Mit Text Mining zum Wettbewerbsvorteil

Marco Nätlitz | *Data Scientist*

areto
CONSULTING. IT WORKS.

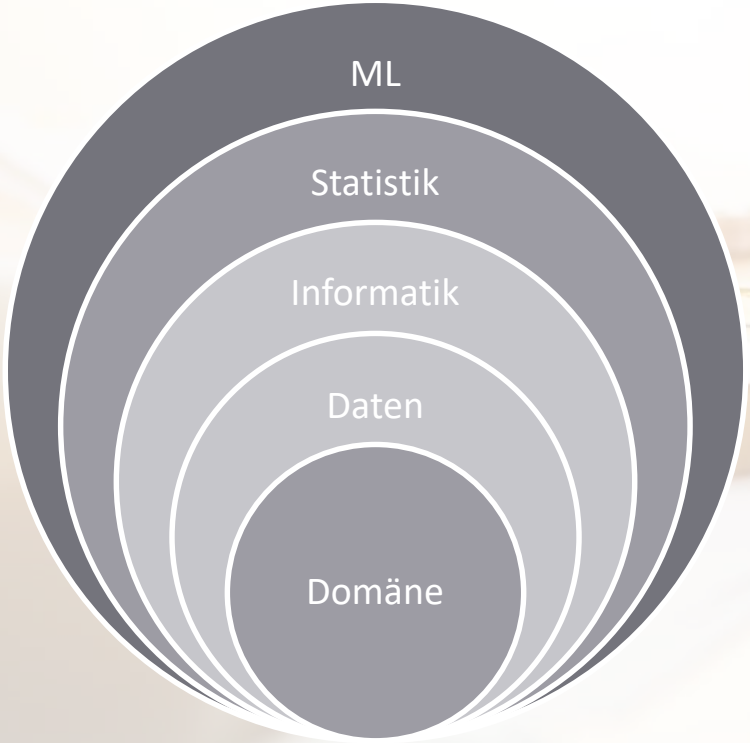
Agenda

Mit Text Mining zum Wettbewerbsvorteil

- Was ist eigentlich Data Science?
- Was ist unter Text Mining zu verstehen?
- Wie kann man aus Text Wissen extrahieren?
- Und zwar mit der Statistikumgebung R ...
- ... am Beispiel von Twitter

Was ist eigentlich Data Science?

Data Science beschreibt die Extraktion von Wissen aus Daten, Begriff geprägt durch Naur (1974)



Warum Data Science?

- knappe Ressourcen
- schnelle Märkte
- Wettbewerbsvorteile

Enabler: Text Mining

Extrahiere Wissen aus Texten

80%

Gentsch, P.; Hänlein, M.: Text Mining. In: WISU 12/99, S. 1646-1653 (1999).

Enabler: Data Science

Warum werden meine Produkte gekauft?

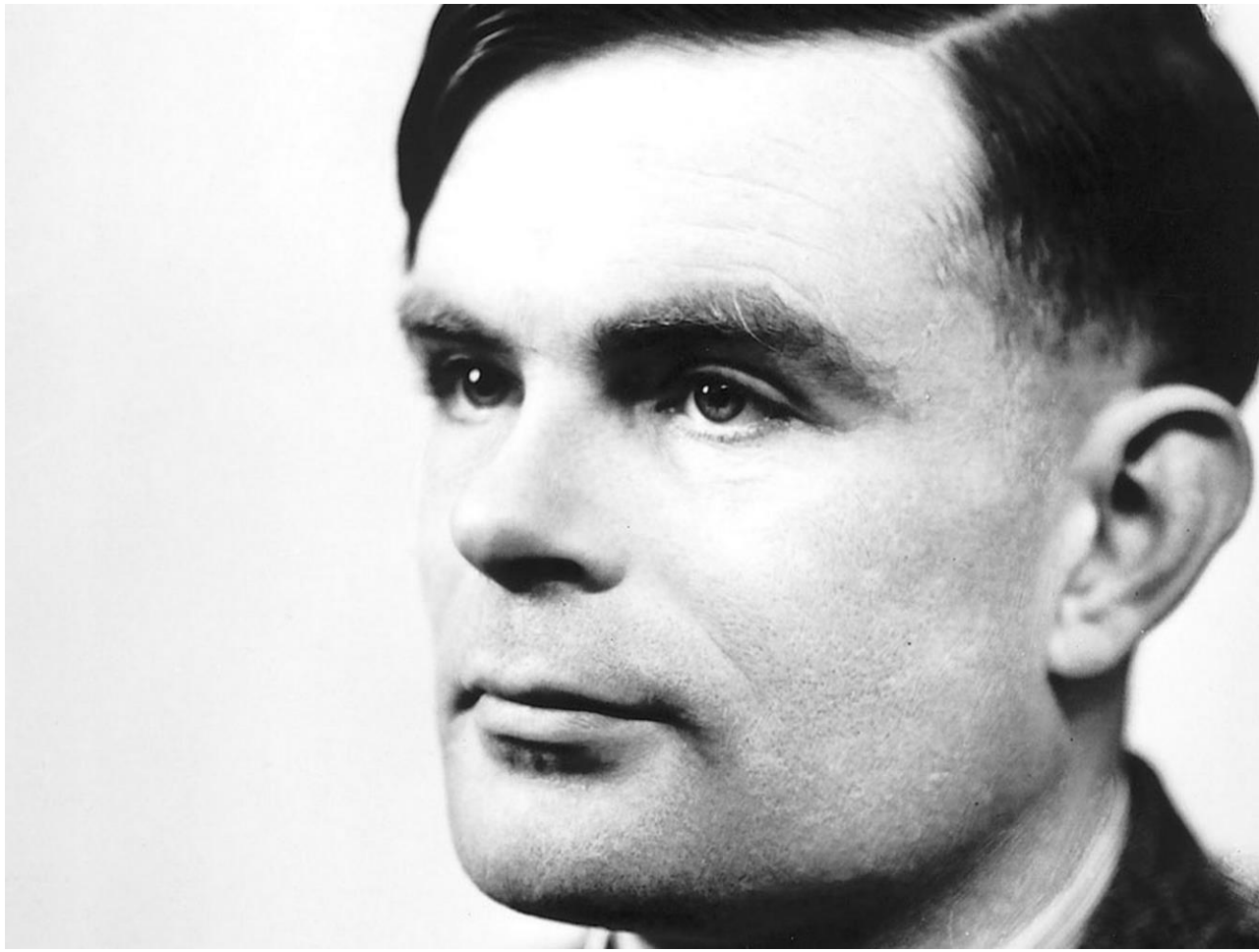
Welcher meiner PKW-Schäden sind Betrugsfälle?

