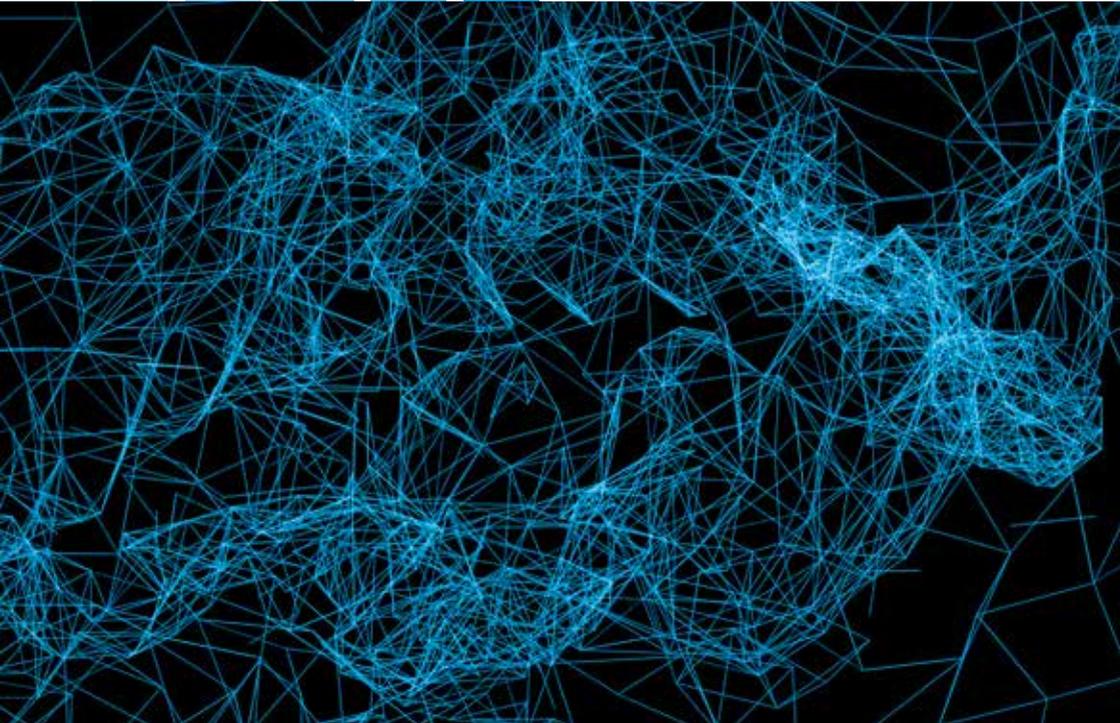


# Modellierung thematischer Nähe in Organisationen durch Machine Learning

Thomas Thiele



# **Modellierung thematischer Nähe in Organisationen durch Machine Learning**

Modeling Thematic Proximity  
in Organizations using Machine Learning

Von der Fakultät für Maschinenwesen der  
Rheinisch-Westfälischen Technischen Hochschule Aachen  
zur Erlangung des akademischen Grades  
eines Doktors der Ingenieurwissenschaften genehmigte Dissertation

vorgelegt von

Thomas David Thiele

Berichter: Außerplanmäßige Professorin Dr. rer. nat. Sabina Jeschke  
Universitätsprofessor Dr.-Ing. Dr. h. c. (UPT) Burkhard Corves

Tag der mündlichen Prüfung: 16. November 2018



Berichte aus der Informatik

**Thomas Thiele**

**Modellierung thematischer Nähe  
in Organisationen durch Machine Learning**

Shaker Verlag  
Aachen 2019

**Bibliografische Information der Deutschen Nationalbibliothek**

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über <http://dnb.d-nb.de> abrufbar.

Zugl.: D 82 (Diss. RWTH Aachen University, 2018)

Copyright Shaker Verlag 2019

Alle Rechte, auch das des auszugsweisen Nachdruckes, der auszugsweisen oder vollständigen Wiedergabe, der Speicherung in Datenverarbeitungsanlagen und der Übersetzung, vorbehalten.

Printed in Germany.

