

# Großputz im CMS!

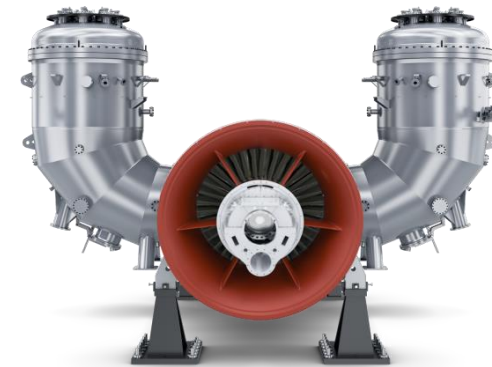
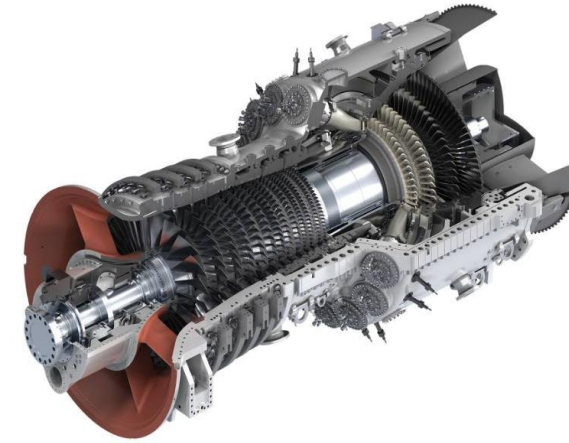
Semantische Ähnlichkeitsanalyse  
für XML-Module

Jan Oevermann (ICMS)  
Timo Fleschutz-Balarezo (SIEMENS)



# Content Management bei Siemens Gas Turbines Berlin

- Dokumentation
  - Produkthandbuch und interne Vertriebsinformationen für Gasturbinen für Großkraftwerke (**13** Mitarbeiter)
  - Aktuell **5** Produktfamilien in verschiedenen Versionen und kundenspezifischen Anpassungen
  - Über **300.000** Textbausteine in mehreren Sprachen



# Content Management bei Siemens Gas Turbines Berlin

## – System

- 2015 Umstieg von Word-basiertem System auf XML-basiertes CMS *TechPub Studio*
- Maschinelle Migration der Word-Dokumente in PI-Mod XML-Module

```
<descriptive nodeid="PI-70006536">  
  <heading>Fuel Gas Requirements</heading>  
  <descriptive_body>  
    <paragraph>This Section defines [...]  
      <table>  
        <row>  
          <entry>  
            <paragraph>Permissible range</paragraph>  
          </entry>  
          <entry>  
            <paragraph>  
              <inlinedata>  
                <si-value>  
                  <number>5</number>  
                  <unit>°C</unit>  
                </si-value>  
              </inlinedata>to  
              <inlinedata>  
                <si-value>  
                  <number>120</number>  
                  <unit>°C</unit>  
                </si-value>  
              </inlinedata>  
            </paragraph>
```

# Auslöser und Folgen von Duplikaten

- Dokumentbasierte Migration
- Multiautorenumgebung
  - Nicht-Finden von existierenden Modulen
- Unkontrollierte Wiederverwendung
  - Copy & Paste
- Negative Auswirkungen auf Information Retrieval
  - Hoher Recall, niedrige Precision
- Höhere Übersetzungskosten
- Höhere Time-to-Market, durch doppelte Erfassung

# Ähnlichkeit

- 1 This device is designed to work with a voltage of 220 V only
- 2 The only voltage this device is designed for is 220 V
- 3 This device works with a voltage of 220 V only

# Ähnlichkeit - Diffing

1 This device is designed to work with a voltage of 220 V only

This ~~only voltage this~~ device is designed ~~for~~ to ~~is~~ work with a voltage of 220 V only.

2 The only voltage this device is designed for is 220 V

~~This~~ The only voltage this device ~~works~~ is ~~with~~ designed ~~a~~ for ~~voltage~~ is ~~of~~ 220 V ~~only~~.

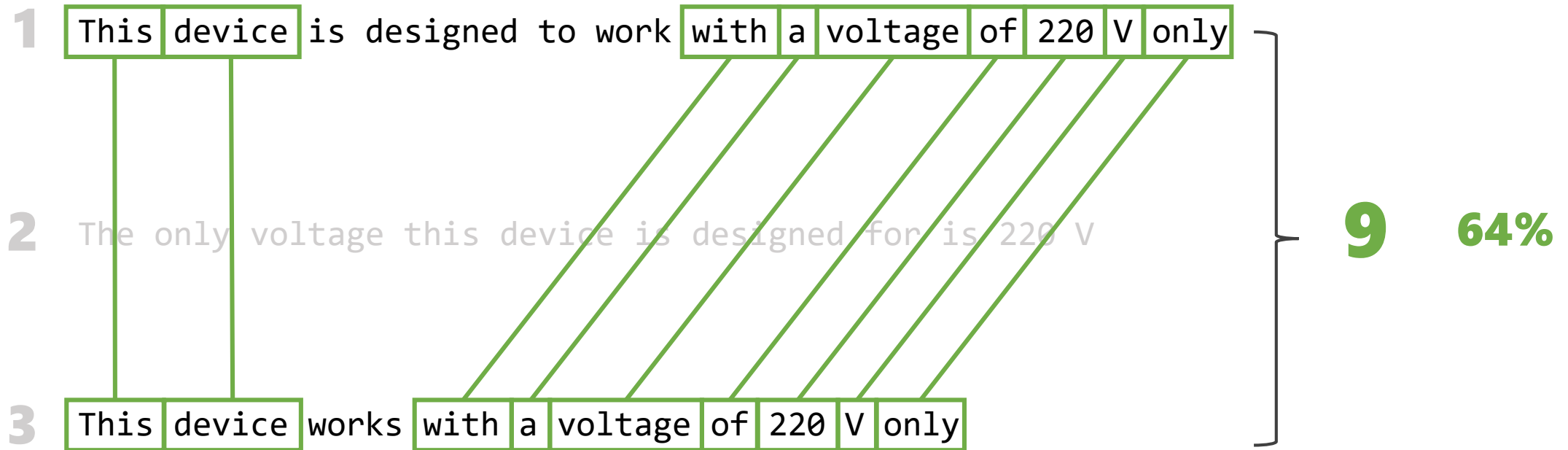
3 This device works with a voltage of 220 V only

