



Ukraine could be 'next  
Afghanistan' for Russia  
if it invades, US senator  
warns

→ Read more

The official pointed  
military intervention  
deployed in eastern  
time.

"It would certainly be an  
increasing request from  
United States, for adding  
place there to ensure the  
face of that kind of aggressive  
that Biden would not be threatened  
most Russian  
equipment being  
ear response this  
would be an  
response from the  
exercises to take  
ern flank allies in the  
al said, but made clear  
military response.



#### Most viewed



'Not great news': US boss  
fires 900 employees on a  
Zoom call



Drake withdraws his two  
2022 Grammy nominations



**Live** Covid news live:  
Omicron likely to become  
dominant variant, UK and  
US experts say

weBlock | 29.03.22

Jannis Baum, Lucas Liebe, Tilman Schütze

Betreuer: Tim Cech

Planung und Konstruktion von AI-basierten interaktiven Systemen mit einer "dunklen Seite der KI"

# Projekt | Motivation

1

- AI-gestützte Zensur von kritischen Informationen & Meinungen

2

- Austausch mit affirmativen Informationen

3

- Einfache Konfigurierbarkeit für Regime aller Art

4

- Tarnung als Adblocker zur Motivation von Nutzern



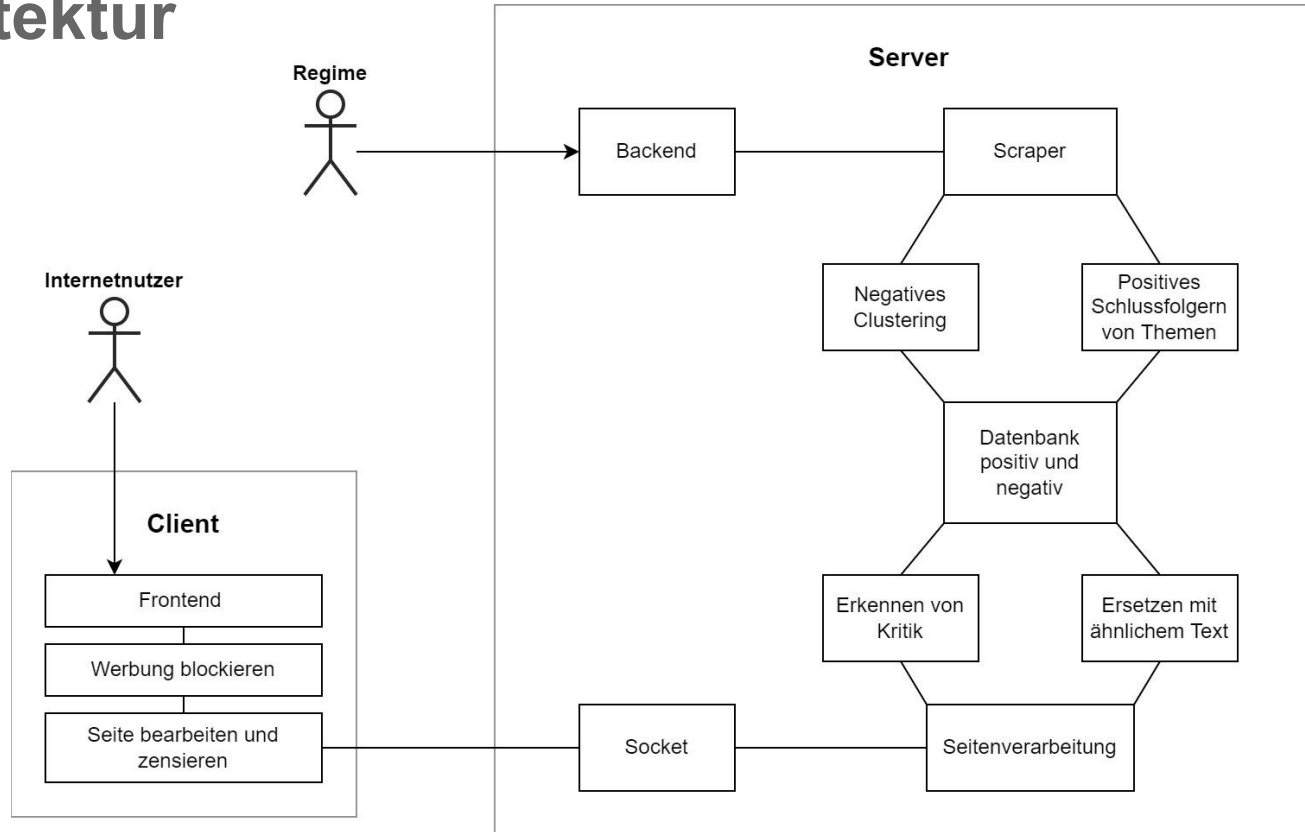
<http://www.schindluder.net/wp-content/uploads/2017/02/zensur.jpg>

# Demonstration

**Demo auf *Human rights Watch***

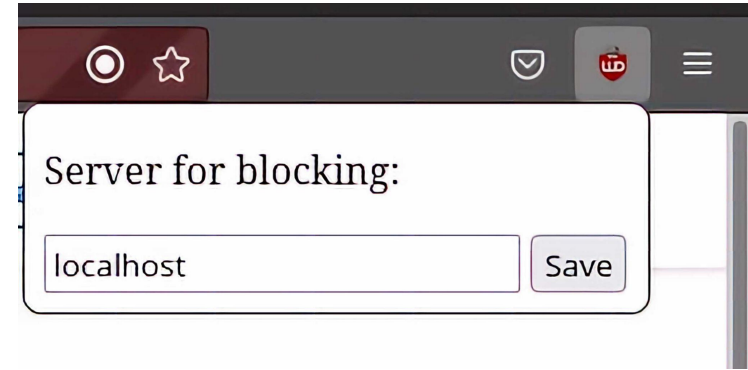


# Architektur



# Client

- ❑ Yarn, Node, Mozilla Firefox
- ❑ Zensur in 2 Schritten
  - ❑ Rotfärbung des als kritisch erkannten Textes
- ❑ Blockieren von Werbung mit CSS Selektoren
  - ❑ Nicht auf dem Level anderer Adblocker
  - ❑ Für viele Nachrichten Webseiten
- ❑ Konfiguration des Zensur Servers



# Server | NLP Einführung

Wie können Algorithmen natürliche Sprache betrachten?

# Server | NLP Einführung

## Wie können Algorithmen natürliche Sprache betrachten?

“Natural language processing (NLP) is [...] concerned with [...] how to program computers to process and analyze large amounts of natural language data.”

([en.wikipedia.org/wiki/Natural\\_language\\_processing](https://en.wikipedia.org/wiki/Natural_language_processing))

# Server | NLP Einführung

## Wie können Algorithmen natürliche Sprache betrachten?

“Natural language processing (NLP) is [...] concerned with [...] how to program computers to process and analyze large amounts of natural language data.”

### ↓ Character filtering & Tokenization

[natural, language, processing, nlp, is, concerned, with, how, to, program, computers, to, process, and, analyze, large, amounts, of, natural, language, data]



# Server | NLP Einführung

## Wie können Algorithmen natürliche Sprache betrachten?

“Natural language processing (NLP) is [...] concerned with [...] how to program computers to process and analyze large amounts of natural language data.”

