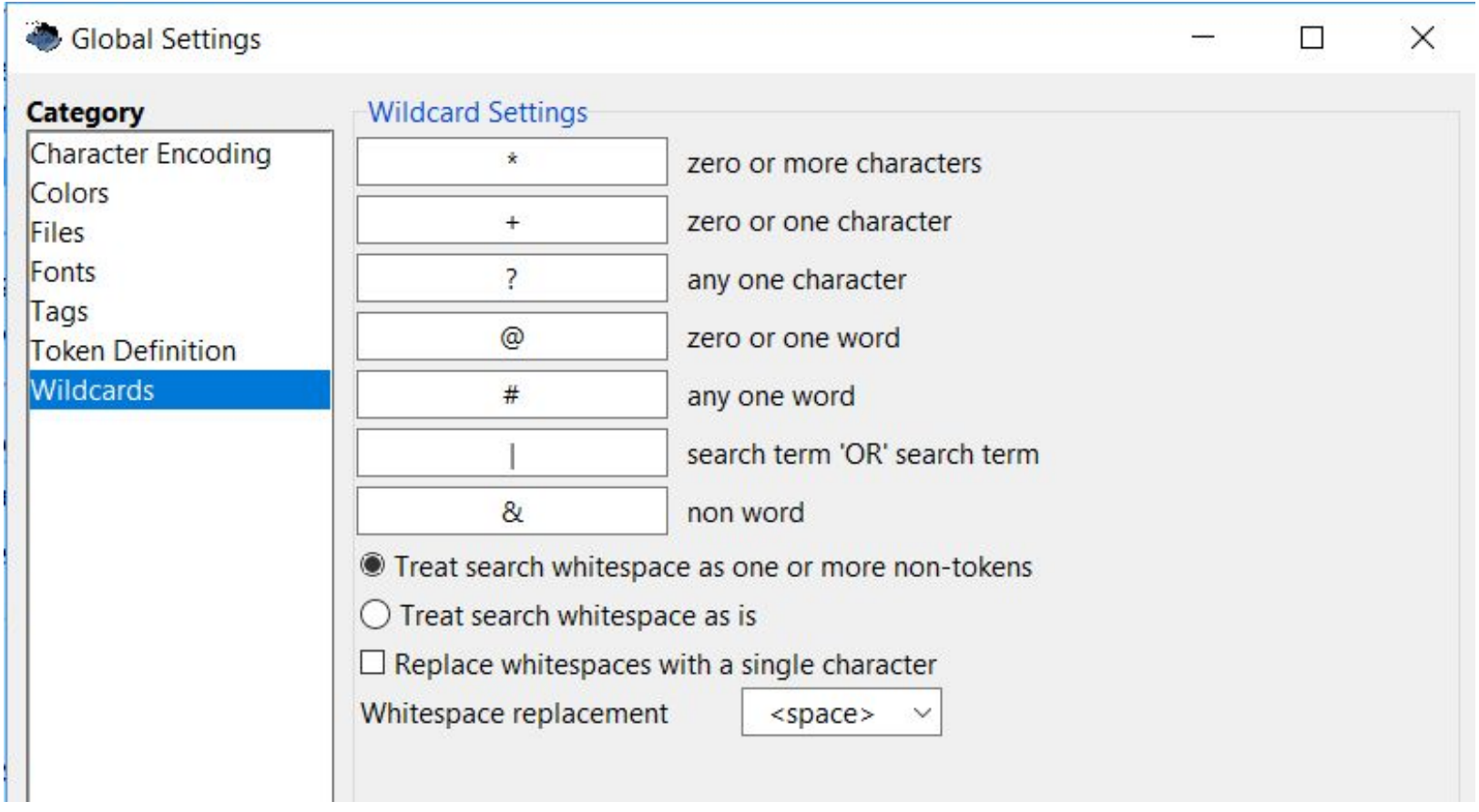
Fortsetzung Antconc

Alle Platzhalter



Reguläre Ausdrücke

- Reguläre Ausdrücke (kurz regex) beschreiben Mengen von Zeichenketten
- In vielen Editoren und in fast allen Programmiersprachen.
- Sogar in MS Word

