

Text Mining mit Python und Power BI

Prof. Dr. Jens Albrecht, Prof. Dr. Roland Zimmermann

<https://github.com/jsalbr/tdwi-2021-text-mining>





Alle Daten zum mitmachen unter:
<https://github.com/jsalbr/tdwi-2021-text-mining>

Zugang zu Google Colab:
<https://colab.research.google.com>



Prof. Dr. Jens Albrecht
TH Nürnberg, Informatik

Data Warehousing, BI,
Data Science, NLP



Prof. Dr. Roland Zimmermann
TH Nürnberg, BWL

BI, Information Design,
NLP, Process Mining



Überblick

<https://github.com/jsalbr/tdwi-2021-text-mining>

<https://www.tdwi-konferenz.de/tdwi-2021/programm/konferenzprogramm.html#item-2976>

Zeit: max. 2:10; verplanen: 1:45

Thema	Dauer	von-bis
Einleitung + Szenario	00:10	16:10
Explorative Datenanalyse	00:20	16:20
Datenvorbereitung	00:10	16:40
Klassifikation	00:15	
Evaluation und Detailanalyse (Power BI)	00:20	17:05
Deep Learning (Embeddings und BERT)	00:15	17:25
Vertiefende Intelligence	00:10	17:40

Herausforderungen mit Text

300
Kundenservice-
Anfragen pro Tag

70.000 Verträge

1.000 Online-
Reviews zum
Produkt

30 Leserkommen-
tare pro Minute

100.000
dokumentierte
Change Requests

... jede einzelne
muss manuell
kategorisiert
werden.

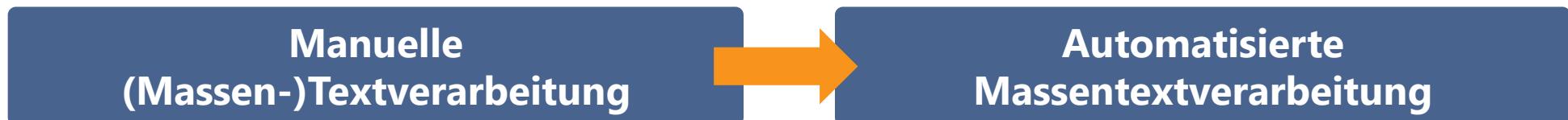
... ohne
automatisiertes
Ablagesystem.

... müssen für die
Einkaufs-
verhandlung alle
der Reihe nach
gelesen werden.

... in der eigenen
Online Community
und der Community
Manager entdeckt
die dringendsten
eher per Zufall.

... und keinen
Überblick, wo die
Kostentreiber
stecken.

Innovation durch Automatisierung



- manuell durchgeführte
 - zeitaufwändige
 - fehlerbehaftete Prozesse
-
- effizienter arbeiten
 - Kosten einsparen
 - Wettbewerbsvorsprung durch neues Wissen erarbeiten
 - Nerven der eigenen Mitarbeiter schonen

Text Analytics Methoden

Bereiche

1

Statistik

2

Unüberwachtes
Machine Learning

3

Überwachtes
Machine Learning

4

Semantische
Analysen

Methoden und Ziele
(Auszug)

KORRELATION

Überblick über Texte,
Datenqualität und
mögl. Bias erkennen

TOPIC MODELING

Identifikation von
Themen in Texten

CLUSTERING

Versteckte Muster in
Content-Archiv
analysieren

KLASSIFIKATION

Neue Texte in bekannte
Kategorien einsortieren

REGRESSION

Trends in Texten
identifizieren

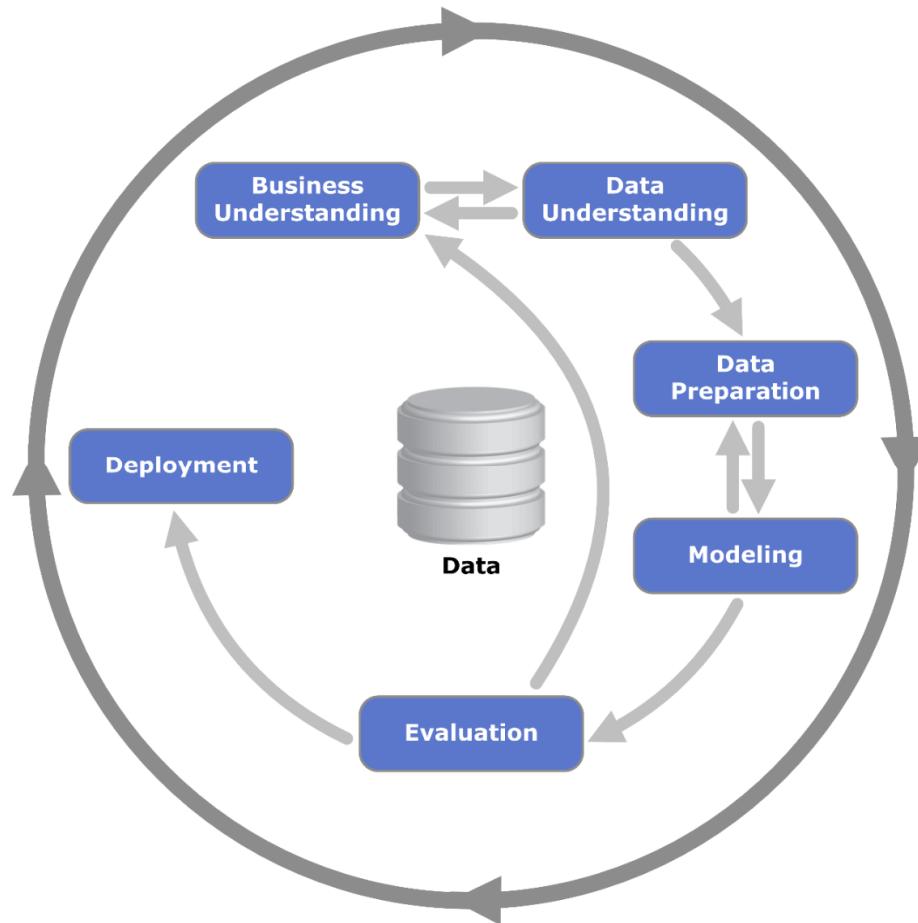
EMBEDDINGS

Semantik, Bedeutung &
Zusammenhänge
extrahieren

QUESTION
ANSWERING

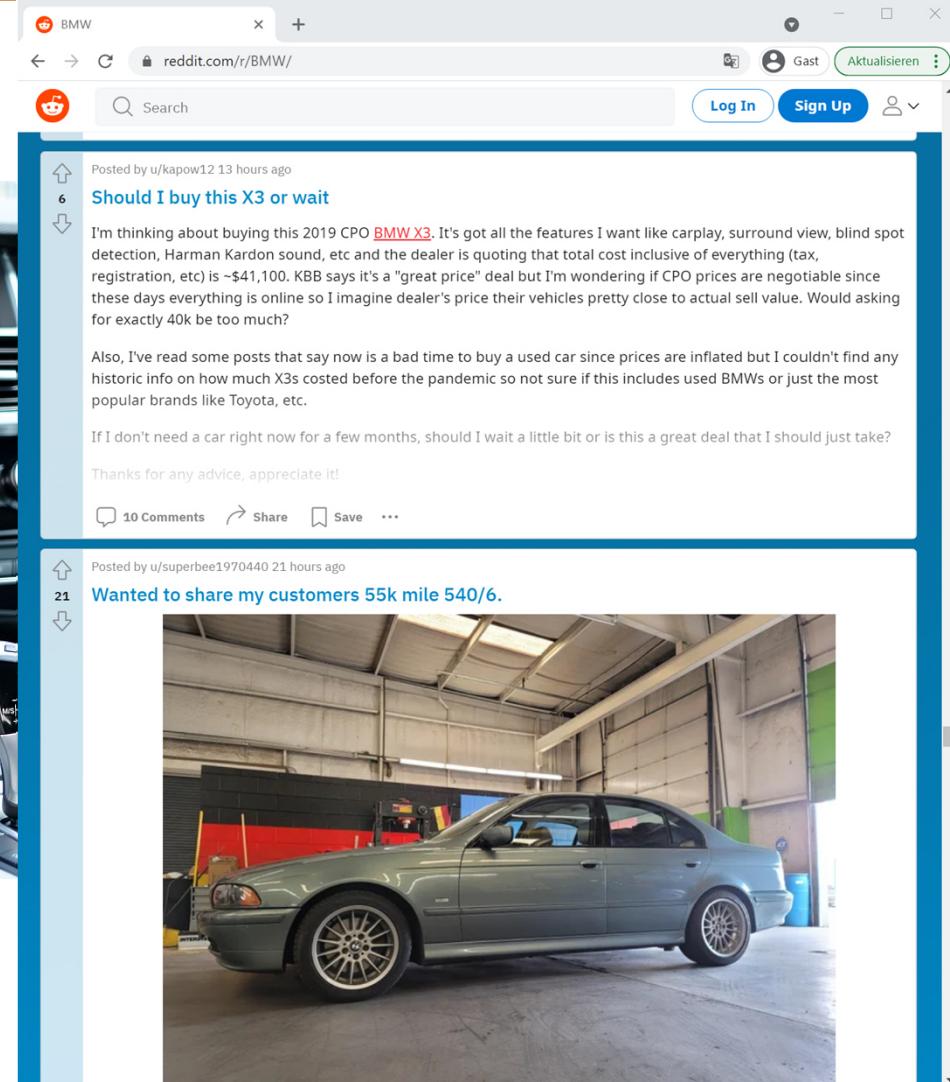
Antworten
zu Fakten-Fragen finden

Crisp DM Prozess



https://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining

Szenario



Should I buy this X3 or wait

I'm thinking about buying this 2019 CPO [BMW X3](#). It's got all the features I want like carplay, surround view, blind spot detection, Harman Kardon sound, etc and the dealer is quoting that total cost inclusive of everything (tax, registration, etc) is ~\$41,100. KBB says it's a "great price" deal but I'm wondering if CPO prices are negotiable since these days everything is online so I imagine dealer's price their vehicles pretty close to actual sell value. Would asking for exactly 40k be too much?