↓ Character filtering & Tokenization

[natural, language, processing, nlp, ~~is~~, concerned, ~~with~~, ~~how~~, ~~to~~, program, computers, ~~to~~, process, ~~and~~, analyze, large, amounts, ~~of~~, natural, language, data]

↓ **Stopword removal** (NLTK)

[natural, language, processing, nlp, concerned, program, computers, process, analyze, large, amounts, natural, language, data]

# Server | NLP Einführung

## Wie können Algorithmen natürliche Sprache betrachten?

“Natural language processing (NLP) is [...] concerned with [...] how to program computers to process and analyze large amounts of natural language data.”

### ↓ Character filtering & Tokenization

[natural, language, processing, nlp, is, concerned, with, how, to, program, computers, to, process, and, analyze, large, amounts, of, natural, language, data]

### ↓ Stopword removal (NLTK)

[natural<sup>al</sup>, language<sup>e</sup>, process<sup>ing</sup>, nlp, concern<sup>ed</sup>, program, comput<sup>ers</sup>, process, analyz<sup>e</sup>, larg<sup>e</sup>, amount<sup>s</sup>, natur<sup>al</sup>, languag<sup>e</sup>, data]

### ↓ Stemming (NLTK)

[natur, languag, process, nlp, concern, program, comput, process, analyz, larg, amount, natur, languag, data]

# Server | NLP Einführung

## Wie können Algorithmen natürliche Sprache betrachten?

“Natural language processing (NLP) is [...] concerned with [...] how to program computers to process and analyze large amounts of natural language data.”

### ↓ Character filtering & Tokenization

[natural, language, processing, nlp, is, concerned, with, how, to, program, computers, to, process, and, analyze, large, amounts, of, natural, language, data]

### ↓ Stopword removal (NLTK)

[natural, language, processing, nlp, concerned, program, computers, process, analyze, large, amounts, natural, language, data]

### ↓ Stemming (NLTK)

[natur, languag, process, nlp, concern, program, comput, process, analyz, larg, amount, natur, languag, data]

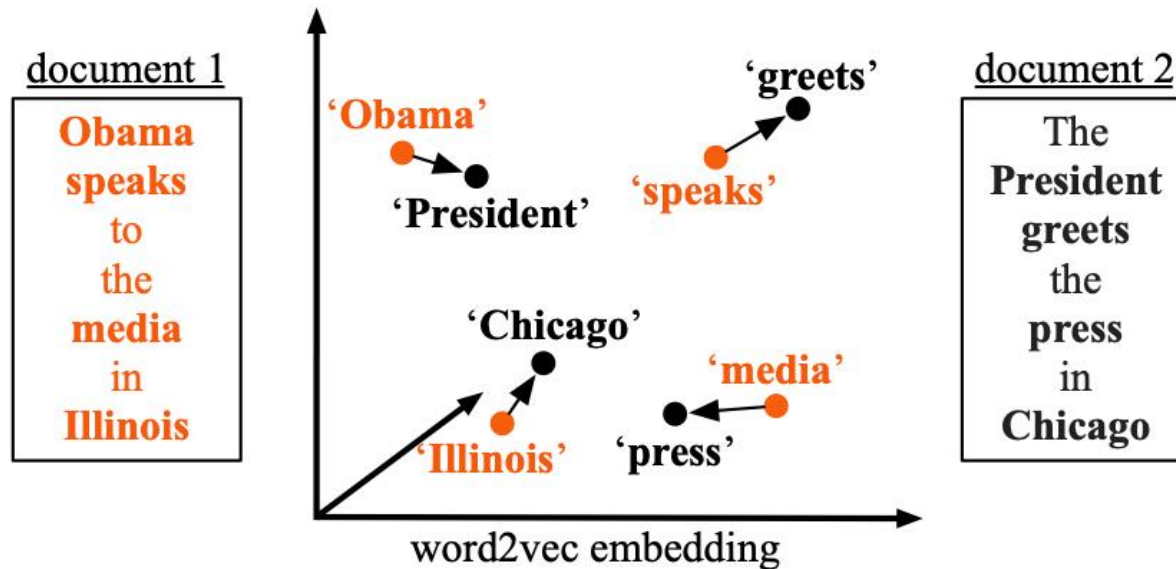
### ↓ Indexing

document: [0, 1, 2, 3, 4, 5, 6, 2, 7, 8, 9, 0, 1, 10], corpus vocabulary: { natur: 0, languag: 1, process: 2, ... }

# Server | Rückblick

## Word Mover's Distance

□ Aufbauend auf Word Embeddings:



<https://towardsai.net/p/nlp/word-movers-distance-wmd-explained-an-effective-method-of-document-classification-89cb258401f4>

# Server | Rückblick

## Word Mover's Distance

- ❑ WMD misst semantische Verschiedenheit
- ❑ Invertiert als Maß für Ähnlichkeit von Absätzen
- ❑ Gensim mit word2vec-google-news-300

# Server | Rückblick

## Sentiment Analysis & Synonyme

- ❏ Weiterhin NLTK Vader Modell zur Sentiment Analysis
  - ➡ Negative/positive Grundstimmung im Text (zwischen 0 und 1)
  
- ❏ NLTK Wordnet Corpus für Zuordnung von Synonymen
  - ➡ Suche nach Synonymen benötigter Begriffe im Text

# Server | Preprocessing Datensammlung

- ☐ Google News als Datenquelle
- ☐ Artikelsuche mit Selenium (Geckodriver, Mozilla Firefox, BeautifulSoup)
- ☐ Datenbank für positive und negative Texte



<https://news.google.com/topstories?hl=en-US&gl=US&ceid=US:en>

# Server | Preprocessing

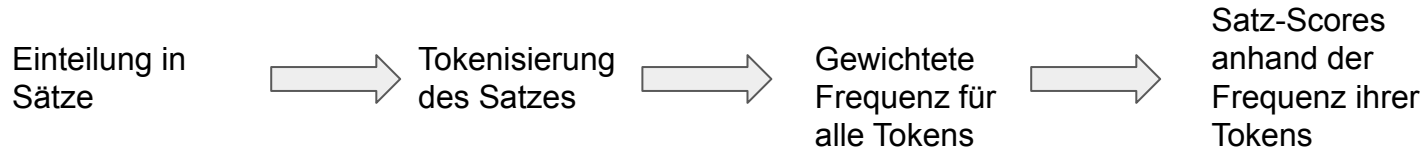
## Konfiguration der Zensurkriterien

- ❑ Konfiguration von unerwünschten Inhalten
- ❑ Extraktive Zusammenfassung der Artikel mit Wortfrequenz
- ❑ Clustering mit Random Search



# Server | Preprocessing

## Zusammenfassung mit Wortfrequenzen



Ergebnis: Top n Sätze sind die Zusammenfassung

# Server | Preprocessing

## Random Search

“The President...”

“On other news...”

“Police have arrested...”

“According to...”

“Breaking news...”

“Rumors from the...”

