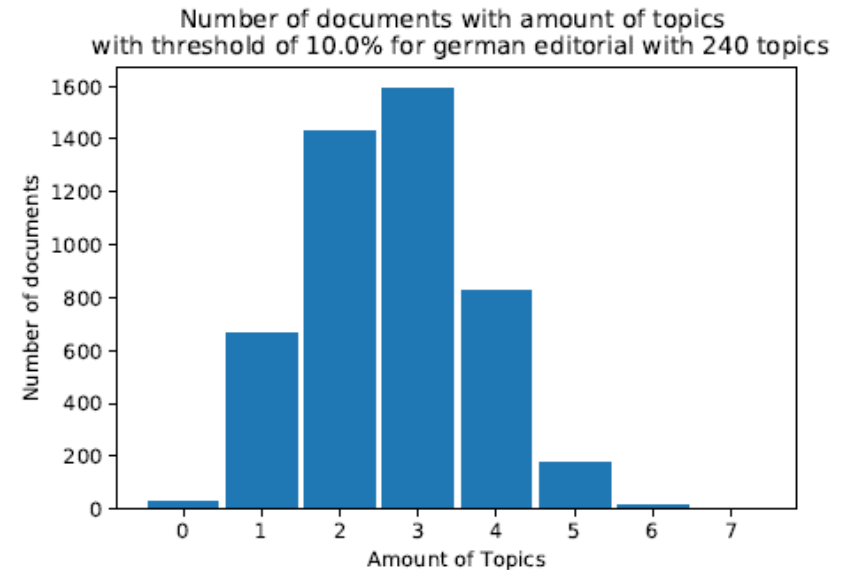
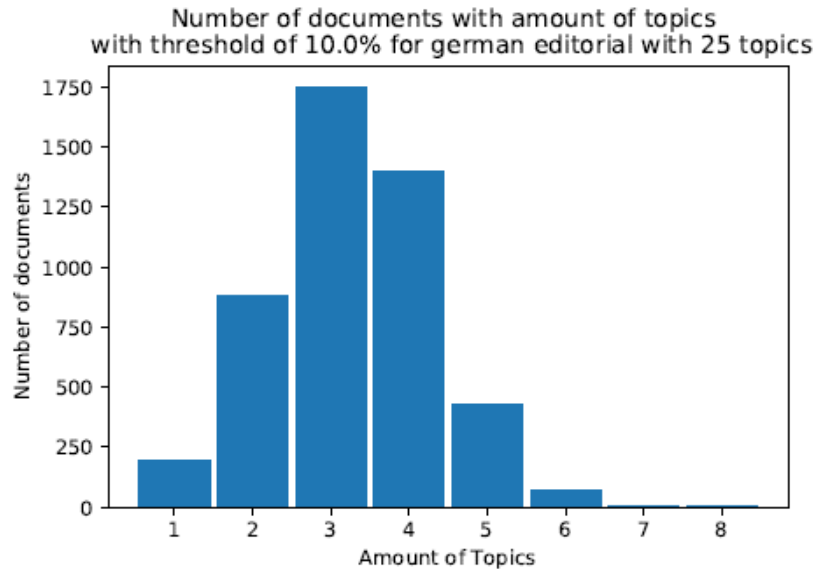
## Jensen Shannon

