Universität Trier
Digital Humanities
Digitale Methoden: Programmieren 2, Maschinelles Lernen
Sommersemester 2023
Seminarleitung: Dr. Thomas Burch

# Open Government Data in Deutschland:
## Eine computergestützte Auswertung der Schlagwörter auf GovData.de

Lisa Nathalie Braune
Matrikelnummer: 1243366
An der Feldport 10
54292 Trier
s2librau@uni-trier.de
Digital Humanities (Master of Science)
Fachsemester: 5

# Inhaltsverzeichnis

# 1. Einleitung

Open Data oder offene Daten sind in den letzten Jahrzehnten immer wichtiger geworden. Gemeint sind damit, Daten, die „von jedermann frei verwendet, nachgenutzt und verbreitet werden können“ (Open Knowledge Foundation Deutschland, 2023). Einschränkungen sind dabei höchstens, dass die Datensätze als Quellen genannt werden müssen sowie weitere Veröffentlichungen auf Basis dieser Daten nur unter genau diesen Prinzipien stattfinden dürfen. Die Bereitstellung dieser Daten kann zu mehr Transparenz verhelfen und ermöglicht überhaupt erst die kommerzielle und nichtkommerzielle Nachnutzung von Akteuren wie Wissenschaftler*innen, Unternehmen oder Privatpersonen (W3C, 2009). So wurden sie beispielsweise genutzt, um verbesserte Sicherheitssysteme zu entwickeln (Abdelrahman, 2022, S. 59) oder um die Zufriedenheit von Telearbeiter*innen während der Pandemie zu ermitteln (Javier de Esteban et al., 2023).

Ein großer Akteur in der Erstellung von Daten ist die öffentliche Verwaltung: „Statistische Daten, Geodaten, Umweltdaten, Wetterdaten, Haushaltsdaten, Forschungsdaten“ und mehr zählen dazu (von Lucke & Gollasch, 2022, S. 58). In Deutschland gibt es neben vielen spezifischen Anbietern wie dem Unfallatlas Deutschland (https://unfallatlas.statistikportal.de/) oder Open Legal Data (https://de.openlegaldata.io/), vor allem auch das zentrale Metadatenportal GovData (GovData, 2023). Hier werden Verlinkungen zu den offenen Datensätzen der öffentlichen Stellen von Bund, Ländern, Kommunen und mehr eingestellt. Während die Datensätze selbst auf den Servern der jeweiligen öffentlichen Stelle liegen, sind auf GovData vor allem die Metadaten veröffentlicht, die diese Datensätze beschreiben. Auf diese Weise sind hier über 86.000 Datensets verknüpft.

Um diese Daten jedoch erst nutzbar zu machen, ist es notwendig, zu wissen, welche Datensätze vorhanden sind. Jedoch ist es bei solch großen Mengen kaum händisch möglich, sich einen Überblick verschaffen zu können. Hier können Prozesse des Natural Language Processing Abhilfe schaffen, also „computergestützte Techniken zur maschinellen Erkennung und Verarbeitung von natürlicher Sprache“ (Fraunhofer-Institut 2019).

Zur Untersuchung von GovData bieten sich vor allem die Schlagwörter an, die zu fast jedem Datensatz hochgeladen werden. Sie sind die einzigen Metadaten auf GovData, deren Anzahl frei festgelegt werden kann, und bieten so im Gegensatz zu der Einteilung in Kategorien, von denen nur 13 angeboten werden, sehr viel mehr Freiraum. Um diese Schlagwörter und Wörter allgemein computergestützt verarbeiten zu können, ist es notwendig, sie in Vektoren umzuwandeln, auch Word Embedding genannt. Mithilfe dessen sollen die Schlagwörter analysiert werden, um ihre Abstände, Zusammengehörigkeiten und Unterschiede untersuchen zu können.

Kapitel 2 beschäftigt sich zu diesem Zweck zuerst mit der Theorie: In Kapitel 2.1 werden die Grundlagen von Open Data und Open Government Data in Deutschland aufgezeigt mit Fokus auf das Metadatenportal GovData. In Kapitel 2.2 wird das Continuous Bag-of-Words-Modell dargestellt, welches zur Vektorisierung der Wörter genutzt wird. Kapitel 3 beschreibt und erläutert das Programm, welches zu diesem Zweck in Python geschrieben wurde. In Kapitel 4 werden die Ergebnisse der Untersuchung dargestellt und Kapitel 5 gibt schließlich eine Zusammenfassung der Studie sowie den Ausblick auf mögliche Anschlussstudien.

# 2. Theoretischer Hintergrund

## 2.1 Open Data in Deutschland

Der Begriff "Open Data" kam erstmals 1995 in den USA auf. Die Scholarly Publishing and Academic Resources Coalition hat damals schon die Vorteile von Open Data angesprochen: Es beschleunige Fortschritt, steigere die Wirtschaft, helfe Personen, keine Durchbrüche zu verpassen und verbessere die Integrität wissenschaftlicher Aufzeichnungen (Jemili & Bouras, 2022, S. 22).

Im Jahr 2009 folgte eine große Bewegung hin zu Open Data, als mehrere Länder verkündeten, dass sie Initiativen starten, um öffentliche Daten frei verfügbar machen. Zu diesen Ländern zählten beispielsweise die USA, Großbritannien und Neuseeland. Dank dieser Initiativen wurde aus der Bewegung für Open Data heraus die Bewegung hin zu Open Government Data gegründet (Jemili & Bouras, 2022, S. 22). In diesem Jahr wurde data.gov ans Netz gebracht, die zentrale Website für Open Government Data der USA. Im Jahr darauf folgte data.gov.uk, das Äquivalent für Großbritannien (Jemili & Bouras, 2022, S. 23). Viele weitere Länder folgten, so auch Deutschland 2013 mit GovData.

Hierzulande regelt seit 2017 zusätzlich §12 des E-Government-Gesetz den Umgang mit offenen Daten, welches 2021 noch ergänzt wurde. Hierdurch sind Ministerien der Bundesverwaltung und Anstalten sowie Körperschaften des öffentlichen Rechts verpflichtet, ihre Verwaltungsdaten in maschinenlesbarer Form über GovData bereitzustellen. Ebenso legt das Datennutzungsgesetz „einheitliche, nichtdiskriminierende Nutzungsbedingungen für Daten des öffentlichen Sektors“ fest (von Lucke & Gollasch, 2022, S. 56).

Diese offenen Verwaltungsdaten, im internationalen Sprachgebrauch vor allem Open Government Data genannt, sind "[s]ämtliche Datenbestände des öffentlichen Sektors, die von Staat und Verwaltung im Interesse der Allgemeinheit der Gesellschaft ohne jedwede Einschränkung zur freien Nutzung, zur Weiterverbreitung und zur freien Weiterverwendung frei zugänglich gemacht werden". Sie beinhalten also alle Daten, die nicht explizit durch Geheimhaltungspflichten geschützt sind,

beispielsweise durch Vorkommen von personenbezogenen Daten (von Lucke & Gollasch, 2022, S. 54).

GovData ist, wie schon beschreiben, das zentrale Metadatenportal für alle öffentlichen offenen Daten. Als solches sind auf der Website Datensätze verknüpft, die für die Veröffentlichung mit Metadaten angereichert wurden. Das Format und der Inhalt dieser Metadaten ist in dem Metadatenmodell DAT-AP.de festgelegt, welches eine deutsche Abwandlung des europäischen Standards DCAT-AP ist, dem Standard für Datenaustausch auf europäischer Ebene. Dieses Modell sieht für Datensätze zwingend vor, einen Titel sowie eine Beschreibung zu beinhalten. Empfohlen sind zudem unter anderem die Bereitstellung von Schlagwörtern, dem räumlichen und zeitlichem Bezug, einer Kontaktmöglichkeit oder einer Einteilung in eine oder mehrere der 13 Kategorien. Sie beinhalten Themengebiete wie Energie, Gesundheit oder Umwelt.

