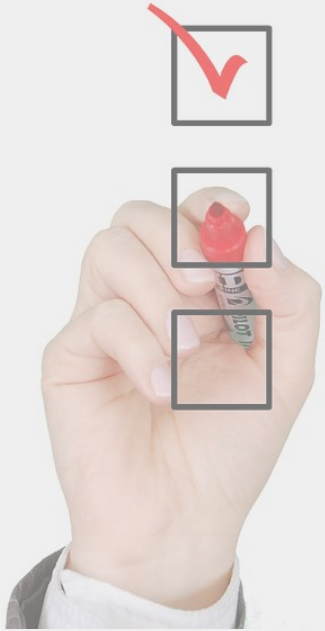
Data Mining mit Twitter

Entwicklung eines prototypischen Python Notebooks zur exemplarischen Extraktion und Auswertung von Daten des Social-Media-Dienstes Twitter anhand von Suchbegriffen oder einem Nutzernamen unter Nutzung der von Twitter zur Verfügung gestellten APIs, allgemein zugänglicher Bibliotheken sowie der Berücksichtigung des Open Source Frameworks TensorFlow

University of Applied Sciences
HOCHSCHULE EMDEN·LEER
Fachbereich Technik
Abteilung Elektrotechnik und Informatik

Gliederung

- Motivation - Ziel - Anforderung
- Grundlagen
- Anwendung
- Zusammenfassung und Ausblick
- Fragerunde und Vorführung



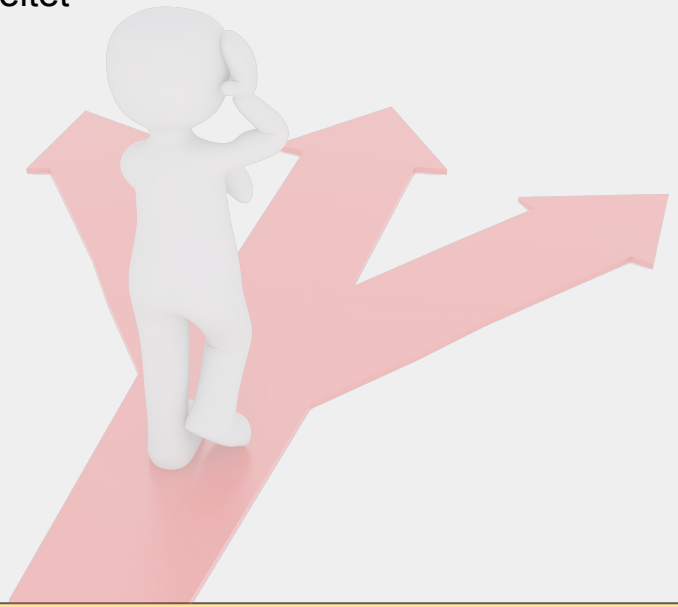
University of Applied Sciences
HOCHSCHULE EMDEN·LEER
Fachbereich Technik
Abteilung Elektrotechnik und Informatik

Motivation

Warum diese Anwendung?

- künstliche Intelligenz und maschinelles Lernen weit verbreitet
- nur wenige Einführungen in deutscher Sprache
- kaum grundlegende Beispiele

⇒ hohe Hürde beim ersten Kontakt für Neulinge



Ziel

Was soll erreicht werden?

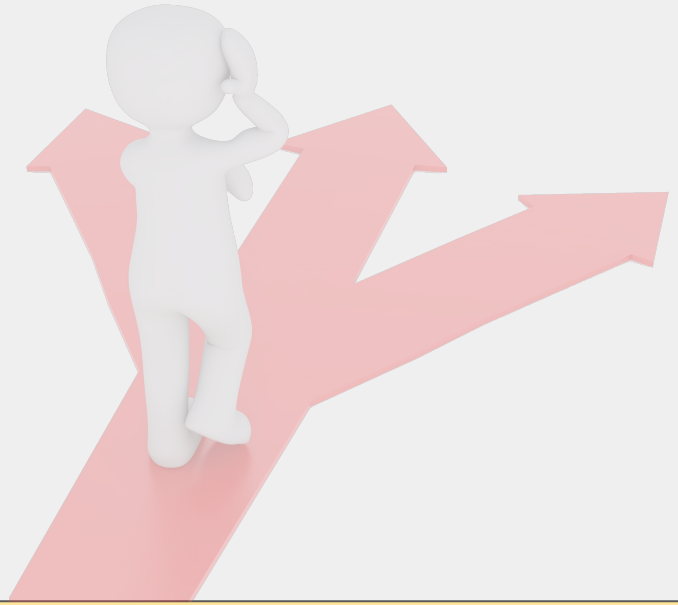
- ausführlich begleitetes Beispiel
- Umsetzung einer konkreten Anwendung
- Installation, Importe, Modellerstellung
- Fokus auf einfachen Code
- wenige Hürden für die eigene Nutzung
- interaktive Möglichkeiten zur eigenen Bearbeitung

Anforderungen

Wie kann dies umgesetzt werden?

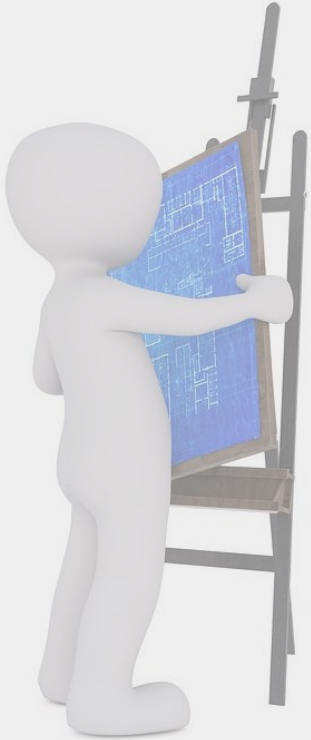
- Umsetzung als frei verfügbares Notebook
- Beschränkung auf einen grundlegenden Workflow
- einfache semantische Analyse
- grundsätzliche Auswertungen anhand von Suchwörtern/Nutzern
- einfache Datenhaltung in Listen und Wörterbüchern

⇒ nicht sinnvoll für den produktiven Einsatz



University of Applied Sciences
HOCHSCHULE EMDEN·LEER
Fachbereich Technik
Abteilung Elektrotechnik und Informatik

Grundlagen

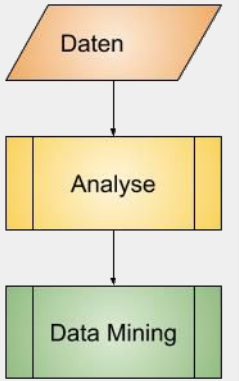


- vom Data Mining zum Social Web Mining
- Twitter
- TensorFlow
- Python

University of Applied Sciences
HOCHSCHULE EMDEN·LEER
Fachbereich Technik
Abteilung Elektrotechnik und Informatik

Grundlagen - vom Data Mining zum Social Web Mining

Data Mining

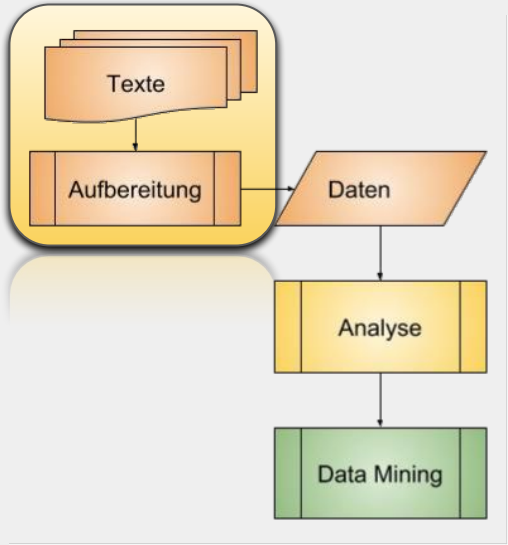


- einer Analyse zugängliche Daten
- Analyse mit Algorithmen
- Data Mining
  - ⇒ Klassifizierung
  - ⇒ Clusterbildung
  - ⇒ ...