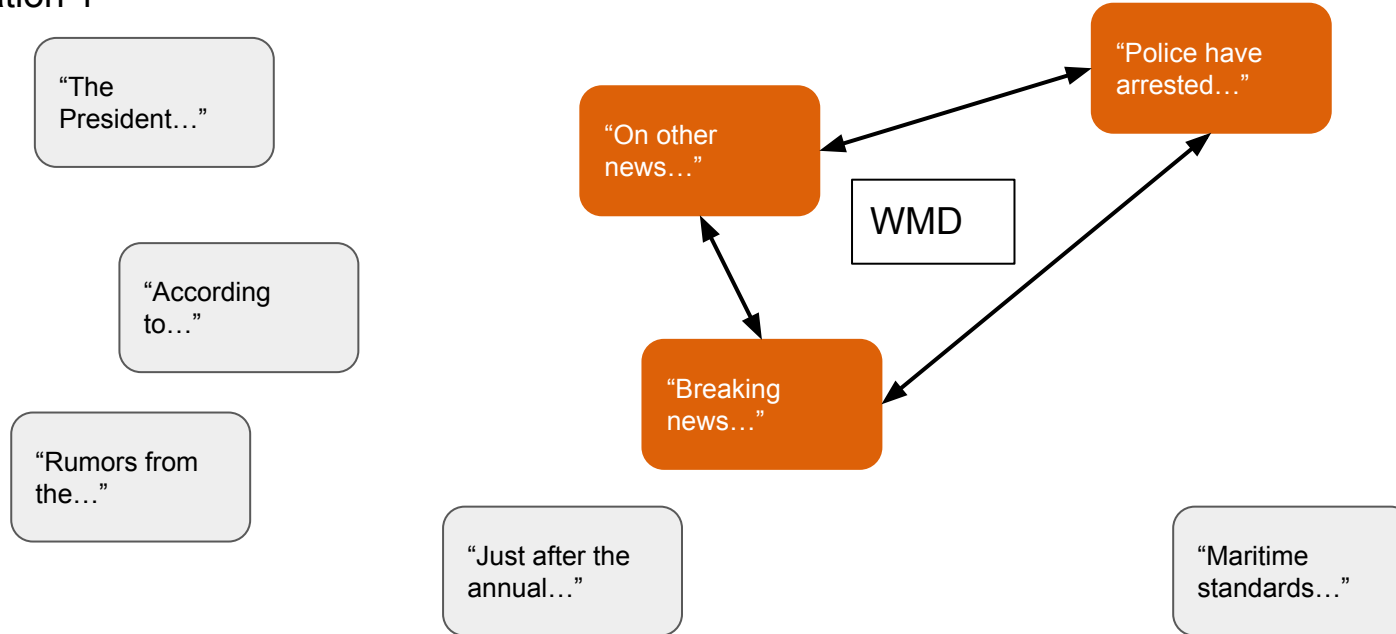
“Just after the annual...”

“Maritime standards...”