ISBN 978-3-8440-6542-8

ISSN 0945-0807

Shaker Verlag GmbH • Postfach 101818 • 52018 Aachen

Telefon: 02407 / 95 96 - 0 • Telefax: 02407 / 95 96 - 9

Internet: [www.shaker.de](http://www.shaker.de) • E-Mail: [info@shaker.de](mailto:info@shaker.de)

## Danksagung

Die vorliegende Dissertation entstand im Rahmen meiner Tätigkeit als wissenschaftlicher Mitarbeiter am Cybernetics Lab der RWTH Aachen University. Besonderer Dank gilt daher Prof. Dr. rer. nat. Sabina Jeschke für die Betreuung der Arbeit und die darüberhinausgehende vortreffliche Zusammenarbeit. Für die Übernahme des Kommissionsvorsitzes danke ich Prof. Dr.-Ing. Christian Brecher sowie Prof. Dr.-Ing. Burkhard Corves in seiner Rolle als zweitem Gutachter. Prof. Dr.-Ing. Tobias Meisen stand mir sowohl fachlich als auch persönlich stets zur Seite, auch ihm gebührt mein herzlichster Dank.

Als wesentlicher Beitrag für das Gelingen der Arbeit empfinde ich die besondere Atmosphäre des Instituts, die maßgeblich von meinen Kolleginnen und Kollegen gestaltet wird. Sie alle haben sich in den letzten Jahren als hervorragende Sparingspartner erwiesen. Die Forschungsgruppe Produktionstechnik hat sich hierbei als ebenso fachliches wie kameradschaftliches Umfeld erwiesen, den Kollegen der Gruppe gilt meine besondere Hochachtung. Ebenso haben der Austausch und die Diskussionen mit den Kollegen des Exzellenzclusters „Integrative Produktionstechnik für Hochlohnländer“ wesentliche Impulse zur Entwicklung der Arbeit gegeben. Mein Dank gilt hierbei im Besonderen Dr. phil. André Calero Valdez. Ebenso Dank gebührt Olivier Pfeiffer und Fabian Scheidt für ihre Unterstützung in der letzten Phase.

Neben dem universitären und beruflichen Umfeld ist der private Rahmen ein ebenso zentraler Bestandteil für das Gelingen der Arbeit. Meine Eltern Renate Jeromin-Thiele und Harald Thiele nehmen hierbei eine wesentliche Rolle ein, sie haben mich von den ersten Ideen bis zur Fertigstellung der Arbeit stets in ganz besonderer Weise ermutigt, unterstützt und gefördert. Ebenso hervorheben möchte ich meine Partnerin Sarah Ginski, die mir besonders im finalen Endspurt trotz gleicher Situation mit viel Energie, Gelassenheit und Unterstützung in allen Lebenslagen beigestanden hat.



## Inhaltsverzeichnis

<b>I ZUSAMMENFASSUNG .....</b>	<b>9</b>
<b>I SUMMARY .....</b>	<b>9</b>
<b>II ABKÜRZUNGSVERZEICHNIS.....</b>	<b>10</b>
<b>III ABBILDUNGSVERZEICHNIS.....</b>	<b>11</b>
<b>IV TABELLENVERZEICHNIS.....</b>	<b>14</b>
<b>V VERWENDETE FORMELZEICHEN.....</b>	<b>15</b>
<b>1 EINFÜHRUNG .....</b>	<b>17</b>
1.1 PROBLEMSTELLUNG UND ZIELE DER ARBEIT.....	17
1.2 AUFBAU DER ARBEIT .....	20
<b>2 DEFINITION ZENTRALER BEGRIFFE.....</b>	<b>23</b>
2.1 ÜBERBLICK.....	23
2.2 THEMATISCHE NÄHE .....	23
2.3 KOOPERATIONEN UND KOOPERATIONSMANAGEMENT .....	25
2.4 ERFASSUNG THEMATISCHER NÄHE IN KOOPERATIONEN DURCH BIBLIOMETRIE .....	28
2.5 ZWISCHENFAZIT .....	33
<b>3 STAND DER FORSCHUNG ZUR VERARBEITUNG TEXTUELLER DATEN.....</b>	<b>35</b>
3.1 ÜBERBLICK.....	35
3.2 INFORMATIONSMODELLIERUNG IM NLP UND TEXT MINING.....	36
3.3 FEATURE ENGINEERING IM KONTEXT DER INFORMATIONSEXTRAKTION AUS TEXTEN .....	41
3.3.1 <i>Mapping von Wörtern und Wortmengen .....</i>	<i>43</i>
3.3.2 <i>Erkennung der Wortbedeutung auf Basis der syntaktischen Umgebung.....</i>	<i>46</i>
3.3.3 <i>Erfassung inhärenter Themen in textuellen Daten .....</i>	<i>48</i>
3.4 VERKNÜPFUNG VON INFORMATIONEN AUS TEXTUELLEN DATEN .....	51
3.4.1 <i>Unsupervised Learning.....</i>	<i>52</i>

