



FOM Hochschule für Oekonomie & Management

Hochschulzentrum Münster

Hausarbeit

im Studiengang Big Data & Business Analytics

im Rahmen der Lehrveranstaltung

Analyse semi- & unstrukturierter Daten

über das Thema

CAPTUM

- Characterisation of Type IIb autoimmune chronic spontaneous urticaria markers -

von

Fiete Ostkamp, Tim Lapstich und Artur Gergert

Betreuer : Prof. Dr. Rüdiger Buchkrämer

Matrikelnummern : 557851, , 562394

Abgabedatum : 16. August 2021

Inhaltsverzeichnis

| | |
|--|------------|
| Abbildungsverzeichnis | III |
| Tabellenverzeichnis | IV |
| Abkürzungsverzeichnis | V |
| Symbolverzeichnis | VI |
| 1 Einleitung | 1 |
| 1.1 Zielsetzung | 1 |
| 1.2 Aufbau der Arbeit | 1 |
| 2 Theoretische Grundlagen der Named Entity Recognition | 2 |
| 2.1 Text Preprocessing | 2 |
| 2.2 Named Entity Recognition | 6 |
| 2.3 Named Entity Linking | 8 |
| 3 Praxis | 9 |
| 3.1 Erstellung eines domänenspezifischen NER Modells | 9 |
| 3.1.1 Labeling von Trainingsdaten | 9 |
| 3.1.2 Annotierungsrichtlinien | 9 |
| 3.1.3 Software | 9 |
| 3.1.4 Training des Modells | 9 |
| 3.1.5 Evaluierung des Modells | 10 |
| 3.2 Markeranalyse | 10 |
| 3.3 Überführung in Tabellenstruktur | 10 |
| 3.4 Markerkorrelationen | 10 |
| 3.5 Aufstellung des Gratingsystems | 10 |
| 4 Fazit | 10 |
| Anhang | 11 |
| Literaturverzeichnis | 12 |

Abbildungsverzeichnis

| | | |
|---|--------------------------------------|---|
| 1 | Text Mining Prozess | 2 |
| 2 | Stemming vs. Lemmatization | 5 |

Tabellenverzeichnis

Abkürzungsverzeichnis

| | |
|------------|--|
| DL | Deep Learning |
| FN | Falsch-Negativ |
| FP | Falsch-Positiv |
| IR | Information Retrieval |
| KDT | “Knowledge Discovery in Textual Databases“ |
| ML | Machine Learning |
| NER | Named Entity Recognition |
| NLP | Natural Language Processing |
| POS | Part-of-Speech |
| TP | Richtig-Positiv |

Symbolverzeichnis

1 Einleitung

1.1 Zielsetzung

Die chronische spontane Urtikaria gehört zu der Gruppe chronischer Urtikaria Erkrankungen. Gekennzeichnet ist diese durch das Wiederauftreten von Quaddeln und/oder Angioödemem über einen Zeithorizont von mehr als sechs Wochen¹. Die geschätzte weltweite Prävalenz chronischer Urtikaria Erkrankten beträgt schätzungsweise 1%. Es liegt lediglich eine geschätzte Prävalenz vor, da es Schwierigkeiten bei der Klassifizierung, der Identifizierung sowie der Diagnose der Erkrankung gibt. Dies ist vor allem auf erhebliche Verzögerungen bei der Diagnose sowie unzureichende Kenntnisse über die chronische Urtikaria zurückzuführen².

Die oben angeführte Problematiken wurden zum Anlass genommen ein Projekt zu initiieren, welches den Auftrag verfolgt einen Beitrag zur bekämpfung der chronischen spontanen Urticaria Krankheit zu leisten. Begleitet wird das Projekt von Ärzten und Spezialisten der Charité in Berlin.

Die vorliegende Hausarbeit behandelt das Teilprojekt „Information Retrieval“. Ziel dieses Teilprojektes war es aus einem Text-Corpus mit insgesamt über 500 medizinische Fachartikel automatisiert Informationen aus den Texten zu extrahieren, um das Wissen über die Krankheit, erfolgreiche Behandlungsmöglichkeiten etc. zu erweitern.