Zum vereinfachten Auslesen der Datensätze bietet GovData unter anderem einen SPARQL-Endpunkt an (GovData, 2023), welcher sich in Python einfach implementieren lässt. SPARQL eine Abfragesprache für Datenbanken, ähnlich wie SQL. Sie fragt jedoch keine relationalen Datenbanken ab, sondern Datenbanken im RDF-Format. RDF steht für Resource Description Framework und bietet eine Alternative zu relationalen Datenbanken. Im Gegensatz zu diesen arbeitet sie nicht mit tabellarischen Daten, sondern die Informationen sind in Tripeln gespeichert, welche aus Subjekt, Prädikat und Objekt bestehen (Wikipedia, 2023).

Um überhaupt ermitteln zu können, welche Datensätze auf GovData vorliegen, können also diese Metadaten, vorzugsweise mittels dem SPARQL-Endpunkt, ausgelesen und analysiert werden. Hierzu bieten sich verschiedene Möglichkeiten. Eine Möglichkeit ist, Titel und Beschreibung zu verarbeiten. Dies würde garantieren, alle Datensätze zu erfassen, da diese beiden Metadaten immer vorhanden sein müssen. Ebenso würde es Sinn ergeben, die 13 vorhandenen Kategorien dazuzuziehen, um die Datensets auszuwerten. Mehr Potential noch sehe ich aber in den Schlagwörtern. Eine Anfrage auf dem auf GovData angebotenen SPARQL-Editor hat ergeben, dass es 1.279 Datensätze gibt, die keine Schlagwörter beinhalten. Diese machen mit rund 1,5% nur einen sehr geringen Anteil aller Datensätze aus.Werden diese Schlagworte gut genutzt, so bieten sie einen hohen Mehrwert, der gut automatisiert nutzbar ist. Werden diese Schlagworte gut vergeben, so können sie einen sehr hohen Mehrwert bieten, der außerdem gut automatisiert nutzbar ist. Sie können den Datensatz thematisch, zeitlich oder örtlich einordnen sowie Akteure, Beteiligte oder andere mit den Daten verbundene Entitäten nennen.

Bei einer ersten manuellen Durchsicht der Schlagworte ließ sich jedoch erkennen, dass die Schlagwortvergabe von Datensatz zu Datensatz stark variieren kann. Dies ist zum einen erkennbar in der Menge der Schlagwörter eines Datensatzes: So gibt es Datensätze komplett ohne Schlagworte

(„Städtische Beteiligungen“, 2023), der Datensatz mit den meisten Schlagworten hat dagegen 720 Schlagworte („Ausländer: Kreise, Stichtag, Geschlecht“, 2018).
Zum anderen ist die Vergabe der Schlagworte sehr unterschiedlich. Einige Schlagworte scheinen für Laien kaum nachvollziehbar und damit auch kaum nutzbar, beispielsweise das Schlagwort "104a abs.1 satz 1 aufenthg ae auf probe" („Ausländer: Deutschland, Stichtag, Geschlecht“, 2018), welches auf §104a im Aufenthaltsgesetz hinweist. Dies ist zwar eine sehr genaue Angabe, jedoch ist die Wahrscheinlichkeit, dass Personen, die nach Datensätzen zu diesem Gesetz suchen, genau diese Formulierung wählen, sehr gering. Das führt auch dazu, dass andere Datensätze kaum genau dieses Schlagwort auch nutzen, da auch hier die Wahrscheinlichkeit dafür sehr gering ist, sofern nicht dieselbe Person dieses Schlagwort vergibt.
Dagegen gibt es einige viel genutzte Schlagwörter, die redundant sind. „Opendata“ ist mit 29.345 Malen in über einem Drittel aller heruntergeladenen Datensets enthalten. Zwar gibt es auf der Website auch Datensätze, die nicht offen sind, weswegen eine Unterscheidung Sinn machen könnte. Jedoch sieht das Metadatenformat für diese Angabe die Kategorie „Verfügbarkeit“ bzw. „availability“ vor, mittels derer sich derartige Informationen weitaus einfacher auslesen lassen können.
Außerdem hat eine SPARQL-Anfrage ergeben, dass viele der Schlagwörter nur einmalig verwendet werden. Von 50.205 insgesamt genutzten Schlagwörtern wurden 20.860 Stück nur ein einziges Mal verwendet. Das entspricht über 40%. Damit bieten auch die Schlagworte keine ideale Datenbasis. Es kann sein, dass das Ergebnis der Vektorisierung und damit auch einer räumlichen Gruppierung der Wörter nur begrenzt nutzbar ist. Dennoch ist es sinnvoll, eine Auswertung dieser Schlagworte anzugehen. Sie sind Teil der größten Menge an Metainformationen über öffentliche offene Daten, die es in Deutschland gibt. Das Ergebnis dieser Arbeit kann nicht nur genutzt werden, die Schlagworte miteinander zu vergleichen, sondern auch die Schlagwortpraxis an sich zu untersuchen und Verbesserungsvorschläge zu bringen.

## 2.2 Das CBOW-Modell

Um die Schlagwörter untersuchen zu können, müssen diese maschinenlesbar verarbeitet werden. Es ist möglich, die Nähe der Wörter untereinander von Hand zu bestimmen, dann müsste jedoch jedes der über 50.000 Schlagwörter mit jedem anderen Wort in Beziehung gesetzt werden, was über zwei Milliarden Prozessen entspräche. In der Praxis ist dies also kein sinnvoller sowie effizienter Weg der Kategorisierung.
Die Lösung für dieses Problem ist das „Word Embedding“. Hierbei wird jedes Wort in einen Vektor mit n Dimensionen verwandelt, um mathematisch Distanzen und Nähen bestimmen zu können. Erst

hiermit werden Prozesse des Natural Language Processing ermöglicht (Rizkalla et al., 2022, S. 233). Hierbei dürfen Vektoren nicht zufällig bestimmt werden, sondern in ihnen muss eine gewisse Bedeutung enthalten sein. Die Vektoren ähnlicher Wörter sollten nah beieinander und die Vektoren sehr weit entfernter Wörter weit auseinander liegen. So sollte "Paris" beispielsweise näher an "Berlin" als an "Frühstück" liegen.

Effiziente Vektoren entstehen mittels Neuronalen Netzen, welche auf einem bestimmten Korpus trainiert werden. Eine Möglichkeit, diese Embeddings zu erhalten, ist die Nutzung vortrainierter Modelle. In diesen sind Wörter aus der deutschen Sprache enthalten, die auf möglichst vielen Texten trainiert wurden und die schon bedeutsame Vektoren enthalten. Diese Vorgehen kann einige Stunden an Rechenzeit ersparen, die ein Neuronales Netz zum Trainieren benötigt. Dies kann gut bei beispielsweise Romanen funktionieren, die in Alltagssprache verfasst sind. Die schon genannten Beispiele der Schlagwörter lassen allerdings schnell erkennen, dass hier ein vorgefertigtes Korpus nicht sinnvoll ist. Während Schlagwörter wie „Umwelt“ und „Verkehr“ noch sinnvoll verortet werden können, sind aufwendigere Schlagwörter wie "104a abs.1 satz 1 aufenthg ae auf probe" gar nicht erst im Wortschatz enthalten. Es bleibt also, ein eigenes Neuronales Netz zu erstellen, welches auf diesem spezifischen Korpus trainiert wird.

Es gibt verschiedene Methoden der Vektorisierung. Zwei im Vergleich zu vorigen Methoden sehr effiziente und Verfahren, die gute Ergebnisse erzielen, sind Skipgram und Continuous Bag-of-Words (CBOW). Diese beiden Modelle wurden 2013 von Google vorgestellt. Im Gegensatz zu vorherigen Ansätzen erstellten sie sehr viel kürzere Vektoren als bisherige Modelle, was die Rechenzeit erheblich verkürzt. Vorige Modelle hatten teilweise Vektoren von einer Länge des gesamten Korpus. Mit den neuen Modellen lässt sich die Länge der Vektoren eigenständig bestimmen (Mikolov et al., 2013).