University of Applied Sciences
HOCHSCHULE EMDEN·LEER
Fachbereich Technik
Abteilung Elektrotechnik und Informatik

Grundlagen - vom Data Mining zum Social Web Mining

Text Mining

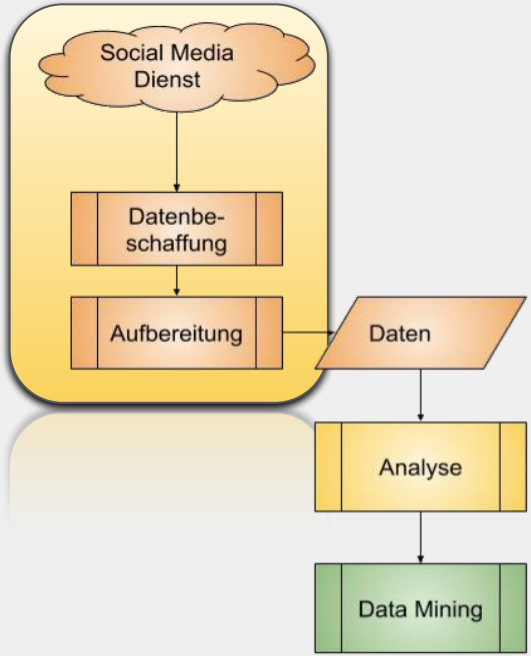


- Umwandlung der Texte in Daten
- einer Analyse zugängliche Daten
- Analyse mit Algorithmen
- Data Mining
  - ⇒ Klassifizierung
  - ⇒ Clusterbildung
  - ⇒ ...

University of Applied Sciences
HOCHSCHULE EMDEN·LEER
Fachbereich Technik
Abteilung Elektrotechnik und Informatik

Grundlagen - vom Data Mining zum Social Web Mining

Social Web Mining



- Ziehen von Inhalten als Text
- Umwandlung der Texte in Daten
- einer Analyse zugängliche Daten
- Analyse mit Algorithmen
- Data Mining
  - ⇒ Klassifizierung
  - ⇒ Clusterbildung
  - ⇒ ...

University of Applied Sciences
HOCHSCHULE EMDEN·LEER
Fachbereich Technik
Abteilung Elektrotechnik und Informatik

Grundlagen - Twitter

Fakten

- April 2016
- Nachrichten als kurze Texte (Tweets)
- User (@) vs. Themen (#)
- Retweet vs. Quote Tweet
- follow vs. follower

Grundlagen - Twitter

Einfluss

- rund 335 Millionen Nutzer
- 68% aus den USA
- 11% Zuwachs in 2018
- 40 Sprachen

Grundlagen - TensorFlow

Fakten

- von Google entwickelt
- Framework zur Entwicklung und Betrieb von Deep Learning Netzwerken
- seit 2015 Open Source

Grundlagen - TensorFlow

Arbeitsweise

- Berechnungsgraph definiert alle Berechnungsschritte
- Optimierung des Graphen vor der Ausführung
- Daten durchlaufen in Form von Tensoren den Graphen (“fließen”)

Grundlagen - Python

Fakten

- seit Anfang 1990er
- Objektorientiert aufgebaut
- Multiparadigmensprache
- Interpreter basierte Sprache
- Windows - Linux - iOS

Grundlagen - Python

Vorteile

- einfacher Aufbau
- dynamische Typisierung
- zahlreiche Erweiterungen
- weite Verbreitung im wissenschaftlichen Umfeld

Grundlagen - Python

Notebook

- baut auf den interaktiven Fähigkeiten auf
- Ausführung über Webprotokoll im Browser
- textbasiert und plattformunabhängig
- fasst Dokumentation und Code zusammen
- Distribution als Textdatei

Grundlagen - Python

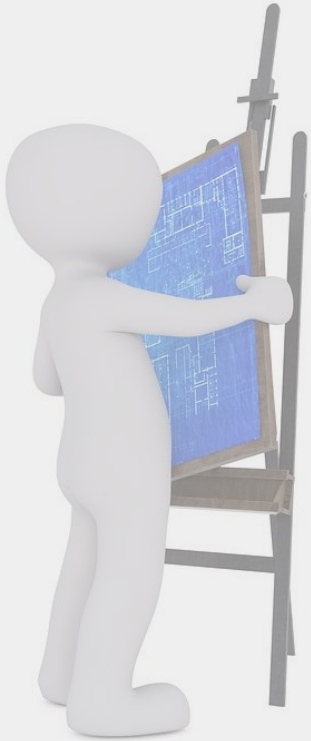
Notebook

- baut auf den interaktiven Fähigkeiten auf
- Ausführung über Webprotokoll im Browser
- textbasiert und plattformunabhängig
- fasst Dokumentation und Code zusammen
- Distribution als Textdatei

Ideal für die einfache Verteilung, Ausführung und Dokumentation

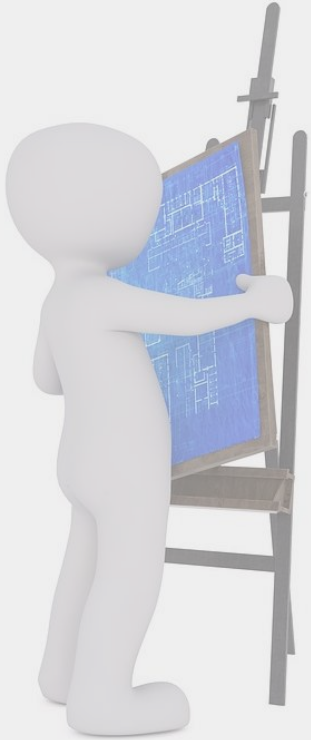
University of Applied Sciences
HOCHSCHULE EMDEN·LEER
Fachbereich Technik
Abteilung Elektrotechnik und Informatik

Anwendung



University of Applied Sciences
HOCHSCHULE EMDEN·LEER
Fachbereich Technik
Abteilung Elektrotechnik und Informatik

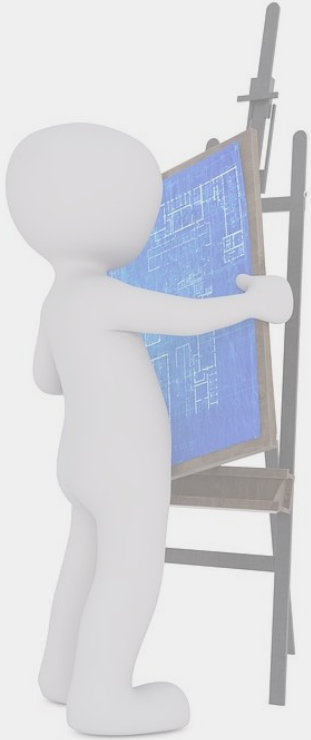
Anwendung



- Ziehen von Texten und Umwandlung in Daten
- semantische Analyse
- Auswertung
- verfügbar machen

University of Applied Sciences
HOCHSCHULE EMDEN·LEER
Fachbereich Technik
Abteilung Elektrotechnik und Informatik

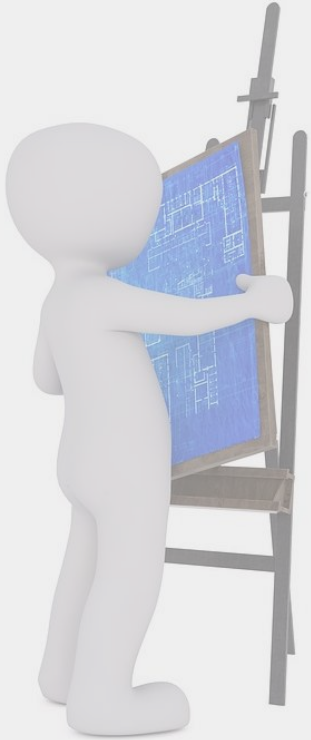
Anwendung - aus Texte Daten generieren



Ziehen von Texten und Umwandlung in Daten

University of Applied Sciences
HOCHSCHULE EMDEN·LEER
Fachbereich Technik
Abteilung Elektrotechnik und Informatik