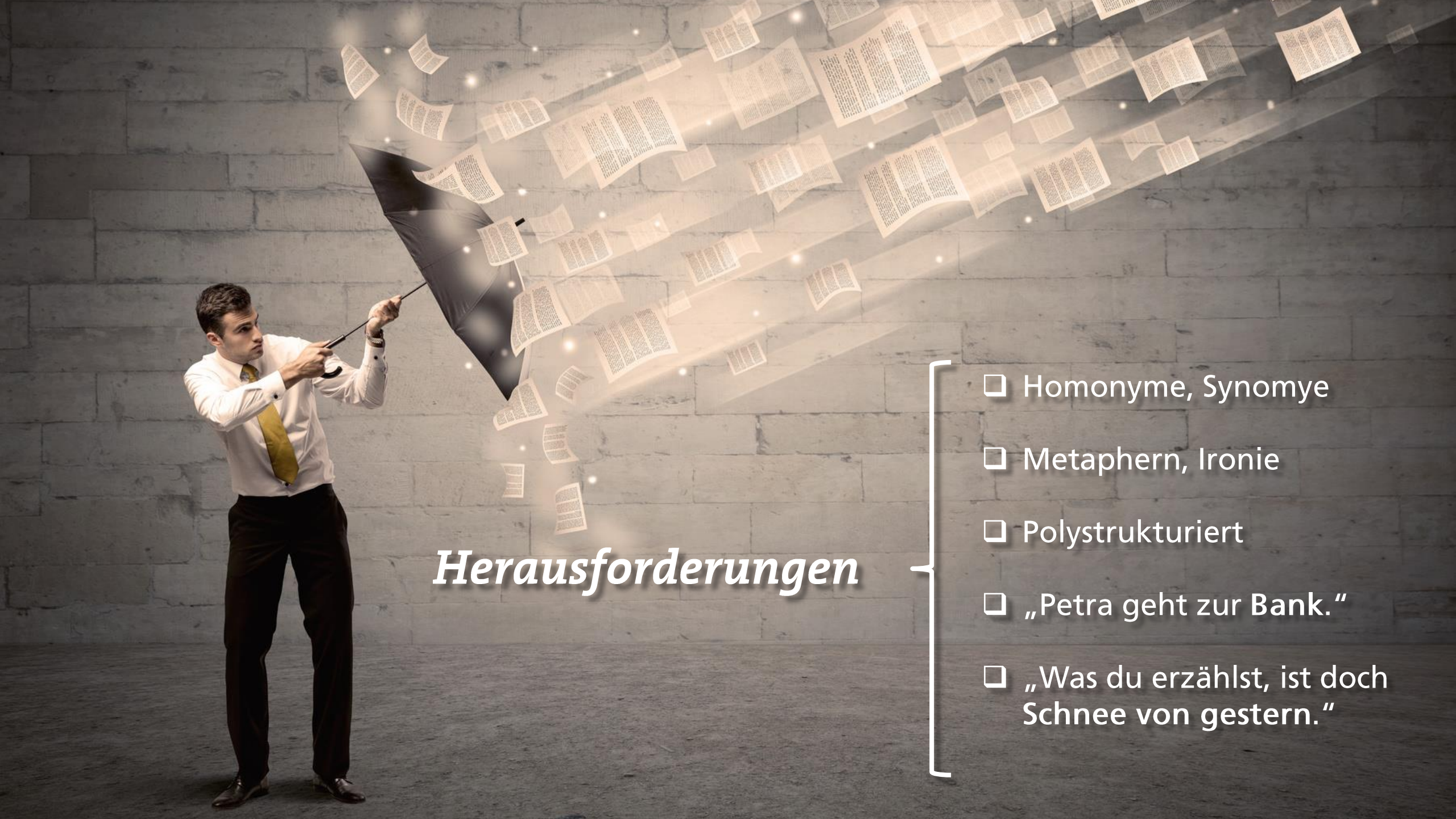
Welche Proteine interagieren miteinander?

Natural Language Processing (NLP)

Was ist unter Text Mining zu verstehen?

- Natural Language Processing (NLP)
- Extrahiere Wissen aus Daten
- Alan Turing (Mathematiker, 1912 – 1954)
- „Computer is able to impersonate a human in a real-time written conversation with a human judge"





Herausforderungen

- Homonyme, Synomye
- Metaphern, Ironie
- Polystrukturiert
- „Petra geht zur **Bank.**“
- „Was du erzählst, ist doch **Schnee von gestern.**“

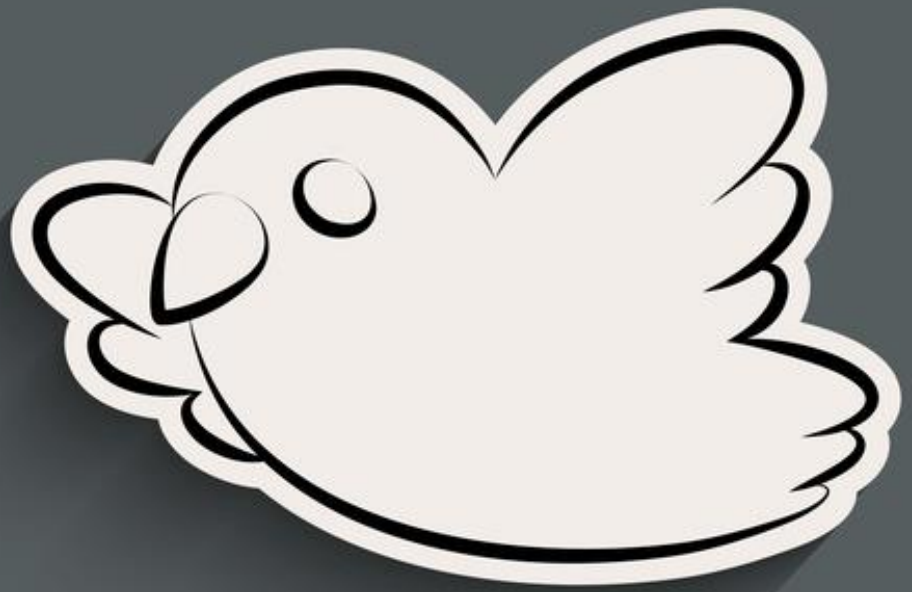
Mit Twitter zu Wettbewerbsvorteilen

Was denken
meine Kunden?

Foto: © Victor Tyakht | fotolia.de

Einnahmen / Jahr $ 665 Mio.

@katyperry hat 84.950.745 Follower



Hash Tag

320 Mio. Nutzer

2006 gegründet

140 Zeichen

twitter

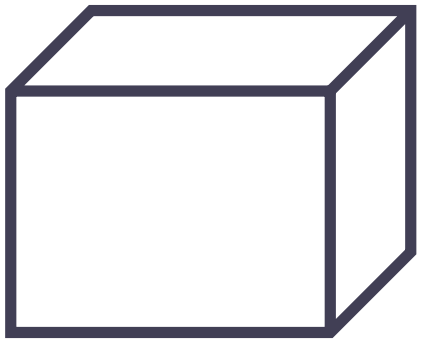
Der Weg zum Sentiment Mining-Modell

Textanalyse mit Twitter-Daten

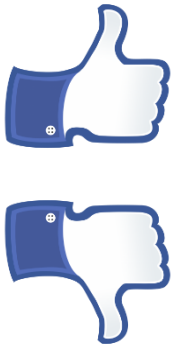
Basis: Tweets

„the iPhone 5c is so beautiful <3 @Apple“

Klassifikations-modell



Ist der Tweet positiv oder negativ?