Bestandteile von regulären Ausdrücken

- Literale
  z.B. 'a' (kleines a)
- Zeichenklassen
  z.B. '.' (hier und im folgenden **ohne** '') = beliebiges Zeichen
- Quantifier
  z.B. '+' = 1 oder beliebig viele des **vorangehenden** Zeichens
  'ab+' matches 'ab' 'abb' 'abbbbb', aber nicht 'abab‘
- Grenzen
  z.B. ' \b' steht für Wortgrenze

z.B. Suche nach "\berst.+" Sucht Wortanfänge mit den Buchstaben erst und mindestens einem weiteren Zeichen, z.B. ‚erste‘, ‚erstaunen‘

Zeichenklassen

- '. ' jedes beliebige Zeichen
- '\d' jede Dezimalzahl (nach Unicode), z.B. 1, 4, 0
  '\D' jedes Zeichen, das keine Dezimalzahl ist
- '\s' jeden whitespace (nach Unicode), z.B. \n \t
  '\S' jedes Zeichen, das kein whitespace Zeichen ist
- '\w' jedes Buchstaben-Zeichen, z.B. A g ö ß 4 é € α И
- [Abc] definiert **1** Zeichen, das entweder A oder b oder c ist.
  zB.: [A-z] **ein** beliebiger Buchstabe zwischen A-z (kein Umlaut, kein ß)

Quantifier

- Quantifier definieren die Häufigkeit des **vorangehenden** Zeichens
- '*' 0 oder häufiger
- '+' 1 oder häufiger
- '?' 0 oder 1
- {n} genau n Zeichen
- {m,n} mindestens m, höchstens mal

Bsp.: .* ein beliebiges Zeichen 0 oder häufiger (nicht ausprobieren!)

[A-D]{3} drei der Großbuchstaben A,B,C,D in Folge, z.B. AAA,

Greediness

- Reguläre Ausdrücke sind von Natur aus ‚greedy‘, d.h. sie versuchen für den Ausdruck einen möglichst langen String zu finden.
- Suche: '\baus.+\b’ findet z.B.

```
ausgewählt phantastischen Ordenstrachten ihrer Herren
überstrahlten. Die gezogenen
aus Ihrem thatenreichen Leben! — Alle Blicke richteten sich
```

- Will man den kürzest möglichen String finden, muss man das durch das ? nach dem Quantifier festlegen: '\baus.+?\b’

Das Fragezeichen hat hier also eine andere Bedeutung

```
ausgewählt phantastischen Ordenstrachten ihrer Her
aus der Ferne, und die Menge verlor sich,
aus und sammelt nichts, und wie reich er
```

Gruppen

Mit Klammern kann man Gruppen bilden, die als Einheit behandelt werden. Z.B.:

'schöne Mädchen|Frau' sucht nach allen Stellen, wo entweder ‚schöne Mädchen‘ oder ‚Frau‘ steht.

'schöne (Mädchen|Frau) ', findet alle Stellen wo entweder ‚schöne Mädchen‘ oder ‚schöne Frau‘ steht

Was macht dieser Ausdruck

- \b(eine[rn]?|die|de[rn])\s[a-zäöüß,]+\sFrau

Was macht dieser Ausdruck

- \b(eine[rn]?|die|de[rn])\s[a-zäöüß,]+\sFrau

1. Wortgrenze: \b
2. eine|einer|einen oder die oder der|den : (eine[rn]?|die|de[rn])
3. Whitespace: \s
4. einer oder mehrere kleine Buchstabe oder Komma: [a-zäöüß,]+
5. Whitespace \s
6. Frau