# Server | Preprocessing

## Random Search

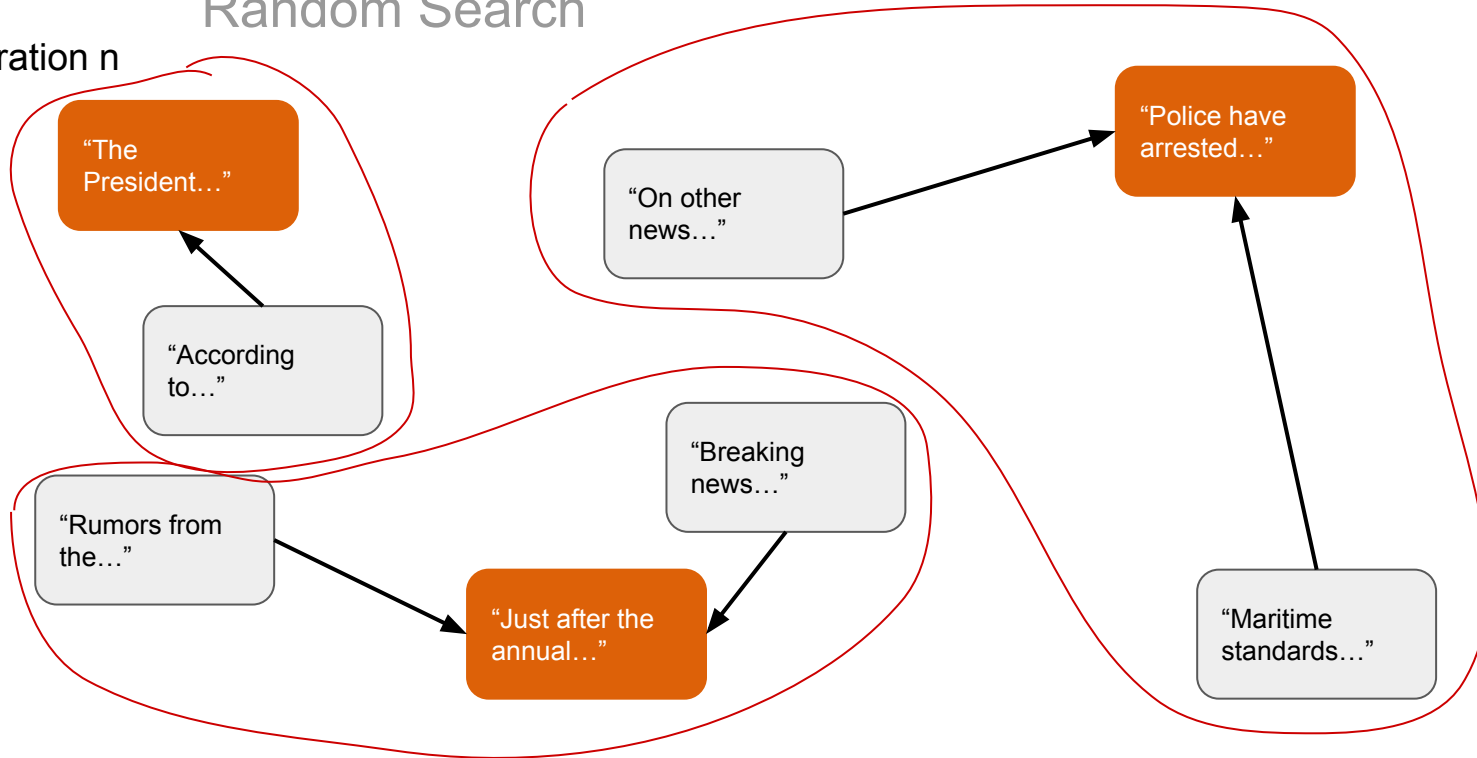
Iteration 1



# Server | Preprocessing

## Random Search

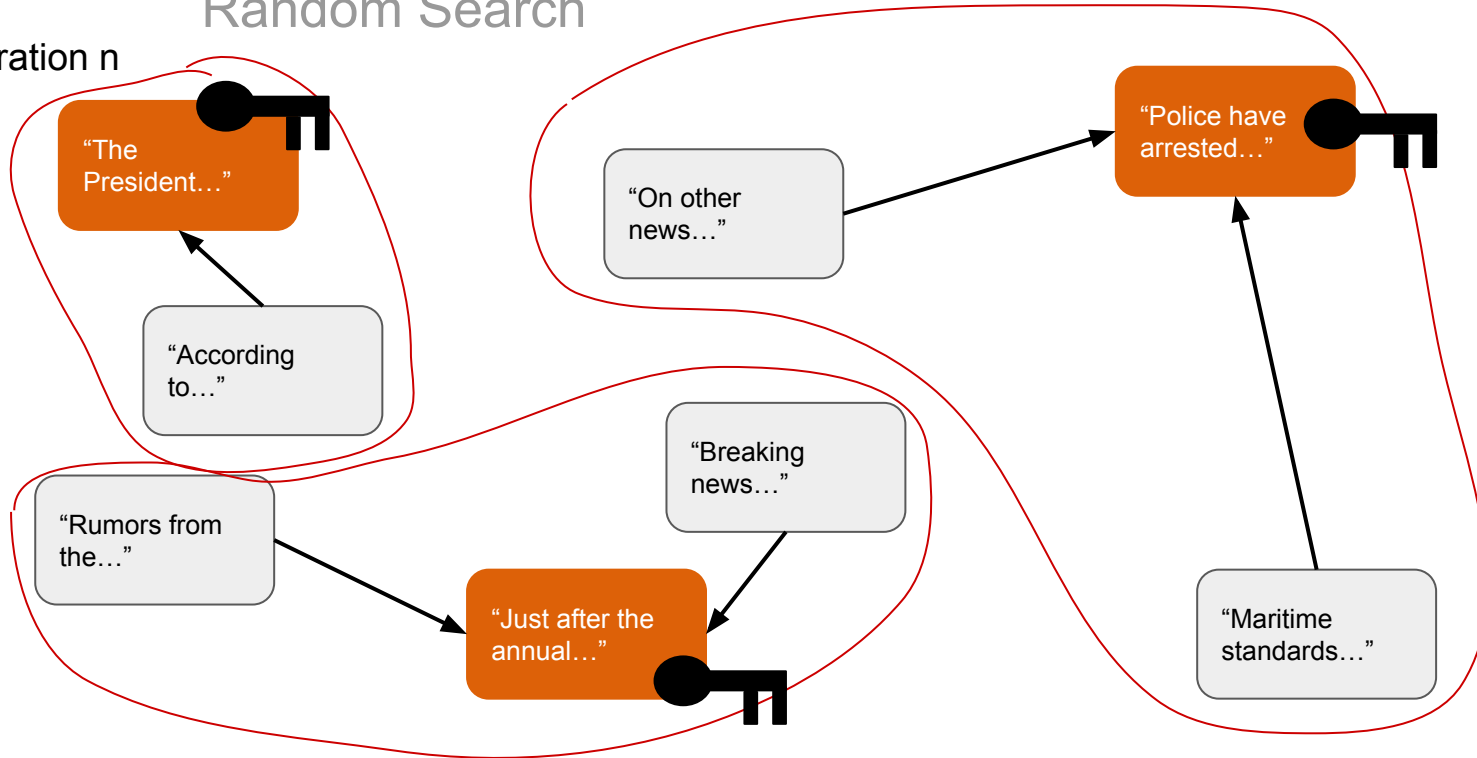
Iteration n



# Server | Preprocessing

## Random Search

Iteration n



# Server | Rückblick

## Austauschen von textuellen Inhalten

- ❑ GPT2: nicht-deterministisches Generieren von (“manchmal ganz gutem”) Text
- ❑ GPT3: Deterministik möglich, Qualität lässt für unsere Anwendung immer noch zu wünschen übrig
- ❑ Völlig anderer Ansatz?

# Server | Preprocessing

## Erkennen von Themen in Paragraphen

- Ähnliches Thema im Wunschparagraphen → guter Austausch-Kandidat?

# Server | Preprocessing

## Erkennen von Themen in Paragraphen

- ❑ Ähnliches Thema im Wunschparagraphen → guter Austausch-Kandidat?

LDA\* weitverbreiteter Topic Modelling-Ansatz

- ❑ betrachtet dokumentweites Auftreten einzelner Wörter
- ❑ schwächer bei kürzeren Texten

*\*Latent Dirichlet Allocation*



# Server | Preprocessing

## Erkennen von Themen in Paragraphen

- ❑ Ähnliches Thema im Wunschparagraphen → guter Austausch-Kandidat?

LDA\* weitverbreiteter Topic Modelling-Ansatz    BTM\* unbekannter, gemacht für kurze Texte

- ❑ betrachtet dokumentweites Auftreten einzelner Wörter
- ❑ schwächer bei kürzeren Texten
- ❑ betrachtet **korpusweites** gemeinsames Auftreten von **Biterms** (Wortpaaren)

*\*Latent Dirichlet Allocation*

*\*Biterm Topic Model*

# Server | Preprocessing

## Biterm Topic Model - Training

### Input

Gegeben

$m$  positiven Paragraphen  $d_j$

bestehend aus Worten  $w$ ,

finde  $n$  Themen  $z_i$

# Server | Preprocessing

## Biterm Topic Model - Training

### Input

Gegeben

$m$  positiven Paragraphen  $d_j$   
bestehend aus Worten  $w$ ,  
finde  $n$  Themen  $z_i$



BTM  
Training

# Server | Preprocessing

## Biterm Topic Model - Training

### Input

Gegeben

$m$  positiven Paragraphen  $d_j$   
bestehend aus Worten  $w$ ,  
finde  $n$  Themen  $z_i$