	TWEET
75	@Apple Is Testing #Technology w/ 'M7' Chip To Help You Figure Out Where You Pa...
76	I have to say, Apple has by far the best customer care service I have ever rec...
77	iOS 7 is so fricking smooth & beautiful!! #ThanxApple @Apple
78	LOVE U @APPLE
79	Thank you @apple, loving my new iPhone 5S!!!!! #apple #iphone5S pic.twitter.c...
80	.@apple has the best customer service. In and out with a new phone in under 10...
81	@apple ear pods are AMAZING! Best sound from in-ear headphones I've ever had!
82	Omg the iPhone 5S is so cool it can read your finger print to unlock your iPho...
83	the iPhone 5c is so beautiful <3 @Apple
84	#AttributeOwnership is exactly why @apple will always be #one! #apple #marketi...
85	Just checked out the specs on the new iOS 7...wow is all I have to say! I can'...
86	I love the new iOS so much!!!!! Thnx @apple @phillydvibing
87	Can't wait to get my #Iphone5S!!! @apple
88	@V2vista Fingerprint scanner: The killer feature of iPhone 5S. This is so bloo...
89	Interesting how so many people seem to be almost willing the demise of @Appl...
90	I LOVE @BNBuzz @NOOKstudy @nookBN and @apple made my life so much easier this ...
91	Just watched the keynote of @apple latest iPhones. I just love the #iPhone5S a...
92	My iPhone wasn't calling correctly so I went to an @apple store (first time) t...
93	Great job @apple on providing best users experience #thinkauto
94	Swapped my #galaxys2 for an #iPhone4S. After one day I'd say I'm an @apple con...
95	Can't wait for my #orange phone upgrade in November :-) #apple iPhone 5s here ...
96	#colored #iphone! The new 5C iphone comes in colors. I love this!! Can't wait ...
97	Whether you're an @apple fan or not, this iPhone 5C video is worth watching, f...
98	@apple Impressive features on the Iphone 5s - fingerprint recognition, now tha...
99	@KuqoGroup @jimmykimmel @BlackBerry @Apple that is unbelievably awesome!!!!! #...
100	EarPods are amazing, thanks @Apple
101	In luv with the iPhone 5s luv the champagne colour and the fingerprint reader ...
102	Used @Apple TV and @explainevrythng to demonstrate student understanding of fa...
103	@brody_knibbs haha mint arent they! top company @apple
104	I'm back yaw, Thanks to @Apple !!!!!

1181 Tweets

Extraktion einer Datenbasis

Der Weg zum Sentiment Mining-Modell

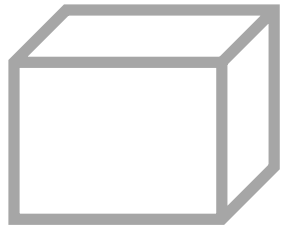
Textanalyse mit Twitter-Daten

Basis:
Tweets

Zielvariable:
Sentiment

Unabhängige
Variablen

Klassifikations-
modell



Amazon Mechanical Turk

Manche Aufgaben lassen sich nur mit menschlicher Intelligenz lösen

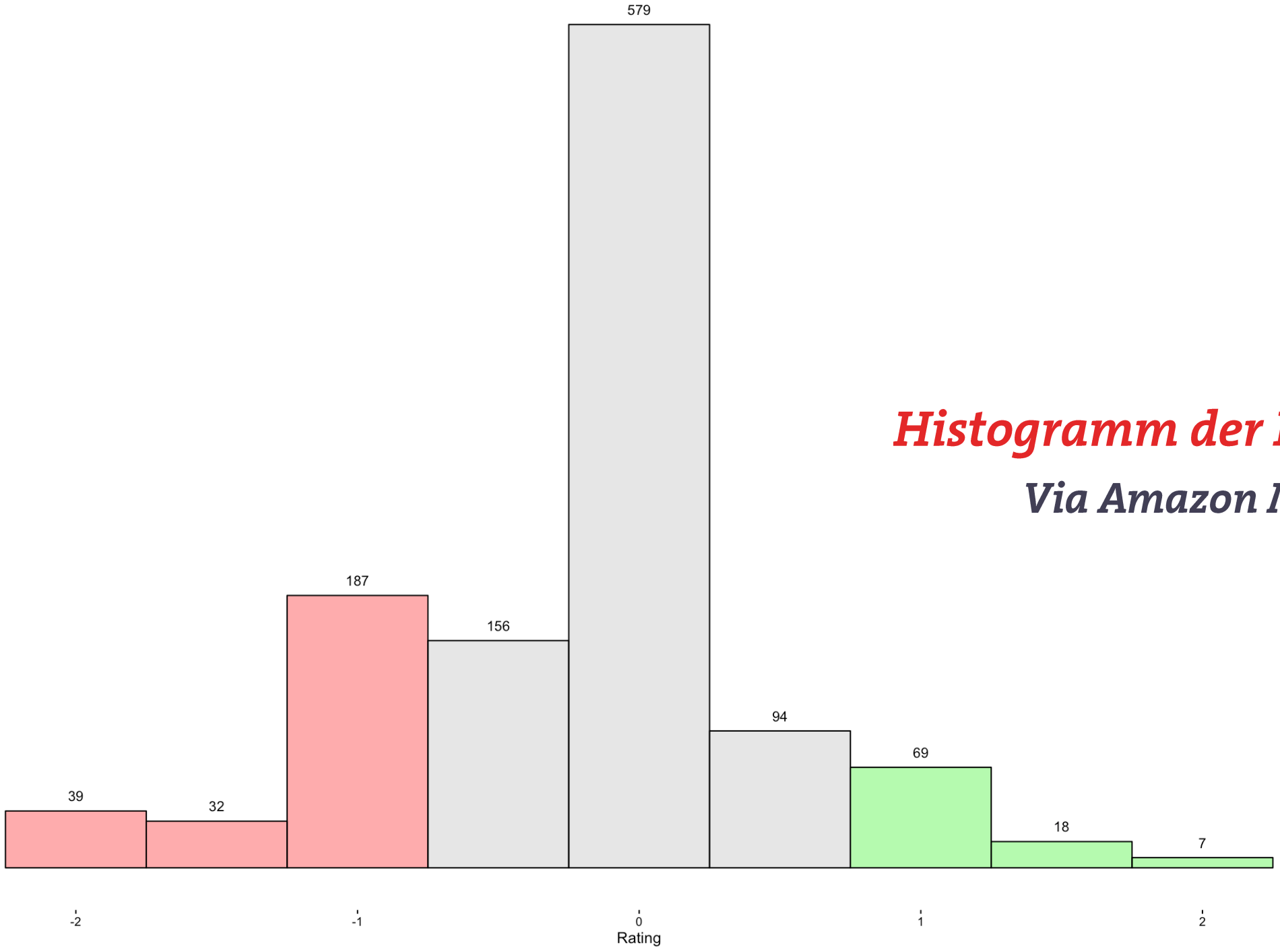
- Crowdsourcing Internet Marketplace
- Tätigkeiten auf Zeit
- Ohne feste Anstellung
- Tätigkeit nicht durch Maschinen machbar
- Sog. Human Intelligence Tasks (HIT)
- 500.000 Auftragssuchende
- In 19 Ländern