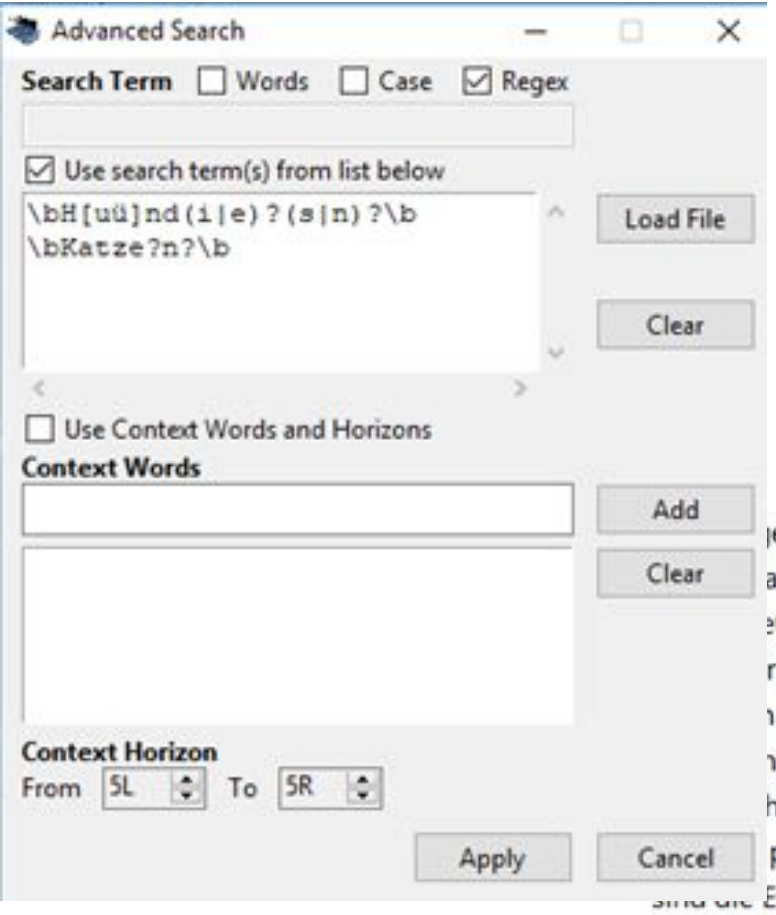
Regex Übungen

- Das letzt Wort eines Satzes
- Datumsangaben mit und ohne Jahr (TT.MM.JJJJ, TT.MM)
- Alle Wörter hinter allen Flexionen/Steigerungen von “gut”
- Wörtliche Rede
- alle XML-Elemente

Regex Tester: https://pythex.org/

Regex Cheatsheet:
https://www.cheatography.com/davechild/cheat-sheets/regular-expressions/

Advanced Search und Regex



…egen dieses Mittel, gerade der Hund des verbrannten Medard war ihr ein Schrecken,
…als Diethelm laut aufschrie: Ein Hund und ein Fuchs ist dein Vater, rathet
…ete und schrie laut auf, daß der Hund bellte. Er hatte einen Schädel mit halbverbra
…r um sich haben wollte als den Hund des verstorbenen Medard, mit dem er oft
…n Worte. Sie sprudelte wie eine Katze. Die häßlichen Kohlen, sagte sie, die waren
…ndet, um auf den verfolgenden Hund zu stürzen; doch sobald sie es wieder
…h glaubten sie das Bellen eines Hundes zu vernehmen, ein Dorf mußte näher sein,
…pflegen; dazwischen bellte der Hund. Jetzt erschienen zwei Personen, voraus lief
…sind die Ewigkeit! — Ich lasse euch den Hund und die Gertrud — mit diesen Worten ging
Knecht mit der Leiter kommen kann, der Hund wird dich führen; Schnappauf, allons! — Der H
…nd wird dich führen; Schnappauf, allons! — Der Hund lief mit Gertruden weg, immer voran; die

Aufgabe

- Welche Texte benutzen die alte Rechtschreibung (Th statt t)?
- Wie müsste eine Suche aussehen, die das Wort ‚Tür‘ unabhängig von der Rechtschreibung sucht (mit allen Fällen, aber ohne Komposita)?