BTM  
Training

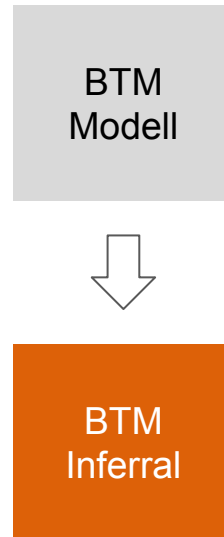


### BTM Modell

Themenverteilung  $P(z)$   
Wortverteilung  $P(w|z)$

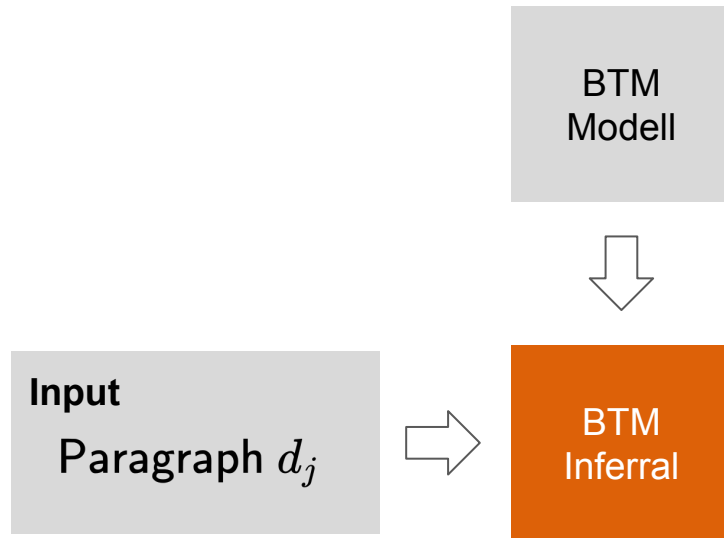
# Server | Preprocessing

## Biterm Topic Model - Inferral



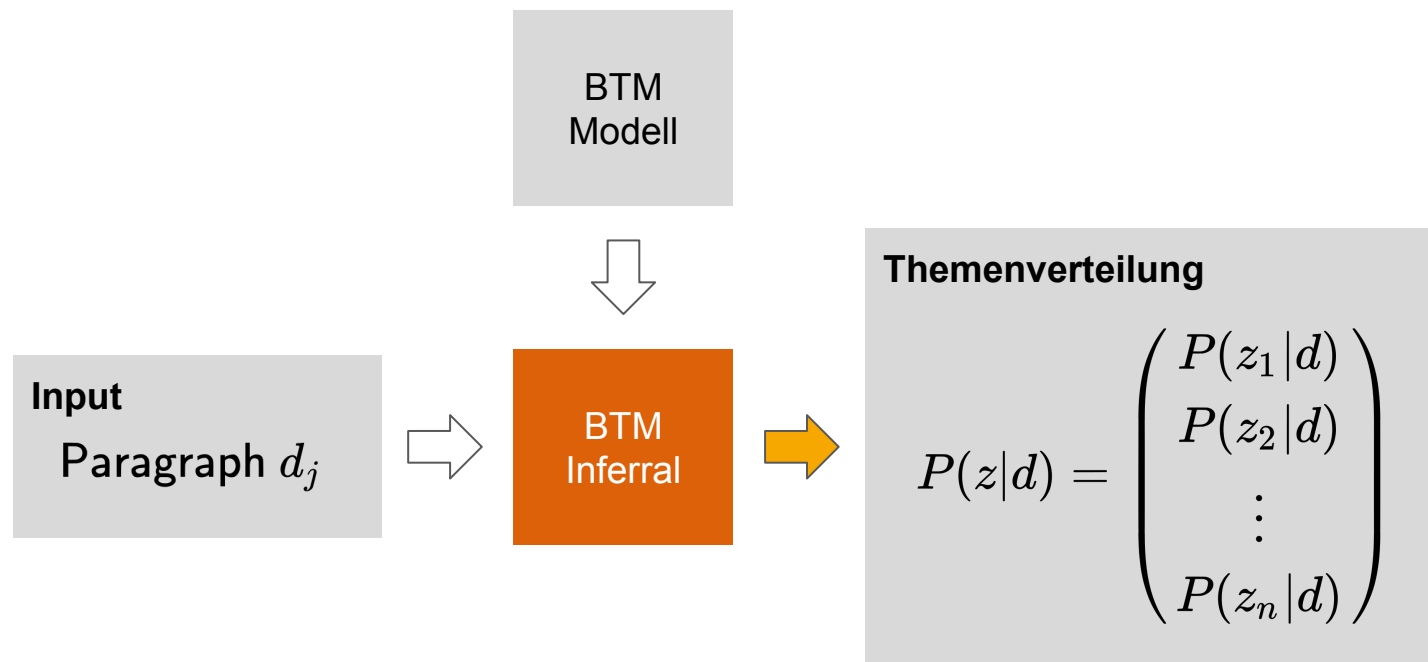
# Server | Preprocessing

## Biterm Topic Model - Inferral



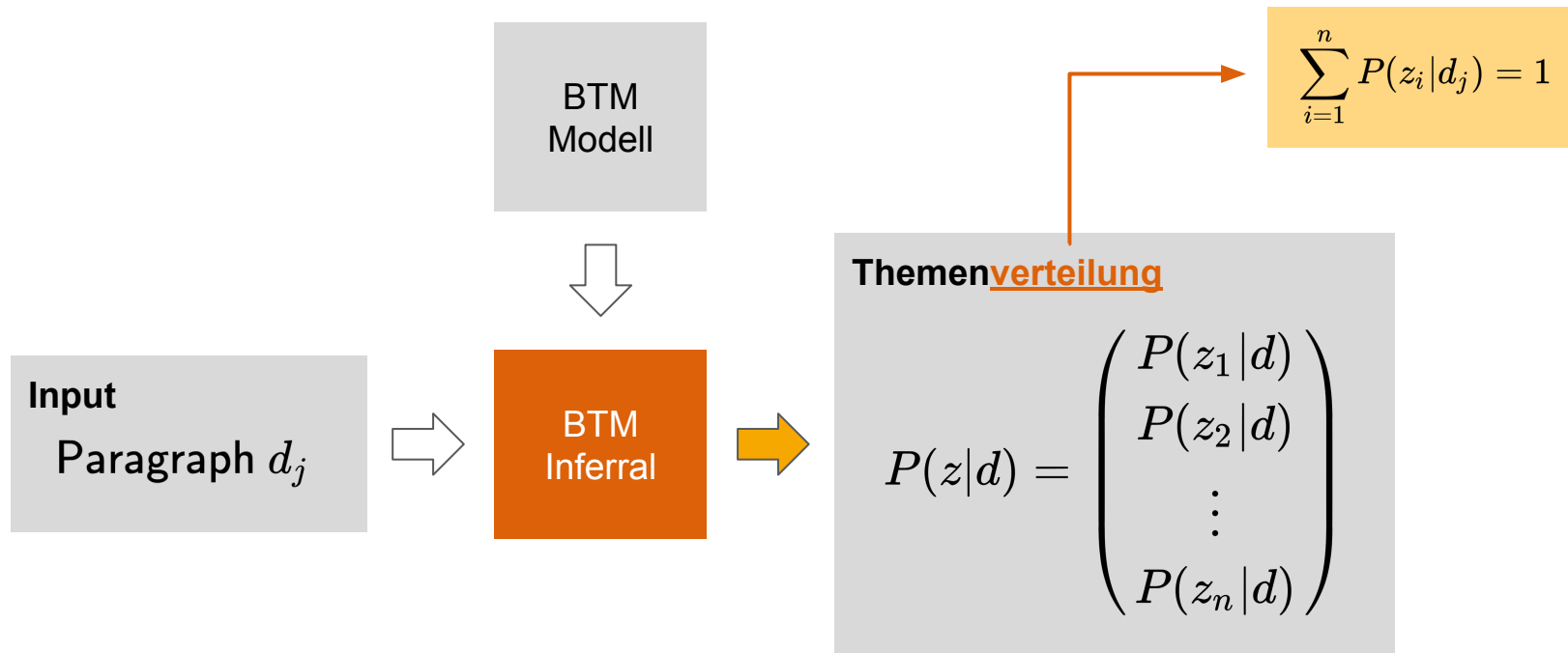
# Server | Preprocessing

## Biterm Topic Model - Inferral



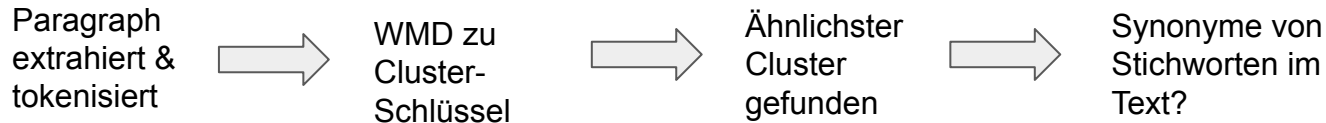
# Server | Preprocessing

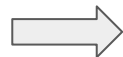
## Biterm Topic Model - Inferred

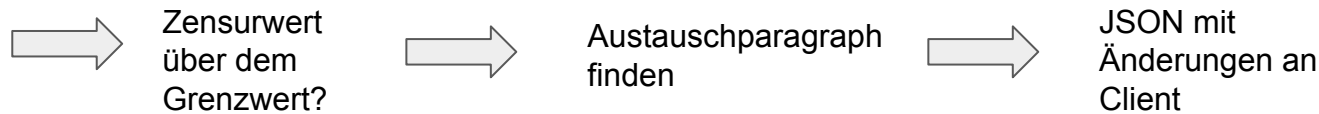




# Server | Zensurprozess




$$\text{score} = (\text{avg}(\text{context-WMD}) + \text{sentiment}) \cdot (\text{sentiment} + 0.01) \cdot 100$$



# Server | Zensurprozess

Zuordnen passender Austauschparagraphen

zu zensierender  
Paragraph  $d^*$

# Server | Zensurprozess

Zuordnen passender Austauschparagraphen

zu zensierender  
Paragraph  $d^*$



$$\begin{pmatrix} P(z_1 | d^*) \\ P(z_2 | d^*) \\ \vdots \\ P(z_n | d^*) \end{pmatrix}$$

# Server | Zensurprozess

Zuordnen passender Austauschparagraphen

positive Paragraphen  $d_j$  aus Scraping

zu zensierender  
Paragraph  $d^*$



$$\begin{pmatrix} P(z_1 | d^*) \\ P(z_2 | d^*) \\ \vdots \\ P(z_n | d^*) \end{pmatrix}$$

# Server | Zensurprozess

Zuordnen passender Austauschparagraphen

positive Paragraphen  $d_j$  aus Scraping



$$\begin{pmatrix} P(z_1|d_1) & P(z_2|d_1) & \cdots & P(z_n|d_1) \\ P(z_1|d_2) & P(z_2|d_2) & \cdots & P(z_n|d_2) \\ \vdots & \vdots & \ddots & \vdots \\ P(z_1|d_m) & P(z_2|d_m) & \cdots & P(z_n|d_m) \end{pmatrix}$$

zu zensierender  
Paragraph  $d^*$



$$\begin{pmatrix} P(z_1|d^*) \\ P(z_2|d^*) \\ \vdots \\ P(z_n|d^*) \end{pmatrix}$$

# Server | Zensurprozess

Zuordnen passender Austauschparagraphen