Via https://www.mturk.com/mturk/welcome

Beurteilen Sie die Einstellung des Verfassers des folgenden Tweets gegenüber Apple Inc.

Histogramm der Beurteilungen

Via Amazon Mechanical Turk



	TWEET	AVG
1	I have to say, Apple has by far the best customer care service I have ever receiv...	2
2	iOS 7 is so fricking smooth & beautiful!! #ThanxApple @Apple	2
3	LOVE U @APPLE	1,8
4	Thank you @apple, loving my new iPhone 5S!!!!! #apple #iphone5S pic.twitter.com/...	1,8
5	.@apple has the best customer service. In and out with a new phone in under 10min!	1,8
6	@apple ear pods are AMAZING! Best sound from in-ear headphones I've ever had!	1,8
7	Omg the iPhone 5S is so cool it can read your finger print to unlock your iPhone ...	1,8
8	Can't wait to get my #Iphone5S!!! @apple	1,6
9	I love the new iOS so much!!!!! Thnx @apple @phillydvibing	1,6
10	Just checked out the specs on the new iOS 7...wow is all I have to say! I can't w...	1,6
11	#AttributeOwnership is exactly why @apple will always be #one! #apple #marketing ...	1,6
12	the iPhone 5c is so beautiful <3 @Apple	1,6
13	@V2vista Fingerprint scanner: The killer feature of iPhone 5S. This is so bloody ...	1,6
14	Just watched the keynote of @apple latest iPhones. I just love the #iPhone5S and ...	1,6
15	I LOVE @BNBuzz @NOOKstudy @nookBN and @apple made my life so much easier this mor...	1,6
16	Interesting how so many people seem to be almost willing the demise of @Apple. ...	1,6
17	My iPhone wasn't calling correctly so I went to an @apple store (first time) told...	1,4
18	Great job @apple on providing best users experience #thinkauto	1,4
19	Swapped my #galaxys2 for an #iPhone4S. After one day I'd say I'm an @apple convert!	1,4
20	Can't wait for my #orange phone upgrade in November :-) #apple iPhone 5s here I c...	1,4
21	EarPods are amazing, thanks @Apple	1,4
22	@KuqoGroup @jimmykimmel @BlackBerry @Apple that is unbelievably awesome!!!!! #cli...	1,4
23	@apple Impressive features on the Iphone 5s - fingerprint recognition, now thats ...	1,4
24	Whether you're an @apple fan or not, this iPhone 5C video is worth watching, for ...	1,4
25	#colored #iphone! The new 5C iphone comes in colors. I love this!! Can't wait til...	1,4

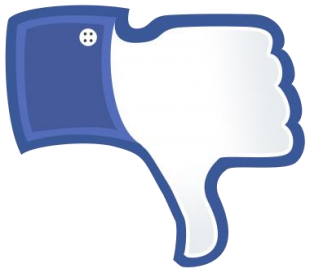
Ø-Rating pro Tweet

Extraktion einer Datenbasis

	TWEET	AVG
1	iphone 5c is ugly as heck what the freak @apple @iphonecompanies	-2
2	@aGounalakis that's nasty! @apple is a nasty brat	-2
3	freak you @apple	-2
4	@APPLE YOU RUINED MY LIFE	-2
5	@apple I hate apple!!!!!	-2
6	@apple worst customer service ever.	-2
7	@Apple YOU COW	-2
8	you guys are cheep little freaks and I hope u get testicular cancer and loose a nut @apple	-2
9	We should boycott @Apple or freakin flame them or something like how they have all this money but can't ...	-2
10	wtf @telstra @apple why would you have pre-order for the 5c the crap phone no one wants and not the 5s i...	-2
11	Hate you @apple	-2
12	freak u @apple	-2
13	freak @apple	-2
14	WHY CANT I freakING SEE PICTURES ON MY TL IM ANNOYED freak YOU @TWITTER @APPLE	-2
15	@APPLE YOU freakING COWS freak YOU	-2
16	@apple I hate you why is my phone not working I'm going to freak out	-2
17	freak YOU @APPLE	-2
18	@verizon @apple I HATE y'all	-1,8
19	I freaking hate you @apple	-1,8
20	freak you @apple my battery is stuff, you should be ashamed, that is all	-1,8
21	They beasted RT @Apple: go freak yourself "@RespectRoyalty5: Apple needs to get that bum butt iPhone 5c ...	-1,8
22	@Apple is trying to force me to update the iOS by crashing every time I try editing a photo and acting l...	-1,8
23	freak you for not having a 5S preorder, @Apple.	-1,8
24	pictures on here won't load freak you @twitter @apple	-1,8
25	freak you @apple @siri	-1,8
26	And this is a reason why I despise @Apple and their "Innovations". We didn't like them then OR NOW pic.t...	-1,8
27	freak you @Apple. Hoping your company goes bankrupt. Yours sincerely, an envious yet loyal customer of e...	-1,8
28	I hate how my phone won't focus when I take a picture with flash wow frick off @apple @iphone	-1,8