Also, I've read some posts that say now is a bad time to buy a used car since prices are inflated but I couldn't find any historic info on how much X3s costed before the pandemic so not sure if this includes used BMWs or just the most popular brands like Toyota, etc.

If I don't need a car right now for a few months, should I wait a little bit or is this a great deal that I should just take?

Thanks for any advice, appreciate it!

Wanted to share my customers 55k mile 540/6.

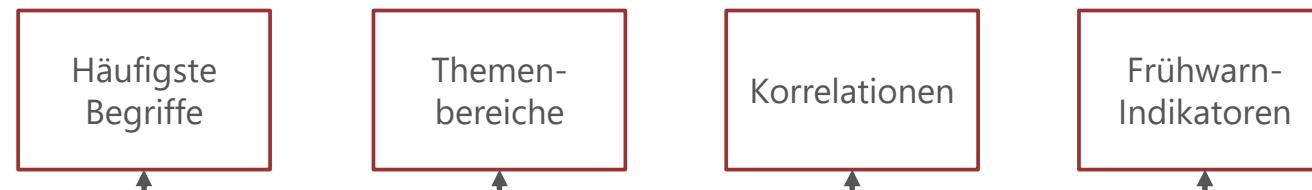


Beispiel: Störungsmeldungen in der Auto-Industrie

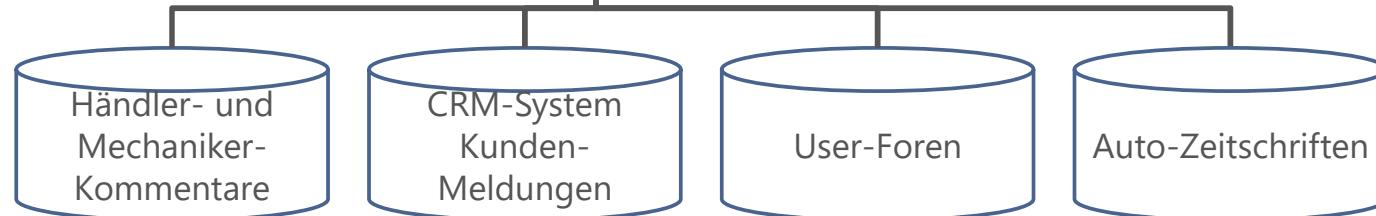
Aktionen



Analysen



Quelldaten



Explorative Datenanalyse mit PowerBI

Power BI als Explorationswerkzeug für Textdaten

- **Datenexploration mit BI-Tool leicht durch Fachanwender**
 - Business und Data Understanding durch Fachabteilung vorbereitbar
 - Arbeitsteilung mit Data-Scientists
- **Texte als Rohdaten anders bzw. ergänzend zu behandeln:**
Ziel: erste fachliche Einschätzungen zu Inhalten erzielen
 - Darstellung von Texten in Tabellenform oft schlecht lesbar:
Thumbnails mit Preview und interaktiver Vergrößerung verbinden
Überblick und Detail für Rohtext-Visualisierungen
 - Rohtexte enthalten erkennbar viele unnütze Wörter:
Sind auch für erste Auszählungen (z.B. Wortwolken) zu ignorieren = Stopwortkonzept relevant
 - Interaktive Filteroptionen vermitteln rasch ersten Überblick über Inhalte der Texte
- Für spätere Phasen des Text-Mining-Prozesses: BI-Tool zur Kommunikation und Bereitstellung der – oftmals komplexen – Text-Mining-Ergebnisse weiternutzen!

Custom Visuals in Power BI für Exploration textueller Daten

Filters

Full Text Search  

Search... 

AND OR

Current Filter

Contains 'hydrogen' or... 

Filter type: Advanced filtering

Show items when the value: contains hydrogen 

And Or

contains electrolysis 

Apply filter 

Custom Visual von Prof. Albrecht

Word Cloud  Create a fun visual from frequent text in your data 

Add

Text Filter  Search across your dataset right from the dashboard 

Add

Tag Cloud - xViz  Tag Cloud helps you get instant insight into the most prominent or prevalent terms in your data 
May require additional purchase

Add

Card Browser  Browse documents using double-sided cards, and click to view in place 

Add

Strippets Browser  A quick way to view document contents 

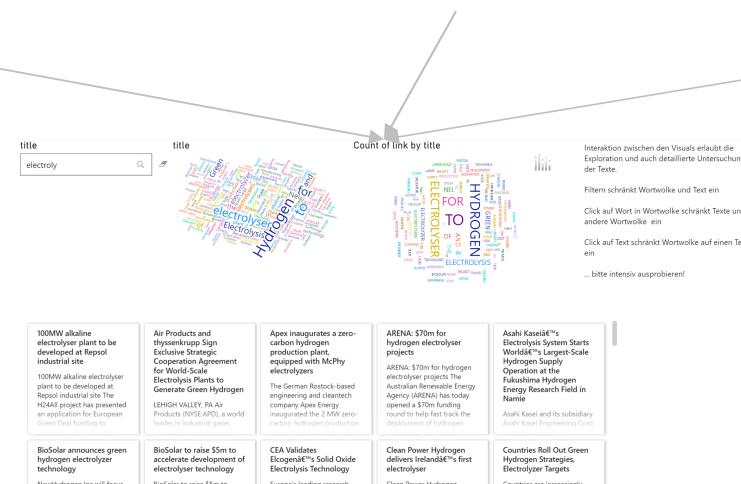
Add

Spezielle Filter für Texte

Wortwolken

Textuelle Card-Darstellungen

Interdependente, interaktive Anwendung entsteht, die individuelle Sichten auf Rohtexte erlaubt

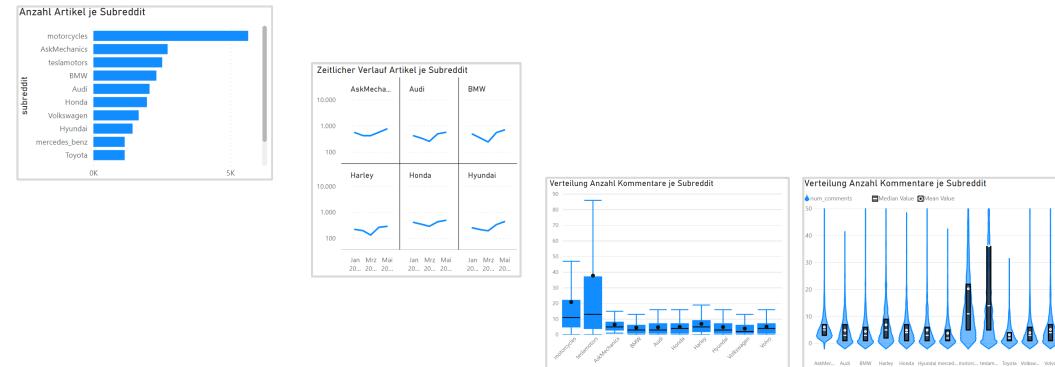


100MW alkaline electrolyzer plant to be developed at Repsol industrial site	Air Products and HydrogenSign Executive Strategic Cooperation Agreement for World-Scale Electrolyzer Plants to Generate Green Hydrogen	Apen inaugurates a zero-carbon hydrogen production plant equipped with McPhy electrolyzers	ARENA: \$70m for hydrogen electrolyser projects	Asahi Kasei's Electrolyzer System Starts World's Largest Scale Hydrogen Supply Operation at the Fukushima Hydrogen Energy Research Field in Name
100MW alkaline electrolyzer plant to be developed at Repsol industrial site	Air Products and HydrogenSign Executive Strategic Cooperation Agreement for World-Scale Electrolyzer Plants to Generate Green Hydrogen	Apen inaugurates a zero-carbon hydrogen production plant equipped with McPhy electrolyzers	ARENA: \$70m for hydrogen electrolyser projects	Asahi Kasei and its subsidiary Asahi Kasei Energy Solutions Co., Ltd. have started the world's largest-scale hydrogen supply operation at the Fukushima Hydrogen Energy Research Field in Name

Visuals und einfaches ETL für Textdaten

Schnelle deskriptive quantitative Analysen

- Balkendiagramme für Mengen
- Liniendiagramme für zeitliche Entwicklungen
- Boxplots oder Violin-Plots für Verteilungen



Word Clouds und einfache Stopwörter

Textfilter

Thumbnails für Text-Details

!!Cv Shaft and Steering Rack Boot leaks!! !!Cv Shaft and Steering Rack Boot leaks!! I've just picked up my 2005	IHELP! P0741 Code - 2006 Accord Lx 4cyl AT 117k IHELP! P0741 Code - 2006 Accord Lx 4cyl AT 117k: Hi all,	... i came out of that corner a different person's story. What's yours? ... i came out of that corner a
"A/C off for engine protection" comes on the dash. 2007 Pontiac Montana. Thermostat issue? "A/C off for engine	"Best" manual BMW? "Best" manual BMW? What are your thoughts on the "best"-manual BMW?	"Company Vehicle - Not for Resale" "Company Vehicle - Not for Resale": Looking at a 2020 A4 Allroad. Originally listed as "compatible"



Full Text Search ?