# Ähnlichkeit

- 1 This device is designed to work with a voltage of 220 V only
- 2 The only voltage this device is designed for is 220 V
- 3 This device works with a voltage of 220 V only



# Ähnlichkeit - Einfach



# Ähnlichkeit - Einfach

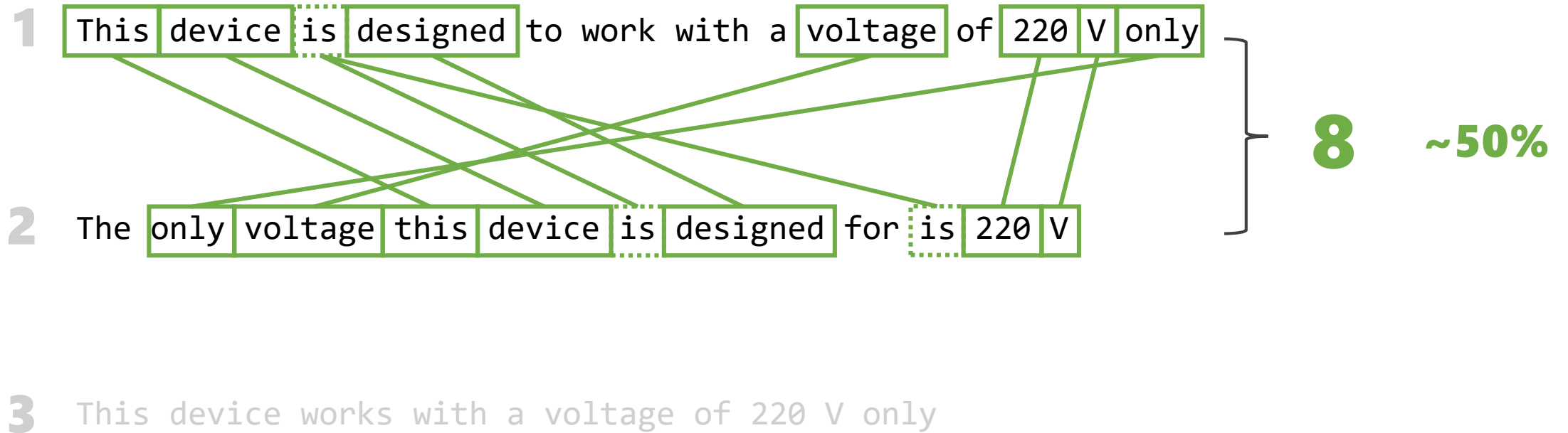
1 This device is designed to work with a voltage of 220 V only

2 The only voltage this device is designed for is 220 V

3 This device works with a voltage of 220 V only

} **6** **40%**

# Ähnlichkeit - Einfach



# Ähnlichkeit - Kosinus

1 This device **is** designed to work with a voltage of 220 V only

2 The only voltage this device **is** designed for **is** 220 V

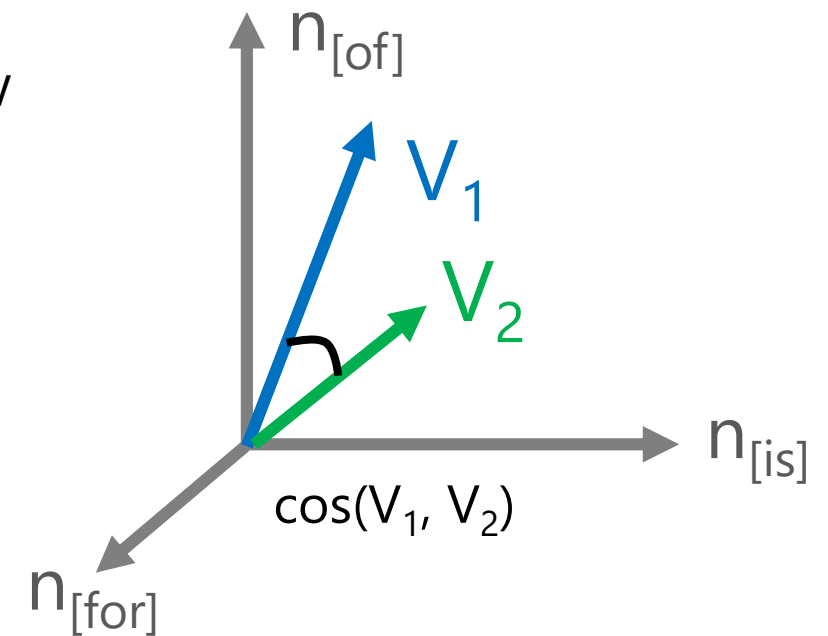
3 This device works with a voltage of 220 V only

# Ähnlichkeit - Kosinus

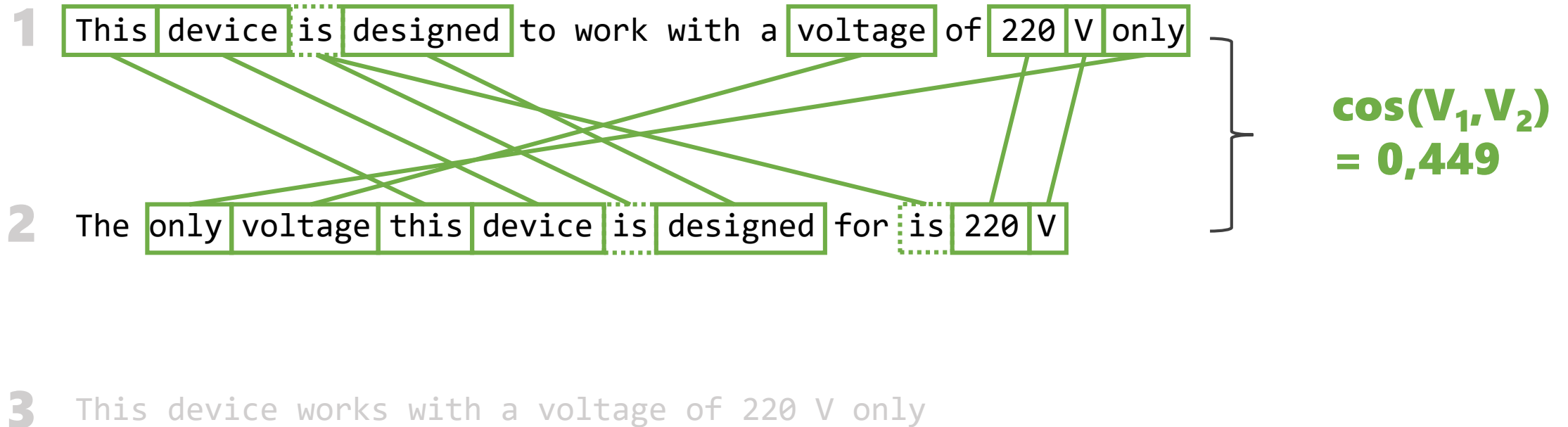
1 This device **is** designed to work with a voltage of 220 V only