1.2 Aufbau der Arbeit

Kapitel 2 enthält die Inhalte des Thesis-Days und alles, was zum inhaltlichen Erstellen der Thesis relevant sein könnte. In Kapitel 3 Praxis findet ihr wichtige Anmerkungen zu \LaTeX , wobei die wirklich wichtigen Dinge im Quelltext dieses Dokumentes stehen (siehe auch die Verzeichnisstruktur in Abbildung ??).

¹ savic.2020.

² savic.2020.

2 Theoretische Grundlagen der Named Entity Recognition

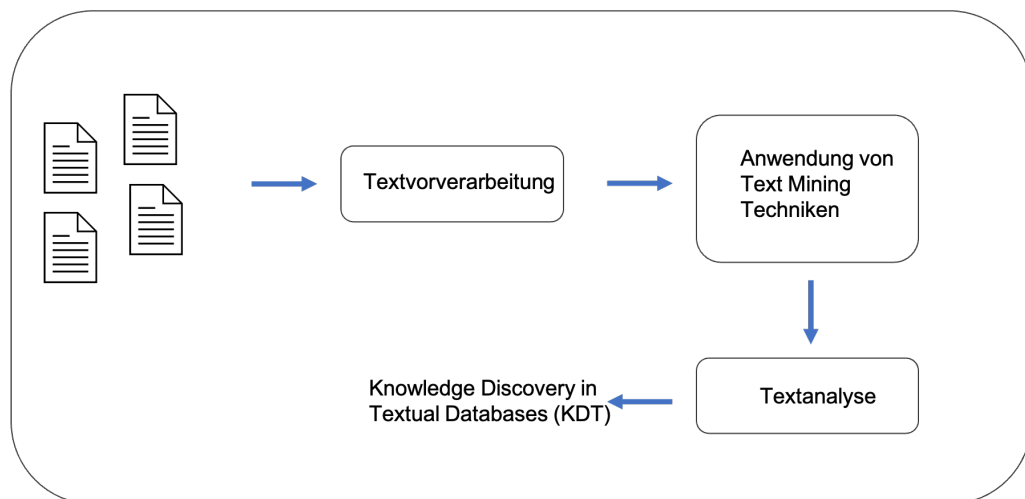
2.1 Text Preprocessing

Text Mining bezeichnet den Prozess, um wesentliche bekannte aber auch unbekannte Informationen aus Textdaten zu generieren³ Die Verarbeitung von unstrukturierten Textdaten wird auch als "Knowledge Discovery in Textual Databases" (KDT) bezeichnet und spielt eine signifikante Rolle in Anwendungsgebieten wie

- Information Retrieval
- Information Extraction
- Natural Language Processing⁴

Im Wesentlichen geht es in allen o. g. Anwendungsgebieten um die Wissen durch das Mining der Texte zu generieren.

Abbildung 1: Text Mining Prozess



Quelle: Eigene Darstellung in Anlehnung an mohan.2015

Wie der Abbildung 1 entnommen werden kann, stellt die Vorverarbeitung von Volltextdaten bei nahe zu jeder Aufgabe im Natural Language Processing (NLP) einen essentiellen und kritischen Schritt dar, da hierbei die fundamentale Basis für die Weiterverarbeitung

³ mohan.2015.

⁴ mohan.2015.

sowie die Entwicklung der Modelle geschaffen wird.⁵ Der Begriff der Textvorverarbeitung umfasst dabei die Anwendung unterschiedlicher Techniken/Methoden, bei denen die Textdokumente für die eigentlichen Zielsetzungen vorbereitet werden. Gängige Techniken für die Vorbereitung der Texte für die nachgelagerten Analysen können folgendermaßen aufgeteilt werden:⁶