Aufgabe

Welche Beschreibungsdimensionen für männliche und weibliche Figuren dominieren? Machen Sie eine Vorhersage und überprüfen Sie ihre These anhand der Sammlung.

NGramme

- N: Eine natürliche Zahl, z.B. 1 oder 2 oder 3…
- Also 1-Gramm, 2-Gramm, 3-Gramm
- Zerlegen wir den folgenden Satz in 2-Gramme:

„Am Morgen des nächsten Tages schneite es.“

„Am Morgen“, „Morgen des“, „des nächsten“, „nächsten Tages“, „Tages schneite“, „schneite es“

- 3-Gramme:

„Am Morgen des“, „Morgen des nächsten“, „des nächsten Tages“

Concordance | Concordance Plot | File View | Clusters/N-Grams | Collocates | Word List | Keyword List

Total No. of N-Gram Types 1353544 **Total No. of N-Gram Tokens** 1594160

Rank	Freq	Range	N-gram
1	171	53	vor sich hin
2	168	52	in der that
3	162	60	in der welt
4	156	58	es war ein
5	144	59	in der hand
6	137	45	nach und nach
7	136	54	daß er sich
8	127	50	ich weiß nicht
9	124	50	in der nacht
10	122	56	in der nähe
11	119	43	in der stadt
12	116	49	hin und her
13	112	48	fuhr er fort
14	103	50	in die höhe
15	101	54	und in der
16	100	49	es ist ein
17	98	44	auf den tisch
18	97	48	und wenn sie
19	96	35	der junge mann
20	96	45	und wenn ich
21	91	40	mit der hand
22	88	49	sich in die
23	86	31	als ob er
24	85	41	aus dem hause