Measure similarity of topics  
High value → high similarity

## Theta

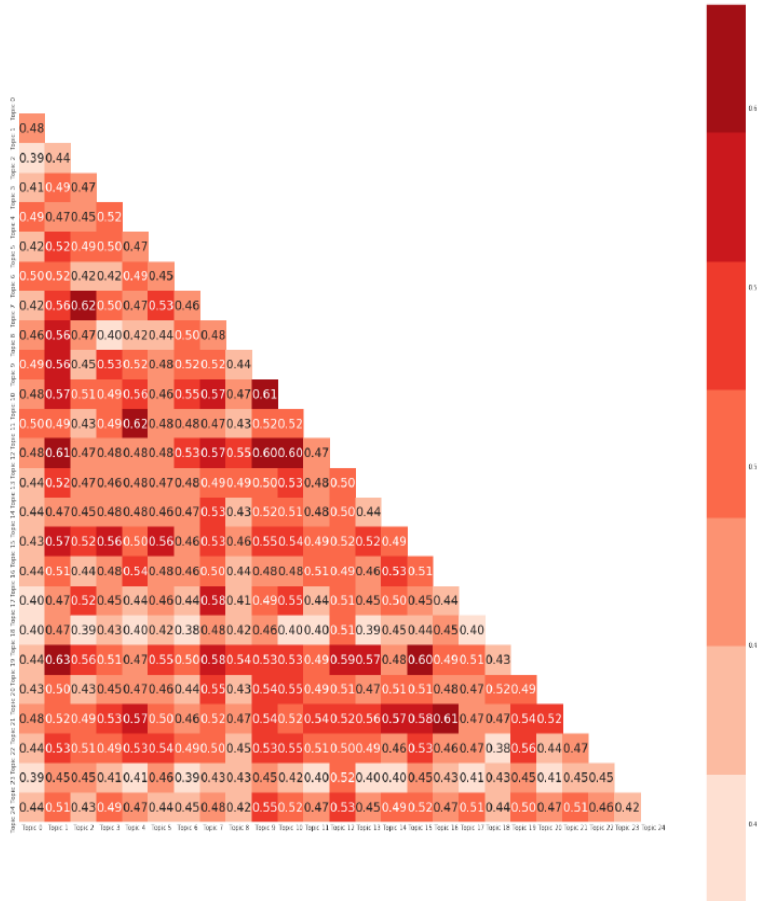
Document topic matrix  
Calculate amount of documents assigned to a topic and amount of topics assigned to document

# Evaluation of Theta



→ higher topic number indicates lower amount of topics in a document

# Jensen Shannon



■ topics inter topic models  
→ similarities are high

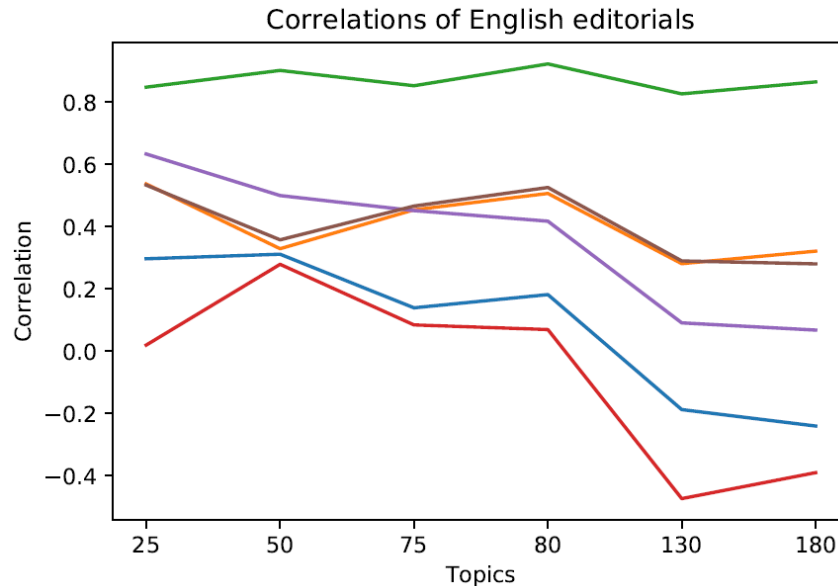
50/75	<p>eiern, fipronil, belasten, niederlande, T9: deutschland, nehmen, verkaufen, betreffen, betrieb, angeben</p>	<p>eiern, fipronil, belasten, niederlande, T57: nehmen, deutschland, verkaufen, betroffen, betreffen, behörde</p>
-------	--	---

■ topics intra topic models  
→ similarities are lower  
→ topic model is not overfitted