- Inhalte Extrahieren und Bereinigen
- Annotationen
- Normalisieren

Zu Beginn der Volltextanalysen stehen häufig die Rohfassungen der Texte zur Verfügung. Diese gilt es im ersten Schritt technisch einzulesen. Hierbei werden auch Daten mit eingelesen, dessen Informationsgehalt gering ist. Beispielshaft zu nennen sind hier HTML tags, Werbung, etc beim Auslesen einer Website⁷ oder Grafiken, ASCII-Codes in PDF-Dateien. Demnach ist das Ziel bei dem **Extrahieren und Bereinigen der Inhalte** die Rohdaten soweit zu säubern, bis sich schließlich die reinen Texte als Resultat ergeben. Nachdem die Texte um die technischen Störfaktoren bereinigt wurden, ist die **Tokenization** eine typische Technik der Textextraktion. In dem Prozess der Tokenisation wird der gesamte zu analysierende Text in einzelne Wörter, Phrasen, Symbole, etc. geteilt. Hierbei wird das Ziel verfolgt die Bedeutung einzelner Wörter innerhalb eines Satzes zu analysieren. Die Tokens dienen nämlich als Eingabewerte für viele weitergehende Prozessschritte.⁸ In jedem Text befinden sich Wörter, die wenig Informationsgehalt bei der Textanalyse bieten. Solche Wörter werden auch als **Stop Words** bezeichnet. Beispiele für solche Stop Words sind Artikel oder Präpositionen wie „der“, „die“, „das“, „ein“, „in“, „mit“, etc. Im Analyseprozess stellt jedes unterschiedliche Wort eine eigene Dimension dar. Durch die Entfernung der Stop Words wird somit die Dimensionshöhe reduziert bei gleichzeitiger Beibehaltung des Informationsgehaltes des jeweiligen Satzes/Textes.⁹ Neben der klassischen Methode, die Stop Words auf Basis einer vordefinierten Liste zu entfernen, sind diverse mathematische und nicht-mathematische Methoden entwickelt worden, um Stop Words in Texten zu identifizieren und zu bereinigen.¹⁰

⁵ **gurusamy.2014.**

⁶ **pahwa.2018.**

⁷ **pahwa.2018.**

⁸ **gurusamy.2014.**

⁹ **mohan.2015.**

¹⁰ Detailliertere Informationen zu unterschiedlichen Methoden für die Entfernung von Stop Words können **mohan.2015** entnommen werden.

Die Annotationen eines Textes sollen die Funktion des jeweiligen Wortes im Kontext des gesamten Satzes identifizieren. Eine gängige Methode ist dabei das so genannte Part-of-Speech (POS)-Tagging. Jener ist ein Prozess, der die Zuweisungen einzelner Wörter zur POS bzw. zu lexikalischen Klassenmarkern wie Nomen, Verben, Adjektiven, usw. vornimmt.¹¹ Bei der Erkennung der Funktion eines Satzes treten folgende Hauptprobleme auf:

- Mehrdeutige Wörter
- Unbekannte Wörter

Ersteres stellt das wichtigste Problem dar. Es gibt Wörter, für die es mehr als einen Tag geben kann. Dieses Problem wird durch einen Fokus des Wortes im Satzkontext gelöst.¹² Andersrum existieren Wörter, die zwar denselben Tag haben, jedoch unterschiedliche Bedeutungen im Satzkontext einnehmen.¹³ Eine Lösung ist hierbei die Betrachtung des einzelnen Wortes, statt dem Kontext. Das menschliche Auge kann solch eine Differenzierung schnell vornehmen, während es für eine Maschine mühselige Arbeit und einen Lernprozess darstellt. Wie zuvor angeführt, ist es für die automatische POS-Erkennung notwendig ein Modell zu trainieren. Hier wurden mit der Zeit diverse Methodiken entwickelt, die sich auf oberster Ebene in überwachte und unüberwachte Methoden aufteilen. Bei den überwachten Ansätzen wird das POS-Modell auf Basis eines Datensatzes trainiert, bei dem die POS-Werte bekannt sind, während das Modell bei den unüberwachten Ansätzen die POS-Werte selbst induziert, da dem Trainings-Datensatz keine bekannten Werte vorliegen.¹⁴

Bei dem **Normalisieren** von Texten wird das Ziel verfolgt ähnliche Wörter zu vereinheitlichen bzw. diese auf einen Standard zu bringen. Dieser Prozess soll vor allem die Dimensionen reduzieren, um die Berechnungen zu vereinfachen und gleichzeitig die Effizienz durch die Standardisierung der Wörter erhöhen.