Search Term Words Case Regex N-Grams

Advanced

Start Stop Sort

Sort by Invert Order

Sort by Freq

Search Term Position On Left On Right

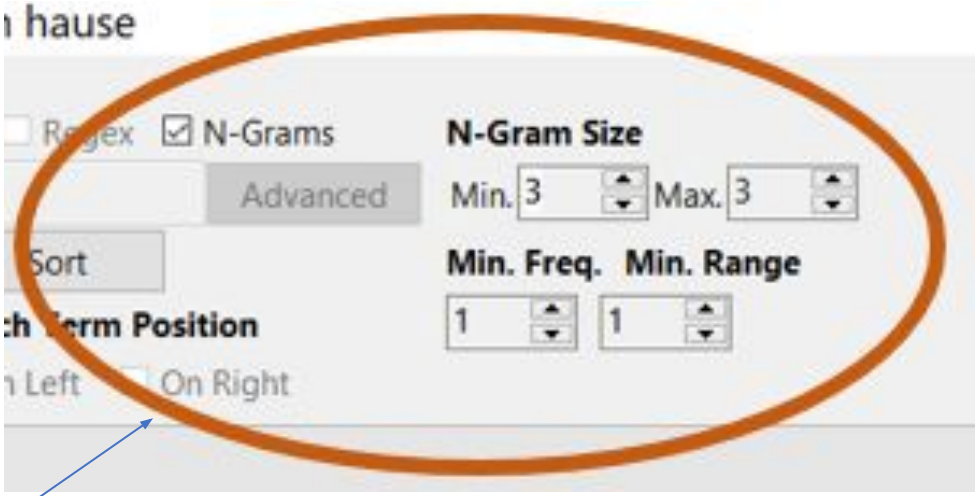
N-Gram Size Min. 3 Max. 3

Min. Freq. Min. Range 1 1

Novellenschatz: 3Gramme

Total No. of N-Gram Types 1353544

Total No. of N-Gram Tokens 1594160

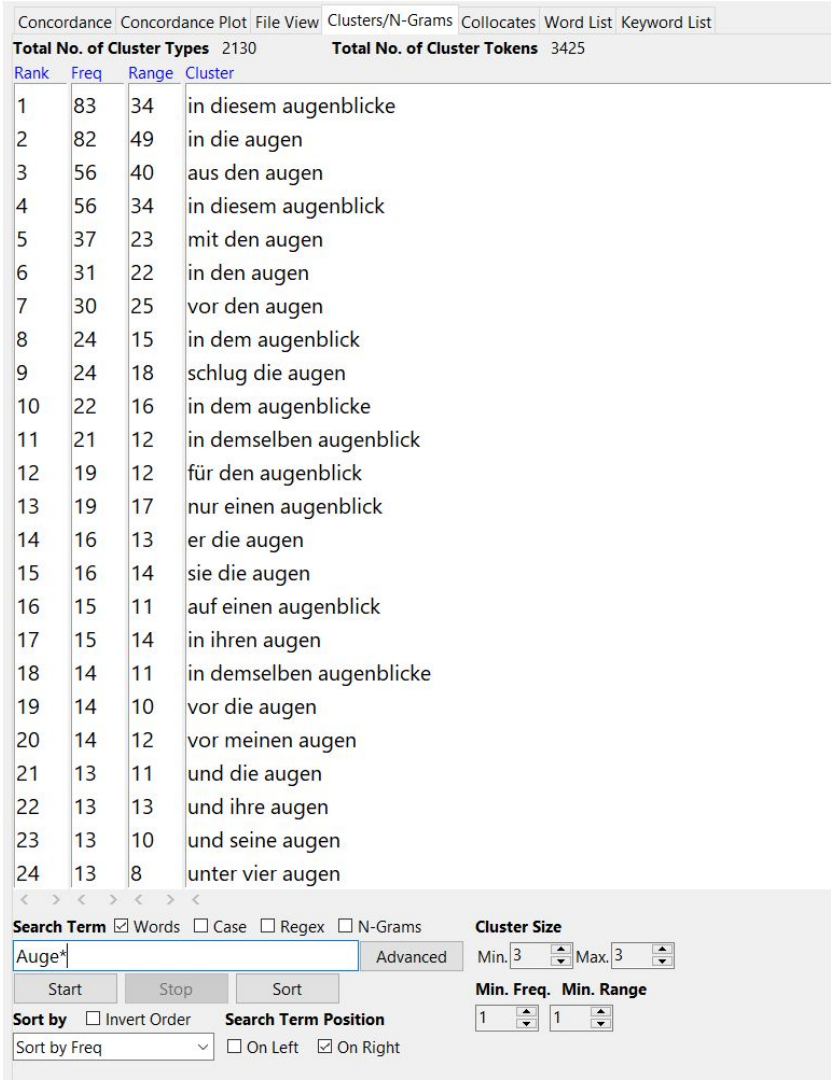


Achtung: Erstellung dauert etwas!

Cluster-Suche

- Cluster-Suche fasst die Ergebnisse der Konkordanz-Suche in Haufen (‚Cluster‘) zusammen
- Ein Cluster ist ein Mehrwortausdruck, z.B. „in diesem Augenblicke“

  Zusätzliche Optionen:
  - Anzahl der Worte im Cluster
  - Range, d.i. in wie vielen Texten kommt das Cluster vor. Freq 5, Range 3 bedeutet also, dass ein Cluster 5mal vorkommt und zwar in 3 verschiedenen Texten
  - Search Term Position: On Left, On Right.: Wo im Cluster steht das Suchwort. Im Deutschen ist ‚On Right‘ häufig fruchtbarer. Allerdings ist bei Mehrwortausdrücken nur ‚On Left‘ sinnvoll, sonst sieht man den Ausdruck nicht.