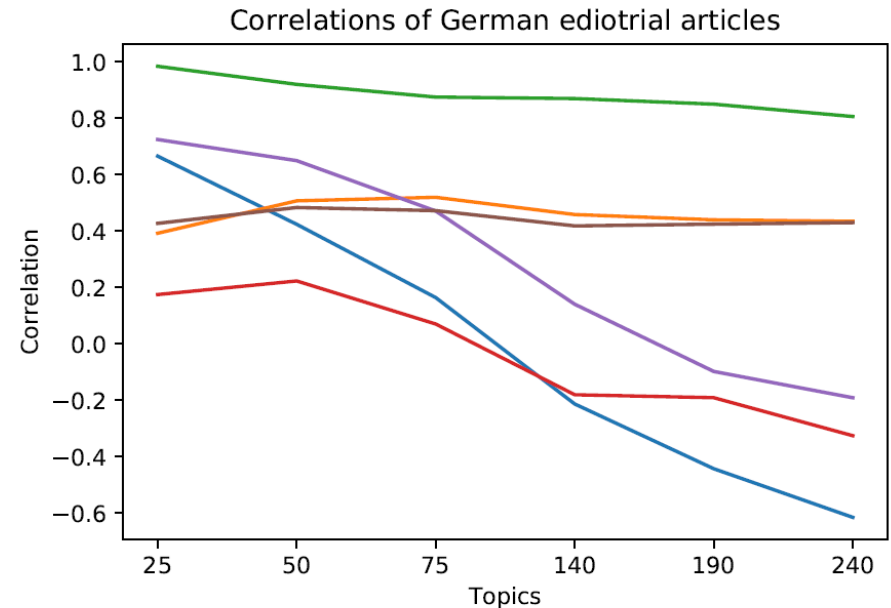
75	<p>bauer, landwirt, euro, T63:preisen, konventionell, biobauer, geld, bekommen, umstellen, ernten</p>	<p>product, verbraucher, kunde, T67:deutschland, <b>preisen</b>, deutsch, prozent, handeln,supermarkt, deutsche</p>
----	---	---



# Correlation



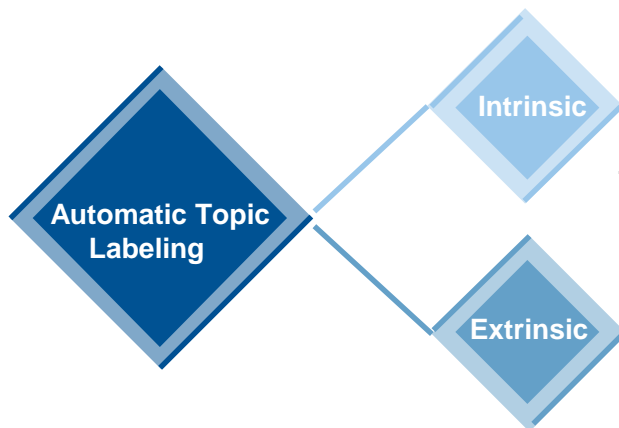
- correlation entropy and alpha
- correlation alpha and coherence
- correlation alpha and documents per topic
- correlation coherence and entropy
- correlation entropy and documents per topic
- correlation coherence and documents per topic



- correlation entropy and alpha
- correlation alpha and coherence
- correlation alpha and documents per topic
- correlation coherence and entropy
- correlation entropy and documents per topic
- correlation coherence and documents per topic

- correlation between alpha and documents per topic
- entropy seems to develop independently from other key figures

# Future work



- Improve the ranking function
- Generate labels based on document titles
- Try out German database or other ontologies

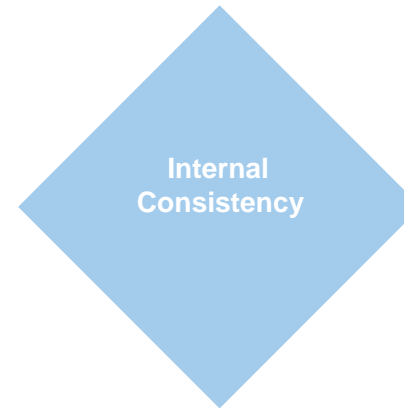


- Evaluate key figures after changing other priors of topic models
- Topic models with automated prior finding

# Conclusion



- extrinsic labeling performed better than the intrinsic labeling
- some labels correspond labeling of Domain experts



- applying key figures does not lead to generic insides
- depending on the nature of the dataset

Backup slides

# Intrinsic labels

	Topic 107	Topic 23
method	waste, compost, use, scrap, material, landfill, ton, environmental, throw, gas	grow, garden, plant, farm, vegetable, seed, year, tomato, produce, farming
intrinsic manual	rahm emanuel waste	hairy vetch homegrown food
	Topic 64	Topic 74
method	milk, raw, dairy, product, cheese, claim, health, cow drink, study	meat, feed, beef, animal, grass, cow, eat, raise, buy, make
intrinsic manual	irritable bowel syndrome dairy product	safran foer animal husbandry