Bei der Normalisierung von Wörtern wird häufig auf die Techniken des **Stemming** oder der **Lemmatization** zurückgegriffen. Das Stemming ist ein Prozess, der zugrundeliegende Wörter auf den Wortstamm herunterbricht.¹⁵ Dieser Wortstamm ist im Ergebnis häufig kein echtes Wort, sondern oftmals eine Buchstabenkombination bzw. ein Präfix, den viele Wörter gemeinsam haben.¹⁶ Es existiert eine Vielzahl an Stemming-Algorithmen, dessen Performance vom jeweiligen Einsatzbereich abhängt, sodass noch kein Standard etabliert

¹¹ kumawat.2015.

¹² gurleenkaursidhu.2013.

¹³ gurleenkaursidhu.2013.

¹⁴ Eine detaillierte Übersicht über vorhandene POS- Methoden sind in kumawat.2015 zu finden.

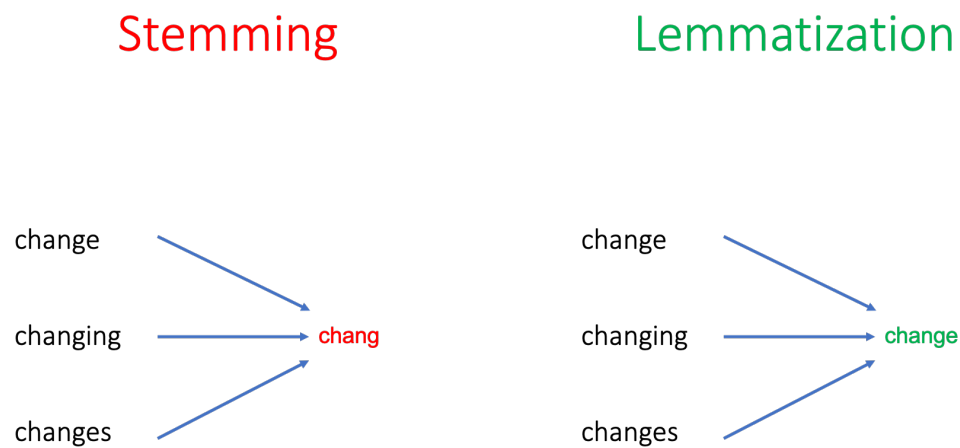
¹⁵ khyani.2021.

¹⁶ khyani.2021.

ist.¹⁷

Die Idee bei der Lemmatization ist das so genannte “Lemma“ oder auch die Vokabularform eines Wortes zu identifizieren.¹⁸ Der Prozess ist ähnlich dem des Stemming, jedoch mit dem Unterschied im Ausgabewert. Während beim Stemming der Ausgabewert der Wortstamm ist und oftmals kein echtes Wort, ist der Ausgabewert bei der Lemmatization das Grundwort aus beispielsweise einem Wörterbuch.¹⁹ Die Abbildung 2 verdeutlicht die Unterschiede der Lemmatization und des Stemming anhand des englischen Wortes “change“ bzw. dessen Abwandlungen.

Abbildung 2: Stemming vs. Lemmatization



Quelle: Eigene Darstellung in Anlehnung an khyani.2021

Ein vorab durchgeführtes POS-Tagging kann die Ergebnisse der Lemmatization optimieren. Wenn das Wort „schloss“ nämlich in einem Text als Nomen durch das POS-Tagging erkannt wird, dann handelt es sich dabei je nach Kontext entweder um ein Schloss zum verriegeln oder um ein Schloss als Gebäude. Wird es dagegen als Verb identifiziert, so wird mit einer hohen Wahrscheinlichkeit zum Lemma „schließen“ umgewandelt.

¹⁷ Eine ausführliche Analyse unterschiedlicher Stemming-Algorithmen kann jivani.2011 entnommen werden.

¹⁸ khyani.2021.

¹⁹ khyani.2021.