3.4.2	<i>Supervised Learning</i> .....	54
3.5	ZWISCHENFAZIT ZUM STAND DER FORSCHUNG .....	59
<b>4</b>	<b>ANFORDERUNGEN AN EIN SYSTEM ZUR ERFASSUNG THEMATISCHER NÄHE</b> .....	<b>61</b>
4.1	ÜBERBLICK .....	61
4.2	DEFINITION DER EINSATZSZENARIEN .....	61
4.3	SPEZIFIKATION DER FUNKTIONALEN ANFORDERUNGEN .....	65
4.3.1	<i>Technologische Anforderungen der Verfahrenselemente</i> .....	66
4.3.2	<i>Technologische Anforderungen des Gesamtsystems</i> .....	71
4.4	SPEZIFIKATION DER NICHT-FUNKTIONALEN ANFORDERUNGEN .....	73
4.5	ZUSAMMENFASSUNG DER ANFORDERUNGEN .....	75
<b>5</b>	<b>SYSTEMENTWURF ZUR ERFASSUNG THEMATISCHER NÄHE</b> .....	<b>77</b>
5.1	ÜBERBLICK .....	77
5.2	KOMPONENTENSPEZIFIKATION DES DATENZENTRIERTEN ANSATZES .....	78
5.3	TRANSFORMATION: ROHDATEN ZU FEATURES .....	80
5.3.1	<i>Datenmodellierung und Pre-Processing</i> .....	80
5.3.2	<i>Topic Model-basiertes Feature Engineering</i> .....	87
5.4	MODELLANSATZ: THEMATISCHE NÄHE DURCH MACHINE LEARNING .....	91
5.5	KLASSIFIKATION UND FEATURE SELEKTION .....	92
5.5.1	<i>Formulierung thematischer Nähe als Klassifikation</i> .....	93
5.5.2	<i>Algorithmenauswahl zur Klassifikation thematischer Nähe</i> .....	99
5.5.3	<i>Feature Transformation für KNN</i> .....	106
5.6	VISUALISIERUNG TEXTUELLER DATEN UND INFORMATIONEN .....	109
5.6.1	<i>Mögliche Visualisierungsformen und Designelemente</i> .....	109
5.6.2	<i>Konzeptionierung der Visualisierung zu thematischer Nähe</i> .....	114
5.7	IMPLEMENTIERUNG DES SYSTEMS .....	119