Ein höherer Wert für ‚Range‘ erzwingt das Finden von Mustern in mehreren Texten

keyness

Unterschied zwischen typisch und distinktiv

Keyness

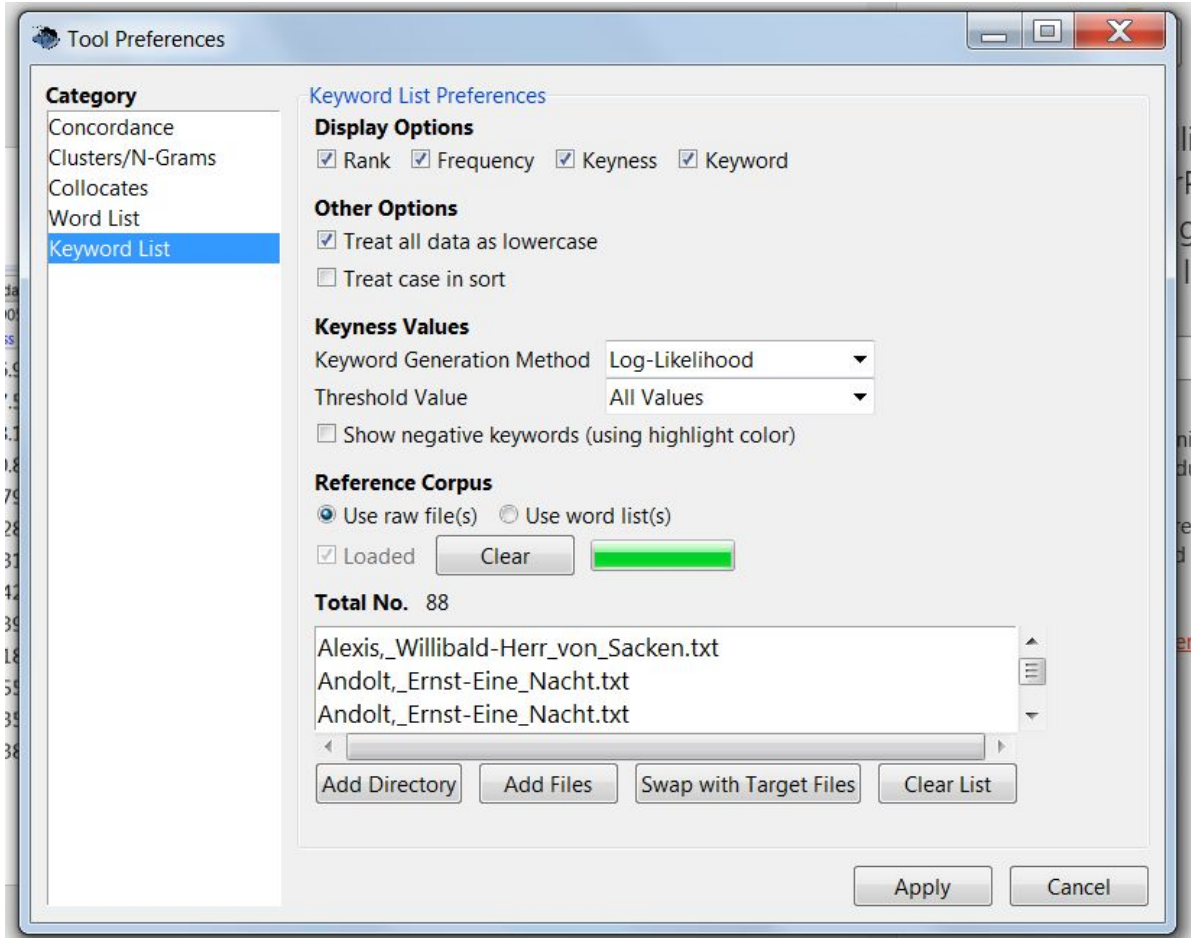
- Ermittelt die Worte, in denen sich eine Textgruppe von einer anderen unterscheidet.
- In AntConc: Die angezeigte Liste zeigt, welche Worte deutlich häufiger in dem untersuchten Korpus (im Vergleich zum Referenzkorpus) auftauchen.
- Zwei Maße:
  - Log likelihood
  - Chi Squared
  - Beide berechnen, ob der Unterschied zwischen den zu erwartenden Werten und den realen Werten so groß ist, dass wahrscheinlich zwei Verteilungen vorliegen