Die Neuerung dieser beiden Modelle war, dass der direkte Kontext des zu untersuchenden Wortes mit einbezogen wurde. Mittels der Fenstergröße lässt sich bestimmen, wie viele Wörter vor und nach dem aktuellen Wort in das Training mit einbezogen werden sollen (Mikolov et al., 2013, S. 4). Bei einer Fenstergröße von zwei werden beispielsweise die zwei Wörter vor und nach dem aktuellen Wort mit einbezogen.

Der Unterschied zwischen diesen beiden Modellen ist, dass CBOW genutzt wird, um aus einem Wort seinen Kontext zu bestimmen, Skipgram bestimmt ein Wort anhand seines Kontexts (Mikolov et al., 2013, S. 4). Für eine Vektorisierung sind beide Modelle geeignet (Hu et al., 2017, S. 1038). CBOW ist jedoch schneller und kann besser syntaktische Informationen speichern, Skipgram dagegen ist für die Erschließung semantischer Informationen besser (Mikolov et al., 2013, S. 7f). Aus Zeitgründen habe ich mich für eine Implementierung mit CBOW entschieden. Dieses Modell

ist schneller als Skipgram, und es wird erwartet, dass das Training dieser Datenmenge einige Stunden in Anspruch nimmt, auch wenn die Datenmengen sich für große Datenmengen noch im kleinen Bereich bewegen. Im Folgenden wird der Aufbau und die Funktionsweise dieses Modells genauer erklärt.

Der erste Schritt für dieses Modell ist die Datenvorbehandlung. Hierzu gehört bei Fließtexten zum Beispiel, Wörter auf ihren Wortstamm zu reduzieren („ging“ zu „gehen“), oder Interpunktion und Zahlen zu entfernen (Hu et al., 2017, S. 1038). Dieser Fall fällt bei Schlagwörtern allerdings weg. Stattdessen werden hier nur Schlagwörter nur „tokenisiert“, das heißt, jedem einzigartigen Wort wird eine einzigartige Identifikationsnummer zugewiesen, welche einer ganzen Zahl entspricht. Das Ergebnis dieses Schrittes ist ein Korpus, welches aus vielen verschiedenen Listen von Zahlen besteht, wobei jede Liste den Schlagwörtern eines bestimmten Datensets entspricht.

Im zweiten Schritt wird das Neuronale Netz erstellt. Dieses ist im Vergleich zu neuronalen Netzen aus vorherigen Modelle vergleichsweise einfach, mit nur drei Ebenen abgesehen von der Input-Ebene (s. Abbildung 1). Die Input-Ebene besteht aus so vielen Neuronen, wie der Wortschatz groß ist. In der zweiten Ebene wurde dieser Input in einen n-dimensionalen Vektor verwandelt, damit sind hier auch genau n Neuronen vorhanden. In der Output-Ebene schließlich, welche wieder die Größe des Korpus hat, wird jedes Neuron mit einem Wert zwischen 0 und 1 belegt, welche addiert genau 1 ergeben. Dies spiegelt die Wahrscheinlichkeit wider, welches der Worte als Output bestimmt werden soll.

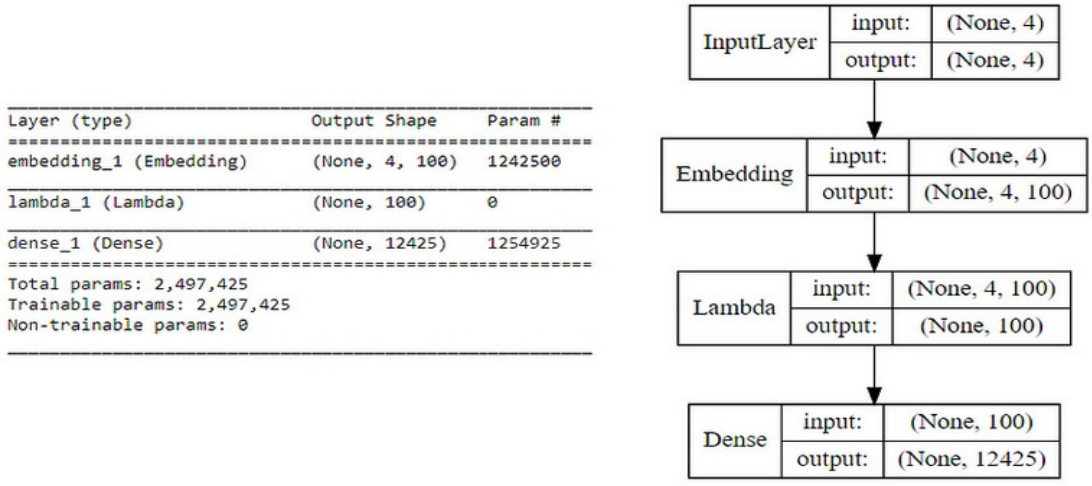


*Abbildung 1: Aufbau eines CBOW-Modells mit 100 Dimensionen und einer Wortschatzgröße von 12.425 Wörtern (Sarkar, 2018)*

Für das Training des Modells wird das gesamte Modell benötigt. Die Informationen für die Vektoren der Worte stecken jedoch in den Gewichtungen der Embedding-Ebene. Die n

Gewichtungen zwischen einem der Inputneuronen (welche dem Wort mit der zugehörigen ID entsprechen) und den n Neuronen der Embedding-Ebene entsprechen genau dem n-dimensionalen Vektor dieses Wortes.
Die veränderbaren Parameter dieses Neuronalen Netzes sind die Fenstergröße, welche oben schon erklärt wurde, die Größe des Wortschatzes (welches entweder auf die Größe des Korpus' oder, für größere Korpora, auf einen darunter liegenden Wert festgelegt wird) und die Vektorengröße, was der Anzahl der Dimensionen entspricht, innerhalb derer die Wörter verortet werden.
Zum Trainieren des Netzes werden Kontext-Wort-Paare verwendet. Diese Paare bestehen aus dem zu betrachtenden Wort und seinem Kontext, also den davor sowie danach liegenden Wörtern, wobei die Fenstergröße deren Anzahl bestimmt. Diese Daten werden in das Netz eingespeist, und mittels Rückpropagierung werden die Vektoren darauf trainiert, nahe an den Vektoren der ihnen nahestehenden Wörter zu sein. Trainiert wird in sogenannten Epochen. Je mehr Epochen, desto besser werden grundsätzlich die Vektoren.
Das Ziel dieser Arbeit ist also, die Datensätze sowie die Schlagworte auf dem deutschem Metadatenportal GovData zu untersuchen. Hierzu verwende ich das CBOW-Modell, um Word Embeddings zu erhalten, welche eine computergestützte Auswertung der Schlagworte ermöglicht. Das Ziel ist, die Praxis der Schlagwortvergabe genauer zu untersuchen. Das Ergebnis dieser Arbeit – ein fertig trainiertes CBOW-Modell – soll sowohl einen Einblick in diese Praxis geben, auf dessen Basis die Schlagwörter evaluiert werden können sowie anschließende Arbeiten ermöglichen wie ein Clustering der Dokumente anhand der Verortung ihrer Schlagwörter. Dementsprechend lauten meine Forschungsfragen:

RQ: Was lässt sich über die Praktik der Schlagwortvergabe auf GovData feststellen?

# 3 Die Methodik

Zur Umsetzung dieser Studie habe ich ein Programm mit Python geschrieben. Dieses ruft sowohl die Daten über den SPARQL-Endpunkt ab, erstellt und trainiert das Neuronale Netz sowie führt anschließende Analysen, insbesondere mit dem Density-Based Spatial Clustering of Applications with Noise (DBSCAN) Algorithmus. Im Folgenden werden die Funktionen dieses Programms erklärt. Zusätzlich zu den hier beschriebenen Abläufen habe ich zudem nach Abschluss der einzelnen Schritte das Abspeichern oder wahlweise Abrufen der Informationen aus vorhandenen Dateien eingebaut, auf diesen Code werde ich hier jedoch nicht weiter eingehen.