Anwendung - aus Texte Daten generieren

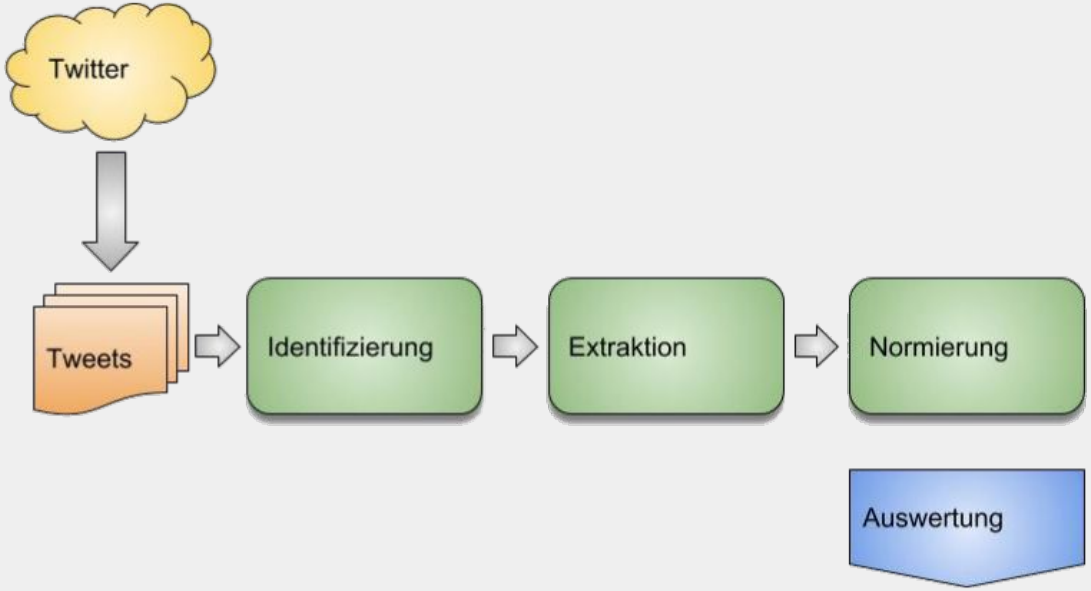


Ziehen von Texten und Umwandlung in Daten

Auswertung mit Methoden des Data Mining

University of Applied Sciences
HOCHSCHULE EMDEN·LEER
Fachbereich Technik
Abteilung Elektrotechnik und Informatik

Anwendung - aus Texte Daten generieren

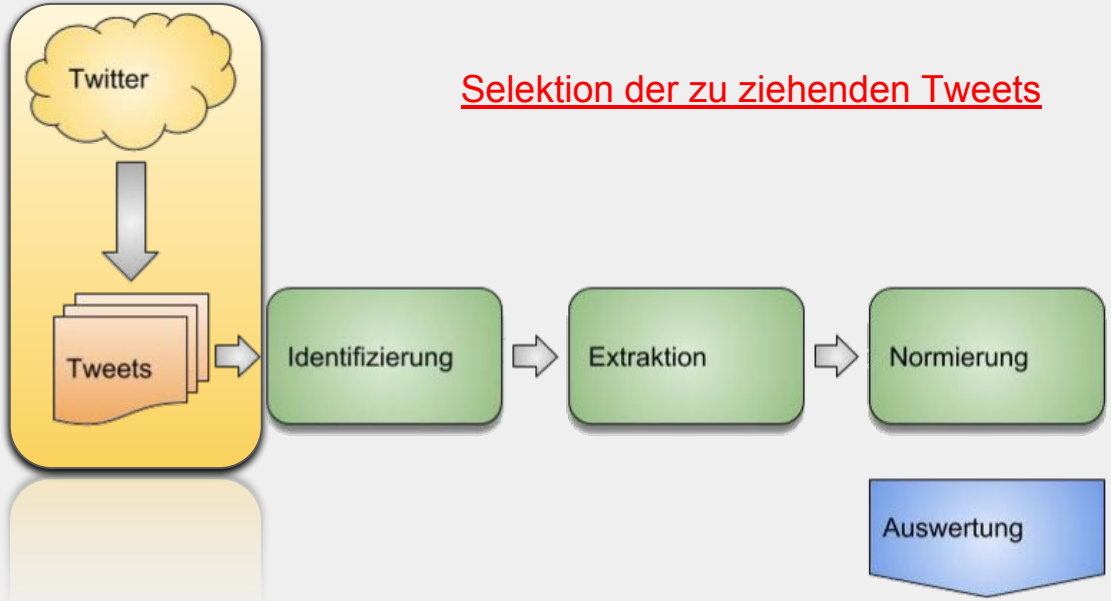


- **Daten ziehen**
- **Identifizierung**
- **Extraktion**
- **Normierung**

University of Applied Sciences
HOCHSCHULE EMDEN·LEER
Fachbereich Technik
Abteilung Elektrotechnik und Informatik

Anwendung - aus Texte Daten generieren

Daten ziehen



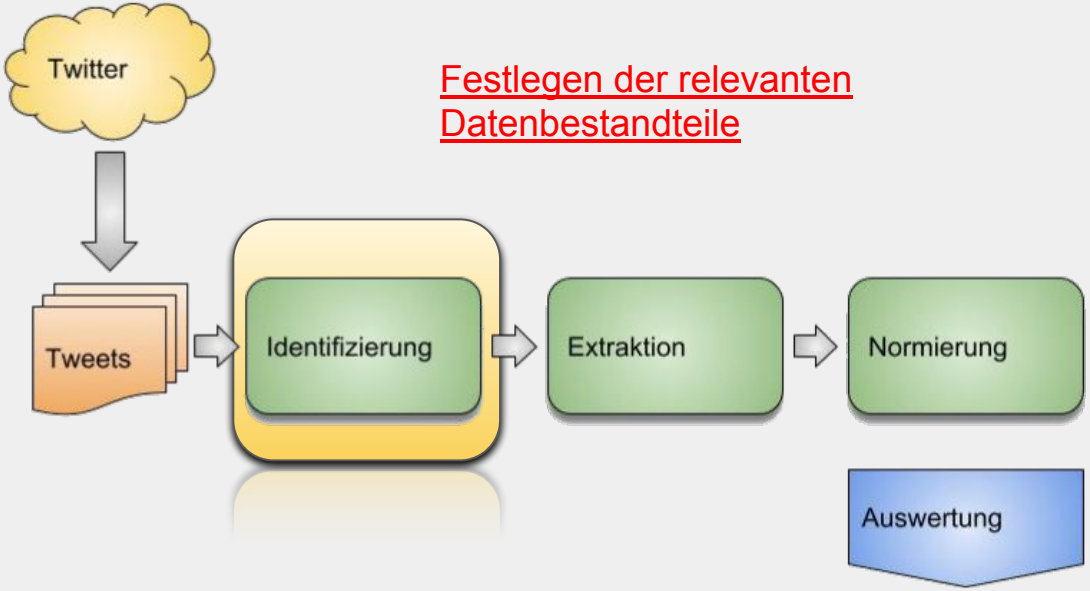
Selektion der zu ziehenden Tweets

- **Daten ziehen**
- **Identifizierung**
- **Extraktion**
- **Normierung**

University of Applied Sciences
HOCHSCHULE EMDEN·LEER
Fachbereich Technik
Abteilung Elektrotechnik und Informatik

Anwendung - aus Texte Daten generieren

Identifizierung



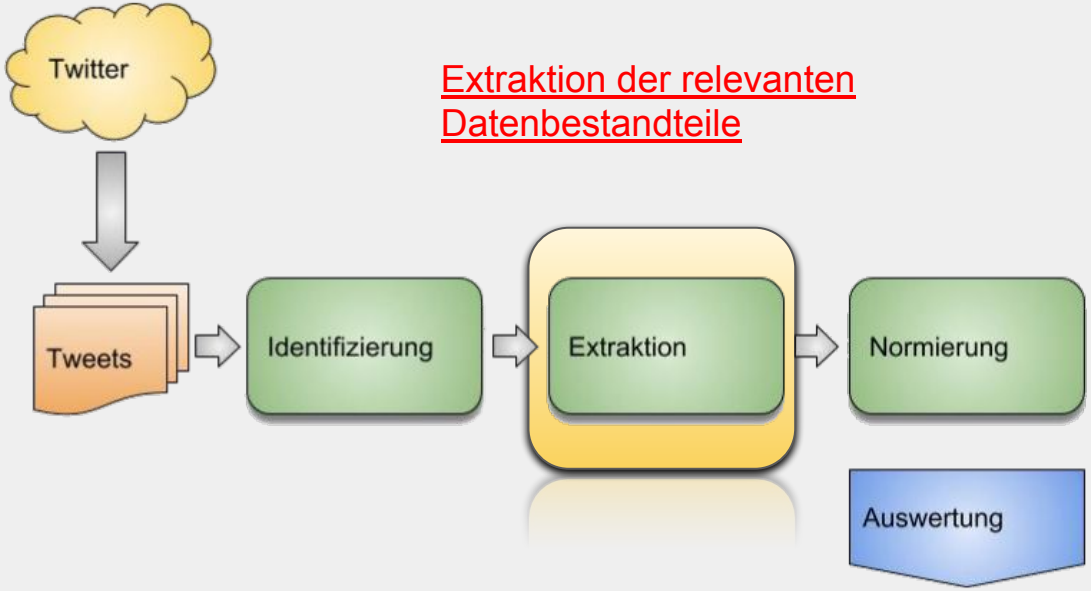
Festlegen der relevanten Datenbestandteile