**Tab. 3.4.:** Topics labeled manually and with intrinsic methods.

# Intrinsic with POS-tags

	Topic 6	Topic 10
with POS-tags	restaurant, fast, chain, meal, say, menu, ingredient, burger, chipotle, mcdonald	child, eat, kid, parent, family, healthy, school, who, health,can
(NN, NN) (JJ, NN) -	music festival hot fudge dunkin donuts	anorexia nervosa premature aging anorexia nervosa
	Topic 23	Topic 37
with POS-tags	meat, beef, feed,animal, grass, cattle,eat, raise, more, pork	carbon, climate, gas, greenhouse, emission, change, reduce, global, industrial, co2
(NN, NN) (JJ, NN) -	sport utility hunted game earl butz	gene splicing interactive map modified organisms

**Tab. 3.5.:** Labeled topics with intrinsic method

# Extrinsic Topics

	Topic 23		Topic 64	
method	grow, garden, plant, farm, vegetable, seed, year, tomato, produce, farming		milk, raw, dairy, product, cheese, cow health, drink, study, claim	
path	entity	produce	abstraction	beverage
ich	entity	produce	abstraction	produce
res	produce	produce	<b>dairy product</b>	beverage
jsn	produce	produce	produce	beverage
lin	produce	produce	beverage	beverage
plg	vegetable	vegetable	<b>dairy product</b>	abstraction
<b>Csf</b>	cultivate	cultivate	nakedness	farm
manual	homegrown food		<b>dairy product</b>	
	Topic 74		Topic 84	
method	meat, feed, beef, grass, eat, raise, cow, buy, make, animal		company, tea, brand, product, drink, honest, new, beverage, consumer, goldman	
path	entity	meat	<b>beverage</b>	<b>beverage</b>
ich	entity	abstraction	physical entity	substance
res	matter	meat	substance	substance
jsn	food	meat	<b>beverage</b>	<b>beverage</b>
lin	matter	meat	<b>beverage</b>	<b>beverage</b>
plg	cattle	physical entity	food	food
<b>Csf</b>	cattle	be	<b>beverage</b>	<b>beverage</b>
manual	animal husbandry		<b>beverage</b>	

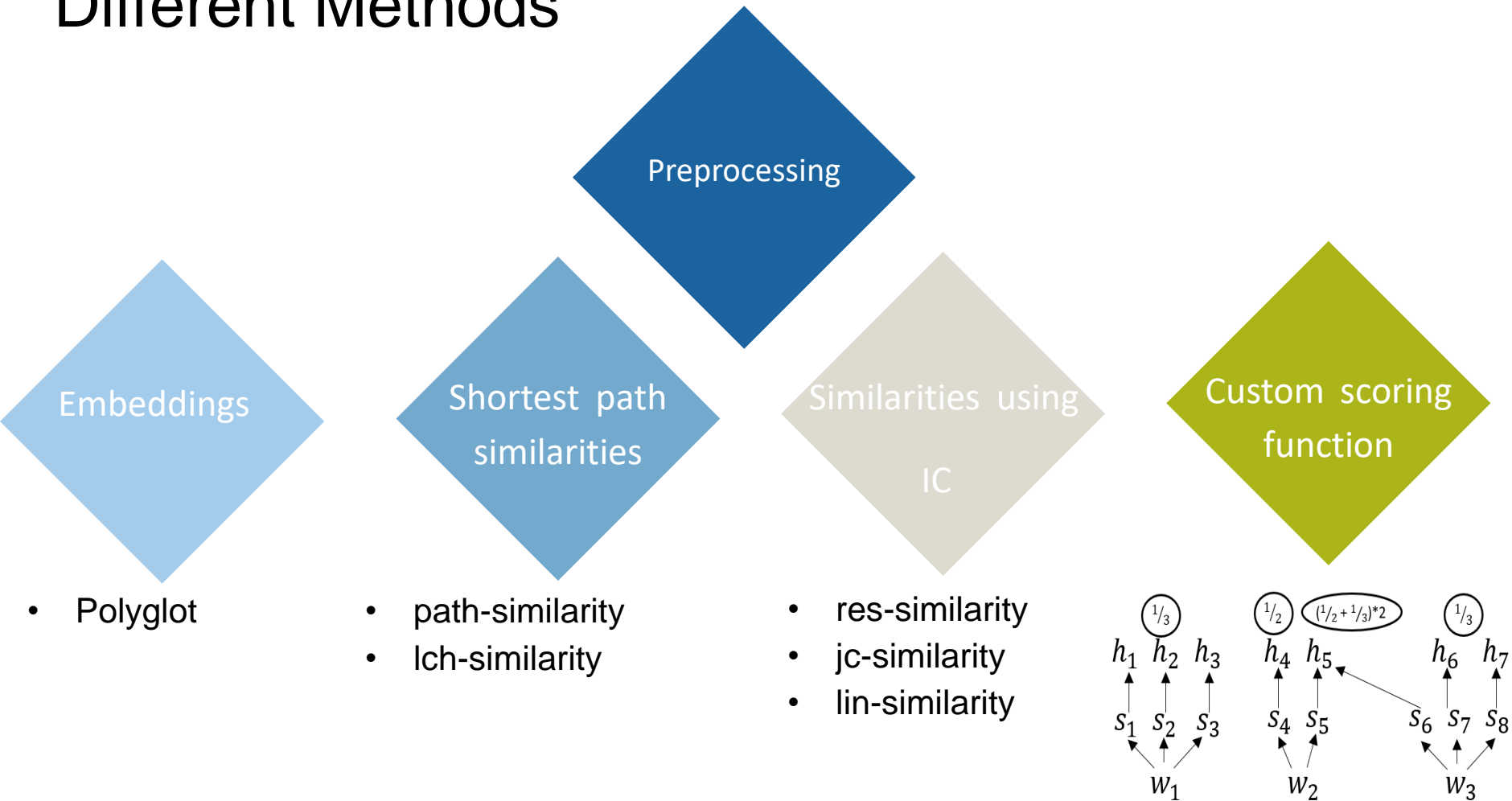
## Extrinsic Ranking according to non-significant words