2 The only voltage this device **is** designed for **is** 220 V

3 This device works with a voltage of 220 V only



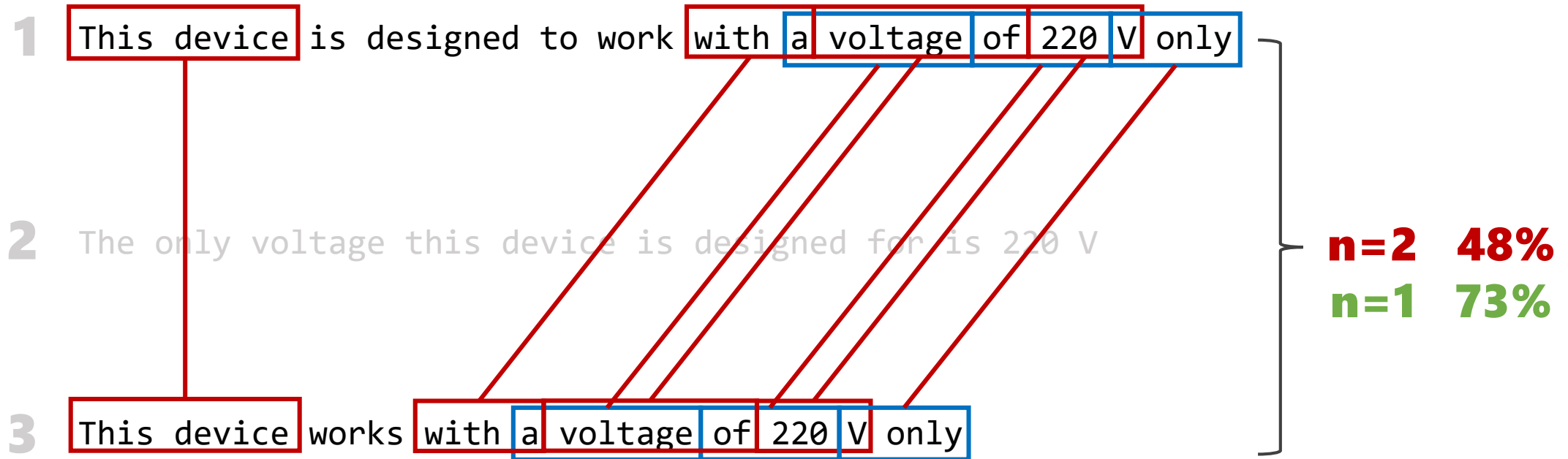
# Ähnlichkeit - Kosinus



# Ähnlichkeit - Wortgruppen

- 1 This device is designed to work with a voltage of 220 V only
- 2 The only voltage this device is designed for is 220 V
- 3 This device works with a voltage of 220 V only
- 
- The diagram illustrates word group similarity between three sentences. Red boxes highlight matching word groups, and red lines connect them across the sentences:
- Sentence 1:** "This device is designed to work with a voltage of 220 V only". Word groups highlighted: "This device", "with a", "voltage of", "220 V".
  - Sentence 2:** "The only voltage this device is designed for is 220 V".
  - Sentence 3:** "This device works with a voltage of 220 V only". Word groups highlighted: "This device", "with a", "voltage of", "220 V".
- Connections shown by red lines:
- A vertical line connects "This device" in Sentence 1 to "This device" in Sentence 3.
  - Three diagonal lines connect the word groups "with a", "voltage of", and "220 V" from Sentence 1 to their corresponding word groups in Sentence 3.

# Ähnlichkeit - Wortgruppen





# Ähnlichkeit - Wortgruppen

- 1 This device is designed to work with a voltage of 220 V only
- 2 The only voltage this device is designed for is 220 V
- 3 This device works with a voltage of 220 V only
- n=2 18%**  
**n=1 67%**
- 
- Detailed description: The diagram illustrates word group similarity between two sentences. Sentence 1 is 'This device is designed to work with a voltage of 220 V only'. Sentence 2 is 'The only voltage this device is designed for is 220 V'. Sentence 3 is 'This device works with a voltage of 220 V only'. Red boxes highlight matching word groups: 'This device is designed' in sentence 1 and 'this device is designed' in sentence 2. Red lines connect these groups. Another red box highlights '220 V' in both sentences, with a red line connecting them. A bracket on the right groups the two sentences, with statistics n=2 (18%) and n=1 (67%). Sentence 3 is faded and not part of the comparison.

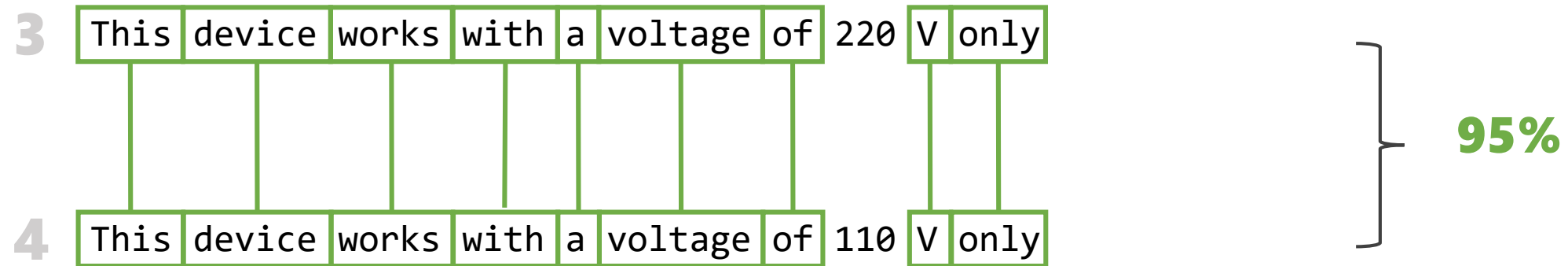
# Semantische Ähnlichkeit

- 1 This device is designed to work with a voltage of 220 V only
- 2 The only voltage this device is designed for is 220 V
- 3 This device works with a voltage of 220 V only
- 4 This device works with a voltage of 110 V only