X

AND OR

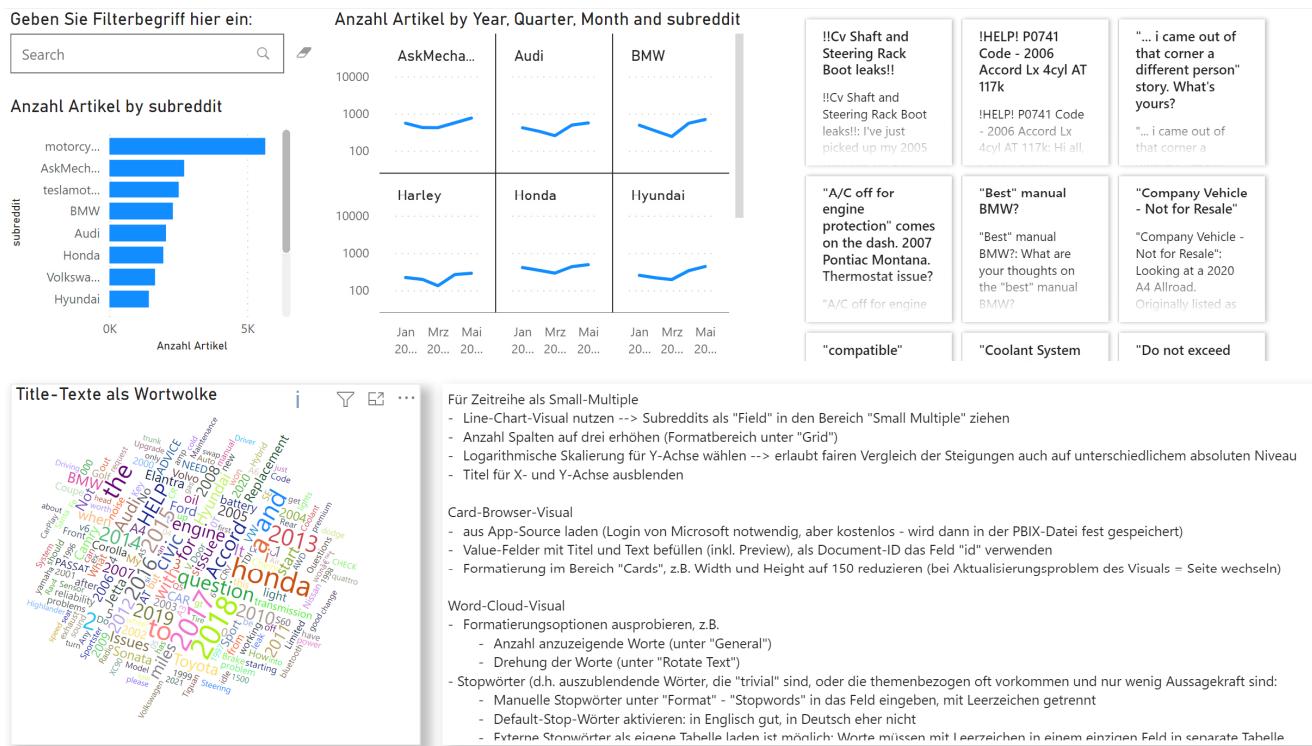
Current Filter

Hier einen Filterbegriff eingeben (für Gesamttexte):

Search


Live-Demo: Übungs- und Lösungsdatei auf Github

<https://github.com/jsalbr/tdwi-2021-text-mining>



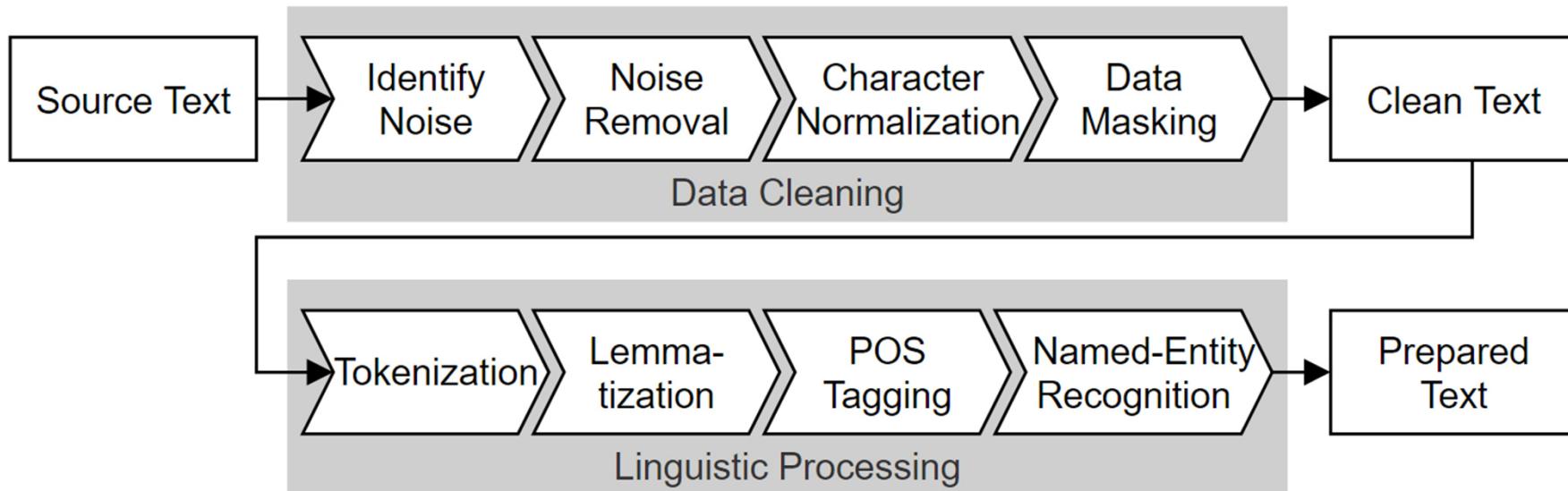
Datenaufbereitung mit Python

Live-Demo

<https://github.com/jsalbr/tdwi-2021-text-mining>

Bitte Notebook "Preprocessing" starten!

Pipeline zur Textaufbereitung



spaCy
<https://spacy.io>

Klassifikation mit Python

Live-Demo

<https://github.com/jsalbr/tdwi-2021-text-mining>

Bitte Notebook "Classification" starten!

Klassifikationsaufgaben

- › Tickets / Service-Requests kategorisieren & automatisiert routen
 - » Kategorie
 - » Priorisierung (Urgency / Impact)
- › Produktstammdaten klassifizieren / zuordnen
- › Sentiment-Analyse
- › Auto-Tagging

Supervised Learning: Trainings-Matrix X und Label-Vektor y

Trainingsmatrix X: $n \times m$ -Matrix

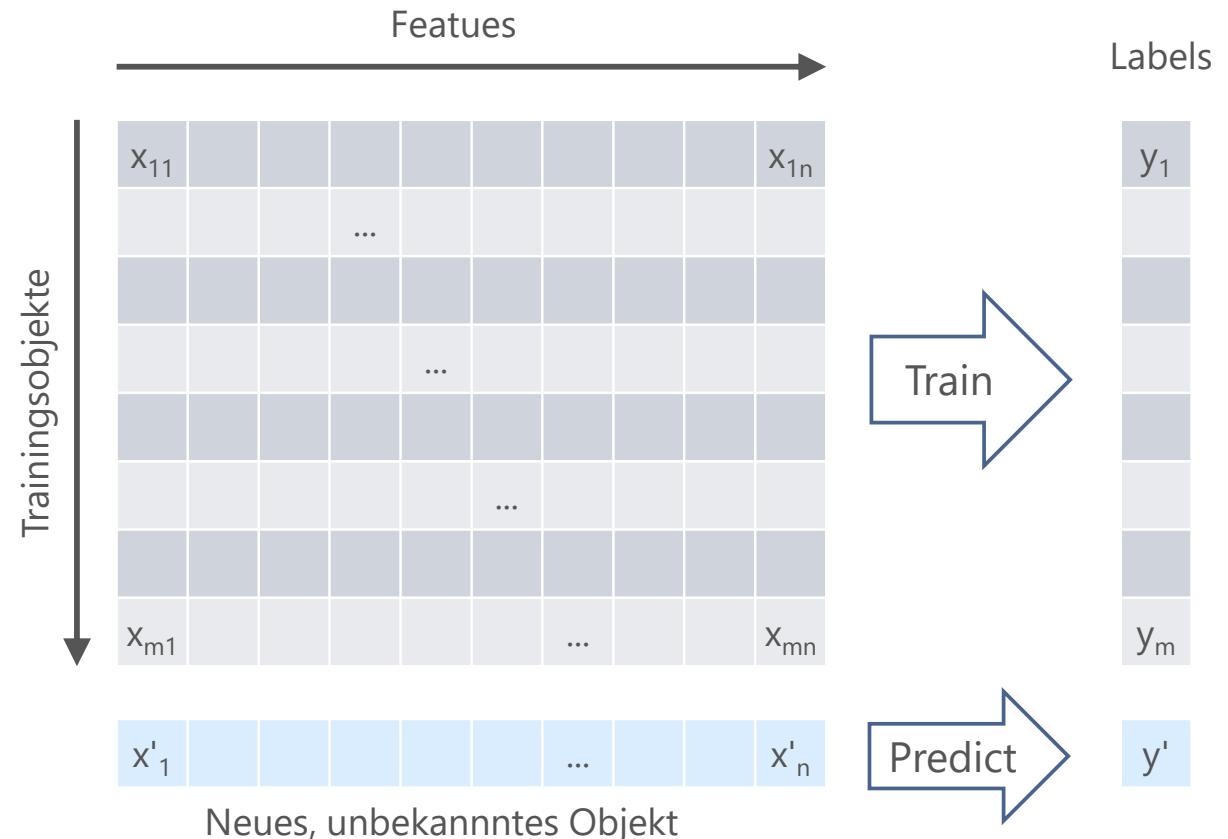
- › n Trainingsobjekte (Samples, Examples) mit jeweils m Features
- › z.B. Kunden, Texte, Bilder, ...