## 3.1 Die Datenabfrage

Die Datenabfrage habe ich mit dem von GovData bereitgestellten SPARQL-Endpunkt durchgeführt, der weiter oben schon beschrieben wurde. Die Funktion „sparql_query“ führt diese Anfrage durch. Mittels des Moduls parse der urllib-Bibliothek werden die Sonderzeichen, die der String mit der Anfrage enthält, durch %xx-Folgen ersetzt, um die Kompatibilität zu gewährleisten (Python 3.12.0 Documentation, 2023). Die requests-Bibliothek dient dazu, die Anfrage an den Server zu stellen und eine Antwort ausgeben zu lassen (Requests 2.31.0 Documentation, 2023). Die Daten werden schließlich in ein Dataframe eingelesen, welches mit der pandas-Bibliothek (Pandas 2.1.1 Documentation, 2023) erstellt wird und sich für die Bearbeitung großer tabellarischer Datenmengen eignet.

Um die Daten weiter verarbeiten zu können, wandelt die Funktion „convert_df_to_lists“ dieses Dataframe in drei Listen um: Eine jeweils für die abgefragten Identifier, die Titel und die Schlagwörter der Datensätze. Die Schlagwörter werden hierbei so gespeichert, wie sie später effizient tokenisiert werden können, also in Listen aus Wörtern, mit Leerzeichen getrennt. Einige Schlagwörter enthalten auch selbst Leerzeichen. Um spätere Probleme bei der Tokenisierung zu vermeiden, werden in diesem Schritt die Leerzeichen innerhalb der Schlagwörter durch Unterstriche ersetzt. Die Schlagworte eines Datensatzes mit den Schlagworten „Stadt Aachen“ und „Finanzen“ würden danach „Stadt-Aachen Finanzen“ lauten.

## 3.2 Das CBOW-Modell

Das zu dem CBOW-Modell gehörige Neuronale Netz wird vor allem mit der Bibliothek keras erstellt (Keras Documentation, 2023), welche weitreichende Funktionen zu Neuronalen Netzen beinhaltet. Das Preprocessing-Modul erlaubt mit der Tokenizer-Klasse eine einfach Tokenisierung des Korpus. Gleichzeitig hiermit erfolgt die Umwandlung aller Großbuchstaben und Kleinbuchstaben. Nach der Tokenisierung werden in den Variablen word2id und id2word jeweils ein Dictionary abgespeichert. Word2id hält die Schlagwörter und ihre zugehörigen Zahlen als Key:Value-Paare, in id2word sind diese andersrum abgespeichert. Das dient der späteren besseren Durchsuchbarkeit.

Die Funktion „build_cbow_model“ baut das Neuronale Netz, welches anschließend für das Training des CBOW-Modells verantwortlich ist. Die Wortschatzgröße wird auf die Menge der einzigartigen Schlagwörter um eins erhöht gesetzt. Da die Länge eines Arrays bei 0 angefangen wird zu zählen, muss hier eins dazu addiert werden. Die Vektorengröße wird auf zwei festgelegt, um sie später grafisch darstellen zu können. Die Fenstergröße schließlich wird auf 3 festgelegt, eine häufig genutzte Fenstergröße.

In der Funktion „train_cbow“ wird das Modell trainiert. Diese Funktion nimmt für jedes Wort im gesamten Korpus seinen Kontext und speist diesen in das Modell ein. Das Ergebnis wird verglichen mit dem Wunschergebnis und der Fehler rückpropagiert, um die Gewichtungen zu ändern und damit langsam die Vektoren sich annähern zu lassen. Die Wort-Kontext-Paare, die hierfür benötigt werden, werden in der Funktion „generate_context_word_pairs“ erstellt. Diese iteriert über den gesamten Korpus und gibt für jedes Wort ein Tupel aus, welches aus den Kontextwörtern und einem Vektor der zu trainierenden Wortes besteht, welches in das Neuronale Netz eingespeist wird. Jede Iteration dieses Training iteriert wiederum über jedes Wort im Korpus, also über rund 1.040.000 Wörter.

## 3.3 DBSCAN-Clustering

Auf Basis der mit dem CBOW-Modell erzielten Vektoren habe ich die so ermittelten zweidimensionalen Vektoren mittels DBSCAN-Clustering untersucht. Bei diesem Verfahren werden Datenpunkte in einem Cluster beziehungsweise einer Gruppe zusammengefasst, die in einer genügend großen Menge nah beieinander liegen. Der Bereich um einen Datenpunkt, der dabei betrachtet wird, wird mit epsilon bestimmt, während die Mindestanzahl an Datenpunkten, die in diesem Bereich liegen muss, mit minPts festgelegt wird. Datenpunkte, die keine ausreichende Nähe zu genügend anderen Punkten haben, werden als „Noise“ festgelegt und keinem Cluster zugeordnet. Auf diese Weise muss der DBSCAN-Algorithmus keine festgelegte Anzahl an Clustern vorgegeben haben, was ideal für die Analyse einer unbekannten Datenmenge ist.

Um ein möglichst ideales epsilon zu ermitteln, wird in der Funktion „get_epsilon“ mithilfe der sklearn-Bibliothek ein k-Nearest-Neighbour-Algorithmus implementiert. Dieser berechnet für jeden Datenpunkt seine k nächsten Nachbarn, in diesem Fall seine 4 nächsten Nachbarn. Diese Distanzen werden aufsteigend sortiert und anschließend in einem Graph ausgegeben. An der Stelle, an der die y-Werte dieses Graphen sprunghaft ansteigen, ist epsilon anzusetzen. Für diesen Datensatz wurde ein epsilon-Wert von etwa 0,3 festgestellt. Dieser Wert wird anfangs für die DBSCAN-Analysen genutzt, ändert sich jedoch, wenn bestimmte Cluster intensiver betrachtet werden. MinPts steht bei 5.

Das DBSCAN-Verfahren wird in der Funktion „cluster_dbscan“ implementiert. Diese Funktion arbeitet mit der DBSCAN-Klasse des sklearn-Moduls „cluster“ (Scikit-learn 1.3.1 Documentation, 2023). Diese Klasse ermöglicht in nur wenigen Zeilen die Implementierung des Algorithmus. Im Anschluss an das DBSCAN-Verfahren werden dessen Ergebnisse mithilfe von plotly, einer Bibliothek zur Visualisierung von großen Datenmengen, dargestellt (Plotly Python Graphing Library, 2023). Plotly eignet sich für diese Art von Daten, da es sowohl interaktiv ist als auch große

Datenmengen schnell bearbeiten kann. Dies führt die Funktion „visualize_cluster“ durch. In der „cluster_information“-Funktion werden zusätzlich einige Informationen über das Cluster ausgegeben, wie die Menge aller verarbeiteten Datenpunkte, die Menge der Punkte, die als Noise klassifiziert wurden, die Anzahl sowie die Länge der Cluster sowie eine Auswertung der Qualität des Clusters. Hierzu wird der Silhouettenkoeffizient genutzt, welcher die Clusterzugehörigkeiten der Datenpunkte untersucht. Der Wert kann von -1 bis zu 1 gehen. -1 bedeutet, dass die Cluster nahe beisammen sind und möglicherweise Datenpunkte falsch zugeordnet wurden. Ein Koeffizient von 0 deutet darauf hin, dass der Abstand zwischen den Clustern insignifikant ist, und ein Koeffizient von 1 deutet auf deutlich auseinander liegende Cluster und korrekte Zuordnungen (Rousseeuw, 1987, S. 46).