- **Daten ziehen**
- **Identifizierung**
- **Extraktion**
- **Normierung**

University of Applied Sciences
HOCHSCHULE EMDEN·LEER
Fachbereich Technik
Abteilung Elektrotechnik und Informatik

Anwendung - aus Texte Daten generieren

Extraktion



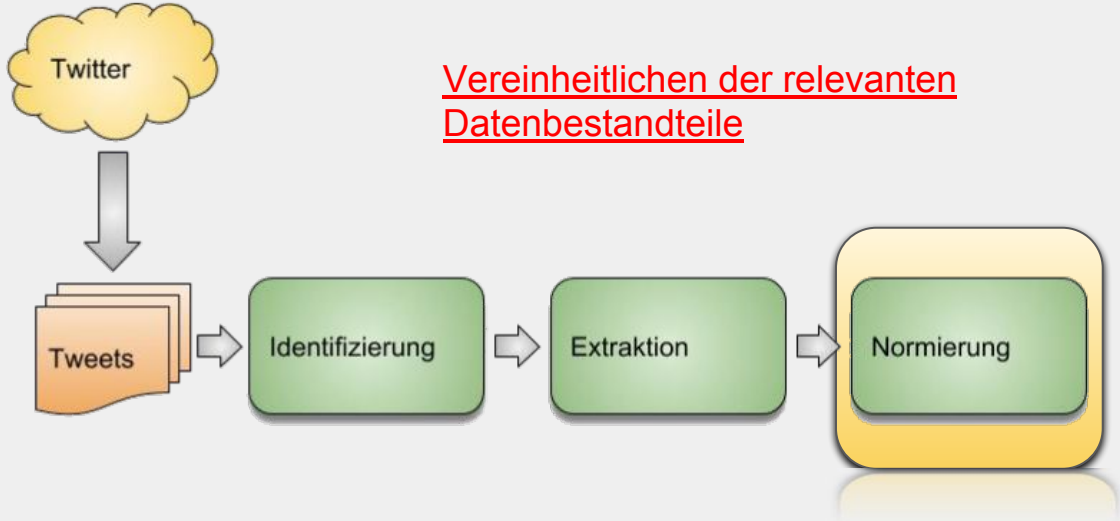
Extraktion der relevanten Datenbestandteile

- **Daten ziehen**
- **Identifizierung**
- **Extraktion**
- **Normierung**

University of Applied Sciences
HOCHSCHULE EMDEN·LEER
Fachbereich Technik
Abteilung Elektrotechnik und Informatik

Anwendung - aus Texte Daten generieren

Normierung



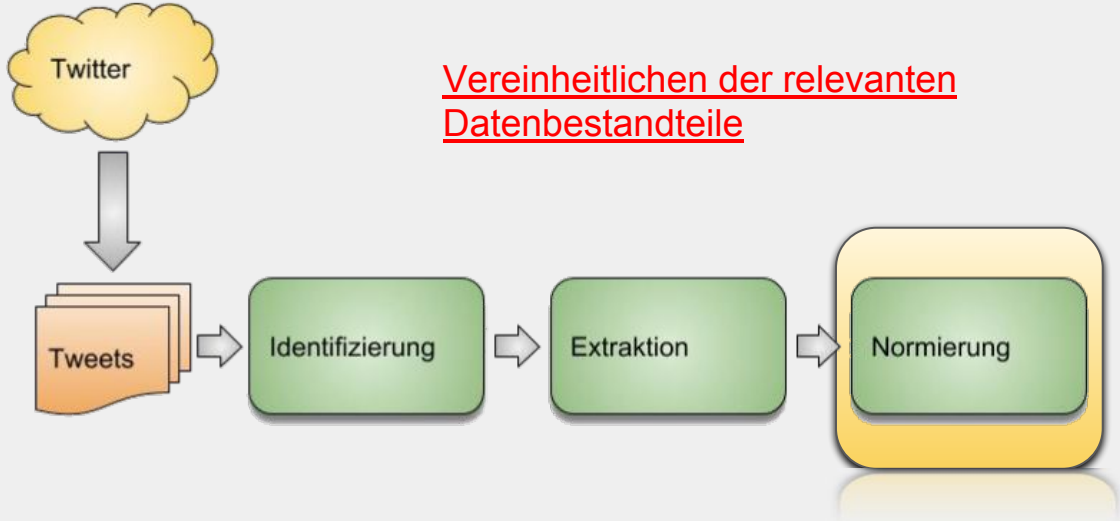
Vereinheitlichen der relevanten Datenbestandteile

- **Daten ziehen**
- **Identifizierung**
- **Extraktion**
- **Normierung**

University of Applied Sciences
HOCHSCHULE EMDEN·LEER
Fachbereich Technik
Abteilung Elektrotechnik und Informatik

Anwendung - aus Texte Daten generieren

Normierung



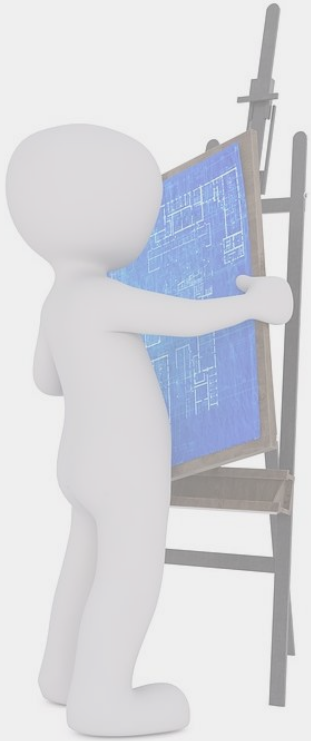
Vereinheitlichen der relevanten Datenbestandteile

- **Daten ziehen**
- **Identifizierung**
- **Extraktion**
- **Normierung**

- Sprachfilterung
- unerwünschter Inhalt, Stopwörter und Interpunktion löschen
- Lemmatisierung und einheitliche Maßangaben
- semantische Analyse

University of Applied Sciences
HOCHSCHULE EMDEN·LEER
Fachbereich Technik
Abteilung Elektrotechnik und Informatik

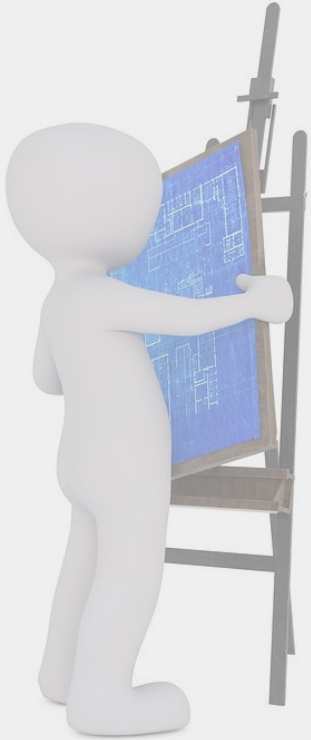
Anwendung - semantische Analyse



Vorhersage der Stimmung eines Textes

University of Applied Sciences
HOCHSCHULE EMDEN·LEER
Fachbereich Technik
Abteilung Elektrotechnik und Informatik

Anwendung - semantische Analyse

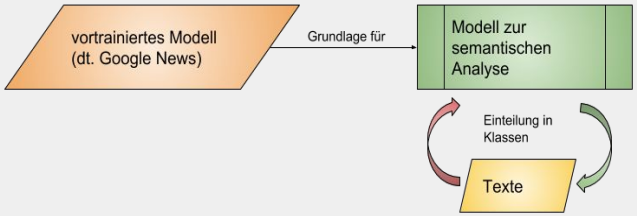


Vorhersage der Stimmung eines Textes

Klassifizierung in positive, neutrale und negative Texte

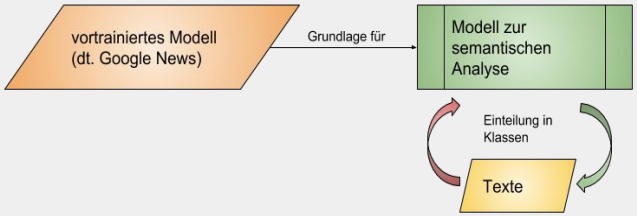
University of Applied Sciences
HOCHSCHULE EMDEN·LEER
Fachbereich Technik
Abteilung Elektrotechnik und Informatik