Label-Vektor y:

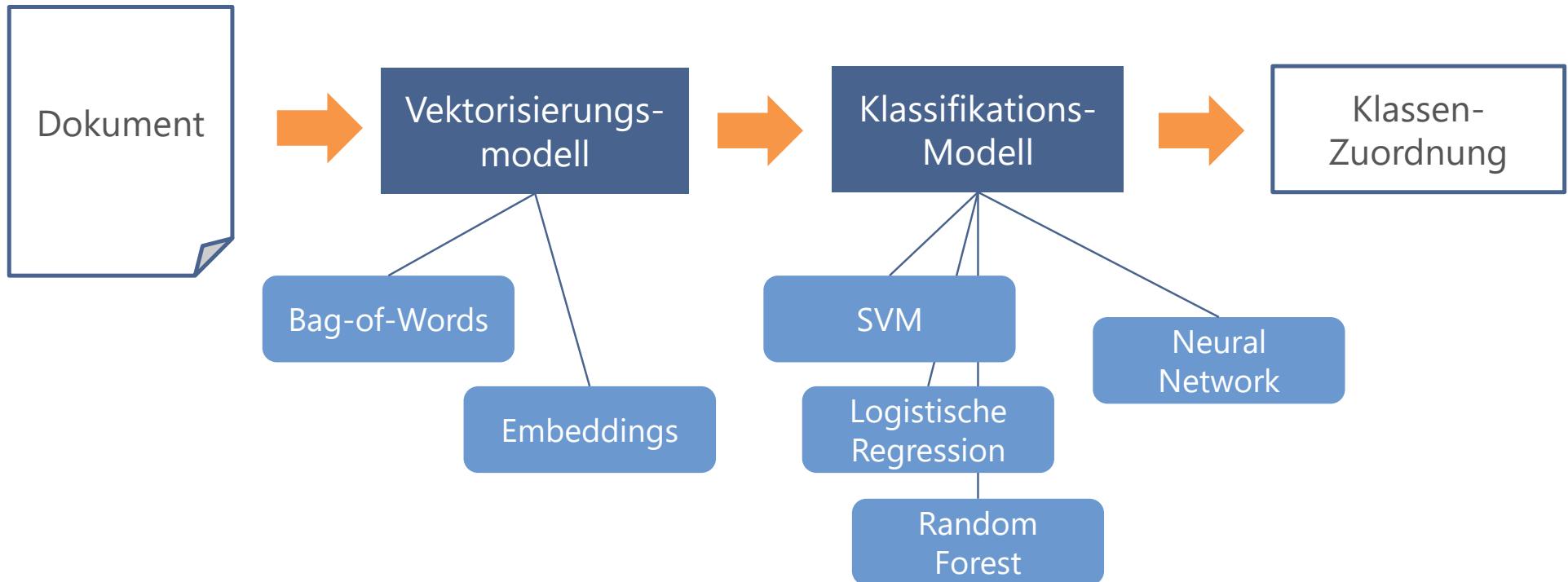
- › Spaltenvektor mit vorherzusagendem Wert

Feature-Modellierung:

- › Überführung der Ausgangsdaten in Vektor-Repräsentation



Text-Klassifikation



Bag-of-Words Vektorisierung

Dokumente

D_1 : „Pete likes London. Pete likes Paris.“

D_2 : „Pete does not like London.“

D_3 : „Pete likes London, but not Paris.“

	Pete	like	London	do	not	but	Paris
D_1	2	2	1				1
D_2	1	1	1	1	1		
D_3	1	1	1		1	1	1

Nur Worthäufigkeiten beachtet

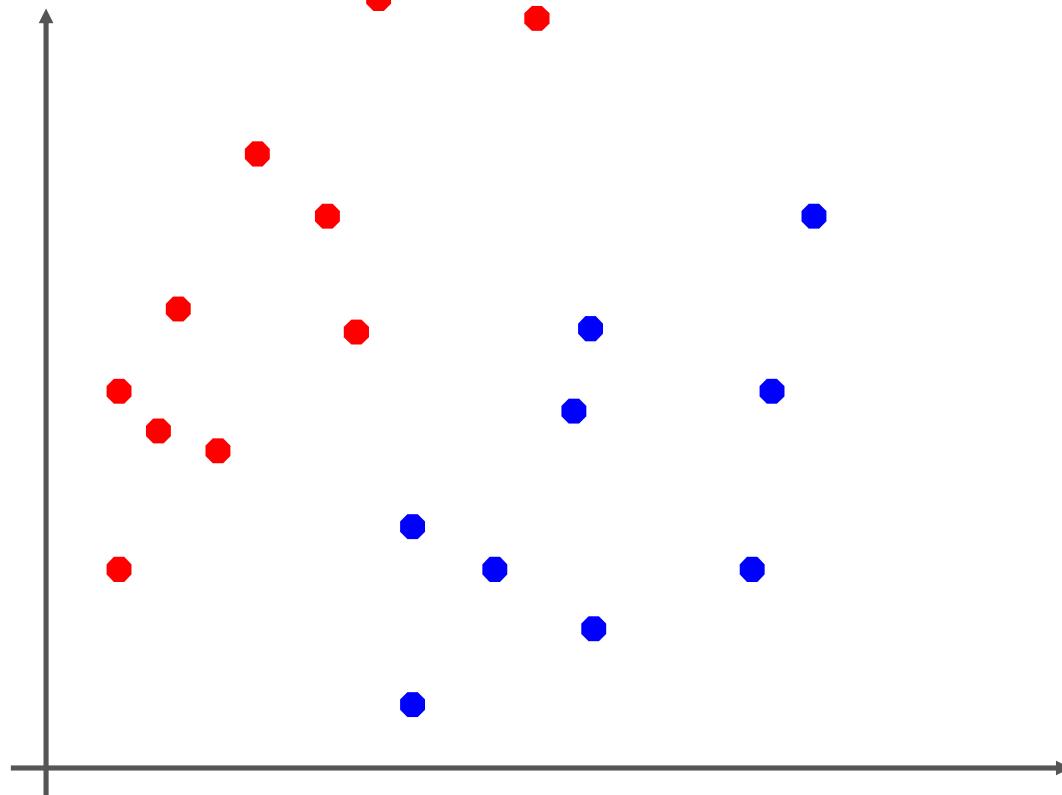
- Term Frequency (TF, TF-IDF)
- Einfach, aber robust
- Basis für viele Algorithmen
(Retrieval, Klassifikation, Topic Modeling)

Nachteile

- Stark vereinfachendes Sprachmodell
- Syntaktische und relationale Informationen gehen verloren