# 4 Ergebnisse

Ich habe am 20.09.2023 die Datenabfrage durchgeführt. Es wurden nur Datensets heruntergeladen, die Titel, Schlagwörter sowie einen Identifier – zur späteren Zuordnung – beinhalten. Insgesamt wurden 83.254 Datensets heruntergeladen, die 50.205 einzigartige und insgesamt 1.040.725 Schlagwörter enthalten. Das CBOW-Modell wurde mit diesen Daten in 10 Epochen trainiert. Jede Epoche dauerte rund 2,5 Stunden. In jeder Epoche wurde auf den gesamten Datensatz trainiert. Insgesamt dauerte das Training also rund über 25 Stunden.

Zur Evaluation des Modells werden oft Wörter ausgewählt, deren nächste Nachbarn ausgegeben und das Ergebnis evaluiert. Zu diesem Zweck habe ich jeweils drei der meistgenutzten Schlagwörter, drei zufällige und drei Schlagwörter ausgewählt, die sehr ähnlich zu anderen Schlagwörtern sind. Das Schlagwort "ausfuhr: Wert" beispielsweise ähnelt sehr stark den Schlagwörtern "einfuhr: gewicht", "einfuhr: wert", "ausfuhr: gewicht", "ausfuhr: volumen", "einfuhr: volumen" und weiteren.

Die ausgewählten Wörter sowie ihre ausgegebenen nächsten Nachbarn lassen sich aus Abbildung 2 auslesen. Einige Wörter wurden recht passend zugeordnet, wie „luftverschmutzung“ zu „verkehr“, „getreide u.ä. erzeugnisse“ zu „natürliche düngemittel“, „zusammenstoß mit and.kfz das einbiegt oder kreuzt“ zu „zusammenstoß fahrzeug fußgänger“. Bei Schlagwörtern 1 bis 3, welche sehr häufig im Korpus vorkommen, erreicht das trainierte Modell eine eher schlechte Qualität. Einzig „luftverschmutzung“ scheint gut zu „verkehr“ zu passen. Die drei zufällig ausgewählten Schlagworte 4 bis 6 haben keine merklich bessere Qualität. Allein zu Schlagwort 4 und 6 passen „getreide u.ä. erzeugnisse“ beziehungsweise „zusammenstoß mit and.kfz das einbiegt oder kreuzt“. Die Schlagworte 7 bis 9 jedoch, welche danach ausgewählt wurden, wie ähnlich sie anderen Schlagworten in ihrer Machart sind, erzielen sehr gute Ergebnisse. Bis auf das Schlagwort

„veräußerte Fläche“ zu dem Schlagwort „ausfuhr: wert“ passen sie alle sehr gut zu ihren Bezugsworten. Daran lässt sich erkennen, dass die Stärke dieses CBOW-Modells in den spezifischen, nicht stark genutzten Schlagwörtern liegt. Gerade für die viel genutzten Schlagworte ist die Qualität eher fragwürdig.

| Nr. | Schlagwort | Nächste drei Schlagwörter |
|---|---|---|
| 1 | vermessung | bottrop krfr. stadt, westerwaldkreis, böblingen landkreis |
| 2 | verkehr | gesundheitsberichterstattung, stichtag, luftverschmutzung |
| 3 | umwelt | lebensmittel, hydrography, mdi-de |
| 4 | natürliche düngemittel | großh.m. keram.erzeugn. glaswaren u.reinigungsm., wirtschaftszweige, getreide u.ä. erzeugnisse |
| 5 | gieleroth | kassettenrecorder, wassertiefen, wintermenggetreide |
| 6 | zusammenstoß fahrzeug fußgänger | zusammenstoß mit and.kfz das einbiegt oder kreuzt, erholung, 600 - 700 |
| 7 | ausfuhr: wert | ausfuhr: wert, einfuhr: wert, veräußerte fläche |
| 8 | erfasste haushalte mit langlebigen gebrauchsgütern | hochgerechnete hh m. langl.gebrauchsgütern in 1000, erfasste haushalte mit langlebigen gebrauchsgütern, erfasste haushalte mit lebensmittelausgaben |
| 9 | drittstaaten zu eg-12 bis 31.12.1994 | eg-10 bis 31.12.1985, drittstaaten zu eu-15 bis 30.04.2004, drittstaaten zu eu-27 seit 01.02.2020 |

*Abbildung 2: Schlagwörter und ihre drei nächsten Nachbarn*

Im auf das Modell aufbauenden DBSCAN-Verfahren wurden nur Schlagwörter genutzt, die mindestens 5 Mal vorkommen, wodurch alleine schon 37.042 Schlagwörter wegfallen, also 74% der gesamten Schlagwörter. Die weiteren Berechnungen wurden auf den verbliebenen 26%, also 13.163 Schlagwörtern, durchgeführt. Das erste Clustering wurde mit epsilon = 0,3 und minPts = 5 durchgeführt. Die Wortvektoren sowie das Ergebnis des Clusterings lässt sich aus Abbildung 3 ablesen. Der Silhouettenkoeffizient dieses Clusterings liegt bei 0,56.

Von allen rund 13.000 Schlagwörtern wurden 1.030, also rund 8%, als Noise erkannt. Die restlichen Datenpunkte teilen sich in 15 Cluster auf, wobei Cluster 1 mit 11.980 Punkten, also rund 91% fast alle Datenpunkte beinhaltet. Dieses Muster zeigt sich auch bei einem niedrigeren Wert für epsilon. Fast immer beansprucht ein Cluster tausende Datenpunkte, während die umliegenden Cluster nur mehrere dutzend Datenpunkte beinhalten. Erst ab einem epsilon-Wert von 0,005 brechen diese Cluster auf. Zu diesem Zeitpunkt ist das größte Cluster nur noch knapp 500 Elemente groß, während es bei epsilon = 0.006 noch 1.473 Datenpunkte sind.

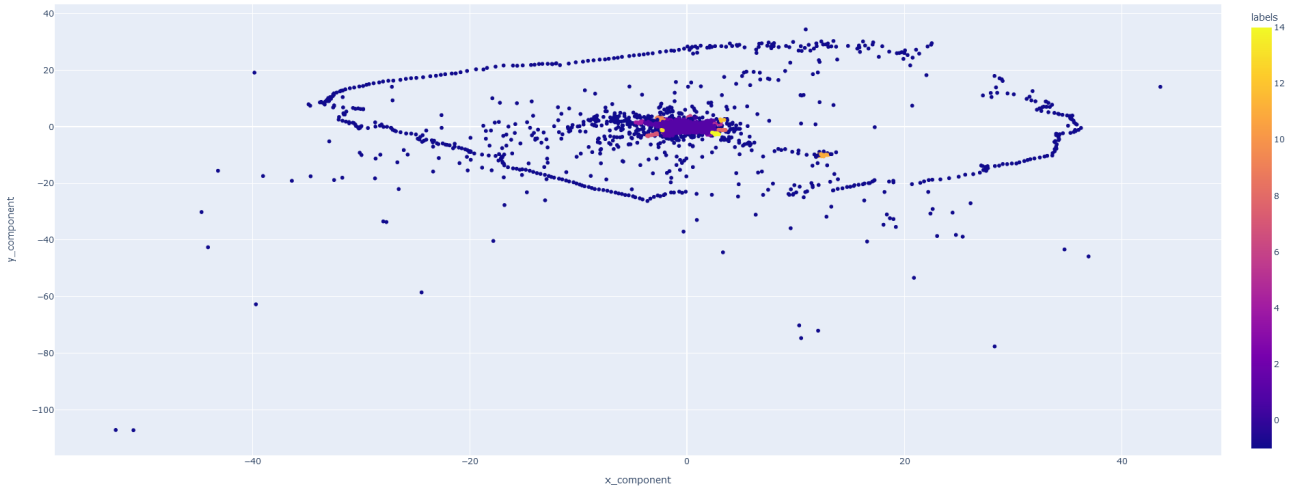


*Abbildung 3: DBSCAN-Clustering mit epsilon = 0,3 und minPts = 5*

Wird der Noise genauer betrachtet, lässt sich schnell erkennen, dass dieser vor allem aus den viel genutzten Schlagwörtern besteht. Von den häufigsten 200 Schlagwörtern werden nur 4 Stück nicht als Noise eingestuft.