Vorgehensweise

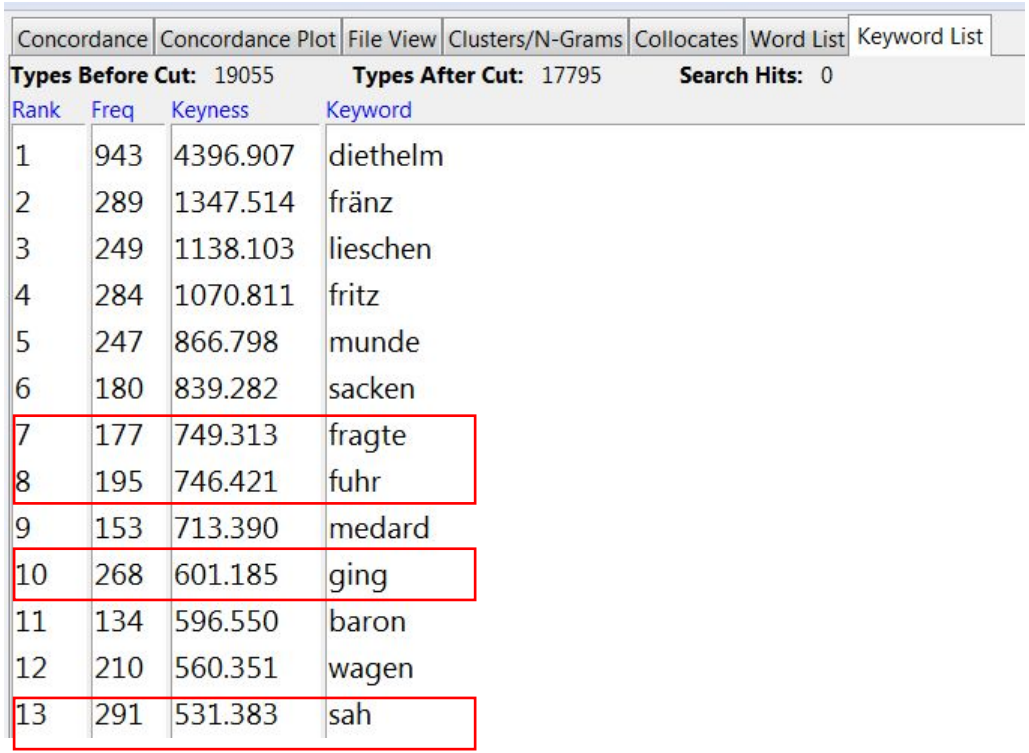
- Festlegung des 1. Korpus
- Festlegung des 2. Korpus
- Wahl eines Keyness-Maßes
- Analyse anstoßen
- Ergebnis der Analyse - die Wortliste - untersuchen

Keyness in AntConc

- Fokuskorpus auswählen wie gewohnt
- Reference Korpus auswählen:
  Tool Preferences->
  Keyword List
- Einstellungen:
- Use raw files -> Load -> Appy