2.2 Named Entity Recognition

Mithilfe der Named Entity Recognition (NER) wird das Ziel verfolgt automatisiert Eigennamen in Texten zu identifizieren, dessen semantische Typen wie beispielsweise Personen, Ort, Organisationen vordefiniert wurden.²⁰ Die NER kann nicht nur ausschließlich für die Extraktion von Informationen aus Texten genutzt werden, viel mehr spielt sie eine wesentliche Rolle in einer Vielzahl von Anwendungen aus dem Gebiet des NLP wie beispielsweise dem Textverständnis, dem Information Retrieval (IR), automatisierter Textzusammenfassungen und Übersetzungen, Fragenbeantwortungen, etc. In der Forschung existiert eine Vielzahl an Definitionen für die zu erkennenden Eigennamen, die hauptsächlich in folgende zwei Kategorien aufgeteilt werden können:

- Generische (z. B. Personen und Ort)
- Domänenspezifische (z. B. Proteine, Enzyme und Gene)²¹

Bei den in der NER angewandten Techniken wird zwischen

- Regelbasierte Ansätze
- Unüberwachte Ansätze
- Merkmalsbasierte überwachte Lernansätze
- Deep-Learning Ansätze

unterschieden, wobei die drei erstgenannten den traditionellen Ansätzen zugehörig sind.²² Regelbasierte Systeme beruhen auf manuell erstellten Regeln. Die zugrundeliegenden Regeln können hierbei z. B. aus domänenspezifischen Ortsverzeichnissen oder syntaktisch-lexikalischen Mustern abgeleitet worden sein.

Das Clustering ist ein typischer Ansatz für unüberwachte NER-Systeme.²³ Auf Basis von Kontextähnlichkeiten werden geclusterte Gruppen generiert aus denen schließlich die Entitäten extrahiert werden. Die Idee bei dieser Technik ist mithilfe eines großen Corpus lexikalische Muster und Statistiken zu berechnen, um daraus auf im Text benannte Entitäten schließen zu können.²⁴

Im Teilbereich der überwachten Ansätze ist NER hauptsächlich eine Klassifikationsaufgabe. Ausgehend von annotierten Datensätzen werden Merkmale entwickelt, um jedes

²⁰ **nadeau.2007.**

²¹ **li.2018.**

²² **li.2018.**

²³ **nadeau.2007.**

²⁴ **li.2018.**

Trainingsbeispiel zu repräsentieren.²⁵ Für die Entwicklung der Modelle kommen dann Algorithmen des Machine Learning (ML) zu Einsatz, um aus den gegebenen annotierten Datensätzen Vorhersagemodelle für noch ungesehene Daten zu erlernen.²⁶ Essentiell in überwachten NER-Systemen ist die Entwicklung der Merkmale. Merkmalsvektoren abstrahieren dabei den Text, bei der ein Wort durch einen oder mehrere boolische, numerische oder nominale Werte dargestellt wird.²⁷

Neben den eben erläuterten traditionellen Methoden für die NER wurden in den letzten Jahren Ansätze im Bereich des Deep Learning (DL) entwickelt, welche sich bewährt haben und Spitzenenergebnisse erzielen.²⁸ Der Einsatz von DL hat wesentliche Vorteile gegenüber den traditionellen Methoden. Zum einen ist es durch die besondere Architektur und den Verarbeitungsmöglichkeiten im Bereich des DL über mehrschichtige künstliche neuronale Netze möglich nicht-lineare Zusammenhänge zu erkennen und zu lernen und zum anderen erleichtern DL-basierte Modelle durch ihre Automation und Selbstständigkeit beim Lernen die Arbeit.²⁹