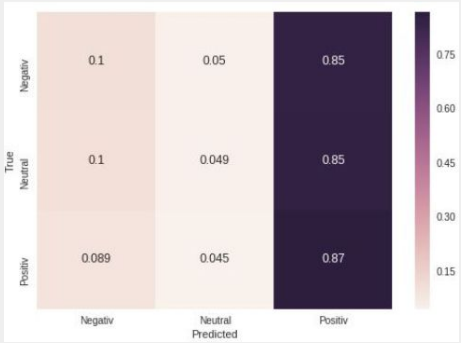
Anwendung - semantische Analyse



- **Vortrainiertes Modell zur Klassifizierung**
- **DNNClassifier als “high end API”**

Anwendung - semantische Analyse





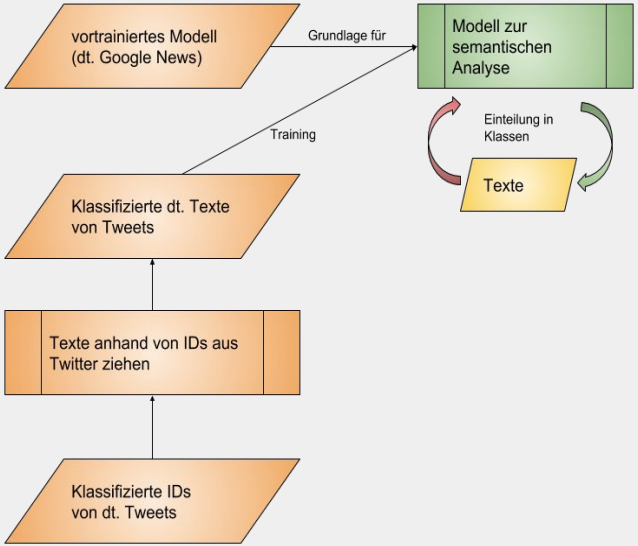
Basis der HeatMap: 10 Tweets mit Stichwort "Weihnachten", gezogen am 27.12.2018. Nach der Vorverarbeitung ergab dies 15 Tweets (inklusive der ursprünglichen Tweets) mit 5 Relationen und 107 unterschiedlichen Wörtern.

- **Vortrainiertes Modell zur Klassifizierung**
- **DNNClassifier bietet Infrastruktur**

⇒ neigt bei Tweets zu einer positiver Bewertung

University of Applied Sciences
HOCHSCHULE EMDEN·LEER
Fachbereich Technik
Abteilung Elektrotechnik und Informatik

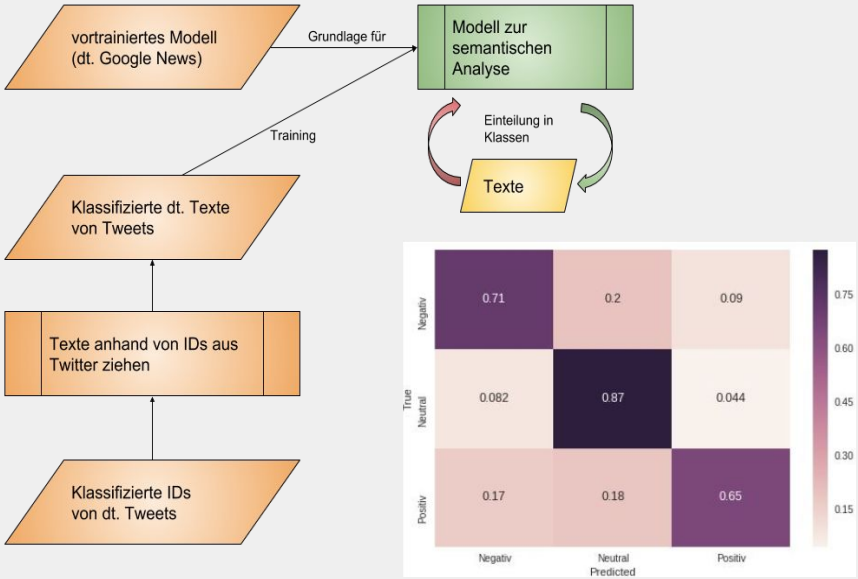
Anwendung - semantische Analyse



- **Vortrainiertes Modell zur Klassifizierung**
- **DNNClassifier bietet Infrastruktur**
- **Zusätzliches Training mit Twitter Korpus**

University of Applied Sciences
HOCHSCHULE EMDEN·LEER
Fachbereich Technik
Abteilung Elektrotechnik und Informatik

Anwendung - semantische Analyse



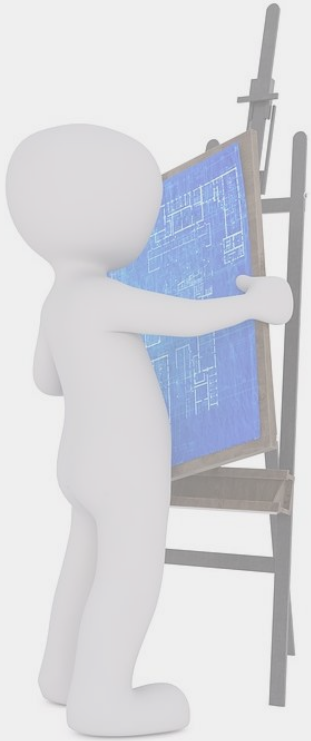
Basis der HeatMap: 10 Tweets mit Stichwort "Weihnachten", gezogen am 27.12.2018. Nach der Vorverarbeitung ergab dies 15 Tweets (inklusive der ursprünglichen Tweets) mit 5 Relationen und 107 unterschiedlichen Wörtern.

- **Vortrainiertes Modell zur Klassifizierung**
- **DNNClassifier bietet Infrastruktur**
- **Zusätzliches Training mit Twitter Korpus**

⇒ deutliche Verbesserung der Genauigkeit bei Tweets

University of Applied Sciences
HOCHSCHULE EMDEN·LEER
Fachbereich Technik
Abteilung Elektrotechnik und Informatik

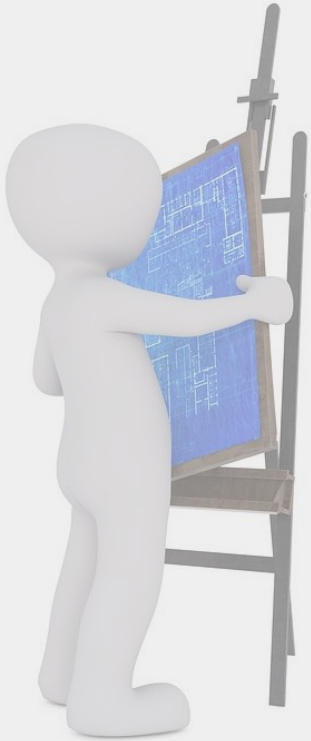
Anwendung - Auswertung



Erkennen von Strukturen in Daten

University of Applied Sciences
HOCHSCHULE EMDEN·LEER
Fachbereich Technik
Abteilung Elektrotechnik und Informatik

Anwendung - Auswertung



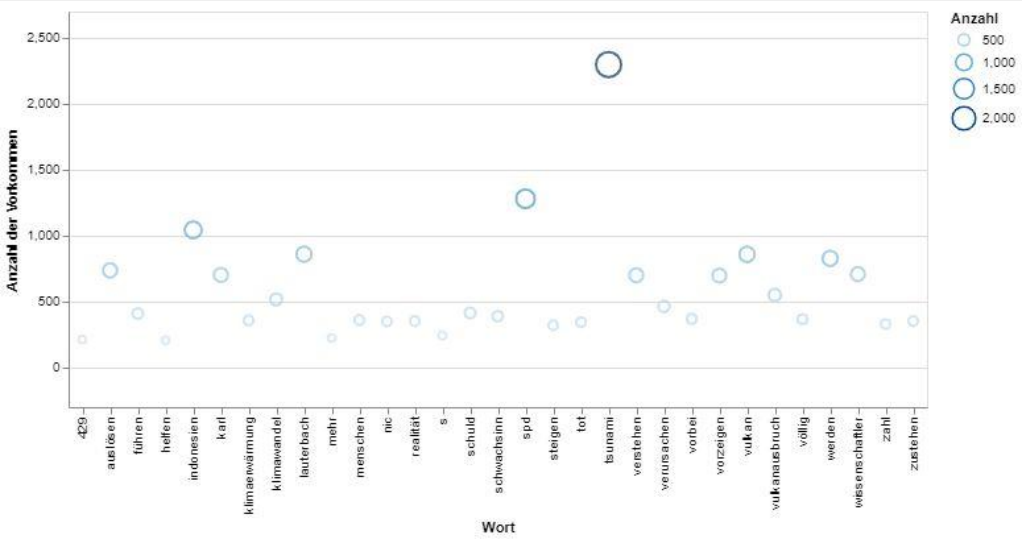
Erkennen von Strukturen in Daten

Vorhersagen für unbekannte Daten

University of Applied Sciences
HOCHSCHULE EMDEN·LEER
Fachbereich Technik
Abteilung Elektrotechnik und Informatik

Anwendung - Auswertung

absolute Häufigkeit eines Wortes ab einer Anzahl von 250



- Größe und Farbe vercoden die Anzahl der Nennungen
- Überblick über die Häufigkeit einzelner Wörter

Basis: 2000 Tweets mit Stichwort "Tsunami", gezogen am 27.12.2018. Nach der Vorverarbeitung ergab dies 3246 Tweets (inklusive der ursprünglichen Tweets) mit 1328 Relationen und 2429 unterschiedlichen Wörtern.