Ø-Rating pro Tweet

Extraktion einer Datenbasis

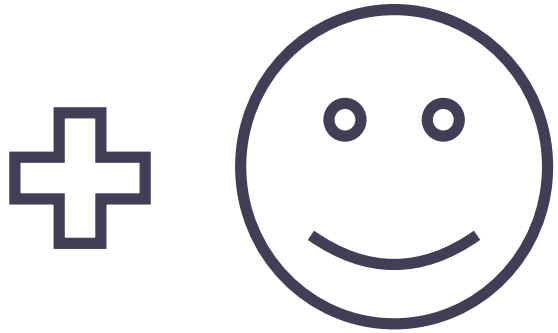


Der Weg zum Sentiment Mining-Modell

Textanalyse mit Twitter-Daten

Basis: Tweets

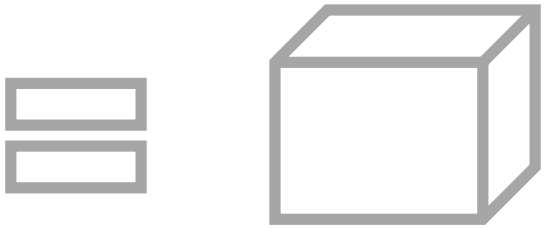
Zielvariable: Sentiment



Unabhängige Variablen

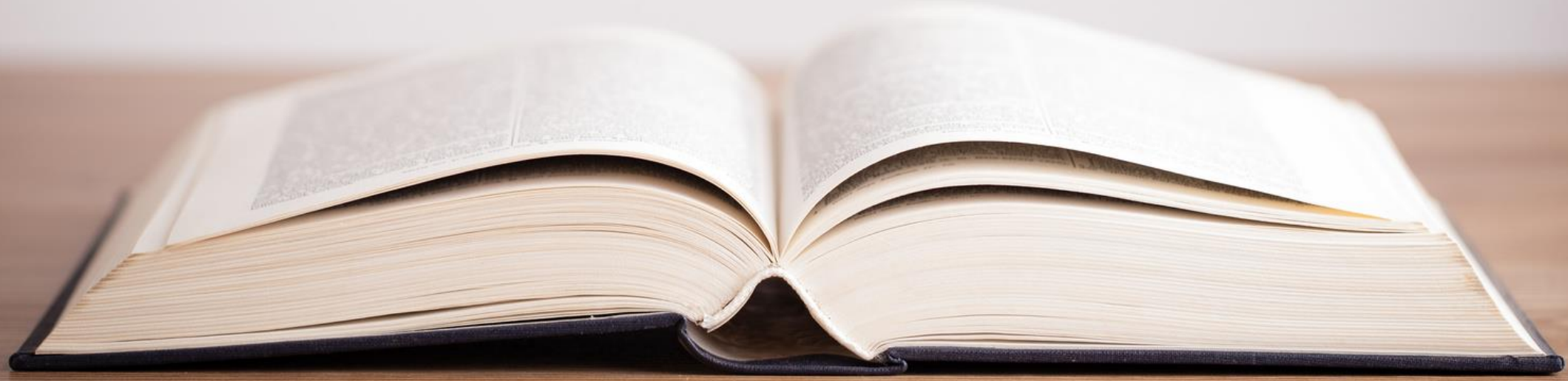
?

Klassifikations-modell



Der „Bag Of Words“-Ansatz

- Unstrukturierter Text wird zu strukturierten Informationen transformiert
- Bereinigung der Groß- und Kleinschreibung, Zeichen, Füllwörter
- Stammformreduktion
- Zähle die Anzahl der Stammformen



Bereinigung von Groß- und Kleinschreibung

Der „Bag Of Words“-Ansatz

1

Omg! the iPhone 5S is so COOL it can read your finger print to unlock your iPhone 5S and to make purchases without a passcode. #Apple @Apple

Kleinbuchstaben

2

omg! the iphone 5s is so cool it can read your finger print to unlock your iphone 5s and to make purchases without a passcode. #apple @apple

Bereinigung der Zeichensetzung

Der „Bag Of Words“-Ansatz

2

omg! the iphone 5s is so cool it can read your finger print to unlock your iphone 5s and to make purchases without a passcode. #apple @apple

Zeichensetzung entfernen

3

omg the iphone 5s is so cool it can read your finger print to unlock your iphone 5s and to make purchases without a passcode apple apple

Entfernung von sogenannten Stoppworten

Der „Bag Of Words“-Ansatz

3

omg the iphone 5s is so cool it can read your finger print to unlock your iphone 5s and to make purchases without a passcode apple apple

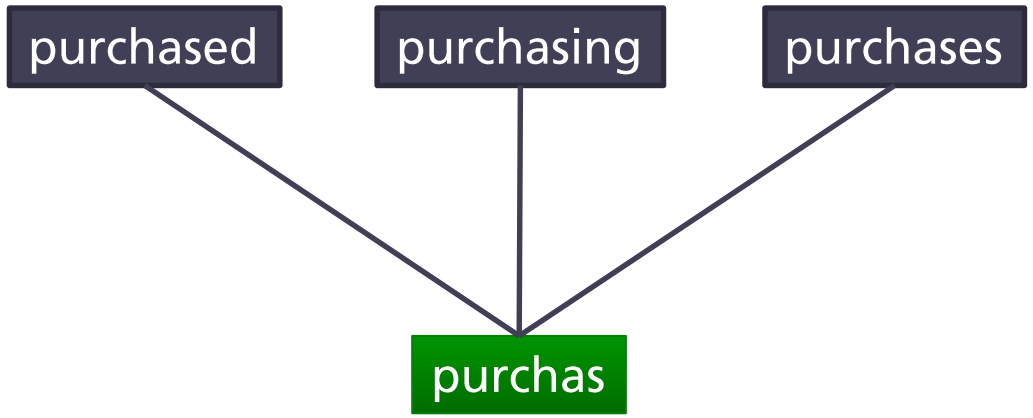
Entfernung von Artikeln, Konjunktionen, Präpositionen, Füllwörtern, etc.

4

omg iphone 5s cool can read finger print unlock iphone 5s make purchases without passcode

Stammformreduktion mit dem Porter-Stemmer-Algorithmus

Der „Bag Of Words“-Ansatz



Porter-Stemmer-Algorithmus:

- Anwendung von Verkürzungsregeln
- Solange Minimalanzahl von Silben erreicht ist

Via http://tartarus.org/martin/index.html und http://tartarus.org/martin/PorterStemmer/

„Stemming“ → Stammformreduktion

Der „Bag Of Words“-Ansatz

4

omg iphone 5s cool can read finger print unlock iphone 5s make purchases without passcode

Stammformreduktion

5

omg iphon 5s cool can read finger print unlock iphon 5s make purchas without passcod