# Semantische Ähnlichkeit

1 This device is designed to work with a voltage of 220 V only

2 The only voltage this device is designed for is 220 V



# Semantische Ähnlichkeit

1 This device is designed to work with a voltage of 220 V only

2 The only voltage this device is designed for is 220 V

3 This device works with a voltage of 220 V only

4 This device works with a voltage of 110 V only

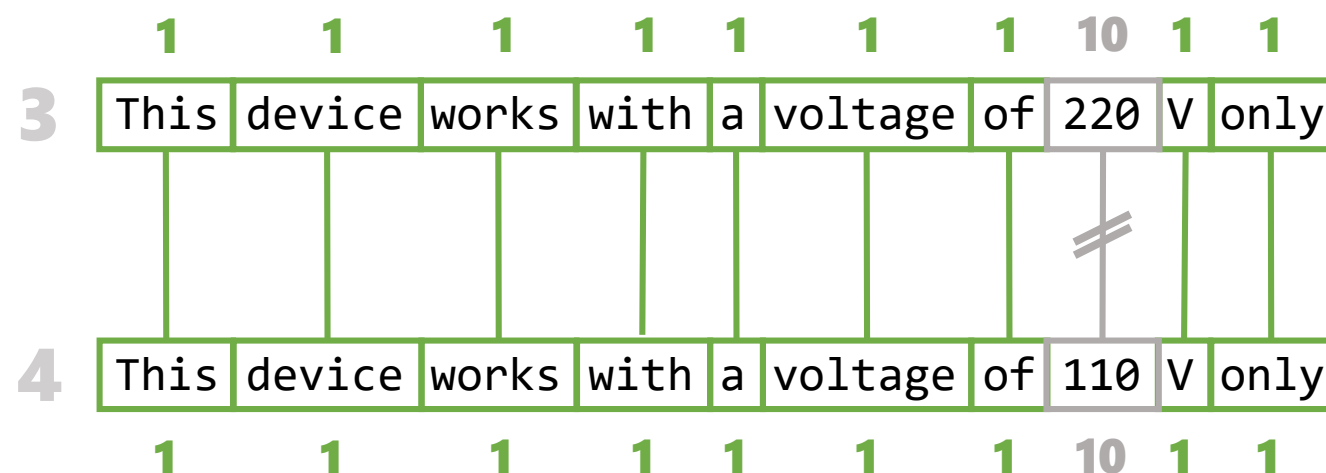
50%

95%

# Semantische Gewichtung

1 This device is designed to work with a voltage of 220 V only

2 The only voltage this device is designed for is 220 V



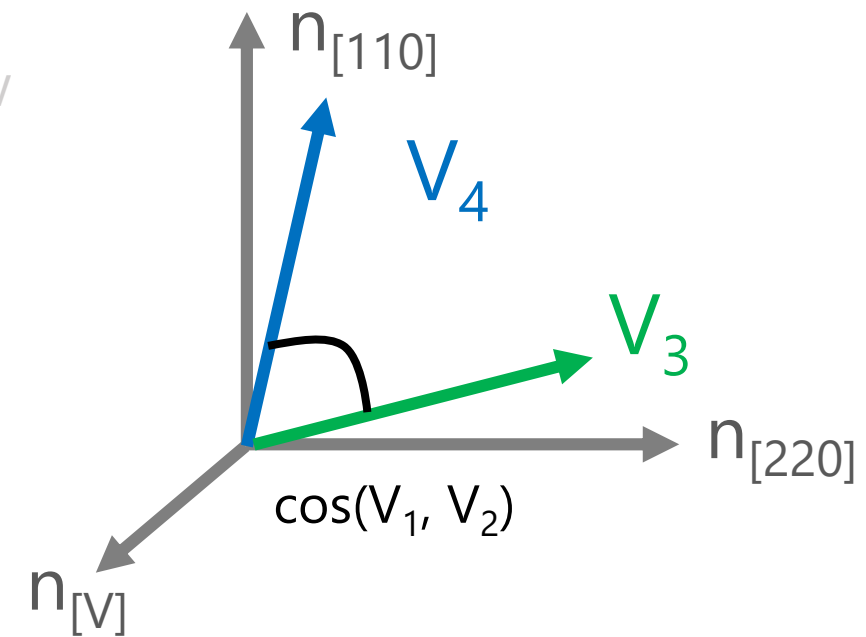
# Semantische Gewichtung

1 This device is designed to work with a voltage of 220 V only

2 The only voltage this device is designed for is 220 V

3 This device works with a voltage of 220 V only

4 This device works with a voltage of 110 V only



# Semantische Gewichtung

1 This device is designed to work with a voltage of 220 V only

2 The only voltage this device is designed for is 220 V

3 This device works with a voltage of 220 V only

4 This device works with a voltage of 110 V only

95%

30%

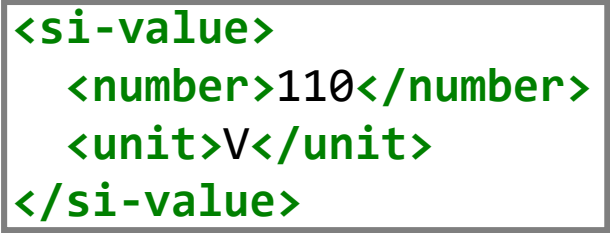
# Semantisches Informationsmodell

```
<paragraph nodeid="Modul-4">  
  This device is designed to work with a voltage of  
    <inlinedata>  
      <si-value>  
        <number>110</number>  
        <unit>V</unit>  
      </si-value>  
    </inlinedata>  
  only.  
</paragraph>
```