University of Applied Sciences
HOCHSCHULE EMDEN·LEER
Fachbereich Technik
Abteilung Elektrotechnik und Informatik

Anwendung - Auswertung

Visualisierung aller vorkommenden Wörter als Wortwolke

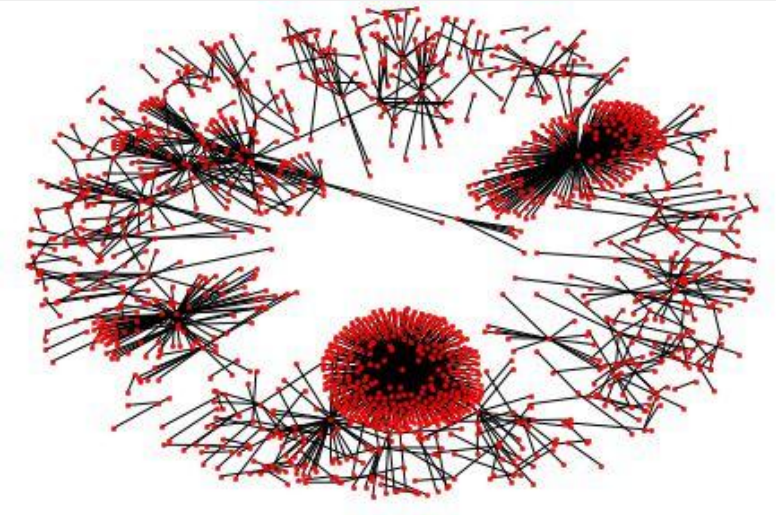
- je Größer umso häufiger
- Färbung unterstreicht die Unterschiede
- verdeutlicht die Häufigkeit von Wörtern

Basis: 2000 Tweets mit Stichwort “Tsunami”, gezogen am 27.12.2018. Nach der Vorverarbeitung ergab dies 3246 Tweets (inklusive der ursprünglichen Tweets) mit 1328 Relationen und 2429 unterschiedlichen Wörtern.

University of Applied Sciences
HOCHSCHULE EMDEN·LEER
Fachbereich Technik
Abteilung Elektrotechnik und Informatik

Anwendung - Auswertung

Relationen von Tweets



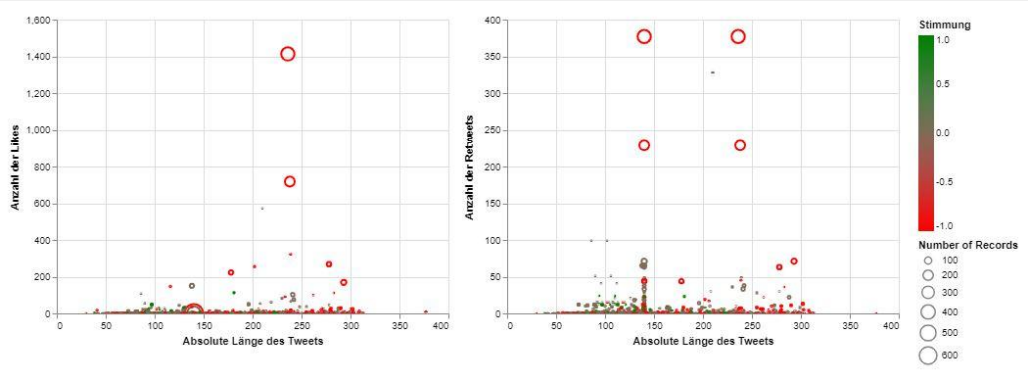
- rote Punkte für Tweets
- Kanten für Verbindungen
- Haufenbildungen zeigen starken Einfluß von Tweets

Basis: 2000 Tweets mit Stichwort “Tsunami”, gezogen am 27.12.2018. Nach der Vorverarbeitung ergab dies 3246 Tweets (inklusive der ursprünglichen Tweets) mit 1328 Relationen und 2429 unterschiedlichen Wörtern.

University of Applied Sciences
HOCHSCHULE EMDEN·LEER
Fachbereich Technik
Abteilung Elektrotechnik und Informatik

Anwendung - Auswertung

absolute Länge von Tweets zur Anzahl der Likes und Retweets



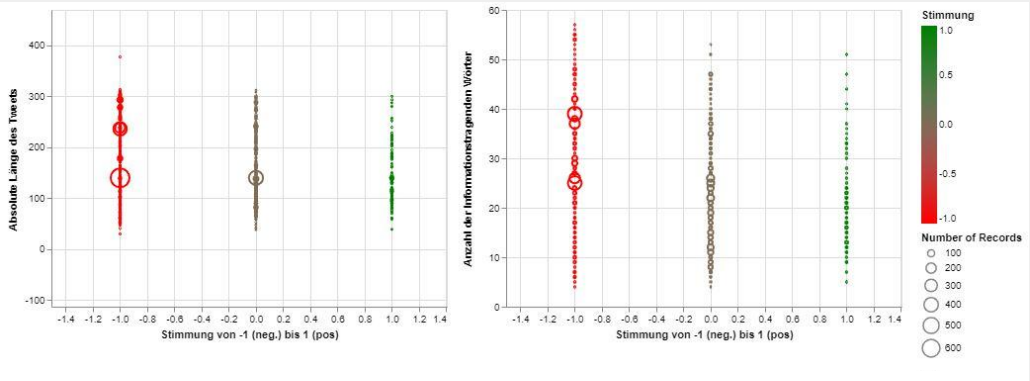
- Färbung zeigt Stimmung
- Kreisgröße zeigt Anzahl
- Zusammenhang von Länge zur Anzahl der Likes und Retweets

Basis: 2000 Tweets mit Stichwort "Tsunami", gezogen am 27.12.2018. Nach der Vorverarbeitung ergab dies 3246 Tweets (inklusive der ursprünglichen Tweets) mit 1328 Relationen und 2429 unterschiedlichen Wörtern.

University of Applied Sciences
HOCHSCHULE EMDEN·LEER
Fachbereich Technik
Abteilung Elektrotechnik und Informatik

Anwendung - Auswertung

Stimmung von Tweets zu deren Länge



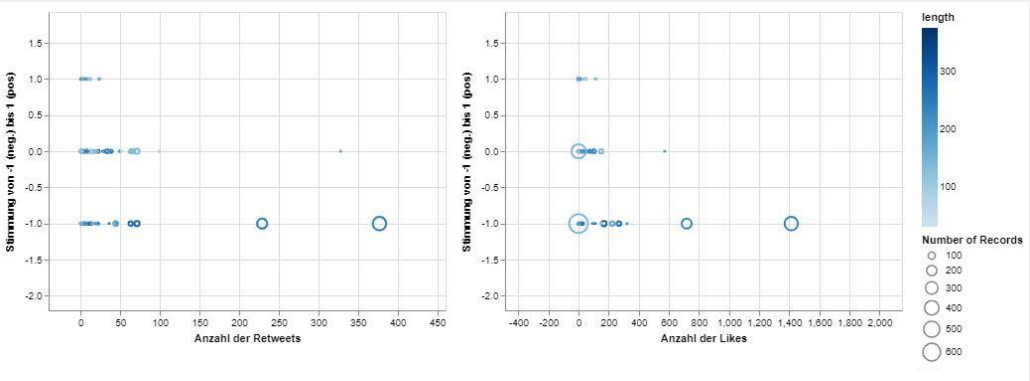
- Färbung zeigt Stimmung
- Kreisgröße zeigt Anzahl
- Zusammenhang von Stimmung zur absoluten Länge sowie der Anzahl informationstragender Wörter

Basis: 2000 Tweets mit Stichwort "Tsunami", gezogen am 27.12.2018. Nach der Vorverarbeitung ergab dies 3246 Tweets (inklusive der ursprünglichen Tweets) mit 1328 Relationen und 2429 unterschiedlichen Wörtern.