Textanalyse mit R

Der „Bag Of Words“-Ansatz

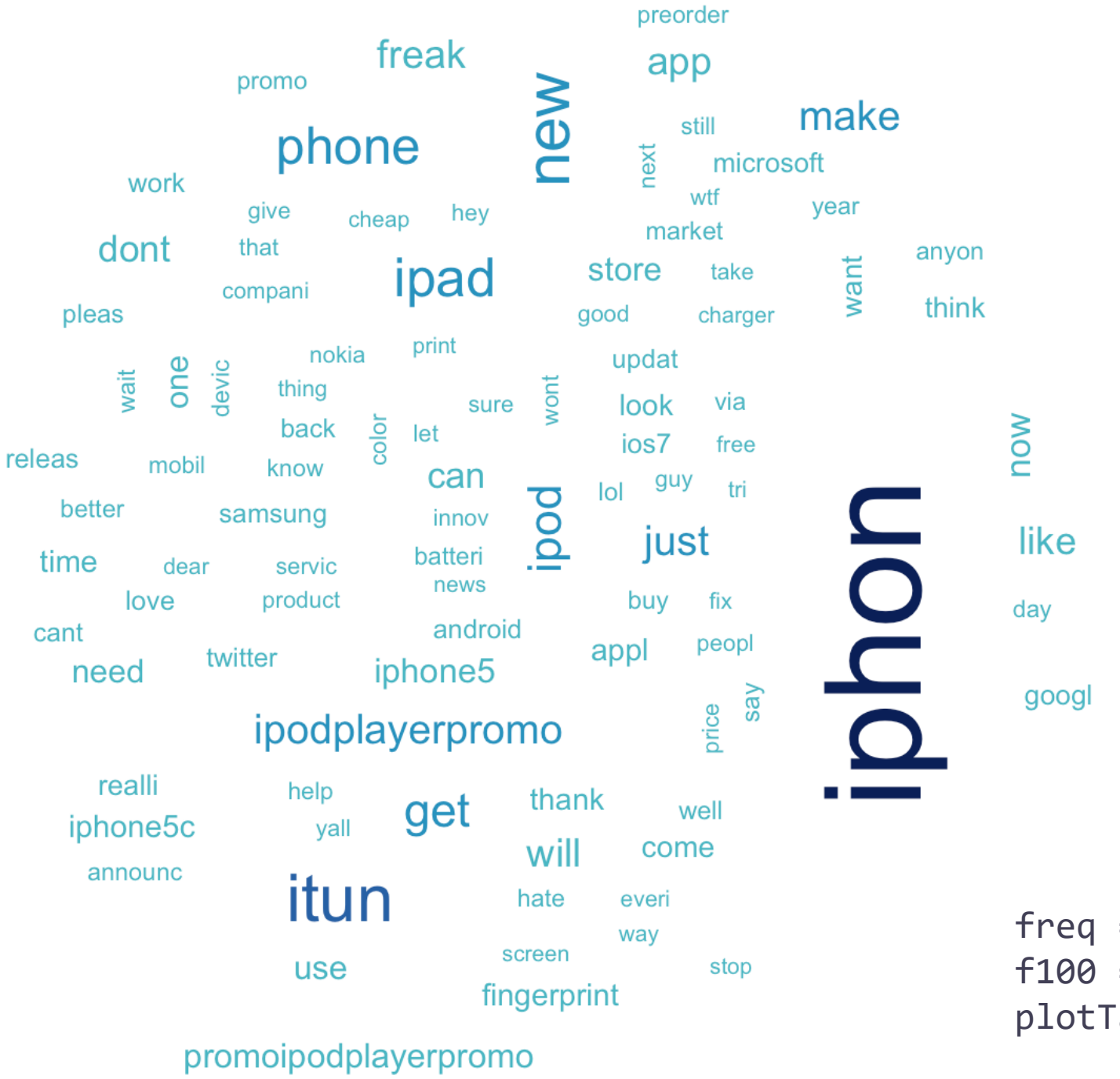
```
library(tm)
library(SnowballC)

tweets = read.csv("tweets_and_rating.csv")

corpus = Corpus(VectorSource(tweets$Tweet))

corpus = tm_map(corpus, tolower)
corpus = tm_map(corpus, PlainTextDocument)
corpus = tm_map(corpus, removePunctuation)
corpus = tm_map(corpus, removeWords, c("apple", stopwords("english")))
corpus = tm_map(corpus, stemDocument)

corpus[[834]]$content  → "omg  iphon 5s   cool  can read  finger print  unlock
                          iphon  5s   make purchas without  passcod"
```



Die 100 häufigsten Terme

Der „Bag Of Words“-Ansatz

```
freq = DocumentTermMatrix(corpus)
f100 = findFreqTerms(freq, lowfreq=100)
plotTagCloud(f100)
```

Zähle die Stammformen je Tweet

Der „Bag Of Words“-Ansatz

1

omg iphon 5s cool can read
finger print unlock iphon 5s
make purchas without passcod

2

want finger print nsa

	iphon	finger	print	omg	want	…
Summe	2	2	2	1	1	…
Tweet 1	2	1	1	1	0	…
Tweet 2	0	1	1	0	1	…
…	…	…	…	…	…	…

	TWEET_ID	IPHON	LIKE	LOVE	HATE	PHONE	ITUN	WANT
1	16	1	0	1	0	0	0	0
2	32	1	0	0	0	0	0	0
3	1	0	0	0	0	0	0	0
4	64	0	0	1	0	0	0	0
5	109	0	0	0	0	0	0	0
6	113	0	0	1	0	0	0	0
7	94	0	0	0	0	0	0	0
8	105	1	0	0	0	0	0	0
9	120	0	0	0	0	0	0	0
10	147	0	0	0	0	0	0	0
11	168	0	0	0	0	0	0	0
12	172	0	1	0	0	0	0	0
13	203	1	1	0	0	0	0	0
14	215	0	0	0	0	0	0	0
15	226	0	0	0	0	0	0	0
16	309	0	0	0	0	0	0	0
17	313	0	0	0	0	0	2	0
18	255	0	0	0	0	0	3	0
19	275	0	0	0	0	0	0	0
20	278	0	1	0	0	0	0	0
21	308	0	0	0	0	0	0	0
22	371	0	0	0	0	0	0	0
23	333	0	0	0	0	0	0	0
24	410	1	0	0	0	0	0	0
25	425	1	0	0	0	0	0	0
26	427	1	0	0	0	0	0	0
27	432	0	0	0	0	0	0	0

Stammformen → Variablen

Die Anzahl zum Wert der Variablen

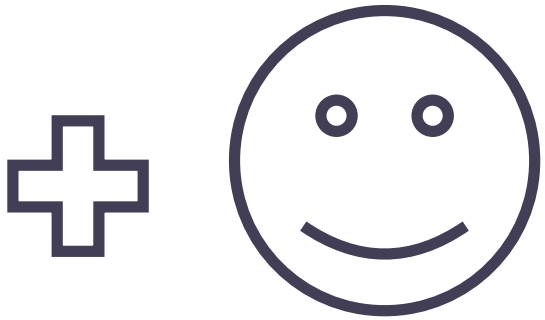
```
sparseDTM = removeSparseTerms(frequencies, 0.995)
     sparse = as.data.frame(as.matrix(sparseDTM))
  colnames(sparse) = make.names(colnames(sparse))
```

Der Weg zum Sentiment Mining-Modell

Textanalyse mit Twitter-Daten

Basis: Tweets

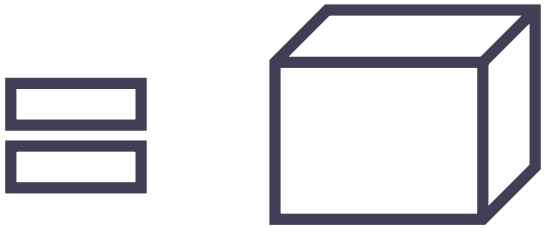
Zielvariable: Sentiment



Unabhängige Variablen

Bag Of Words

Klassifikations-modell



Vereinfachung der Zielvariablen

Aus Zahlen wird eine binäre Entscheidung

LOVE U @APPLE → Rating = 2,0 → “Positive”