*Abbildung 4: DBSCAN-Clustering mit epsilon = 0,3 und minPts = 5, ohne Noise*

Um die bestehenden Cluster näher zu untersuchen, zeigt Abbildung 4 das Clustering-Ergebnis ohne die Datenpunkte, die als Noise klassifiziert wurden. Auf einige Cluster soll hier genauer eingegangen werden. Welche Schlagwörter zu den hier genannten Clustern zugeordnet wurden, lässt sich in Anhang 1: „Clusterzugehörigkeiten von Cluster 0 und 7 bis 14“ einsehen. Cluster 7 und 11 beinhalten einige nahe Schlagwörter. So sind zu Cluster 7 acht Datenpunkte zugeordnet, die „Drittstaaten zu eg-12 bis 31.12.1994“ und ähnlich lauten. Ebenso existieren hier vier Schlagwörter, die mit „Gebiet der ehemaligen“ beginnen, beispielsweise „Gebiet der ehemaligen Sowjetunion“.

Auch in Cluster 14 und 8 lassen sich solche offensichtlich nahe beieinander liegenden Schlagworte erkennen. In allen diesen Clustern sind aber immer auch andere Datenpunkte zugeordnet, die nicht zu den anderen Schlagworten zu passen scheinen. Beispielsweise beinhaltet Cluster 7 auch Worte wie „Schifffahrt“ oder „Trinkwasserschutzgebiet“. Viele Cluster, darunter Cluster 0, 9, 10, 12 und 13, scheinen aber insgesamt kein kohärentes Thema zu finden.

Im Anschluss daran habe ich die Datenpunkte des größten Clusters, Cluster 1, weiter untersucht. Zu diesem Zweck wurde epsilon auf 0,15 gesetzt. Auch hier lässt sich wieder ein großes Cluster mit 11.541 Datenpunkten erkennen, das entspricht 96% der gesamten Datenpunkte. Die meisten der kleineren Cluster beinhalten maximal 11 Datenpunkte. Cluster 2, 8, 10 und 11 jedoch sind hier besonders hervorzuheben. Nicht nur enthält Cluster 2 sogar 47 Punkte, alle drei enthalten Länder als Schlagwörter: Thailand, Angola, Zypern, Philippinen, Türkei, Gambia, Kuba, Peru und viele weitere Länder sind hier aufgeführt. Fraglich ist, wieso Cluster 2, obwohl es genauso Länder enthält, sehr weit weg von Clustern 8, 10 und 11 liegt. Die gesamten Schlagwörter, die diesen vier Clustern zugeordnet werden, lassen sich in Anhang 2: „Clusterzugehörigkeiten von Cluster 2, 8, 10 und 11“ einsehen.

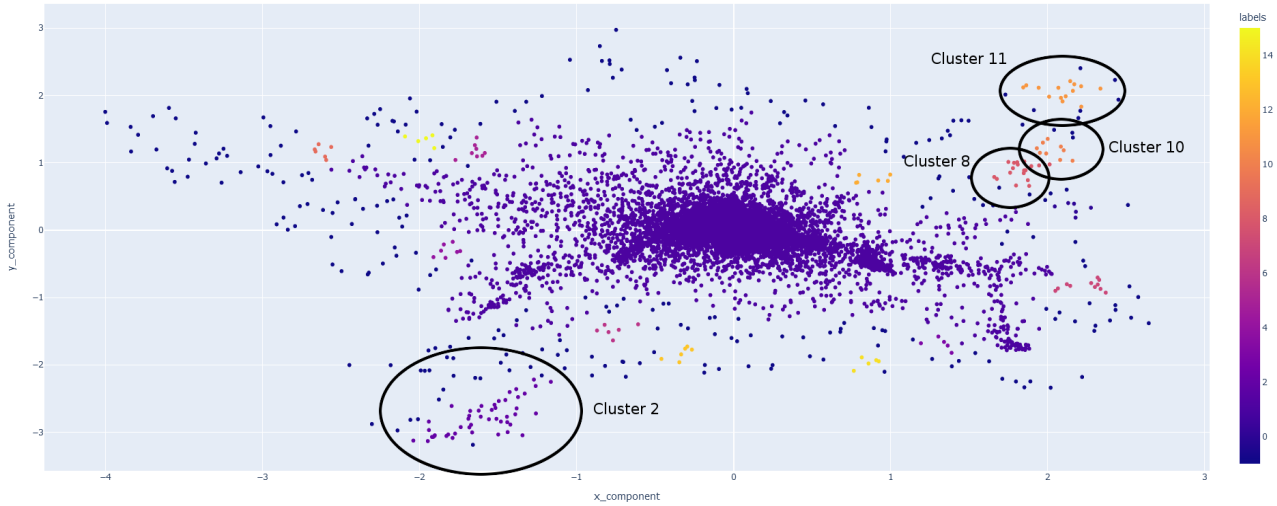


*Abbildung 5: DBSCAN-Clustering des großen initialen Clusters, epsilon = 0,15 und minPts = 5*

Je weiter epsilon hier verfeinert wird, desto mehr wird das große Cluster aufgebrochen. Wie im obigen Beispiel lassen sich immer wieder kleinere sinnvolle Cluster erkennen. Insbesondere in Clustern mit mehr als 10 Punkten lassen sich zusammenhängende Schlagworte erkennen. So gibt es beispielsweise bei epsilon = 0,05 ein Cluster, welches mindestens 20 Schlagworte in der Art von „Belgien ab 1999“ enthält.

Ein einheitliches epsilon zu nutzen, funktioniert also nicht für diesen Datensatz. Während es einige Datenpunkte gibt, die extrem weiter auseinander liegen – vor allem die am häufigsten genutzten –

konzentrieren sich die allermeisten Punkte um den Nullpunkt herum. Je näher am Nullpunkt die Vektoren liegen, desto kleiner sind die Abstände zwischen den Punkten.
Hieraus lassen sich einige Schlüsse für die Praxis der Schlagwortvergabe auf GovData ziehen. Viele der genutzten Schlagworte werden kaum wiederverwendet. Nur 26% aller Schlagwörter wurden überhaupt mehr als fünf Mal verwendet. Wahrscheinlich werden innerhalb einer öffentlichen Behörde immer sehr ähnliche Schlagwörter benutzt, die aber von der nächsten Behörde nicht wiederverwendet werden. Diese These wird durch eine weitere SPARQL-Anfrage unterstützt: Fast die Hälfte aller Schlagwörter wird nur von einer einzigen Institution genutzt. Nur 6.422 Schlagwörter wurden von mehr als drei Behörden genutzt. Das würde auch erklären, wieso sich eigentlich zusammengehörige Cluster (wie die Länderbezeichnungen weiter oben) an verschiedenen Orten im Raum befinden. Jede Institution vergibt, abgesehen von den viel genutzten Schlagwörtern, ihr ganz eigenes Set an Schlagwörtern, das von anderen Behörden kaum oder gar nicht genutzt wird. Ebenso sind starke Unterschiede in der Anzahl der Schlagworte zu erkennen. Die Anzahl der Schlagworte liegt zwischen 0 und 720.
Diese Unterschiede deutet darauf hin, dass je nach Behörde unterschiedliche Personen für die Metadaten- und damit auch die Schlagwortvergabe zuständig sind. Damit verbunden sind starke Unterschiede in der Praxis der Schlagwortvergabe, welche sich negativ auf die Auffindbarkeit und maschinelle Verarbeitung dieser Metadaten auswirken.