Die NER umfasst zwei Teilaufgaben: Typenidentifikation und Grenzerkennung. Die Bewertung eines entwickelten NER-Systems wird in der Regel durch den Vergleich mit den menschlich getätigten Annotationen vorgenommen. Der Vergleich kann entweder über eine exakte oder durch eine partielle Übereinstimmungsevaluation vorgenommen werden.³⁰ Bei der exakten Übereinstimmungsevaluation wird geprüft, ob das System sowohl den richtigen Typen als auch die Grenzen korrekt identifiziert.³¹ Bei der partiellen Übereinstimmungsevaluation genügt es, wenn, unabhängig von den Grenzen, eine korrekte Identifikation des Typen vom System vorgenommen wird, solange es eine Überschneidung zwischen den ermittelten Grenzen und den wahren Grenzen gibt. Bewertungskennzahlen für die Qualität des NER-Systems sind Precision, Recall und der F-Score. Um diese zu ermitteln, wird vorab die Anzahl der Falsch-Positiv (FP), Falsch-Negativ (FN) und Richtig-Positiv (TP) Zuordnungen ermittelt.

- FP: Eine Entität wurde vom System erkannt, welche in Wirklichkeit keine ist
- FN: Eine Entität wurde nicht vom System erkannt, welche in Wirklichkeit jedoch eine darstellt
- TP: Eine Entität wurde vom System korrekt erkannt

²⁵ li.2018.

²⁶ li.2018.

²⁷ nadeau.2007.

²⁸ li.2018.

²⁹ li.2018.

³⁰ li.2018.

³¹ tjongkimsang.2003.

Die Precision stellt sich dar als

$$Precision = \frac{TP}{(TP + FP)} \quad (1)$$

und kann als Verhältnis von richtig erkannten Entitäten zur Gesamtheit an identifizierten Entitäten interpretiert werden.

Der Recall

$$Recall = \frac{TP}{(TP + FN)} \quad (2)$$

bezieht die richtig erkannten Entitäten auf die Gesamtheit aller möglich gewesenen Entitäten.

Der ausgeglichene F-Score vereint die Precision und den Recall zu einem harmonischen Mittel:

$$F - Score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \quad (3)$$

2.3 Named Entity Linking

3 Praxis

3.1 Erstellung eines domänenspezifischen NER Modells

Wie im Kapitel 2.2 erläutert lässt sich zwischen allgemeinen und domänenspezifischen NER-Modellen unterscheiden. Domänenspezifische Modelle sind dabei solche, die für einen dedizierten Themenbereich erstellt wurden und daher in diesem Kontext besonders gute Ergebnisse erzielen. Es gibt bereits spezialisierte Modelle im Bereich der Medizin die mit einem medizinischen Korpus trainiert wurden. Dazu zählen zum Beispiel BioBERT, ScispaCy oder Y³². Im folgenden beschreiben wir, wie wir ein eigenes Modell für das Themengebiet der Urtikaria-Forschung erstellt haben.

3.1.1 Labeling von Trainingsdaten

Das Annotieren von Trainingsdaten hat einen zentralen Anteil bei der Erstellung eines eigenen Modells. Akkurate Daten sind essentiell für die Genauigkeit des resultierenden Modells. Neben der Menge der zum Training zur Verfügung stehenden Daten ist es unerlässlich, dass diese widerspruchsfrei sind.

3.1.2 Annotierungsrichtlinien

33

3.1.3 Software

34

3.1.4 Training des Modells

Für das Training des Modells werden die annotierten Daten aus den vorherigen Schritten verwendet. Für das Training des Modells wurde die Natural Language Processing Bibliothek spaCy³⁵ verwendet. Diese wurde im Hinblick auf ihre umfangreichen Funktionalitäten und benutzerfreundlichkeit ausgewählt.

³² Li, J. et al., 2020, S.12.

³³ vgl. Neves, M., 2014.

³⁴ vgl. Neves, M., Leser, U., 2014.

³⁵ Honnibal, M., Montani, I., 2017.

- Warum SpaCy
- Modellarchitektur ()
- Trainingsparameter

Die Parameter für das Training des Modells orientieren sich im Wesentlichen an den vor-eingestellten Werten. Es wurden keine starken Änderungen vorgenommen, da bei den beteiligten Personen bisher noch nicht die notwendige Erfahrung vorhanden war um ein Parametertuning vorzunehmen. Bei der zugrundeliegenden Architektur handelt es sich so-mit um einen sogenannten Transition Based Parser. Dieser funktioniert indem...