Rating ≤ -1 → “Negative”

I hate how my phone won’t focus when I take a picture with flash @Apple! → Rating = -1,8 → “Negative”

	TWEET_ID	RATING	IPHON	LIKE	LOVE	HATE	PHONE	ITUN	WANT
1	777	Positive	2	0	0	0	0	0	
2	1163	Negative	1	0	0	1	0	0	
3	1140	Negative	1	0	0	0	0	0	
4	1110	Negative	1	0	0	0	0	0	
5	1077	Negative	1	0	0	0	0	0	
6	1035	Negative	1	0	0	0	0	0	
7	795	Positive	1	0	0	0	0	0	
8	557	Positive	1	0	0	0	0	0	
9	292	Positive	1	0	0	0	0	0	
10	169	Positive	1	0	0	0	0	0	
11	439	Positive	0	0	0	0	0	0	
12	107	Positive	0	0	0	0	0	0	
13	544	Positive	0	0	0	0	0	0	
14	1157	Negative	0	1	0	0	0	0	
15	569	Positive	0	1	0	0	0	0	
16	622	Positive	0	0	0	0	0	0	
17	672	Positive	0	0	0	0	0	0	
18	700	Positive	0	0	0	0	0	0	
19	703	Positive	0	0	0	0	0	0	
20	1123	Negative	0	0	0	1	0	0	
21	115	Positive	0	0	0	0	0	0	
22	886	Positive	0	0	0	0	0	0	
23	928	Positive	0	0	0	0	0	0	
24	899	Positive	0	0	0	0	0	0	
25	901	Positive	0	0	0	0	1	0	
26	946	Positive	0	0	0	0	0	0	
27	965	Positive	0	0	0	0	0	0	
28	976	Positive	0	0	0	0	0	1	
29	1017	Negative	0	0	0	0	0	0	
30	1063	Negative	0	0	0	0	0	0	
31	1174	Negative	0	0	0	1	1	0	

Binäre Zielvariable

Bildung des Klassifikationsmodells

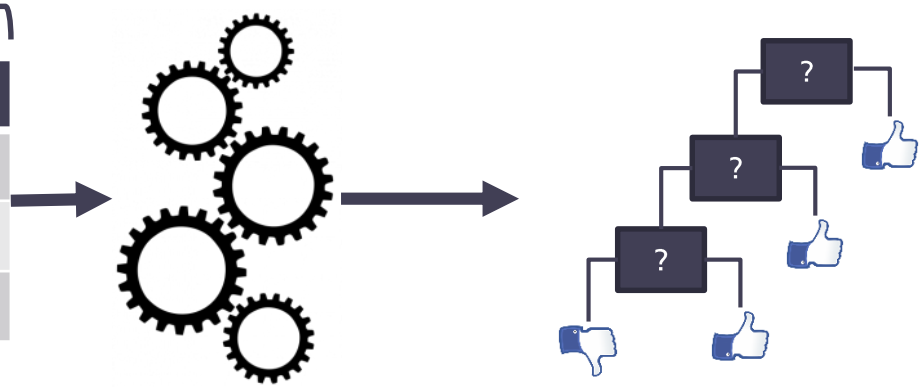
```
                                    sparse$Rating = 'Negative'
sparse[tweets$Avg >= -1,]$Rating = 'Positive'
```

Klassifikationsmodell

Ist ein Tweet negativ oder positiv?

Anzahl der Stammformen je Tweet erklären die Zielvariable

	iphon	finger	print	omg	want	…	Rating
Tweet 1	1	1	1	1	0	…	👍
Tweet 2	0	1	1	0	1	…	👎
…	…	…	…	…	…	…	…