1. <b>res-similarity</b>	2. <b>lin-similarity</b>	3. polyglot embeddings (plg)
4. res-similarity	5. jsn-similarity	6. lin-similarity
7. <b>path-similarity</b>	8. <b>polyglot embeddings</b>	9. jsn-similarity
10. <b>ich-similarity</b>	11. path-similarity	12. ich-similarity

**Tab. 3.8.:** Ranked similarity functions. **Bold** similarities denote the similarities, which were applied on preprocessed topics.



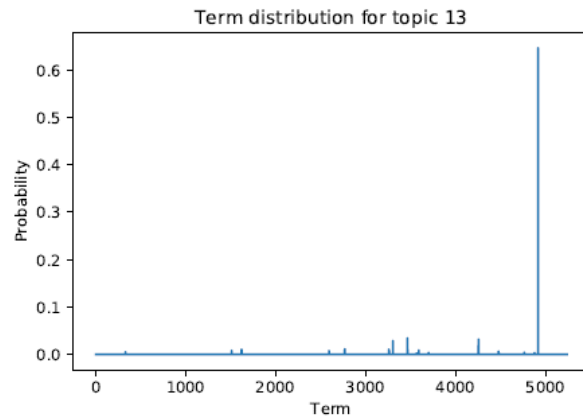
# Different Methods



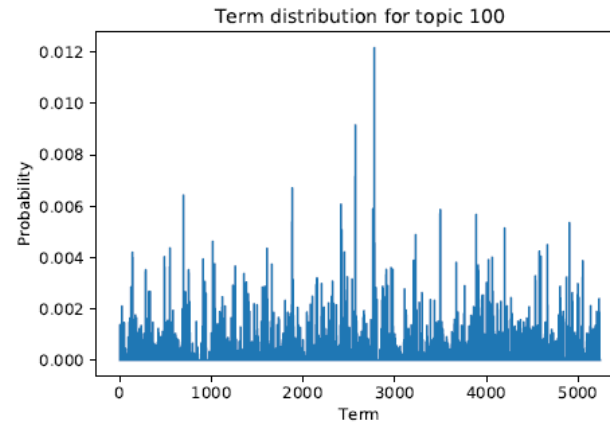
# Internal consistency: Entropy



(a) Plotted entropy for German editorial articles