---

5.8	ZWISCHENFAZIT SYSTEMENTWURF .....	121
<b>6</b>	<b>ANWENDUNGSFALL: ÜBERGREIFENDE VERNETZUNG EINES EXZELLENZCLUSTERS ..</b>	<b>123</b>
6.1	ÜBERBLICK .....	123
6.2	ORGANISATIONSRAHMEN UND THEMATISCHE NÄHE IM EXZELLENZCLUSTER .....	123
6.3	BESCHREIBUNG DES DATENKORPUS UND DER DATENSTRUKTUR .....	126
6.4	APPLIKATION DES SYSTEMENTWURFS .....	128
6.4.1	<i>Pre-Processing der Daten</i> .....	129
6.4.2	<i>Applikation und Evaluation der LDA</i> .....	132
6.4.3	<i>Training und Evaluation des Matchmakers</i> .....	137
6.4.4	<i>Eigenschaften der Klassifikationsmodelle</i> .....	143
6.5	EXEMPLARISCHE ERGEBNISVISUALISIERUNG.....	147
6.6	ZWISCHENFAZIT .....	150
<b>7</b>	<b>ÜBERTRAGBARKEIT UND AUSBLICK.....</b>	<b>153</b>
7.1	ÜBERBLICK.....	153
7.2	KRITISCHE WÜRDIGUNG.....	153
7.2.1	<i>Validierung mittels qualitativer Testdaten</i> .....	153
7.2.2	<i>Systementwurf im Vergleich mit bibliometrischen Verfahren</i> .....	157
7.3	DISKUSSION DER METHODISCHEN LIMITIERUNGEN .....	160
7.4	AUSBLICK .....	162
7.4.1	<i>Methodische Weiterentwicklungspotentiale</i> .....	162
7.4.2	<i>Übertragbarkeit auf weitere Anwendungsszenarien</i> .....	166
<b>8</b>	<b>RESÜMEE .....</b>	<b>171</b>

<b>9</b>	<b>ANHANG .....</b>	<b>173</b>
9.1	BESTIMMUNG DER ANZAHL THEMEN IN PROJEKTBEZOGENEN TOPIC MODELLEN .....	173
9.2	ERGEBNISSE DES TOPICS MODELLINGS .....	179
<b>10</b>	<b>LITERATURVERZEICHNIS .....</b>	<b>191</b>

## I Zusammenfassung

Komplexe Problemstellungen nicht nur als disziplinäre Fragestellung zu begreifen, sondern über Fachgrenzen hinaus Lösungen zu entwickeln, erweist sich nicht nur als Trend, sondern auch als Notwendigkeit. Die Identifikation geeigneter Kooperationspartner und die Suche nach gemeinsamen Themen ist jedoch oftmals zeitaufwändig.

In der vorliegenden Arbeit wird daher ein System konzeptioniert und entwickelt, welches die Modellierung thematischer Nähe in Organisationen durch einen Machine Learning Ansatz erlaubt. Grundlage hierfür sind textuelle Daten, aus welchen zunächst mittels eines generativen Verfahrens inhärente Themen extrahiert werden. Danach werden diese Themen einem diskriminierenden Verfahren unterzogen, welches ein Matchmaking zwischen Themen unterschiedlicher organisationaler Entitäten ableitet. Die durch diese Verfahrenskette generierten Ergebnisse werden dann für den Nutzer in Form eines graphenbasierten Ansatzes visualisiert, es entsteht eine Landkarte verknüpfter Themen auf Basis eines automatisierten Prozesses.

## I Summary

To understand complex problems not only as a disciplinary question, but to develop solutions beyond the boundaries of disciplines, proves to be not only a trend, but also a necessity. The identification of suitable cooperation partners and the search for common topics is often time-consuming.

In the present work, a system is therefore conceived and developed which allows the modelling of thematic proximity in organisations through a Machine Learning approach. This is based on textual data, from which inherent topics are extracted using a generative procedure. Thereafter, these topics are subjected to a discriminatory procedure which derives a matchmaking between topics of different organisational entities. The results generated by this process chain are then visualized for the user in the form of a graph-based approach, resulting in a map of linked topics on the basis of an automated process.