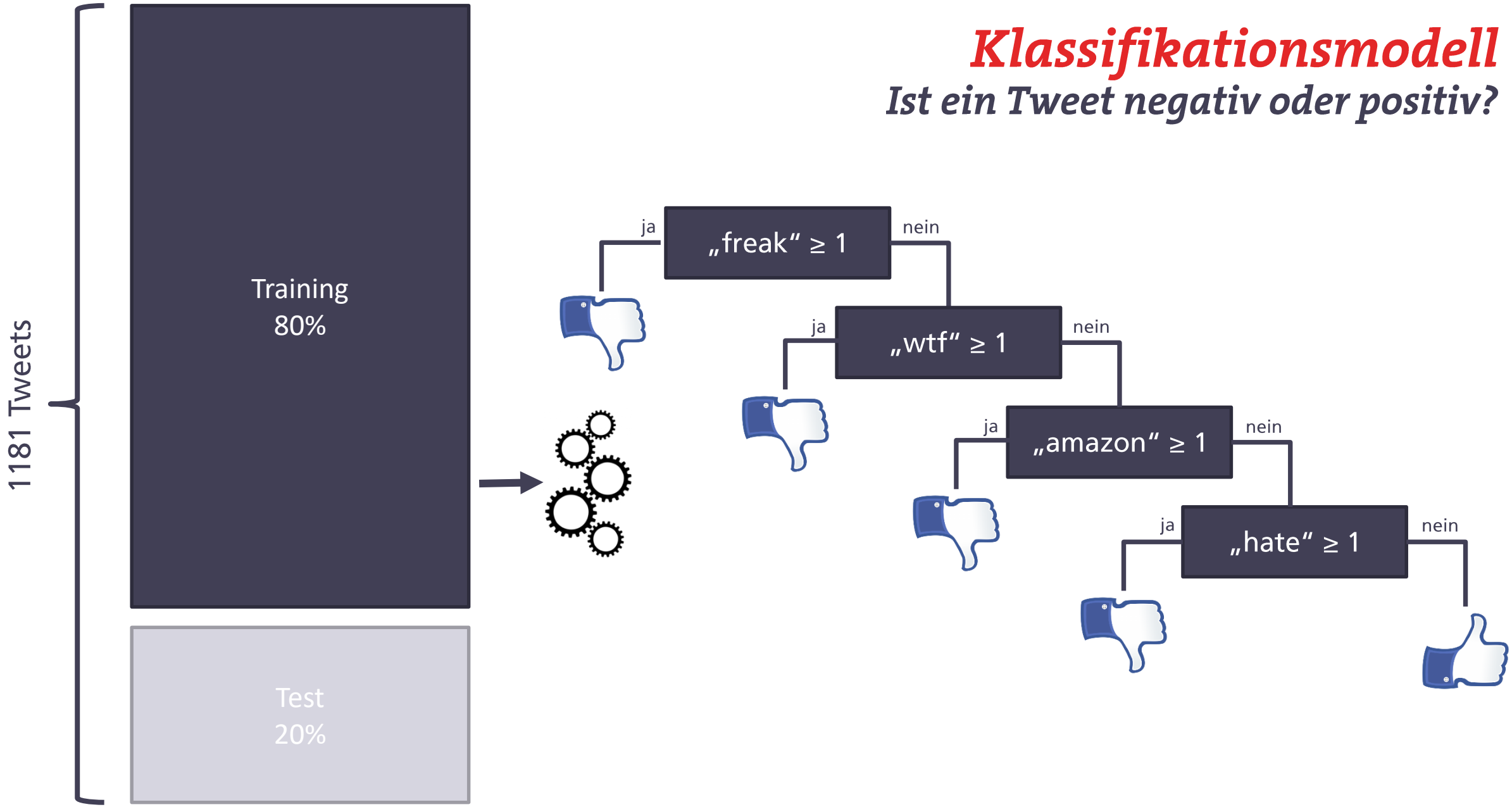


Klassifikation

Die Klassifikation erfolgt anhand eines **Entscheidungsbaumes.**

Klassifikationsmodell

Ist ein Tweet negativ oder positiv?



Berechnung des Klassifikationsmodells mit R

Ist ein Tweet negativ oder positiv?

```
library(caTools)
library(rpart)

set.seed(123)

split = sample.split(sparse$Rating, SplitRatio = 0.8)

trainSparse = subset(sparse, split == TRUE)
testSparse = subset(sparse, split == FALSE)

decTree = rpart( Rating ~ . -TWEET_ID, data=trainSparse, method="class")

pred = predict(decTree, newdata=testSparse, type="class")
```

Evaluation des Modells

Ist ein Tweet negativ oder positiv?

```
(…)

predictCART = predict(tweetCART, newdata=testSparse, type="class")
table(testSparse$Negative, pred)
```

	Positiv	Negativ
Positiv (real)	18	37
Negativ (real)	6	294

$$\rightarrow \quad \frac{294 + 18}{294 + 6 + 37 + 18} \quad \rightarrow \quad 84{,}5\%$$

Fazit & Ausblick

Referenzmodell in Textanalyse-Projekten

„Bag Of Words“

Social Media als Datenbasis

Die Häufigkeit eines jeden Wortes wird gezählt

80% Text

Extraktion von Wissen aus Texten

Text Mining

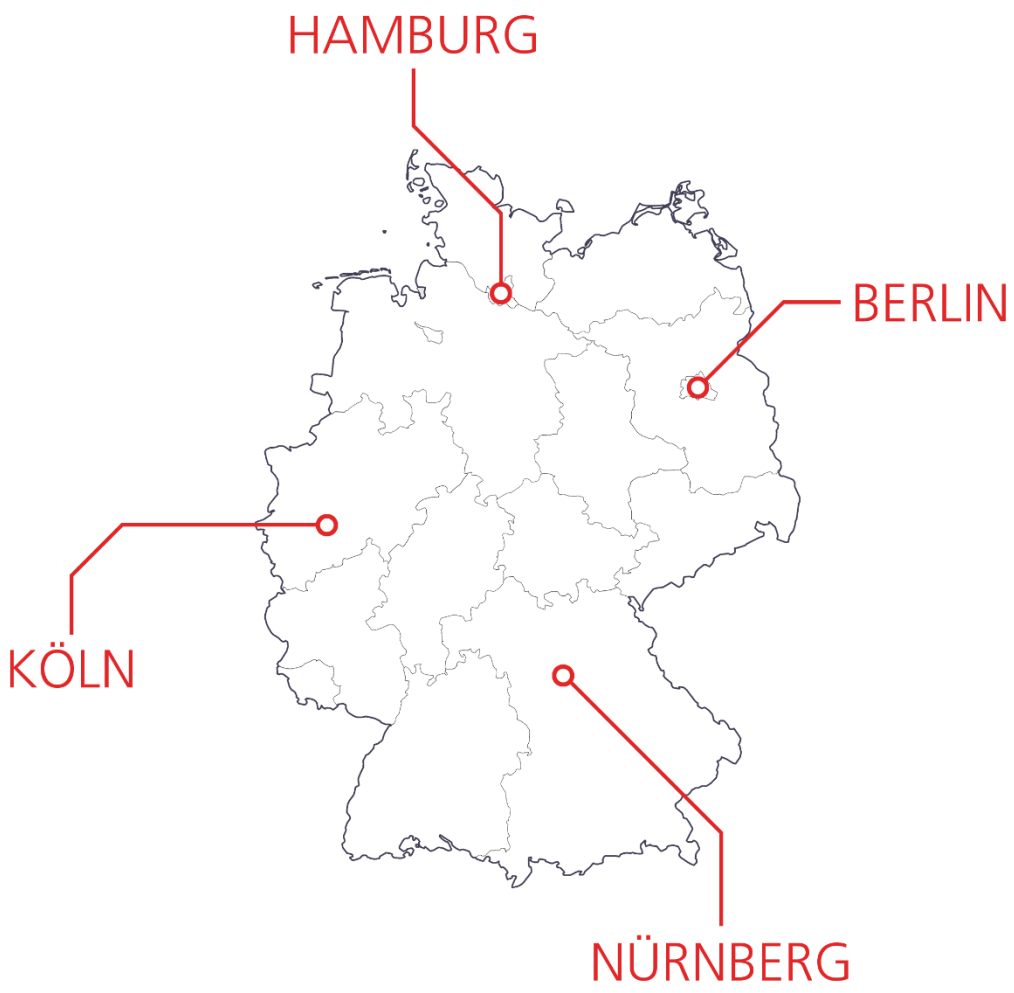
Hohe Genauigkeit mit einfachen Mitteln

Menschliche Intelligenz nicht ersetzbar

Marco Nätlitz
BI-Spezialist & Data Scientist
mna@areto.de

areto consulting gmbh
Data Warehouse
Business Intelligence

areto
CONSULTING. IT WORKS.



Visit us at our Branches:

KÖLN

Schanzenstraße 6-20
51063 Köln
Telefon: +49 221 66 95 75-0
Telefax: +49 221 66 95 75-99
E-Mail: info@areto.de

HAMBURG

Rothenbaumchaussee 27
20148 Hamburg
Telefon: +49 40 22 86 53 20
Telefax: +49 40 22 86 53 23
E-Mail: info@areto.de

NÜRNBERG

Neumeyerstraße 24
90411 Nürnberg
Telefon: +49 9 11 14 88 66 77
Telefax: +49 9 11 14 88 66 79
E-Mail: info@areto.de

BERLIN

Weichselstraße 34a
10247 Berlin
Telefon: +49 30 54 90 87 51
Telefax: +49 30 54 90 87 52
E-Mail: info@areto.de

Henning Schwass
Leiter Marketing
henning.schwass@areto.de

areto consulting gmbh
Data Warehouse
Business Intelligence

areto
CONSULTING. IT WORKS.