Verbesserung durch n-Gramme

Klassifikation mit SVM

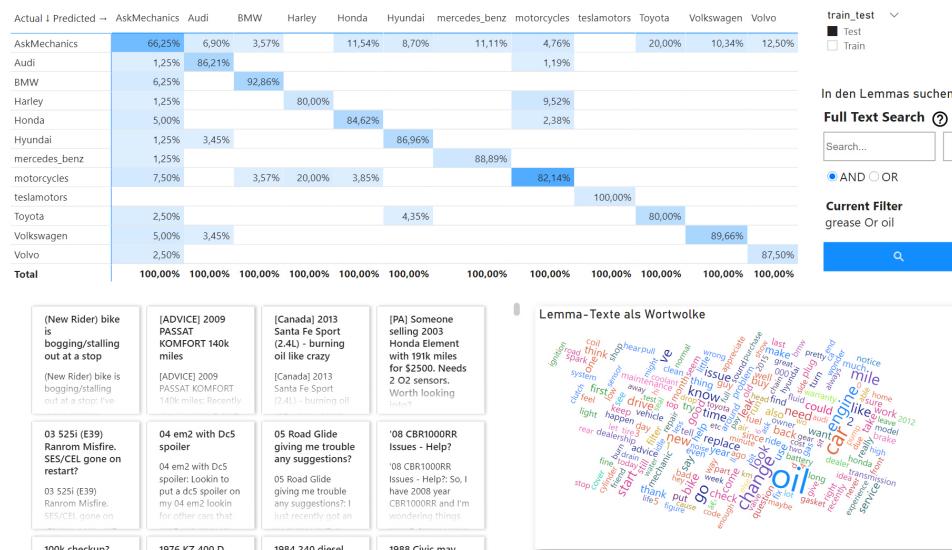


Evaluation mit Power BI

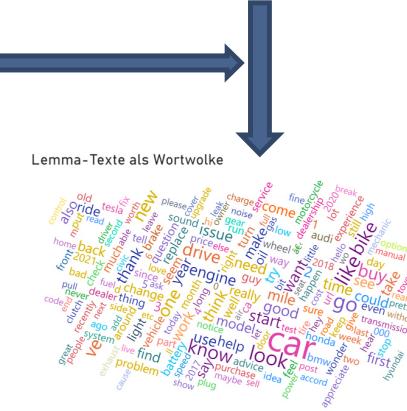
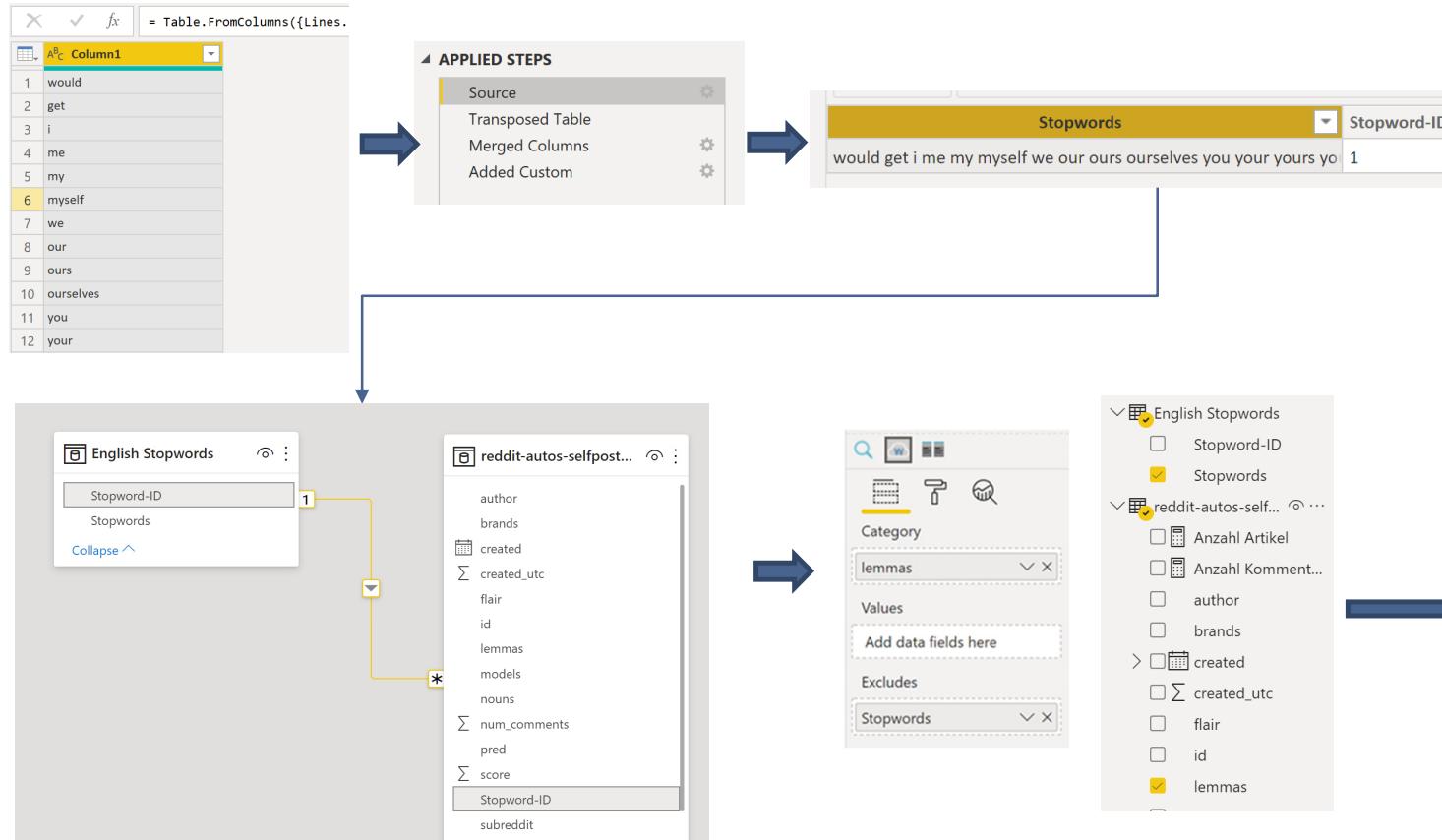
Text-Mining-Ergebnisse darstellen und inhaltlich prüfen

Aufbau einer interaktiven „Confusion Matrix“

- Matrix-Visual – erlaubt Filterung aus der Confusion-Matrix auf Texte und Wortwolken
- Anteil an Spalten (Prediction-Spalten)
→ Wieviel ist je Kategorie richtig vorhergesagt worden?
- Filter (Slicer) für Train vs. Test

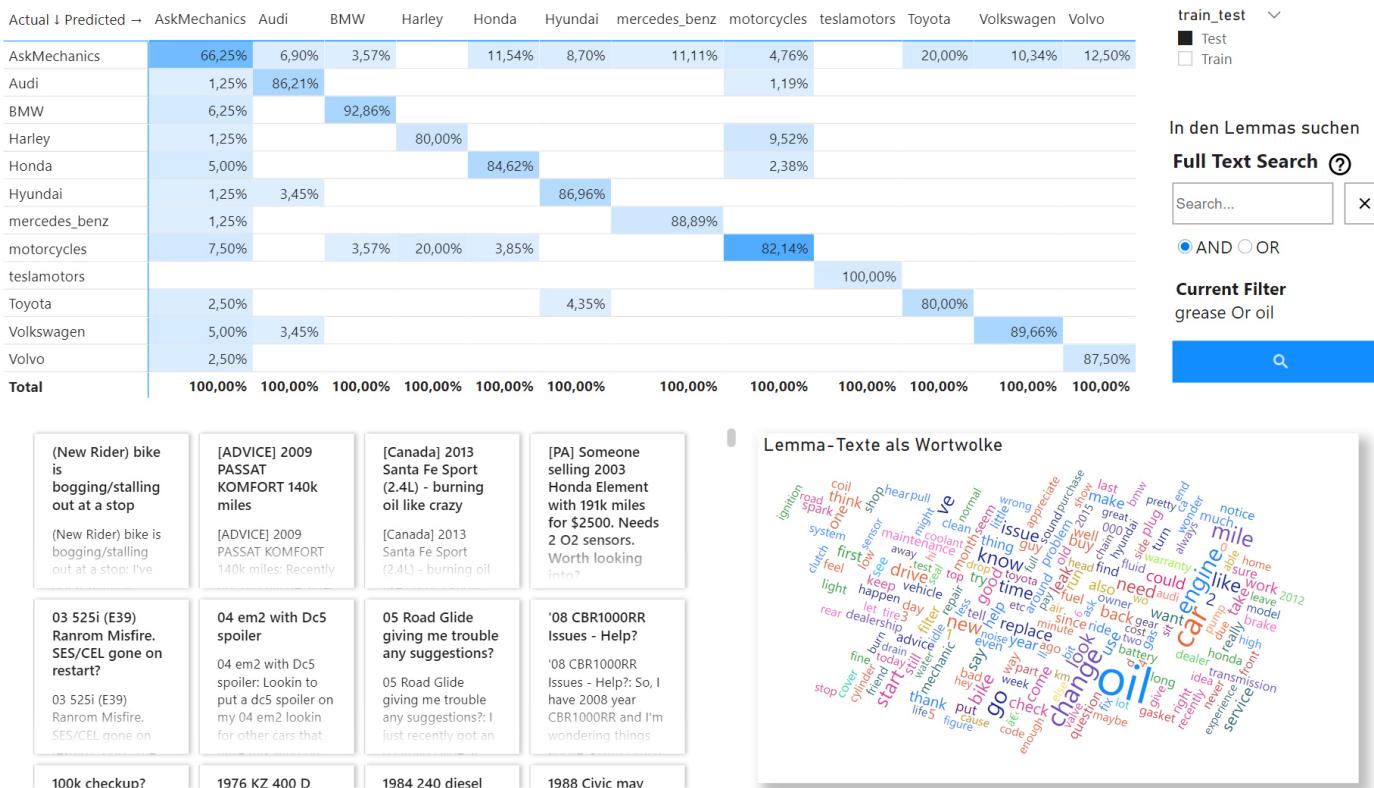


Datenmodell mit StopWort-Tabelle



Live-Demo: Übungs- und Lösungsdatei auf Github

<https://github.com/jsalbr/tdwi-2021-text-mining>



Stop-Wörter über ETL-Funktionen anbinden

Aufgabe für „Datentransformation“, d.h. Query Editor in Power BI

Stopwort-Tabelle suchen/erzeugen, z.B. von <https://gist.github.com/sebleier/554280> und in Power-BI als neue Query laden (schon vorhanden in Demo-PBIX)

1. Transpose-Transformation auf die lange Tabelle
2. Spalten zu einer Spalte verschmelzen über „Merged Columns“ mit Seperator „Space“ → damit als ein Feld über Relation an NLP-Ergebnisse joinbar → für WordCloud-Visual
3. Neue Spalte ergänzen mit fixem Wert, z.B. „1“ → für Verlinkung der Tabellen (Primärschlüssel) → gleiche Pseudospalte auch an NLP-Ergebnisse hängen mit Wert „1“ = Fremdschlüssel
4. Daten laden und Relation zwischen Tabellen über die Spalte herstellen im Model-View
5. In Word-Cloud-Visual unter „Excludes“ einfügen,