3.1.5 Evaluierung des Modells

36

3.2 Markeranalyse

3.3 Überführung in Tabellenstruktur

3.4 Markerkorrelationen

3.5 Aufstellung des Gratingsystems

4 Fazit

Wünsche Euch allen viel Erfolg für das 7. Semester und bei der Erstellung der Thesis. Über Anregungen und Verbesserung an dieser Vorlage würde ich mich sehr freuen.

³⁶ Tsai, R. T.-H. et al., 2006.

Anhang

Anhang 1: Beispielanhang

Dieser Abschnitt dient nur dazu zu demonstrieren, wie ein Anhang aufgebaut sein kann.







Anhang 1.1: Weitere Gliederungsebene

Auch eine zweite Gliederungsebene ist möglich.

Anhang 2: Bilder

Auch mit Bildern. Diese tauchen nicht im Abbildungsverzeichnis auf.

Abbildung 3: Beispielbild

| Name | Änderungsdatum | Typ | Größe |
|---|------------------|--------------------|-------|
|  abbildungen | 29.08.2013 01:25 | Dateiordner | |
|  kapitel | 29.08.2013 00:55 | Dateiordner | |
|  literatur | 31.08.2013 18:17 | Dateiordner | |
|  skripte | 01.09.2013 00:10 | Dateiordner | |
|  compile.bat | 31.08.2013 20:11 | Windows-Batchda... | 1 KB |
|  thesis_main.tex | 01.09.2013 00:25 | LaTeX Document | 5 KB |

Literaturverzeichnis

Honnibal, Matthew, Montani, Ines (2017): spaCy 2: Natural Language Understanding with Bloom Embeddings, Convolutional Neural Networks and Incremental Parsing, o. O., 2017

Neves, M., Leser, U. (2014): A Survey on Annotation Tools for the Biomedical Literature, in: Briefings in Bioinformatics, 15 (2014), Nr. 2, S. 327–340, [Zugriff: 2021-08-14]

Neves, Mariana (2014): An Analysis on the Entity Annotations in Biological Corpora, in: F1000Research, 3 (2014), S. 96, [Zugriff: 2021-08-10]

Tsai, Richard Tzong-Han, Wu, Shih-Hung, Chou, Wen-Chi, Lin, Yu-Chun, He, Ding, Hsiang, Jieh, Sung, Ting-Yi, Hsu, Wen-Lian (2006): Various Criteria in the Evaluation of Biomedical Named Entity Recognition, in: BMC Bioinformatics, 7 (2006), Nr. 1, S. 92, [Zugriff: 2021-08-14]

Internetquellen

Li, Jing, Sun, Aixin, Han, Jianglei, Li, Chenliang (2020): A Survey on Deep Learning for Named Entity Recognition, arXiv: 1812.09449 [cs], <<http://arxiv.org/abs/1812.09449>> (2020-03-18) [Zugriff: 2021-08-08]

Ehrenwörtliche Erklärung

Hiermit versichere ich, dass die vorliegende Arbeit von mir selbstständig und ohne unerlaubte Hilfe angefertigt worden ist, insbesondere dass ich alle Stellen, die wörtlich oder annähernd wörtlich aus Veröffentlichungen entnommen sind, durch Zitate als solche gekennzeichnet habe. Ich versichere auch, dass die von mir eingereichte schriftliche Version mit der digitalen Version übereinstimmt. Weiterhin erkläre ich, dass die Arbeit in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde/Prüfungsstelle vorgelegen hat. Ich erkläre mich damit **einverstanden/nicht einverstanden**, dass die Arbeit der Öffentlichkeit zugänglich gemacht wird. Ich erkläre mich damit einverstanden, dass die Digitalversion dieser Arbeit zwecks Plagiatsprüfung auf die Server externer Anbieter hochgeladen werden darf. Die Plagiatsprüfung stellt keine Zurverfügungstellung für die Öffentlichkeit dar.

Münster, 16.8.2021

(Ort, Datum)

A handwritten signature in black ink, consisting of a large, stylized 'H' followed by a series of loops and a final flourish.

(Eigenhändige Unterschrift)