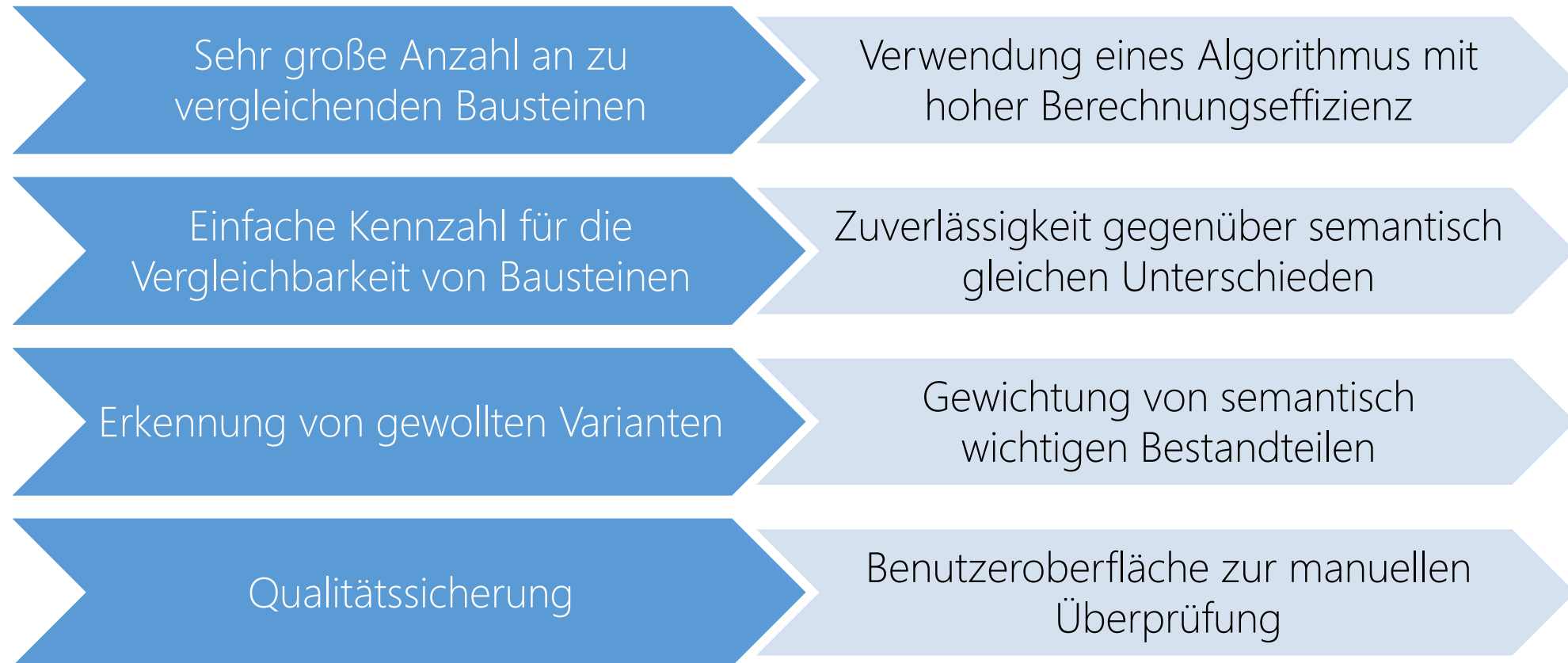
# Semantisches Informationsmodell

```
<paragraph nodeid="Modul-4">  
  This device is designed to work with a voltage of  
  <inlinedata>  
    <si-value>  
      <number>110</number>  
      <unit>V</unit>  
    </si-value>  
  </inlinedata>  
  only.  
</paragraph>
```