University of Applied Sciences
HOCHSCHULE EMDEN·LEER
Fachbereich Technik
Abteilung Elektrotechnik und Informatik

Anwendung - Auswertung

Stimmung von Tweets zu der Anzahl an Likes und Retweets

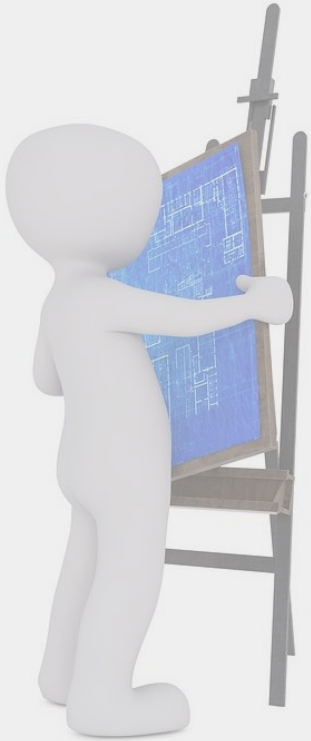


- Färbung zeigt Länge
- Kreisgröße zeigt Anzahl
- Zusammenhang von Stimmung zur Anzahl der Retweets und Likes

Basis: 2000 Tweets mit Stichwort "Tsunami", gezogen am 27.12.2018. Nach der Vorverarbeitung ergab dies 3246 Tweets (inklusive der ursprünglichen Tweets) mit 1328 Relationen und 2429 unterschiedlichen Wörtern.

University of Applied Sciences
HOCHSCHULE EMDEN·LEER
Fachbereich Technik
Abteilung Elektrotechnik und Informatik

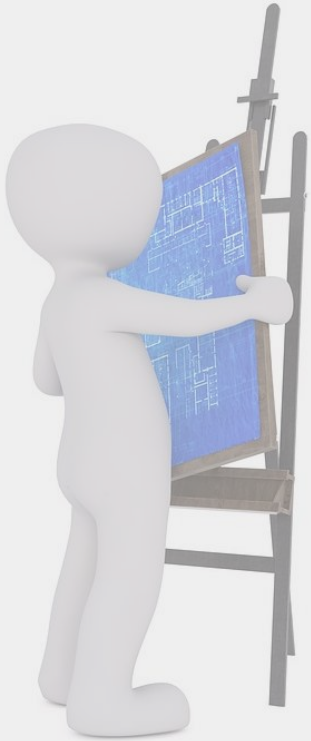
Anwendung - verfügbar machen



Schritte zur Verwendung der Anwendung

University of Applied Sciences
HOCHSCHULE EMDEN·LEER
Fachbereich Technik
Abteilung Elektrotechnik und Informatik

Anwendung - verfügbar machen

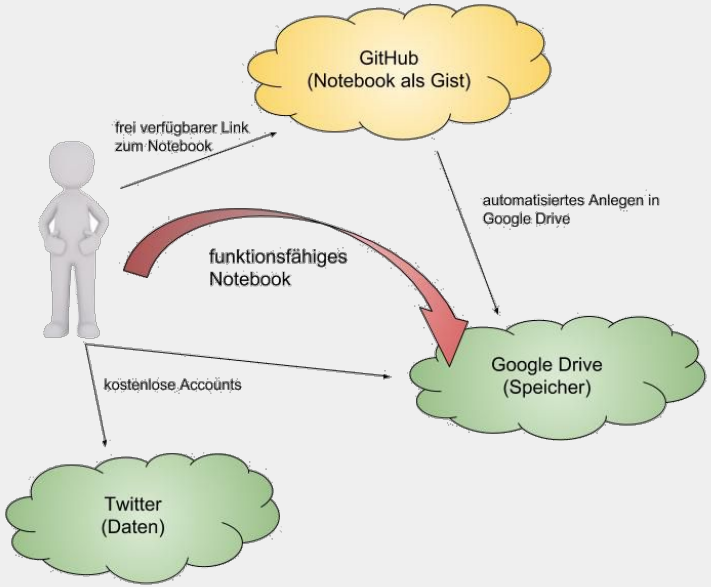


Schritte zur Verwendung der Anwendung

Einfacher Zugang durch wenige Anforderungen

University of Applied Sciences
HOCHSCHULE EMDEN·LEER
Fachbereich Technik
Abteilung Elektrotechnik und Informatik

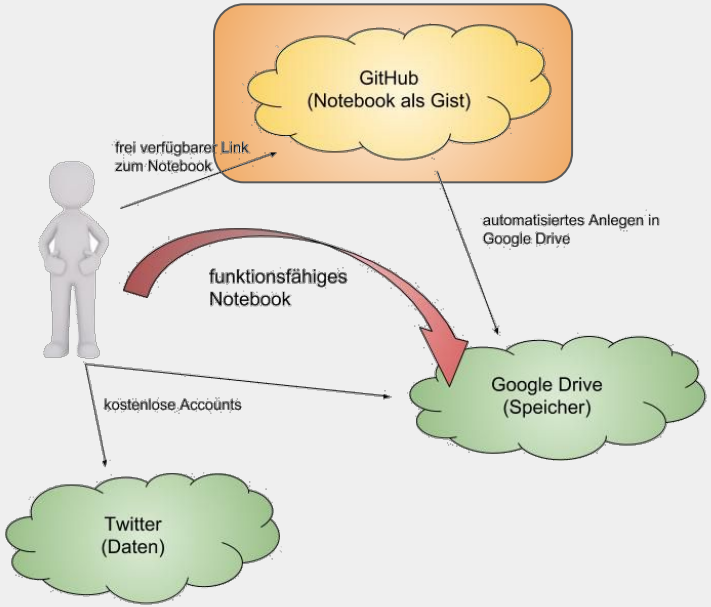
Anwendung - verfügbar machen



- **Einfach**
- **ohne Installation**
- **möglichst geführt**

University of Applied Sciences
HOCHSCHULE EMDEN·LEER
Fachbereich Technik
Abteilung Elektrotechnik und Informatik

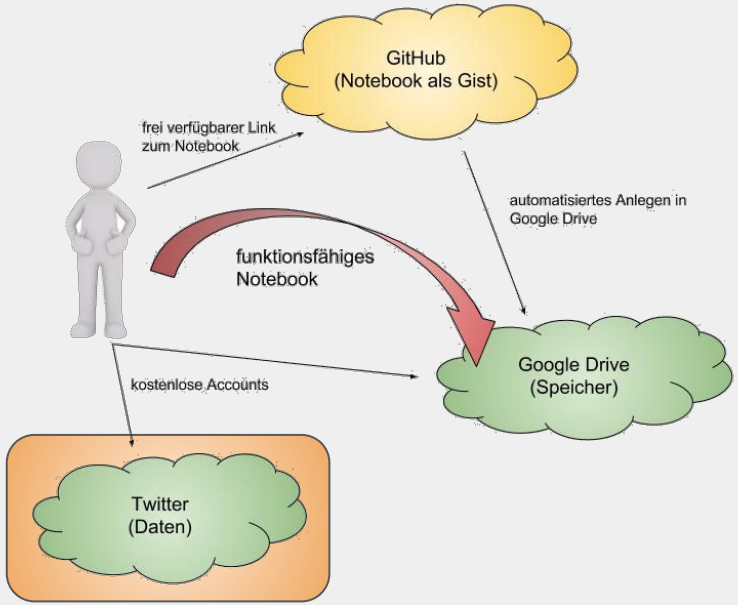
Anwendung - verfügbar machen



- **Link zum Notebook**

University of Applied Sciences
HOCHSCHULE EMDEN·LEER
Fachbereich Technik
Abteilung Elektrotechnik und Informatik

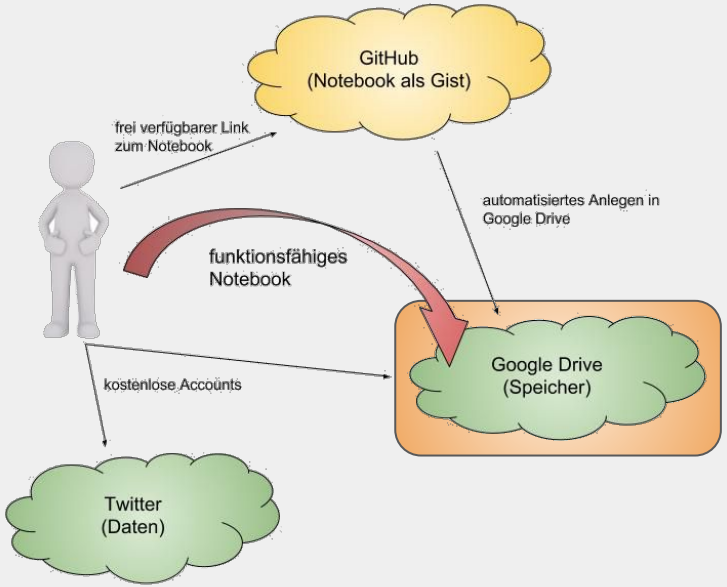
Anwendung - verfügbar machen



- **Link zum Notebook**
- **Account bei Twitter**

University of Applied Sciences
HOCHSCHULE EMDEN·LEER
Fachbereich Technik
Abteilung Elektrotechnik und Informatik