# 5 Fazit

Ich habe in dieser Arbeit mittels eines CBOW-Modells, welches auf einem Neuronalen Netz beruht, ein Word Embedding aller Schlagworte des Metadatenportals GovData vorgenommen. Diese Website veröffentlicht öffentliche Daten für Deutschland, die allermeisten davon unter einer offenen Lizenz. Das Modell sowie die darauf aufbauende Auswertung mittels DBSCAN sollte einen Eindruck von der Praxis der Schlagwortvergabe für GovData geben.
Das CBOW-Modell zeigt einige Schwierigkeiten. Insbesondere die viel genutzten Schlagworte werden nicht gut verortet. Die wenig genutzten und spezielleren Schlagworte erzielen dabei eine sehr viel bessere Qualität. Dementsprechend sind die Ergebnisse der darauf basierenden Clustering-Analyse nur begrenzt aussagekräftig.
In dieser Analyse hat sich ergeben, dass die Schlagwörter sehr unterschiedliche Dichten haben, was die Auswertung weiter erschwert. Es gibt einige wenige Schlagwörter, die sehr weit vom Zentrum entfernt sind, während der allergrößte Teil der Schlagwörter sich zwischen (-1, -1) und (1, 1) bewegt. Je näher die Punkte am Nullpunkt sind, desto enger sind sie verteilt. Während die kleineren Cluster mit maximal 10 Datenpunkten oft keinen Sinnzusammenhang erkennen lassen, enthalten

viele der größeren Cluster eindeutig zusammenhängende Schlagwörter wie Ländernamen. Dieses Muster zieht sich auch weiter, je geringer epsilon gewählt wird.

All das, und die Tatsache, dass ein großer Teil der Schlagworte nur von jeweils einer Behörde genutzt wird, zeigt, dass die für Metadaten verantwortlichen Personen sowohl von Behörde zu Behörde unterschiedlich sind, da sie keine einheitlichen Schlagwörter nutzen. Das verschlechtert die Auffindbarkeit von Datensätzen, welche ein wichtiges Kriterium und Ziel offener Daten ist.

Basierend auf diesen Ergebnissen können einige Verbesserungsvorschläge erbracht werden. Beispielsweise scheint es angebracht, die Schlagworte vorher intensiver vorzubereiten. So könnten besonders lange Schlagwörter aufgeteilt werden in ihre Bestandteile und Schlagworte, die keinen Sinn ergeben, weggelassen werden, wie das Schlagwort „05166“. Besonders wenig genutzte Schlagwörter könnten weggelassen werden, da die Vermutung besteht, dass gerade diese so speziell sind, dass sie für ein Clustering nicht förderlich sind und damit die Rechenzeit verkürzt werden kann.

Ebenso kann es sich lohnen, die Parameter des CBOW-Modells zu ändern, sowohl die Fenstergröße wie auch die Größe der Vektoren. Eine Überlegung wäre es, die Fenstergröße auf 719 festzulegen, um zu gewährleisten, dass ein Schlagwort immer mit allen anderen Schlagworten desselben Datensatzes trainiert wird. Auch eine getrennte Modellbildung von häufigen und seltenen Schlagwörtern und darauf aufbauende Analysen können sinnvoll sein, da diese Arbeit gezeigt hat, dass diese beiden Gruppen sehr unterschiedliche Abstände vorweisen.

Aufbauend auf dieser Arbeit kann eine thematische Sortierung der Datensätze entstehen. Die Dokumente bestehen dabei aus den addierten Vektoren ihrer Datensätze. Auf diese Weise lassen sich Dokument-Vektoren erstellen, deren Zusammenstellung im Raum idealerweise thematische Gruppierungen wiedergeben kann.

# Literaturverzeichnis


Abdelrahman, Omer Hassan (2022): Open Government Data: Development, Practice and Challenges. In Vijayalakshmi Kakulapati (Hrsg.): *Open Data*. London: IntechOpen, 59-72.

„Ausländer: Deutschland, Stichtag, Geschlecht“ (2018), *GovData*, https://www.govdata.de/web/guest/suchen/-/details/auslander-deutschland-stichtag-geschlechtaufenthaltstitel

„Ausländer: Kreise, Stichtag, Geschlecht“ (2018), *GovData*, https://www.govdata.de/web/guest/suchen/-/details/auslander-kreise-stichtag-geschlecht-aufenthaltstitellandergruppierungen

Curiel, Javier de Esteban/Antonovica, Arta/Morales, Maria del Rosario Sánchez (2023): Inductive open data study on teleworking dissatisfaction in Span during the Covid-19-pandemic. In: *International Journal of Manpower* ahead-of-print.

Fraunhofer-Institut (2019): Natural Language Processing. In: *Europäische Sicherheit und Technik*.

GovData (2023): *GovData: Das Datenportal für Deutschland*. https://www.govdata.de/

Hu, Kai/et al. (2017): A domain keyword analysis approach extending Term Frequency-Keyword Active Index with Google Word2Vec model. In: *Scientometrics* 114, 1031-1068.

Jemili, Farah/Bouras, Hajer (2022): Intrusion Detection Based on Big Data Fuzzy Analytics. In Vijayalakshmi Kakulapati (Hrsg.): *Open Data*. London: IntechOpen, 59-72.

Keras Documentation (2023): *Keras: Deep Learning library for Theano and TensorFlow*. https://faroit.com/keras-docs/1.2.0/

Mikolov, Tomas/Chen, Kai/Corrado, Greg/Dean, Jeffrey (2013): Efficient Estimation of Word Representations in Vector Space. In: Proceedings of Workshop at ICLR. https://browse.arxiv.org/pdf/1301.3781.pdf

Open Knowledge Foundation Deutschland (2023): *Offene Daten*. https://okfn.de/themen/open_data/

Pandas 2.1.1 Documentation (2023): *Pandas documentation*. https://pandas.pydata.org/docs/

Plotly Python Graphing Library (2023): *Plotly Open Source Graphing Library for Python*. https://plotly.com/python/

Python 3.12.0 Documentation (2023): *Urllib.parse – Parse URLs into components*. https://docs.python.org/3/library/urllib.parse.html

Requests 2.31.0 Documentation (2023): *Requests: HTTP for Humans*. https://requests.readthedocs.io/en/latest/

Rizkalla, Sandra/Atiya, Amir F./Shaheen, Samir (2022): A Polarity Capturing Sphere for Word to Vector Representation. In Massimo Esposito/Giovanni Luca Masala/Aniello Minutolo/Marco Pota (Hrsg.): *Natural Language Processing: Emerging Neural Approaches and Applications*. Basel: MDPI, 233-254.

Rousseeuw, Peter J. (1987): Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. In: *Journal of Computational and Applied Mathematics* 20, 53-65.

Sarkar, Dipanjan (2018): *Implementing Deep Learning Methods and Feature Engineering for Text Data: The Continuous Bag of Words (CBOW)*. https://www.kdnuggets.com/2018/04/implementing-deep-learning-methods-feature-engineering-text-data-cbow.html

Scikit-learn 1.3.1 Documentation (2023): *Sklearn.cluster.DBSCAN*. https://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html

„Städtische Beteiligungen“ (2023), *GovData*, https://www.govdata.de/web/guest/suchen/-/details/stadtische-beteiligungen-zuschusse-an-stadtische-tochterunternehmen

Von Lucke, Jörn/Gollasch, Katja (2022): *Open Government: Offenes Regierungs- und Verwaltungshandeln – Leitbilder, Ziele und Methoden*. Wiesbaden: Springer Gabler.