(b) Topic coverage for the topic with the lowest entropy

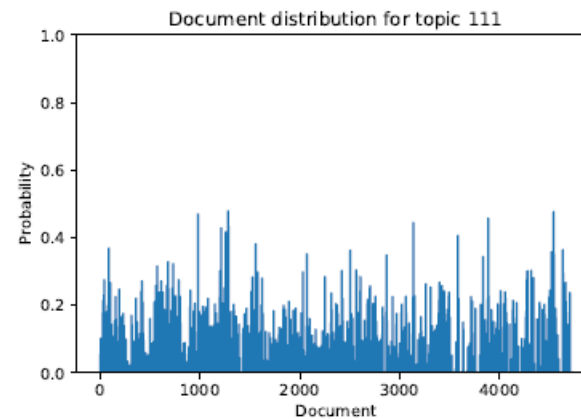
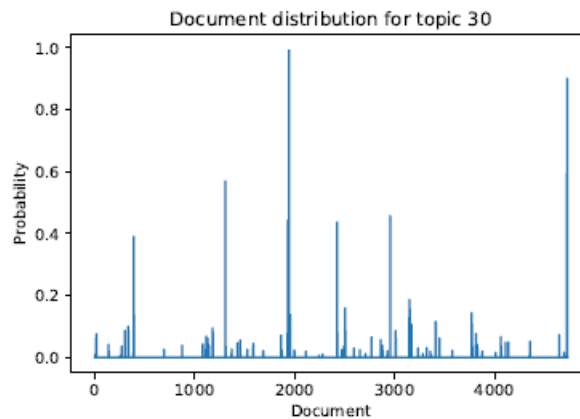


(c) Topic coverage for the topic with the highest entropy

# Internal consistency: Alpha



(a) Plotted alphas for German editorial articles



(b) Document coverage for the topic with the lowest alpha value (c) Document coverage for the topic with the highest alpha value

# Internal consistency: Theta

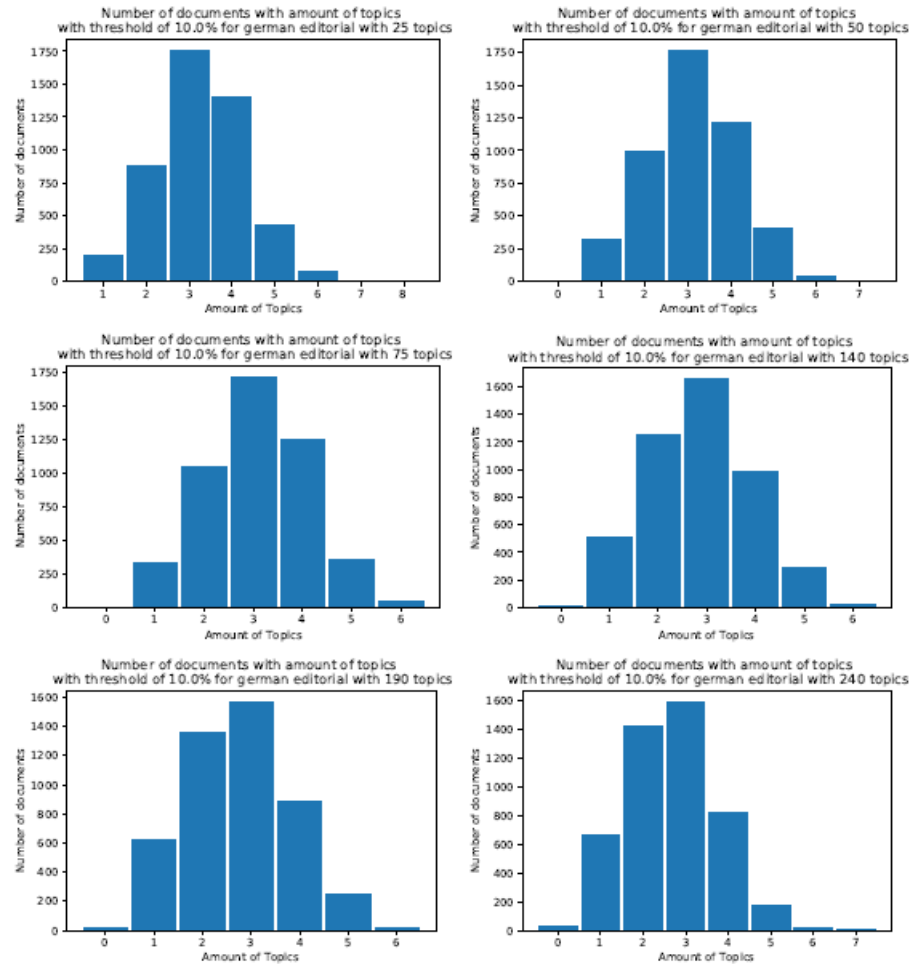


Fig. 4.18.: Amount of topics in documents over a threshold of 10% for German editorial articles