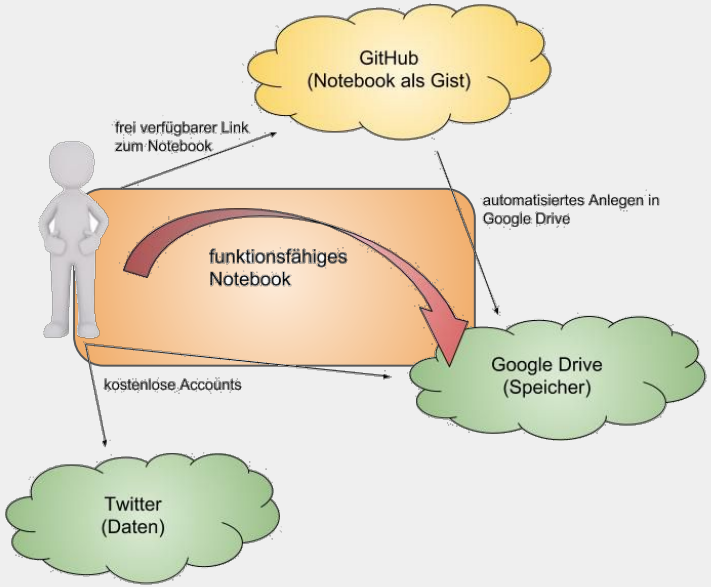
Anwendung - verfügbar machen



- **Link zum Notebook**
- **Account bei Twitter**
- **Account bei Google Drive**

University of Applied Sciences
HOCHSCHULE EMDEN·LEER
Fachbereich Technik
Abteilung Elektrotechnik und Informatik

Anwendung - verfügbar machen



- **Link zum Notebook**
- **Account bei Twitter**
- **Account bei Google Drive**

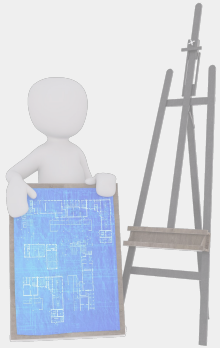
⇒ funktionsfähiges Notebook

University of Applied Sciences
HOCHSCHULE EMDEN·LEER
Fachbereich Technik
Abteilung Elektrotechnik und Informatik

Zusammenfassung

Ergebnis des Praxisprojekts

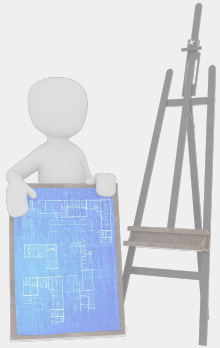
- **leicht verteil- und ausführbare Anwendung verfügbar**
- **kommentierter Workflow zur Vorverarbeitung von Texten**
- **Aufbau und Einsatz eines neuronalen Netzes zur semantischen Analyse mit Tensor Flow**
- **Überblick über den Einsatz etablierte Bibliotheken und Module**
  - Natural Language Toolkit
  - Tweepy
  - ...



Zusammenfassung

Ergebnis des Praxisprojekts

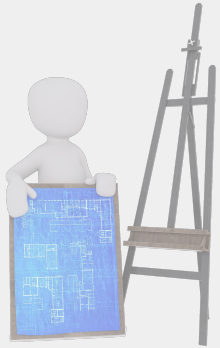
- **Einschränkend ist anzumerken:**
  - Motivation ist die Einführung
  - einfache Hinführung vor Effizienz
  - nicht für den realen Einsatz



Zusammenfassung

Ergebnis des Praxisprojekts

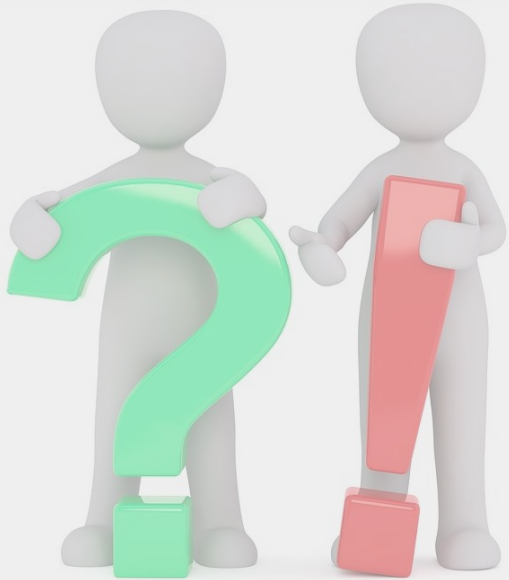
- **Einschränkend ist anzumerken:**
  - Motivation ist die Einführung
  - einfache Hinführung vor Effizienz
  - nicht für den realen Einsatz
  - jede Aufgabenstellung definiert den notwendigen Workflow neu



Ausblick

Wie könnte es weitergehen?

- Evaluierung des Trainings bezüglich
  - des vortrainiertes Modells
  - des verwendeten Twitterkorpuses
- Evaluierung des Workflows
  - der Daten
  - der semantischen Analyse
- Evaluierung der Verständlichkeit
  - des begleitenden Textes
  - des Beispielcodes



Fragerunde und Vorführung

Vielen Dank für Ihre Aufmerksamkeit!

Haben Sie noch Fragen?

Gerne beantworte ich Ihnen diese vor oder nach der Vorführung...

University of Applied Sciences
HOCHSCHULE EMDEN·LEER
Fachbereich Technik
Abteilung Elektrotechnik und Informatik

Marken und Rechte

Auf Grund der besseren Lesbarkeit wurden für geschützte Begriffe, Warennamen, Marken usw. keine Angaben zu den Rechten Dritter gemacht. Die Verwendung in dieser Arbeit berechtigt daher nicht zu der Annahme, dass diese frei von Rechten Dritter sind.

Bildnachweis

Für alle Bilder der Plattform Pixabay (https://pixabay.com/) gelten deren Lizenzbedingungen. Diese kann unter https://pixabay.com/de/service/terms/#license eingesehen werden.

https://pixabay.com/de/checkliste-liste-%C3%BCberpr%C3%BCfen-marke-1919328/	(Folie 2)
https://pixabay.com/de/richtung-weg-entscheidung-ziel-2320124/	(Folie 3,4,5)
https://pixabay.com/de/staffelei-schulung-bildung-training-2714167/	(Folie 6, 18, 19, 20, 21, 28, 29, 34, 35, 42, 43)
https://pixabay.com/de/twitter-tweet-twitter-vogel-312464/	(Folie 10, 11)
https://pixabay.com/de/tafel-staffelei-architekt-ingenieur-2714168/	(Folie 49, 50, 51)
https://pixabay.com/de/fragezeichen-frage-hilfe-antwort-2314115/	(Folie 52)
https://pixabay.com/de/fragezeichen-frage-antwort-1019993/	(Folie 53)

https://de.wikipedia.org/wiki/Python_(Programmiersprache)#/media/File:Python_logo_and_wordmark.svg (Folien 12, 13)
By TensorFlow - vectors combined, edited - Begoon - https://github.com/tensorflow/tensorflow
https://github.com/valohai/ml-logos/blob/master/tensorflow-text.svg
https://github.com/valohai/ml-logos/blob/master/tensorflow-tf.svg, Apache License 2.0,
https://commons.wikimedia.org/w/index.php?curid=65268375

https://de.wikipedia.org/wiki/TensorFlow#/media/File:TensorFlowLogo.svg (Folie 14, 15, 16, 17)
By www.python.org - https://www.python.org/community/logos/, GPL,
https://commons.wikimedia.org/w/index.php?curid=34991637/

Für alle sonstigen Abbildungen liegen die Rechte beim Autor

University of Applied Sciences
HOCHSCHULE EMDEN·LEER
Fachbereich Technik
Abteilung Elektrotechnik und Informatik