W3C (2009): *Publishing Open Data*. https://www.w3.org/TR/gov-data/

Wikipedia (2023): *Resource Description Framework*. https://de.wikipedia.org/wiki/Resource_Description_Framework

# Anhang

*Anhang 1: Clusterzugehörigkeiten von Cluster 0 und 7 bis 14*

| **Schlagwort** | **Label** |
|---|---|
| niederlande | 14 |
| italien | 14 |
| veräußerte_fläche | 14 |
| weißenthurm | 14 |
| nichtdeutsch | 14 |
| st | 14 |
| bundesstraßen | 14 |
| geschichte | 14 |
| bauwerk | 14 |
| ausfuhr:_wert | 14 |
| einfuhr:_gewicht | 14 |
| einfuhr:_wert | 14 |
| human_health_and_safety | 14 |
| plis | 13 |
| israel | 13 |
| interessenvertr._kirchl._u._sonst._vereinigungen | 13 |
| serbien_einschl._kosovo__03.06.2006-16.02.2008 | 13 |
| präkambrium | 13 |
| llur | 12 |
| grundsicherung_im_alter_und_bei_erwerbsminderung | 12 |
| jura | 12 |
| personalstandstatistik_des_öffentlichen_dienstes | 12 |
| werbung_und_marktforschung | 12 |
| petrographie | 11 |
| niedersächsisches_küstenmeer | 11 |
| geologische_karte | 11 |
| metallerzeugung_und_-bearbeitung | 11 |
| bodengesellschaft | 11 |
| effektive_durchwurzelungstiefe | 11 |
| bewirtschaftungsgebiete/schutzgebiete/ geregelte_gebiete_und_berichterstattungseinheiten | 11 |
| bodenhorizont | 11 |
| bodenfeuchtigkeit | 11 |
| tertiär | 11 |
| bodenmineralogie | 11 |
| bodengestaltung | 11 |
| bodenchemie | 11 |
| fahrrad | 11 |
| bodendekontamination | 11 |
| krankenhäuser | 11 |
| webanalyse | 10 |
| abwasser | 10 |
| handwerksordnung | 10 |
| remote_sensing | 10 |
| sn | 10 |

| | |
|---|---|
| a7 | 9 |
| adv | 9 |
| schule | 9 |
| msrl | 9 |
| empfänger_von_grundsicherung | 9 |
| malta | 9 |
| seeverkehr_des_hafens_hamburg | 9 |
| grundwasserkoerper | 9 |
| bevölkerungsstand-nach-altersgruppen | 8 |
| biologie | 8 |
| erfasste_haushalte_mit_langlebigen_gebrauchsgütern | 8 |
| erfasste_haushalte_mit_immobilienvermögen | 8 |
| erfasste_haushalte_mit_lebensmittelausgaben | 8 |
| hochgerechnete_haushalte_in_1000 | 8 |
| hochgerechnete_hh_m._lebensmittelausgaben_in_1000 | 8 |
| hochgerechnete_hh_m._langl.gebrauchsgütern_in_1000 | 8 |
| erfasste_haushalte_mit_einnahmen_und_ausgaben | 8 |
| miesmuschelkulturen | 8 |
| betriebe | 7 |
| verkehr_und_lagerei | 7 |
| schifffahrt | 7 |
| b432 | 7 |
| trinkwasserschutzgebiet | 7 |
| a24 | 7 |
| ederfugl | 7 |
| drittstaaten_zu_eg-12_bis_31.12.1994 | 7 |
| drittstaaten_zu_eg-9_bis_31.12.1980 | 7 |
| drittstaaten_zu_eu-15_bis_30.04.2004 | 7 |
| drittstaaten_zu_eu-25_bis_31.12.2006 | 7 |
| drittstaaten_zu_eu-27_bis_30.06.2013 | 7 |
| drittstaaten_zu_eu-27_seit_01.02.2020 | 7 |
| drittstaaten_zu_eu-28_bis_31.01.2020 | 7 |
| drittstaaten_zu_ewg-6_bis_31.12.1972 | 7 |
| eg-10_bis_31.12.1985 | 7 |
| eg-12_bis_31.12.1994 | 7 |
| eg-9_bis_31.12.1980 | 7 |
| eu-15_bis_30.04.2004 | 7 |
| eu-25_bis_31.12.2006 | 7 |
| eu-27_bis_30.06.2013 | 7 |
| eu-27_seit_01.02.2020 | 7 |
| eu-28_bis_31.01.2020 | 7 |
| ewg-6_bis_31.12.1972 | 7 |
| gastarbeiterländer | 7 |
| gebiet_der_ehemaligen_sowjetunion | 7 |
| gebiet_der_ehemaligen_tschechoslowakei | 7 |
| gebiet_des_ehemaligen_jugoslawien | 7 |
| gebiet_des_ehemaligen_serbien_und_montenegro | 7 |
| ländergruppierungen | 7 |
| nordafrika | 7 |
| ost-_und_zentralasien | 7 |
| ostafrika | 7 |
| süd-_und_südostasien | 7 |

| | |
|---|---|
| vereinigtes_königreich_einschl.brit.überseegebiete | 7 |
| vorderasien | 7 |
| mfund-projekt_limbo | 7 |
| umwelt_&_klima | 7 |
| wasser | 0 |
| salzmarschen | 0 |
| bodenmechanik | 0 |
| sonstige_wirtschaftliche_dienstleistungen | 0 |
| grünstrom | 0 |

*Anhang 2: Clusterzugehörigkeiten von Cluster 2, 8, 10 und 11*

| **Schlagwort** | **Label** |
|---|---|
| abwasserentsorgung | 11 |
| australien_und_ozeanien | 11 |
| thailand | 11 |
| angola | 11 |
| argentinien | 11 |
| indonesien | 11 |
| benin | 11 |
| ecuador | 11 |
| eswatini | 11 |
| bahrain | 11 |
| belize | 11 |
| botsuana | 11 |
| barbados | 11 |
| brunei_darussalam | 11 |
| zypern | 10 |
| dominikanische_republik | 10 |
| jemen | 10 |
| drittstaaten_zu_eg-10_bis_31.12.1985 | 10 |
| mauritius | 10 |
| mosambik | 10 |
| guatemala | 10 |
| grenada | 10 |
| heime_ohne_erholungs-_und_ferienheime | 10 |
| bevölkerungsstatistik | 8 |
| türkei | 8 |
| china | 8 |
| eisrandlagen | 8 |
| philippinen | 8 |
| togo | 8 |
| gambia | 8 |
| marokko | 8 |
| nepal | 8 |
| niger | 8 |
| saudi-arabien | 8 |
| gabun | 8 |
| ruanda | 8 |
| el_salvador | 8 |
| fidschi | 8 |

| | |
|---|---|
| oman | 8 |
| nicaragua | 8 |
| st._vincent_und_die_grenadinen | 8 |
| planung | 2 |
| südafrika | 2 |
| salzgitter | 2 |
| gifhorn | 2 |
| lanuv | 2 |
| ausgaben | 2 |
| geologie_und_geobasisdaten | 2 |
| landau | 2 |
| produzierendes_gewerbe_ohne_baugewerbe | 2 |
| rp | 2 |
| vereinigte_arabische_emirate | 2 |
| originalwerte | 2 |
| rechtsform | 2 |
| jordanien | 2 |
| kamerun | 2 |
| neuseeland | 2 |
| vietnam | 2 |
| mexiko | 2 |
| ernährungsgewerbe_und_tabakverarbeitung | 2 |
| kuba | 2 |
| kuwait | 2 |
| peru | 2 |
| singapur | 2 |
| estland | 2 |
| sierra_leone | 2 |
| sri_lanka | 2 |
| grunddaten_der_krankenhäuser | 2 |
| katar | 2 |
| kolumbien | 2 |
| komoren | 2 |
| lesotho | 2 |
| seychellen | 2 |
| simbabwe | 2 |
| suriname | 2 |
| tonga | 2 |
| trinidad_und_tobago | 2 |
| tschad | 2 |
| vanuatu | 2 |
| bosnien_und_herzegowina | 2 |
| panama | 2 |
| sao_tome_und_principe | 2 |
| st._kitts_und_nevis | 2 |
| st._lucia | 2 |
| tuvalu | 2 |
| personengesellschaften | 2 |
| wohnort | 2 |
| arbeiter | 2 |