positive Paragraphen  $d_j$  aus Scraping



$$\begin{pmatrix} P(z_1|d_1) & P(z_2|d_1) & \cdots & P(z_n|d_1) \\ P(z_1|d_2) & P(z_2|d_2) & \cdots & P(z_n|d_2) \\ \vdots & \vdots & \ddots & \vdots \\ P(z_1|d_m) & P(z_2|d_m) & \cdots & P(z_n|d_m) \end{pmatrix}$$

zu zensierender  
Paragraph  $d^*$



$$\begin{pmatrix} P(z_1|d^*) \\ P(z_2|d^*) \\ \vdots \\ P(z_n|d^*) \end{pmatrix}$$

# Server | Zensurprozess

Zuordnen passender Austauschparagraphen

positive Paragraphen  $d_j$  aus Scraping



$$\begin{pmatrix} P(z_1|d_1) & P(z_2|d_1) & \cdots & P(z_n|d_1) \\ P(z_1|d_2) & P(z_2|d_2) & \cdots & P(z_n|d_2) \\ \vdots & \vdots & \ddots & \vdots \\ P(z_1|d_m) & P(z_2|d_m) & \cdots & P(z_n|d_m) \end{pmatrix}$$

zu zensierender  
Paragraph  $d^*$



$$\begin{pmatrix} P(z_1|d^*) \\ P(z_2|d^*) \\ \vdots \\ P(z_n|d^*) \end{pmatrix}$$

$$\sum_{i=1}^n P(z_i|d_j) = 1$$

# Server | Zensurprozess

Zuordnen passender Austauschparagraphen

positive Paragraphen  $d_j$  aus Scraping



zu zensierender  
Paragraph  $d^*$



$$\begin{pmatrix} P(z_1|d_1) & P(z_2|d_1) & \cdots & P(z_n|d_1) \\ P(z_1|d_2) & P(z_2|d_2) & \cdots & P(z_n|d_2) \\ \vdots & \vdots & \ddots & \vdots \\ P(z_1|d_m) & P(z_2|d_m) & \cdots & P(z_n|d_m) \end{pmatrix} \cdot \begin{pmatrix} P(z_1|d^*) \\ P(z_2|d^*) \\ \vdots \\ P(z_n|d^*) \end{pmatrix} = \begin{pmatrix} \sim_{d_1} \\ \sim_{d_2} \\ \vdots \\ \sim_{d_m} \end{pmatrix}$$



# Server | Zensurprozess

Zuordnen passender Austauschparagraphen

positive Paragraphen  $d_j$  aus Scraping



$$\begin{pmatrix} P(z_1|d_1) & P(z_2|d_1) & \cdots & P(z_n|d_1) \\ P(z_1|d_2) & P(z_2|d_2) & \cdots & P(z_n|d_2) \\ \vdots & \vdots & \ddots & \vdots \\ P(z_1|d_m) & P(z_2|d_m) & \cdots & P(z_n|d_m) \end{pmatrix}$$

zu zensierender  
Paragraph  $d^*$



$$\begin{pmatrix} P(z_1|d^*) \\ P(z_2|d^*) \\ \vdots \\ P(z_n|d^*) \end{pmatrix}$$