*Hinweis: Bei Ergänzung neuer Wörter in der Stopwort-Quelle müssen die dadurch erzeugten neuen Spalten in der Rohtabelle auch über die Spaltenverschmelzung berücksichtigt werden
→ ETL-Schritt ggf. löschen und neu einfügen oder über Editor ergänzen*

Explainable AI

Entscheidungen des Modells erkläbar machen

Spezielle Verfahren

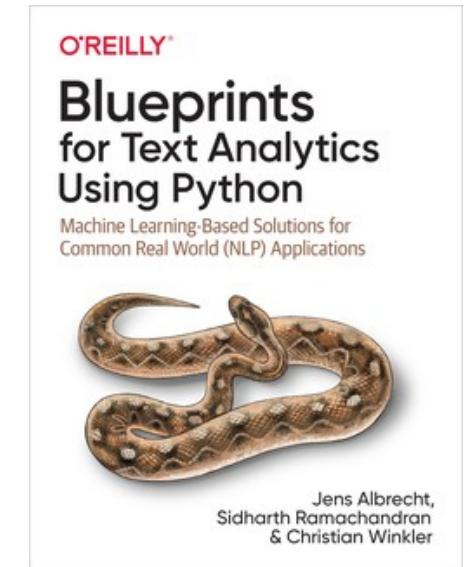
- › Lime
- › Anchor
- › SHAP

Buchhinweis: Blueprints for Text Analytics Using Python

<https://github.com/blueprints-for-text-analytics-python/blueprints-text>

Freier Download von Kapitel 7: "How to Explain a Text Classifier"

https://get.oreilly.com/ind_blueprints-for-text-analysis-using-python.html



Semantische Analysen mit Deep Learning

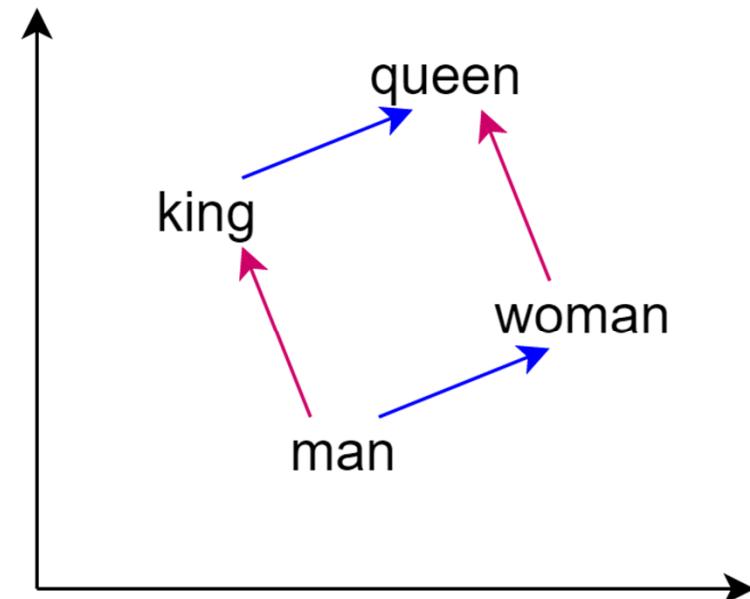
Live-Demo

<https://github.com/jsalbr/tdwi-2021-text-mining>

Bitte Notebook "Advanced" starten!

Idee von Word Embeddings

	"Royalty"	"Femininity"	"Animality"	"Livelyness"	"Housyness"
Queen	0.94	0.88	0.01	0.97	0.01
King	0.92	0.17	0.09	0.92	0.03
Woman	0.08	0.94	0.03	0.94	0.05
Cat	0.35	0.72	0.97	0.92	0.04
Dog	0.03	0.24	0.92	0.95	0.04
Palace	0.62	0.46	0.03	0.08	0.98

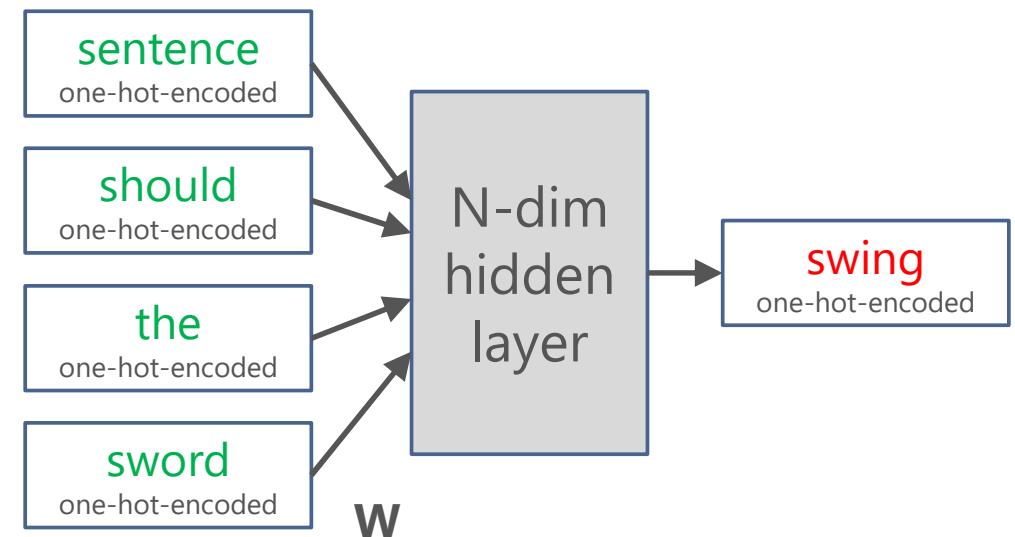


Word2Vec (Continuous Bag-of-Words)

"The man who passes the sentence should swing the sword." – Ned Stark

(Un-)Supervised Training:

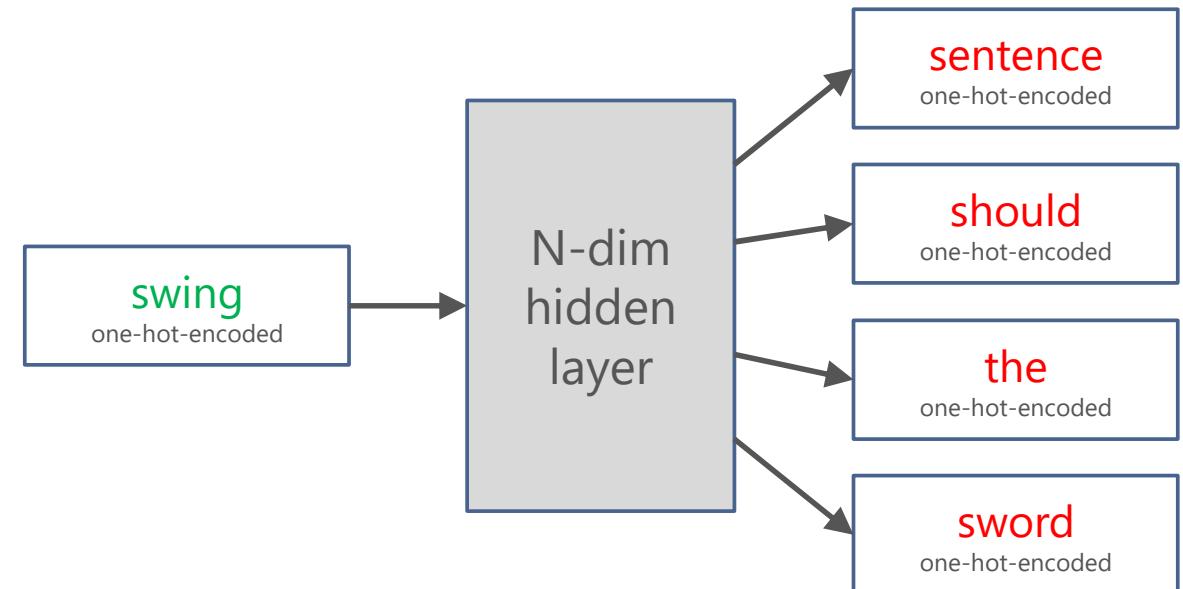
- › Vorhersage von Target-Wort (rot) zu gegebenem Kontext (grün)
- › Kontextgröße (hier 4) ist Hyperparameter
- › Reihenfolge im Kontext wird ignoriert (bag of words)
- › Zeilen der Weight-Matrix W liefern N-dim. Wort-Vektoren