## II Abkürzungsverzeichnis

ACM: Association for Computing Machinery

ASCII: American Standard Code for Information Interchange

BOW: Bag-of-words Modell

CBOW: Continous Bag of Words

CMS: Content Management Systeme

CoE: Exzellenzcluster

CRISP-DM: Cross Industry Standard Process for Data Mining

CSP: Cross-Sectional Processes

csv: Comma separated value

DFG: Deutsche Forschungsgemeinschaft

ECM: Entprise Content Management Systeme

h-Index: Hirsch-Index

HTTP: Hypertext Transfer Protocoll

ICD: Integrative Cluster Domain

IEEE: Institute of Electrical and Electronics Engineers

JSON: JavaScript Object Notation

kNN: k-nearest Neighbor

KNN: Künstliche Neuronale Netze

LDA: Latent Dirichlet Allocation

LSI: Latent Semantic Indexing

NLP: Natural Language Processing

OCR: Optical Character Recognition

pdf: Portable Document Format

POS-Tagging: Part-of-Speech-Tagging

relu: Rectified Linear Units

SVM: Support Vector Machine

tanh: Tangens Hyperbolicus

TDM: Term-Document Matrix

tf-idf: Term Frequency – Inverse Document Frequency

txt: Text Dateien

UCS: Universal Coded Character Set

XML: Extensible Markup Language

### III Abbildungsverzeichnis

Abbildung 1: Forschungsfragen der Arbeit und funktionale Ebenen.....	19
Abbildung 2: Struktur und Aufbau der Arbeit .....	21
Abbildung 3: Strömungen in der bibliometrischen Analyse nach Bellis (2014) sowie Nacke (1979), Nalimov und Mulchenko (1971) und Pritchard (1969).....	30
Abbildung 4: Überblick zu Kapitel 2 und 3 .....	35
Abbildung 5: Zusammenhang Zeichen, Daten, Informationen und Wissen nach Krcmar (2015) .....	37
Abbildung 6: Konvertierungsprozess unstrukturierter Textdaten .....	40
Abbildung 7: Darstellung des nested Chinese Restaurant Process nach Blei et al. (2010).....	52
Abbildung 8: Ergebnis eines hierarchischen Topic Modellings (Blei et al. 2010).....	53
Abbildung 9: Feed-Forward KNN für Textklassifikation .....	58
Abbildung 10: Use Case Diagramm zu Einsatzszenario I .....	64
Abbildung 11: Use Case Diagramm zu Einsatzszenario II .....	65
Abbildung 12: Zusammenfassung der Prozesse in der Datenakquisition und - aufbereitung .....	67
Abbildung 13: Zusammenfassung der Prozesse in der Analyse.....	69
Abbildung 14: Komponentendiagramm des Grobentwurfs .....	79
Abbildung 15: Datenmodell des Pre-Processings.....	81
Abbildung 16: Aktivitätsdiagramm der Komponenten PreProcessor .....	86
Abbildung 17: Schema der Klassifikation von Themen (Rauten).....	95
Abbildung 18: Beispielhafter Gesamtprozess der Klassifikation .....	99
Abbildung 19: Klassifikationsgenauigkeit über Datensatzumfang eines Naive Bayes Klassifikators im Vergleich mit einem Decision Tree C4.5 nach Domingos und Pazzani (1997).....	102

Abbildung 20: Schematische Darstellung zum Einfluss von überlappenden Daten im Trainingsset nach Aggarwal (2018).....	103
Abbildung 21: Schematische Darstellung der Schemata zu Word2Vec Grundarchitekturen nach Mikolov et al. (2013).....	108
Abbildung 22: Eigene Darstellung des Topic Model Visualisiers nach Chaney und Blei (2012).....	110
Abbildung 23: Matrixdarstellung von Thema-Begriff Beziehungen nach Chuang et al. (2012a).....	111
Abbildung 24: Graph zu Themen in Forschungsanträgen nach Gretarsson et al. (2012).....	112
Abbildung 25: Eigene Schemadarstellung des verwendeten Graphen.....	117
Abbildung 26: Schemadarstellung des Graphen mit Highlighting .....	118
Abbildung 27: Korpusstatistik zur Anzahl der Begriffe je Projektkorpus .....	128
Abbildung 28: Evaluation des word2vec Modells (Trainingsiterationen).....	131
Abbildung 29: Bestimmung der Themenanzahl am Beispiel des Projekts CSP1.....	135
Abbildung 30: Ergebnisse der Grid-Search zu unterschiedlichen Aktivierungsfunktionen .....	138
Abbildung 31: Ergebnisse der Grid-Search zur unterschiedlichen Anzahl von Hidden Layern .....	139
Abbildung 32: Ergebnisse der Grid-Search zu unterschiedlichen Neuronenkonfigurationen .....	140
Abbildung 33: Einfluss der Top-Worte auf Basis der a-posteriori Wahrscheinlichkeiten der LDA.....	141
Abbildung 34: Trainings-Accuracy in Abhängigkeit zum Projekt.....	143
Abbildung 35: (1) Heatmap der Prädiktionsvektoren als max-Funktion (vgl. Abschnitt 5.5).....	144
Abbildung 36: (2) Heatmap der Prädiktionsvektoren als Summenfunktion (vgl. Abschnitt 5.5) .....	144