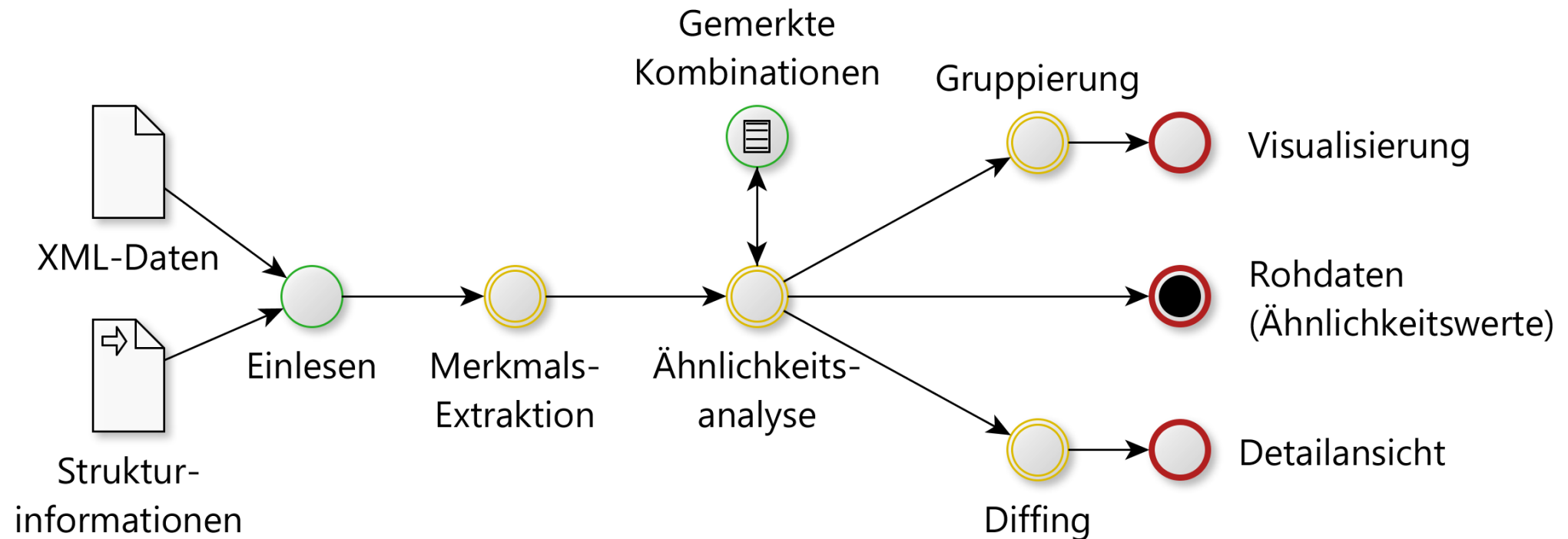


*Gewichtet*

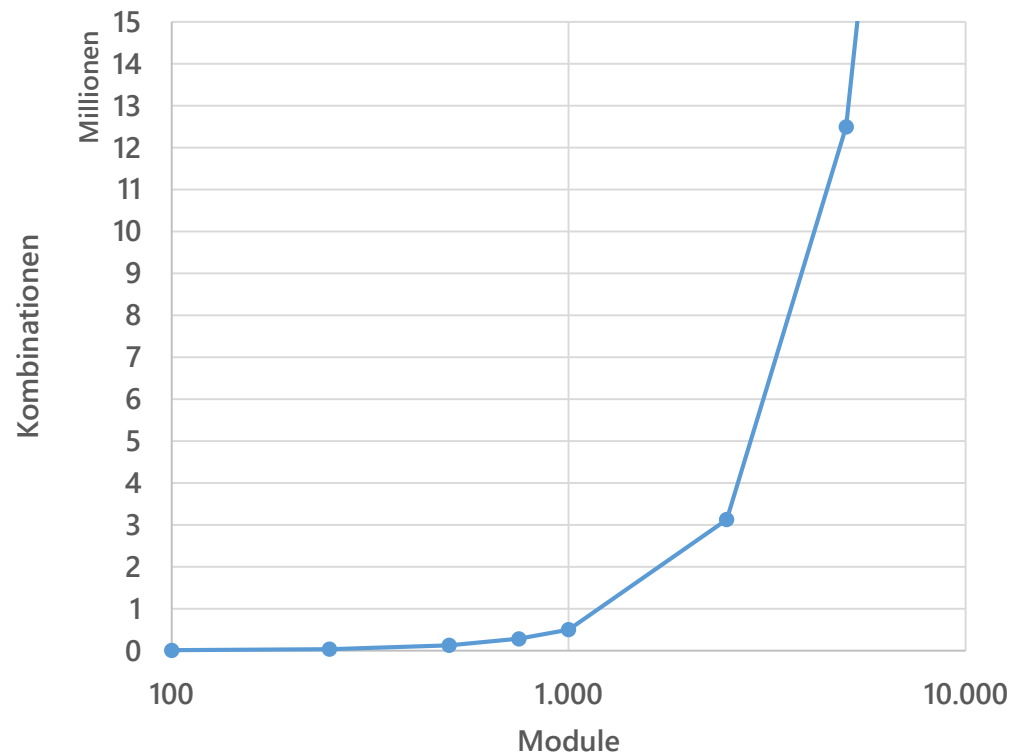
# Projektanforderungen & Lösungsansätze



# Architektur

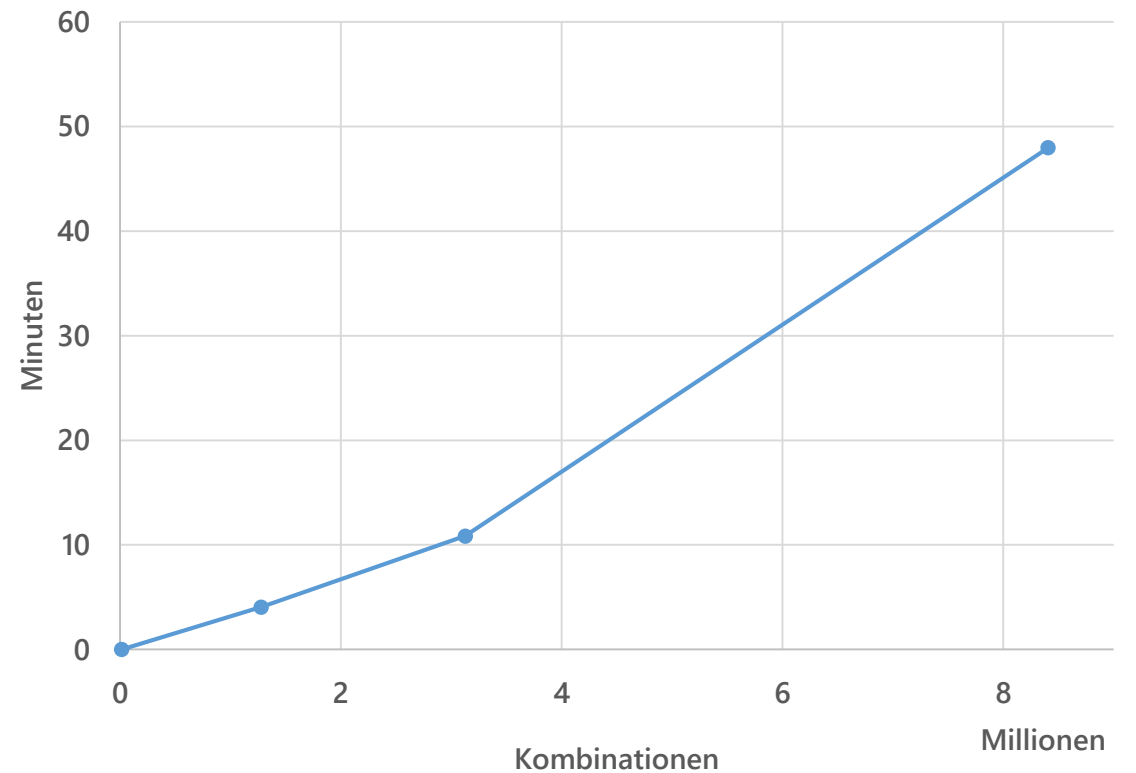
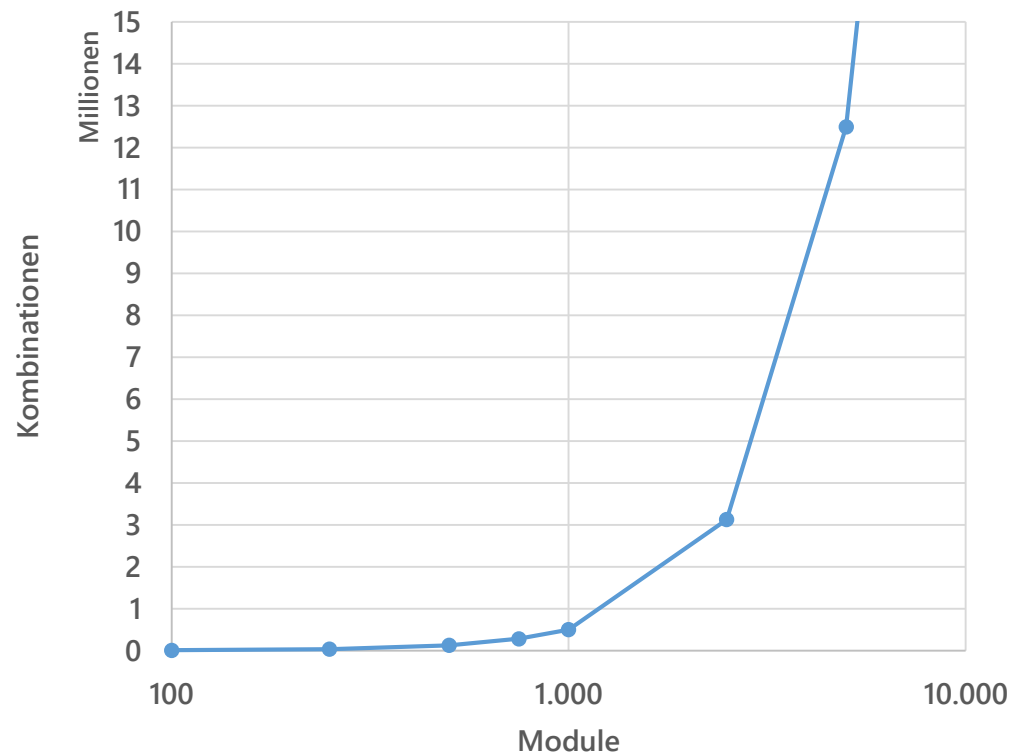