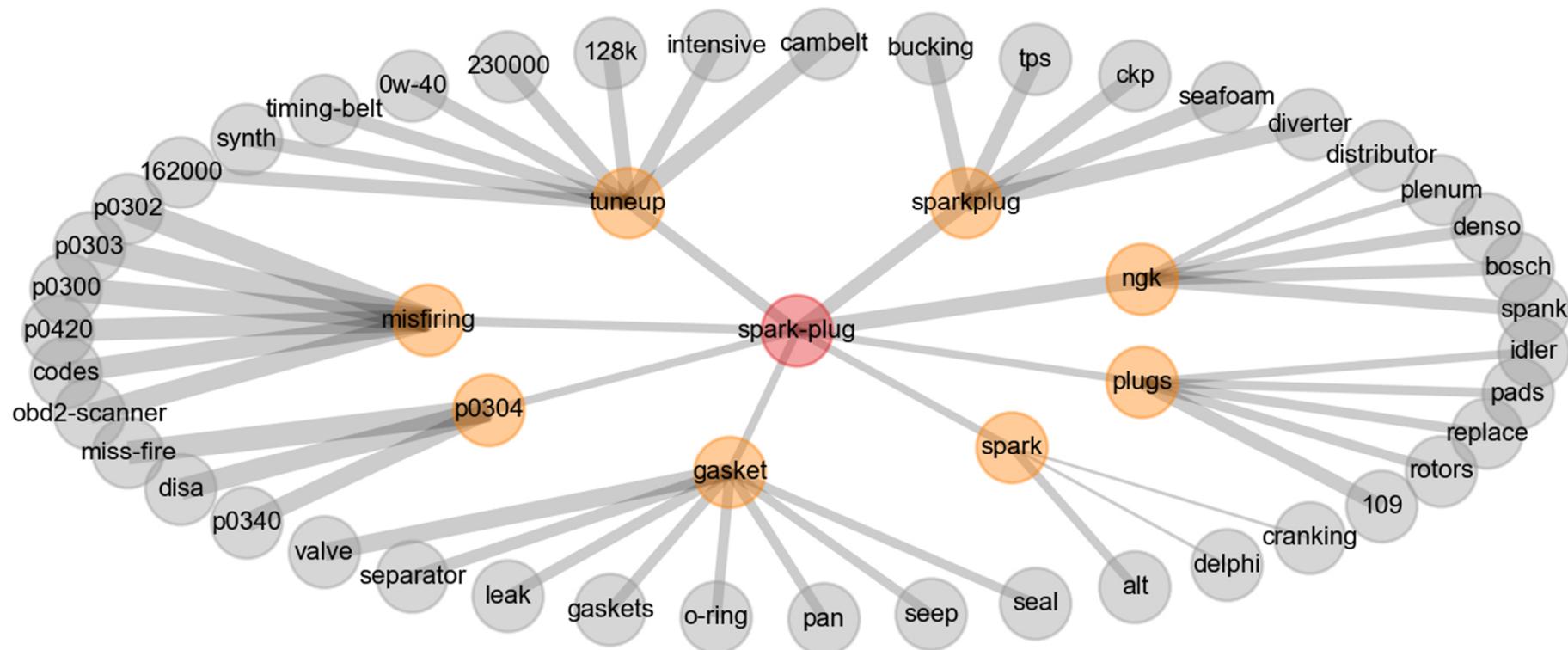
Word2Vec (Skip-Gram)

*"The man who passes the **sentence** should swing the sword." – Ned Stark*

- Vorhersage von Kontext (rot) zu gegebenem Fokus-Wort (grün)
- Trainiert langsamer als CBOW, aber i.d.R. bessere Ergebnisse

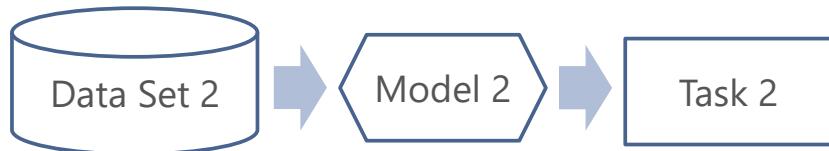
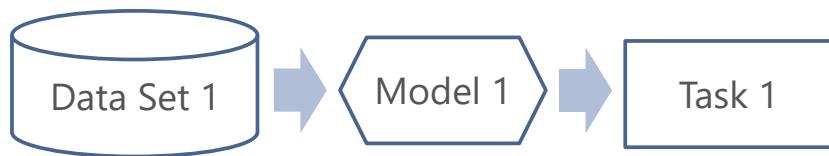


Word Embeddings zur Erschließung von Fachvokabular



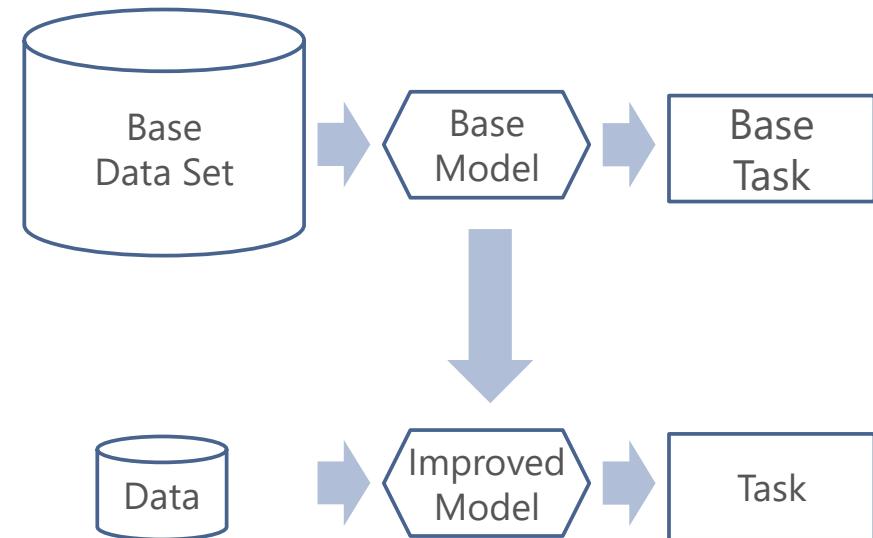
Transfer Learning mit BERT & Co.

Klassisches ML



Ein Modell wird für genau eine Aufgabe bei Null beginnend trainiert.
Es werden immer viele Trainingsdaten benötigt.

Transfer Learning



Ein Basis-Modell, dass auf einem großen Datensatz trainiert wurde, wird für eine spezifische Aufgabe angepasst.

Question Answering

```
context = """
Doktor Leonard H. McCoy (genannt „Pille“) dient 27 Jahre unter Captain James T. Kirk
als Erster Medizinischer Offizier auf der USS Enterprise (NCC-1701)
und der USS Enterprise (NCC-1701-A).

Er selbst bezeichnet sich sehr bescheiden gerne als einfacher Landarzt,
obwohl er seine Laufbahn bei der Sternenflotte als brillanter,
junger Mediziner mit magischen Händen beginnt und später bis ins
hohe Alter von 137 Jahren den Rang eines Admirals bekleidet.
Zu seinen Fachgebieten zählen: Psychologie, außerirdische Medizin und Exobiologie.
"""
```

```
predict_estimator(context, "Unter wem diente McCoy?")
```

executed in 12.6s, finished 17:42:45 2019-05-07

Running Prediction...

Predictions to Single Prediction

Question: Unter wem diente McCoy?

Answer: Captain James T. Kirk

time

12.600003957748413



Question Answering

Named Entity Recognition

Document Classification

Question Answering

Modern NLP models can understand questions in natural language and find the answer in a text passage.
Try it yourself!

Let's use the model.

Now enter your own text or use an example:

Passage

Paste your passage here

Remaining chars: 15000 / 15000

Question answering can be performed on larger corpus, this is a demo.

Ask me something
and get an answer

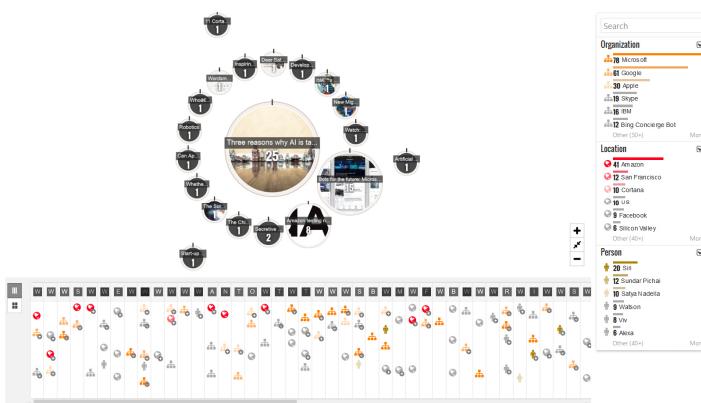
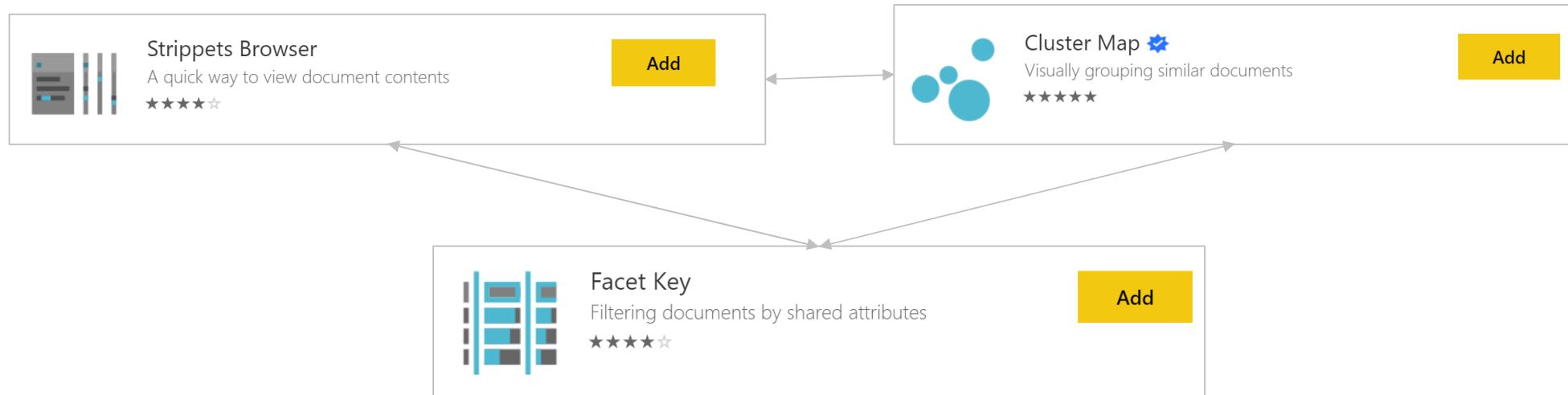
Question