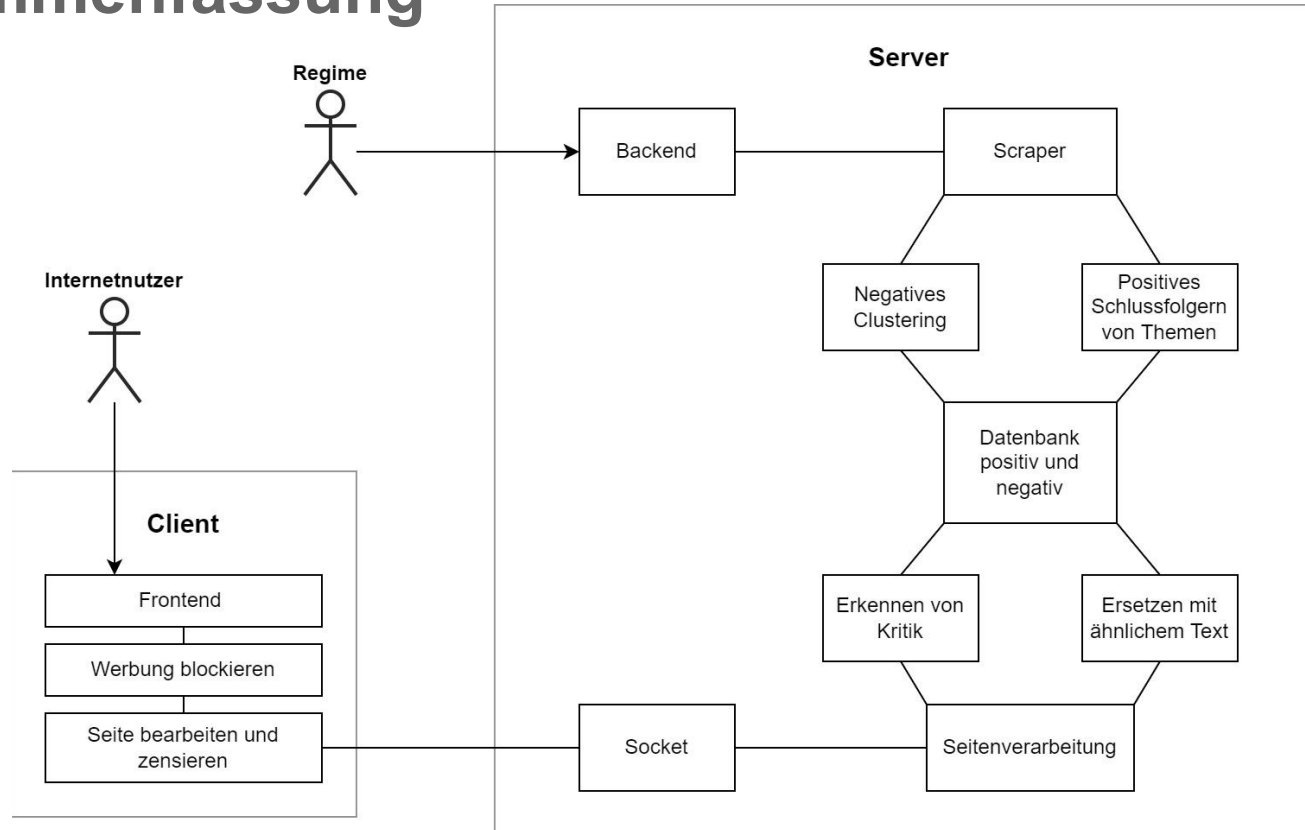
=

$$\begin{pmatrix} \sim_{d_1} \\ \sim_{d_2} \\ \vdots \\ \sim_{d_m} \end{pmatrix}$$

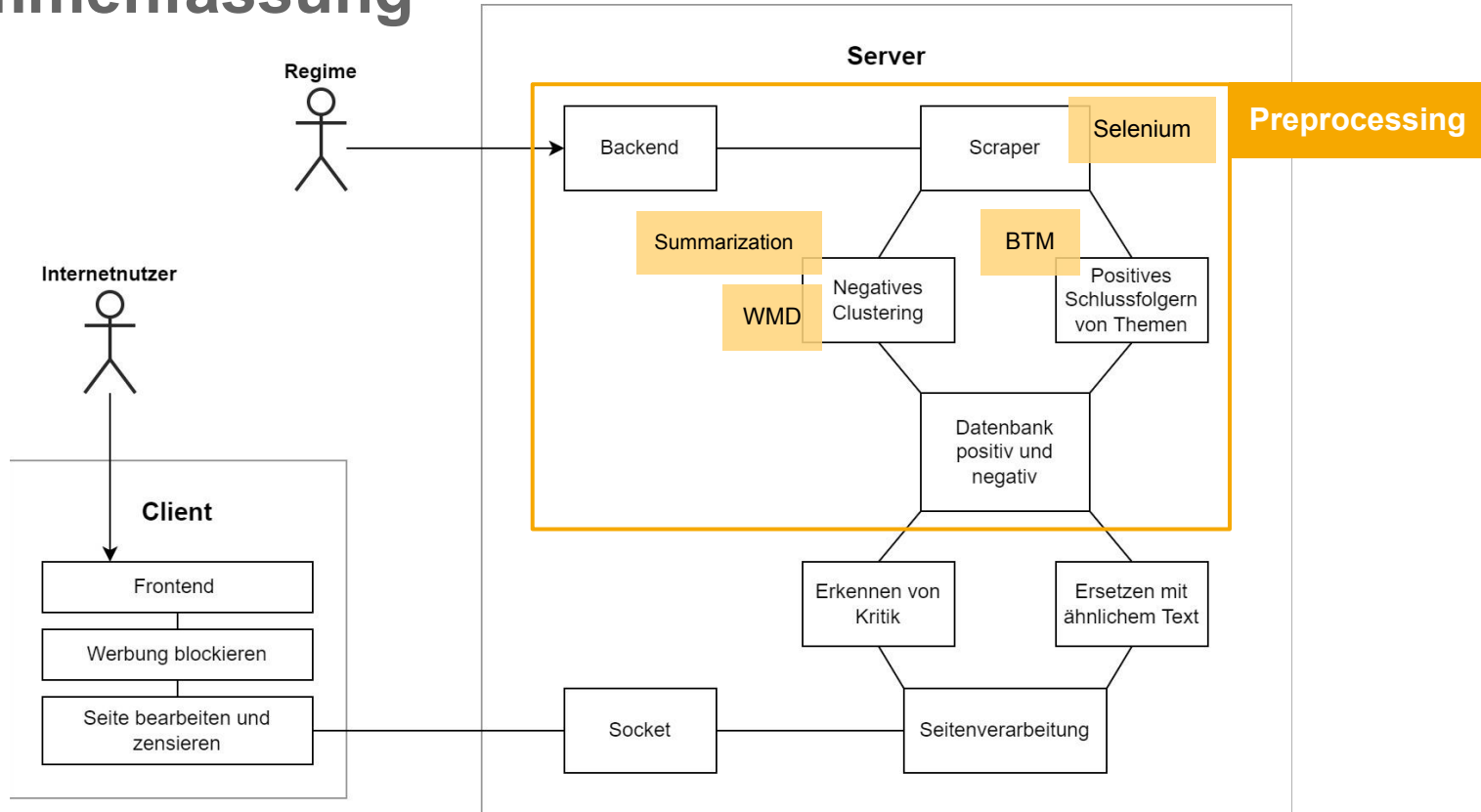


$\max(\sim_d)$  korrespondiert mit dem Paragraphen  $d'$ , der ähnlichste Themenverteilung zu  $d^*$  hat