# Skalierung



$$|C| = \frac{n * (n - 1)}{2}$$

# Skalierung



# Technologie & Forschung

- Implementiert in JavaScript
- Berechnungen im Browser (lokal)
- Möglichkeiten für Speichern / Export
- Prototyp verfügbar: <http://semsim.fastclass.de>
- Paper:  
OEVERMANN J. / LÜTH C. (2018): Semantically Weighted Similarity Analysis for XML-based Content Components. Proceedings of the 18th ACM Symposium on Document Engineering. DocEng 2018, Halifax, Canada.

# Live Demo

*fastclass*

## SIEMENS Demo 1

[Ergebnisse sichern ▾](#)[Neue Analyse](#)

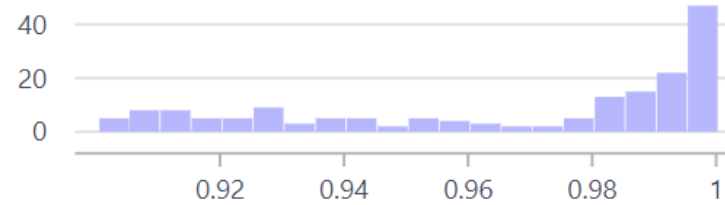
## Duplikate

Ähnlichkeitsbeziehungen

173

bei 13695 Kombinationen

## Ähnlichkeiten



## Zeit

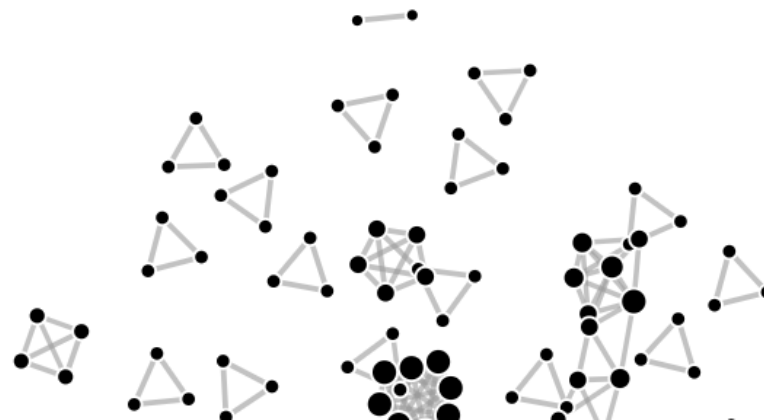
Benötigte Zeit

0.61s

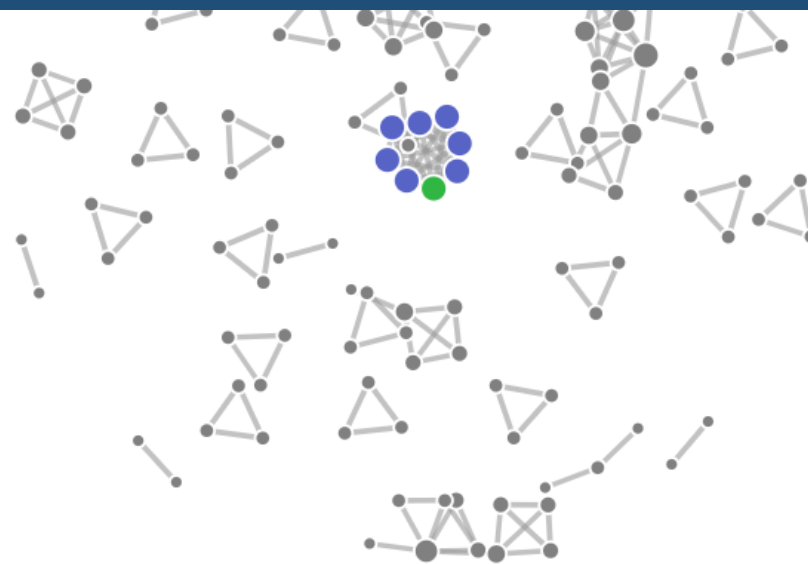
für 166 analysierte Objekte

## Objektähnlichkeiten

Cluster von ähnlichen oder duplikaten Objekten.







## Detailansicht Ähnlichkeitswerte

Tabellarische Übersicht der Ähnlichkeiten

PI-70116789 ✕

#	Modul-ID	Modul-ID	Ähnlichkeit	Aktion
59	PI-70006797	PI-70116789	98.61%	<a href="#">Vergleiche</a>
105	PI-70163367	PI-70116789	98.88%	<a href="#">Vergleiche</a>
138	PI-70148781	PI-70116789	99.13%	<a href="#">Vergleiche</a>

## Detailansicht Ähnlichkeitswerte

Tabellarische Übersicht der Ähnlichkeiten

PI-70116789 ✕

#	Modul-ID	Modul-ID	Ähnlichkeit	Aktion
59	PI-70006797	PI-70116789	98.61%	<a href="#">Vergleiche</a>
105	PI-70163367	PI-70116789	98.88%	<a href="#">Vergleiche</a>
138	PI-70148781	PI-70116789	99.13%	<a href="#">Vergleiche</a>
142	PI-70127224	PI-70116789	100%	<a href="#">Vergleiche</a>
151	PI-70119052	PI-70116789	91.8%	<a href="#">Vergleiche</a>
161	PI-70119042	PI-70116789	99.02%	<a href="#">Vergleiche</a>
169	PI-70116789	PI-70120142	99.02%	<a href="#">Vergleiche</a>