---

Abbildung 37: Beispiel zu nicht-kommutativen Eigenschaften der Prädiktionen.....	145
Abbildung 38: Detailvergleich Fall (1) und (2) zur Berechnung der finalen Prädiktion .....	147
Abbildung 39: Exemplarischer Graph zu CSP1.....	148
Abbildung 40: Detaildarstellung von Matches zu CSP1.....	149
Abbildung 41: Prädizierte Matches auf Basis beidseitiger Nennungen in Tabelle 6	155
Abbildung 42: Prädizierte Matches auf Basis einseitiger Nennungen in Tabelle 6	156
Abbildung 43: Graphen-Darstellung der Autorenanalyse im CoE unter Verwendung von TIGRS (nach Dietze et al. 2016) .....	158
Abbildung 44: Ordnungsrahmen der Übertragbarkeit entlang des Bezugsrahmens und der Daten .....	167
Abbildung 45: Bestimmung der Themenanzahl am Beispiel des Projekts A2.....	173
Abbildung 46: Bestimmung der Themenanzahl am Beispiel des Projekts A3.....	173
Abbildung 47: Bestimmung der Themenanzahl am Beispiel des Projekts B1.....	174
Abbildung 48: Bestimmung der Themenanzahl am Beispiel des Projekts B2.....	174
Abbildung 49: Bestimmung der Themenanzahl am Beispiel des Projekts C2 .....	175
Abbildung 50: Bestimmung der Themenanzahl am Beispiel des Projekts C3 .....	175
Abbildung 51: Bestimmung der Themenanzahl am Beispiel des Projekts D1 .....	176
Abbildung 52: Bestimmung der Themenanzahl am Beispiel des Projekts D2 .....	176
Abbildung 53: Bestimmung der Themenanzahl am Beispiel des Projekts D3 .....	177
Abbildung 54: Bestimmung der Themenanzahl am Beispiel des Projekts CSP1....	177
Abbildung 55: Bestimmung der Themenanzahl am Beispiel des Projekts CSP2....	178
Abbildung 56: Bestimmung der Themenanzahl am Beispiel des Projekts CSP3....	178

#### IV Tabellenverzeichnis

Tabelle 1 Übersicht zu Varianten der Gewichtungskomponenten des tf-idf nach Hu und Liu (2012), Feinerer (2017) und Manning et al. (2008b).....	47
Tabelle 2: Übersicht der verwendeten Bibliotheken.....	120
Tabelle 3: Evaluation des <i>word2vec</i> Skip-gram-Modells (Accuracy) .....	130
Tabelle 4: Referenz-Accuracy auf Basis des Google analogy test sets nach Mikolov et al. (2013).....	132
Tabelle 5: Übersicht zu den im Projekt CSP1 ermittelten Themen.....	136
Tabelle 6: Übersicht Nennungen "Welche Teilprojekte des Exzellenzclusters sollten vermehrt miteinander kooperieren?" .....	155
Tabelle 7: Übersicht zu den im Projekt A2 ermittelten Themen.....	179
Tabelle 8: Übersicht zu den im Projekt A3 ermittelten Themen.....	180
Tabelle 9: Übersicht zu den im Projekt B1 ermittelten Themen.....	181
Tabelle 10: Übersicht zu den im Projekt B2 ermittelten Themen.....	182
Tabelle 11: Übersicht zu den im Projekt C2 ermittelten Themen.....	183
Tabelle 12: Übersicht zu den im Projekt C3 ermittelten Themen.....	184
Tabelle 13: Übersicht zu den im Projekt D1 ermittelten Themen.....	185
Tabelle 14: Übersicht zu den im Projekt D2 ermittelten Themen.....	186
Tabelle 15: Übersicht zu den im Projekt D3 ermittelten Themen.....	187
Tabelle 16: Übersicht zu den im Projekt CSP1 ermittelten Themen.....	188
Tabelle 17: Übersicht zu den im Projekt CSP2 ermittelten Themen.....	189
Tabelle 18: Übersicht zu den im Projekt CSP3 ermittelten Themen.....	190