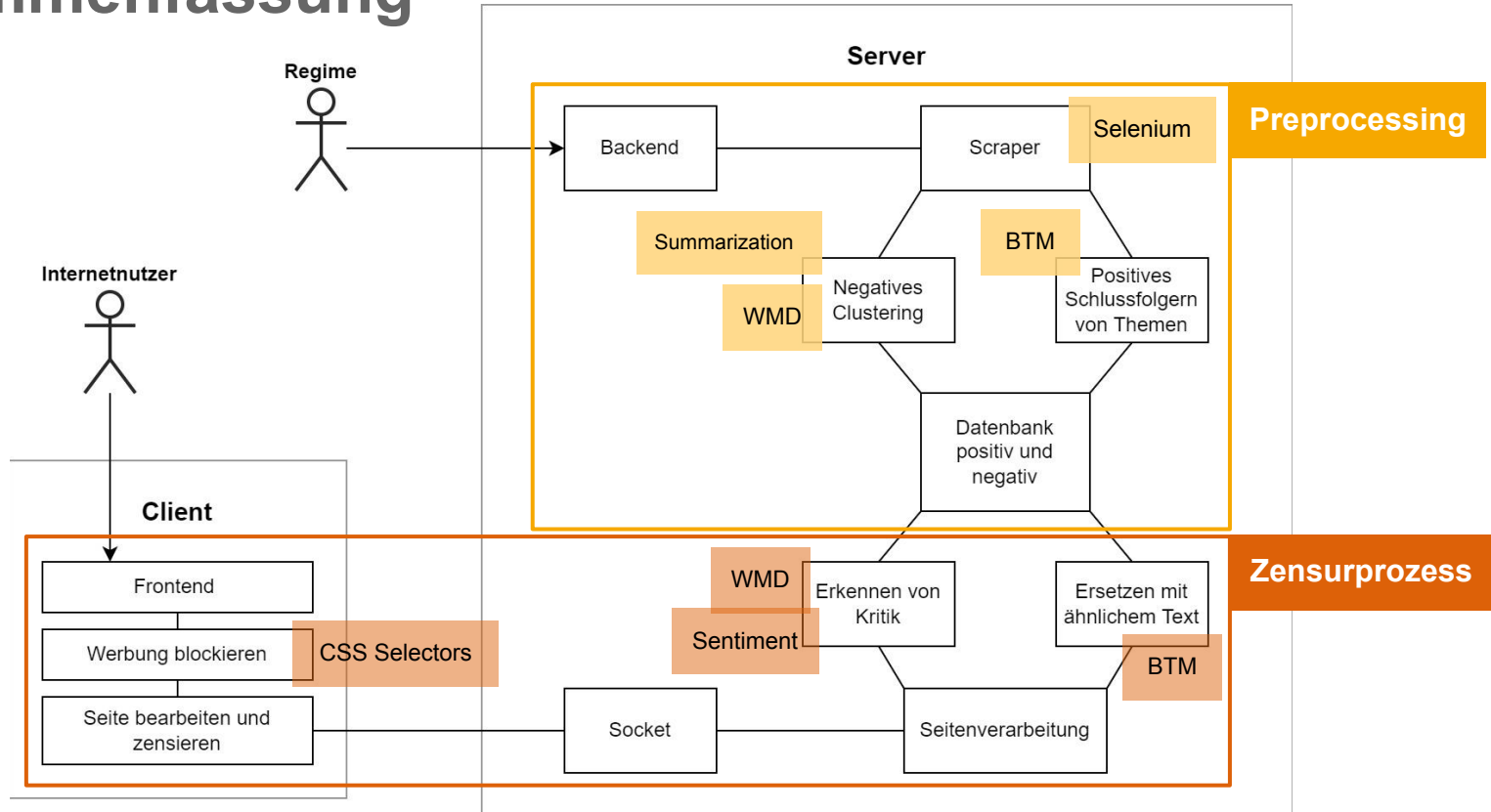
# Zusammenfassung



# Zusammenfassung



# Zusammenfassung



# Fazit | Skalierbarkeit

- ❑ “Datenbank” & Beschränkungen durch RAM-Größe
- ❑ Server & Umgang mit Anfragen
- ❑ Fokus auf Leistungsoptimierung
- ❑ Zensur von anderen als textuellen Medien (z.B. Bilder, Videos, Dokumente)

# Fazit | Anwendbarkeit

- ❑ Per Hand bearbeitete Datenbank
  - ❑ Sehr viele Artikel benötigt
  - ❑ Sollte über längeren Zeitraum wachsen und gesäubert werden
  
- ❑ Requirements sorgen erst für Genauigkeit
  
- ❑ Mit kleinen Anpassungen erschreckend gut

# Quellen | Word Mover's Distance

- **Supervised Word Mover's Distance**

Gao Huang\*, Chuan Guo\*, Matt J. Kusner, Yu Sun, Kilian Q. Weinberger, Fei Sha  
Neural Information Processing Systems (NeurIPS), 2016. Oral Presentation.

- **Fast and robust Earth Mover's Distances**

O. Pele and M. Werman,  
2009 IEEE 12th International Conference on Computer Vision, 2009, pp. 460-467, doi: 10.1109/ICCV.2009.5459199.

- <https://github.com/RaRe-Technologies/gensim-data>
- [https://radimrehurek.com/gensim/auto\\_examples/tutorials/run\\_wmd.html#computing-the-word-mover-s-distance](https://radimrehurek.com/gensim/auto_examples/tutorials/run_wmd.html#computing-the-word-mover-s-distance)
- <https://radimrehurek.com/gensim/models/keyedvectors.html>
- [https://en.wikipedia.org/wiki/Earth\\_mover%27s\\_distance](https://en.wikipedia.org/wiki/Earth_mover%27s_distance)

# Quellen | Topic Modelling

- **A Bitern Topic Model for Short Texts**  
Xiaohui Yan, Jiafeng Guo, Yanyan Lan, Xueqi Cheng  
Institute of Computing Technology, CAS  
  
<https://github.com/xiaohuiyan/BTM>
- **Latent Dirichlet Allocation**  
David M. Blei, Andrew Y. Ng, Michael I. Jordan  
Journal of Machine Learning Research 3 (2003) 993-1022
- **Introduction to the Dirichlet Distribution and Related Processes**  
Bela A. Frigyik, Amol Kapila, Maya R. Gupta  
Department of Electrical Engineering, University of Washington  
  
<https://builtin.com/data-science/dirichlet-distribution>



# Quellen | Weitere

## Sentiment Analysis

- [https://www.researchgate.net/publication/330880816\\_Sentiment\\_Analysis\\_of\\_News\\_Articles\\_A\\_Lexicon\\_based\\_Approach](https://www.researchgate.net/publication/330880816_Sentiment_Analysis_of_News_Articles_A_Lexicon_based_Approach)
- <https://medium.com/@b.terryjack/nlp-pre-trained-sentiment-analysis-1eb52a9d742c>

## Weitere Quellen im Zusammenhang mit NLP

- <https://www.nltk.org>
- <https://monkeylearn.com/blog/what-is-tf-idf/>
- <https://towardsdatascience.com/understand-text-summarization-and-create-your-own-summarizer-in-python-b26a9f09fc70>

## Scraping

- <http://www.pavantestingtools.com/p/selenium.html>
- <https://beautiful-soup-4.readthedocs.io/en/latest/>
- <https://github.com/mozilla/geckodriver>

## Client

- <https://developer.mozilla.org/en-US/docs/Web/HTTP/CORS>
- <https://developer.mozilla.org/en-US/docs/Mozilla/Add-ons/WebExtensions>

# Quellen | Unsere Repositories

- **Hauptprojekt** <https://github.com/jannis-baum/weBlock/>
- **BTM-Fork & Python-Wrapper** <https://github.com/jannis-baum/bitern-topic-model>