## Vergleichsansicht

Module: **PI-70119047** ↔ **PI-70120137**

Titel: **Fuel Gas Constituents and Properties**

Ähnlichkeit: 93.8038%

Unterschiede: +14 -24

Gewichtet: si-value

reactions occurring in the premix piping which could completely destroy the burners. **5** In the case of fuels with **increased** higher hydrocarbon (**C n H m | CnHm**, n = 2) **contents | content greater than 10vol.%** and/or **a** hydrogen content **greater than 10vol.%,** there is a risk of **increased | greater** combustion instability even in diffusion mode; this is due to formation of condensation and/or higher flame velocity. **This requires consultation with Siemens Power Generation, Gas Turbine Order Implementation. Ethane (C 2 H 6 ) contents of up to a maximum of 15vol.% are permissible and do not require special approval. 6** Sulfur and hydrogen sulfide Elemental sulfur is not permissible. Sulfur content as referred to in this document defines the sum of sulfur derived from the sulfur-containing compounds occurring in the fuel gas (mercaptans (R-SH), hydrogen sulfide (H 2 S), odorizing agents (e.g., THT), carbonyl sulfide (COS), and others). With a total sulfur content 20mg/kg, provision for preheating of the fuel gas to at least 60°C (with a maximum operational tolerance of ±10K) must be made. At higher sulfur contents, additional operating restrictions and adaptations to the fuel gas system

# Anpassungen im Projektverlauf

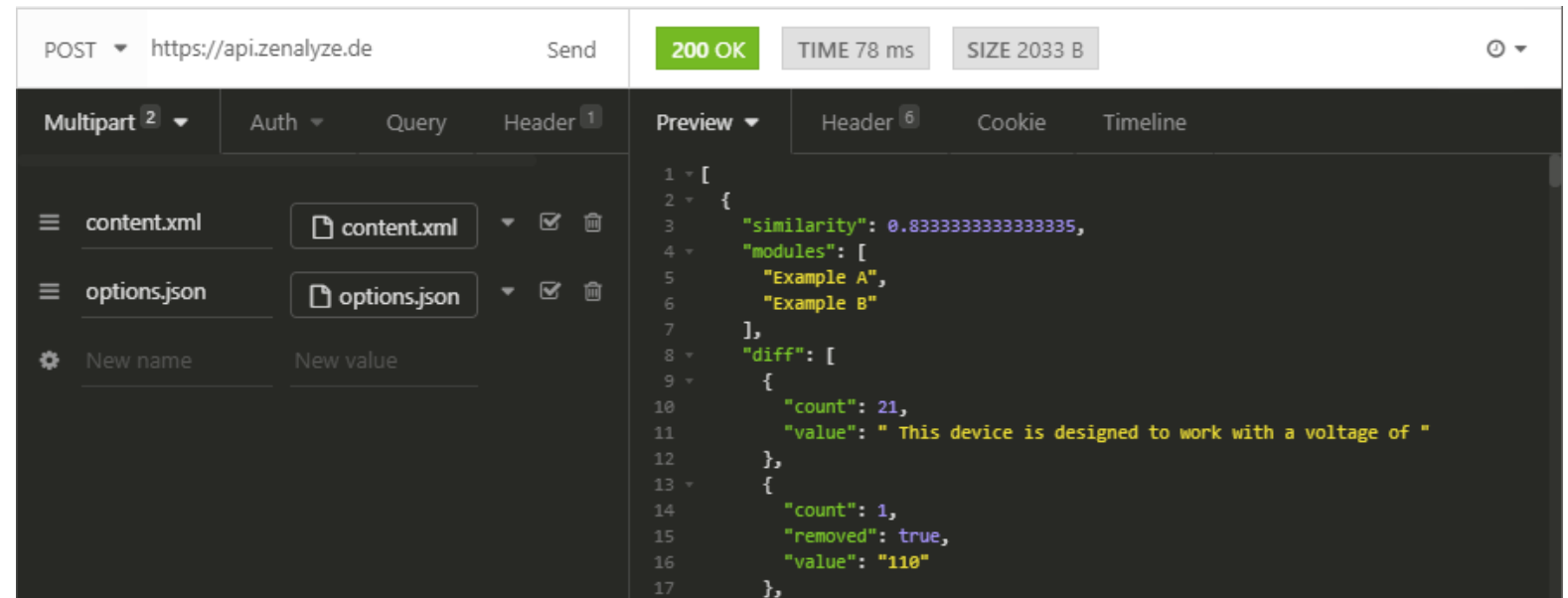
- Filter zum Import von XML-Daten
- Integration eines Textvergleichs zur schnellen Überprüfung
- Dynamische Filterung und Markierung der gewählten Bausteine
- Vorschau der Überschrift in der Übersichtstabelle
- Selektive Zusammenstellung der Quelltexte für den Vergleich auf Basis der inhaltlichen Struktur

# Ausblick

- Optimierungspotenziale
  - Erweiterung zum Abgleich neuer Texte mit bestehenden Bausteinen
  - Integration von Metadaten / Klassifizierungen zur Unterstützung der Entscheidungsfindung
- Anwendungsszenarien
  - Überarbeitung von bestehender Dokumentation zur Reduzierung der Textbausteine
  - Versionsvergleich von Dokumentationen
  - Prognose Übersetzungskosten

# Ausblick

- Weiterentwicklung
  - Alternative Dateiformate (iiRDS, DITA, etc.)
  - Direkte Systemintegration
  - API als Webservice



Vielen Dank für Ihre Aufmerksamkeit



**Jan Oevermann**

jan.oevermann@icms.de



**Dr. Timo Fleschutz-Balarezo**

timo.fleschutz@siemens.com

**Besuchen Sie uns an Stand G07 in Halle C2**

Ihre Meinung ist uns wichtig! Sagen Sie uns bitte, wie Ihnen der Vortrag gefallen hat. Wir freuen uns auf Ihr Feedback per Smartphone oder Tablet unter:

<http://in11.honestly.de>

oder scannen Sie den QR-Code.

Das Bewertungstool steht Ihnen auch noch nach der Tagung zur Verfügung!