Ask your question here

<https://demos.deepset.ai>

Vertiefende „Intelligence“ mit Power BI

Komplexere Visualisierungsoptionen für z.B. NER-Ergebnisse



Erläuterung des Konzeptes zum Nachlesen hier:

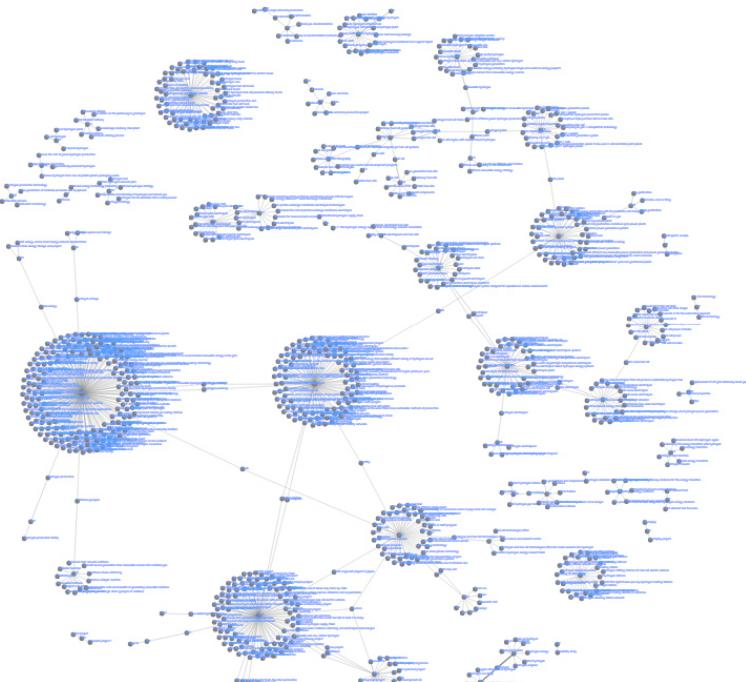
<https://powerbi.microsoft.com/de-de/blog/new-power-bi-custom-visuals-for-browsing-and-analyzing-collections-of-text/>

Live-Demo-Zugang über PowerBI-Web:

<https://app.powerbi.com/view?r=eyJrIjoiMDYwZjMwODMtZTAyOS00YWZmLWEzNTgtYTJiNjZjYjg5MDIliwidCI6jA0NmEyNDriLWZhZGQtNDA1My04OWVjLTU5Y2IxMTJiMzJhOClsImMiOjZ9>

Visualisierung von Topics und NER-Ergebnissen als Netzwerke

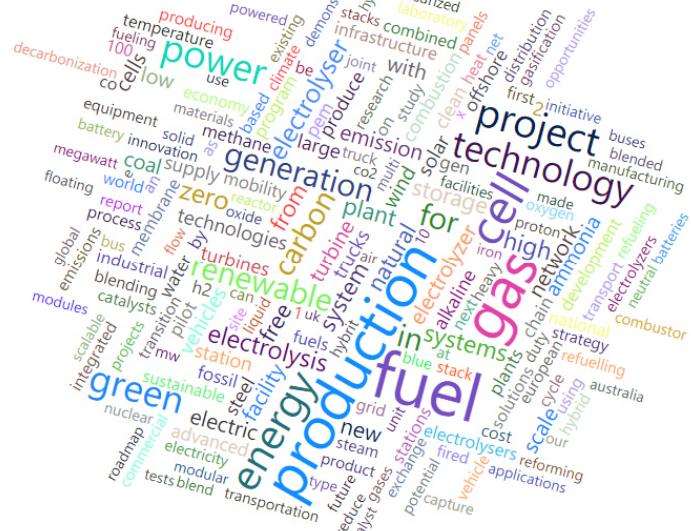
What topics are covered?



organisations
No. of Paragraphs

doe	80
itm power	80
air liquide	72
plug power	59
toyota	58
linde	56
hyundai	54
hydrogen council	44
nel	43
engie	42
siemens	41
uniper	41
arena	40
shell	39
Total	1845

What is the object of cooperation?

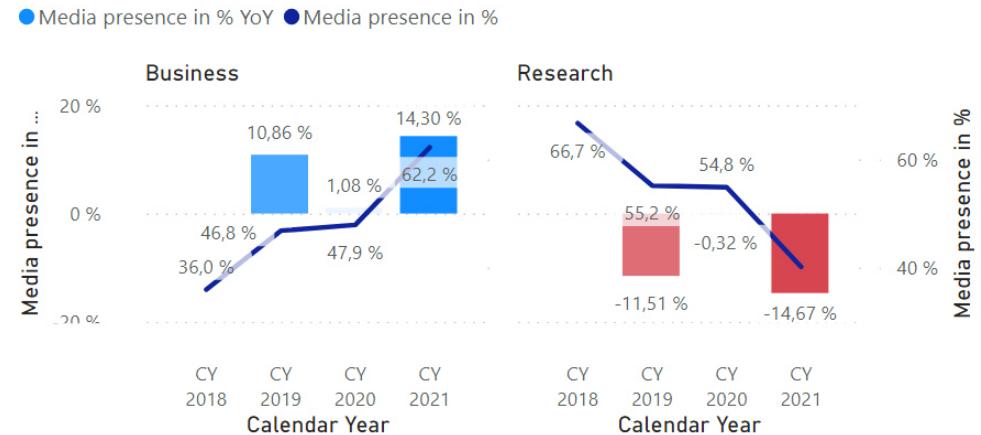


Veränderungen mit Time-Intelligence-Funktionen berechnen

Medienpräsenz und Änderung bei Topics

Calendar Year	CY 2021	No of paragraphs (coop)	Media presence in %	Media presence in % YoY
top_50_words				
production, hydrogen, gas, facility, ...	81	17,5 %	7,02 %	
hydrogen, ccus, boilers, bond, engi...	42	9,1 %	3,31 %	
zero, carbon, emission, project, free,...	37	8,0 %	1,71 %	
green, hydrogen, investment, blue, ...	28	6,0 %	2,51 %	
scale, large, project, production, exp...	23	5,0 %	-0,53 %	
hydrogen, nitrogen, polymers, socie...	22	4,8 %	-3,36 %	
natural, gas, network, hydrogen, ble...	18	3,9 %	-0,30 %	
Total	463	100,0 %	0,00 %	

Media presence in % YoY and Media presence in % by Calendar Year and Business or Research



Fazit



Prof. Dr. Jens Albrecht
TH Nürnberg, Informatik



Data Warehousing, BI,
Data Science, NLP
jens.albrecht@th-nuernberg.de
Für Beratungsprojekte:
jens.albrecht@data-knowhow.de

**Sprechen Sie uns an, für Ihre
akuten Text-Mining-Fragen!**



Prof. Dr. Roland Zimmermann
TH Nürnberg, BWL

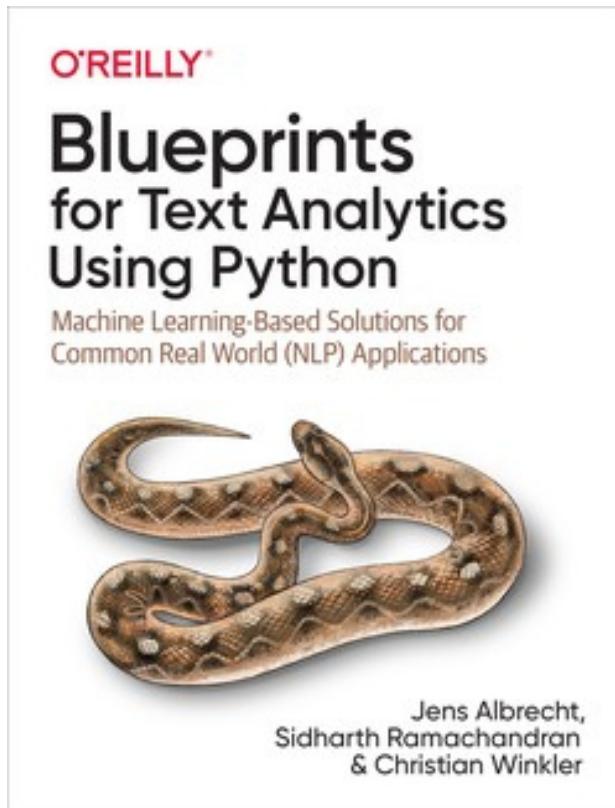


BI, Information Design,
NLP, Process Mining
roland.zimmermann@th-nuernberg.de
Für Beratungsprojekte:
zimmermann@architecting-analytics.com

Blueprints for Text Analytics Using Python

(O'Reilly, Dez. 2020)

Jens Albrecht, Sidharth Ramachandran, Christian Winkler



<https://github.com/blueprints-for-text-analytics-python/blueprints-text>

Freier Download von Kapitel 7: "How to Explain a Text Classifier"

https://get.oreilly.com/ind_blueprints-for-text-analysis-using-python.html