## V Verwendete Formelzeichen

$1_A(x)$	Funktion	Indikatorfunktion nach Chaney und Blei (2012)
$a_j$	Maßzahl	Aktivierung eines KNN
$\alpha$	Hyperparameter	Initialer Hyperparameter der LDA
$ave$	Funktion	Durschnittswert (z.B. Mittelwert oder Median)
$\beta_{iv}$	Verteilung	Topicverteilung des Themas $i$ aus Topic Modell $v$
$\gamma$	Maßzahl	Gewichtungsfaktor für byte size basierte Normalisierung
$CharLength^Y$	Funktion	Funktion zur Ermittlung der Zeichenlänge eines Wortes oder Dokuments mit Gewichtungsfaktor $\gamma$
$d_i$	Mengenelement	Dokument in Menge $D$
$df_t$	Häufigkeit	Dokumenthäufigkeit aller Dokumente mit Begriff $t$
$e_k$	Menge	Menge der Begriffe einer Entität $E_k$
$\xi_{i,j}$	Maßzahl	Ähnlichkeit zwischen zwei Themen nach Chaney und Blei (2012)
$f(\cdot)$	Funktion	Allgemeine Klassifikationsfunktion
$GesamtPred_{E_j m}$	Maßzahl	Ergebnismatrix einer Prädiktion des $m$ -ten Themas der Entität $E_j$
$h$	Maßzahl	Hirsch Index
$h(a_j)$	Funktion	Aktivierungsfunktion von $a_j$ eines KNN
$idf$	Bezeichnung	Inverse Document Frequency
$k$	Hyperparameter	Anzahl der Themen einer LDA
$\mathcal{L}$	Menge	Menge der Label

$N$	Menge	Anzahl der Dokumente in der Dokumentensammlung
$n_{t_i}$	Häufigkeit	Auftrittshäufigkeit des Wortes $t_i$ im Korpus
$n_{t_i,d}$	Häufigkeit	Auftrittshäufigkeit des Wortes $t_i$ im Dokument $d$
$\mathbb{N}$	Menge	Menge der natürlichen Zahlen
$P_{(t_i)}$	Wahrscheinlichkeit	Auftrittswahrscheinlichkeit eines Begriffes $t_i$
$Pred_{E_j k}$	Maßzahl	Ergebnis der Klassifikation des $k$ -ten Begriffes der Entität $E_j$
$p_j$	Mengenelement	Person in Menge $P$
$T_E$	Menge	Menge aller Themen einer Entität $E$
$t_i$	Mengenelement	Wort in Wortmenge $T$
$t_i^n$	Menge	Wortsequenz bestehend aus $n$ Worten
$tf$	Bezeichnung	Wortfrequenz
$T_E$	Menge	Topicraum einer Entität $E$
$V$	Menge	Menge aller $word2vec$ Vektoren
$w$	Variable	Beliebiges Häufigkeitsmaß in einer Kosinusnormalisierung, z.B. $tf$ oder $tf - idf$
$w_{ji}^{(1)}$	Hyperparameter	Gewicht zwischen dem $j$ -ten Neurons und der $i$ -ten Inputvariable im (1)-sten Layer eines KNN
$w_{j0}^{(1)}$	Hyperparameter	Bias des $j$ -ten Neurons im (1)-sten Layer eines KNN
$x_i$	Variable	Inputvariable eines KNN
$\vec{y}_{E_j k}$	Vektor	Ergebnisvektor zur Prädiktion des $k$ -ten Begriffes der Entität $E_j$