Vergleich von Dramen und Novellen

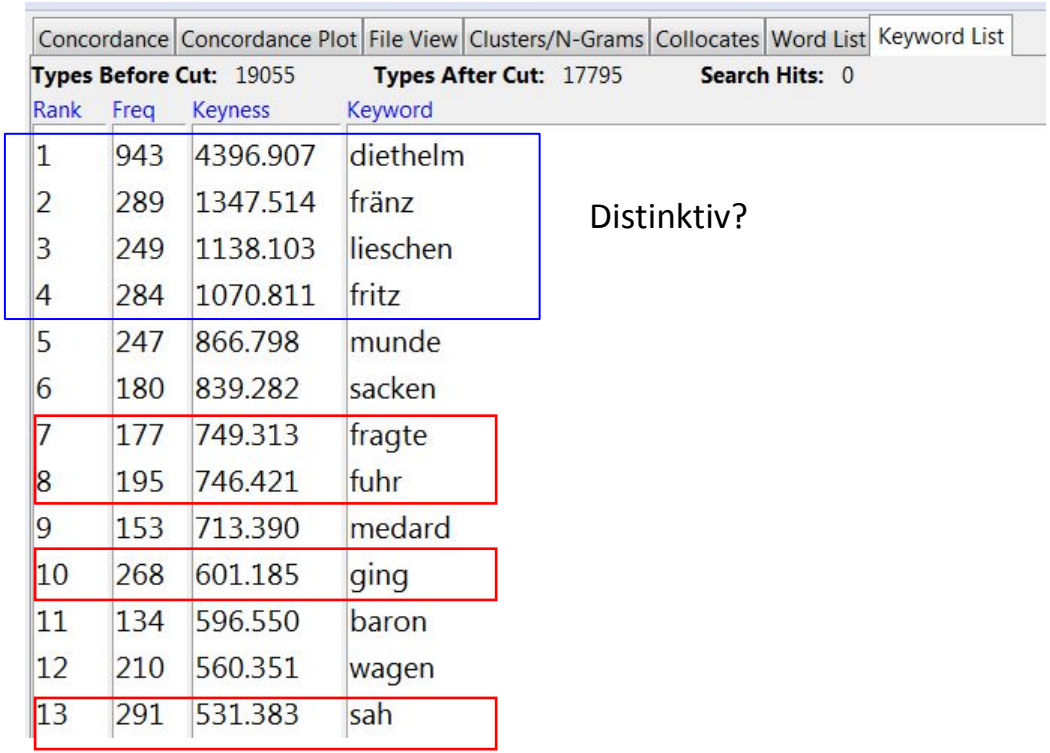


Concordance | Concordance Plot | File View | Clusters/N-Grams | Collocates | Word List | Keyword List

Types Before Cut: 19055 **Types After Cut:** 17795 **Search Hits:** 0

Rank	Freq	Keyness	Keyword
1	943	4396.907	diethelm
2	289	1347.514	fränz
3	249	1138.103	lieschen
4	284	1070.811	fritz
5	247	866.798	munde
6	180	839.282	sacken
7	177	749.313	fragte
8	195	746.421	fuhr
9	153	713.390	medard
10	268	601.185	ging
11	134	596.550	baron
12	210	560.351	wagen
13	291	531.383	sah

Vergleich von Dramen und Novellen



Concordance | Concordance Plot | File View | Clusters/N-Grams | Collocates | Word List | Keyword List

Types Before Cut: 19055 **Types After Cut:** 17795 **Search Hits:** 0

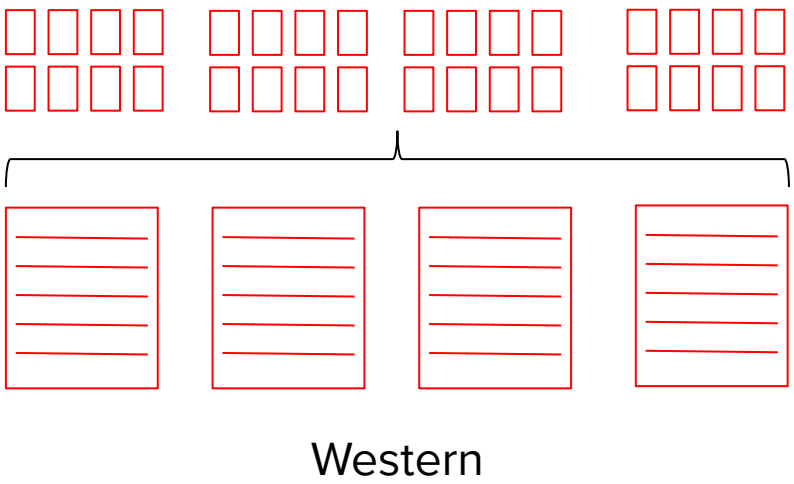
Rank	Freq	Keyness	Keyword
1	943	4396.907	diethelm
2	289	1347.514	fränz
3	249	1138.103	lieschen
4	284	1070.811	fritz
5	247	866.798	munde
6	180	839.282	sacken
7	177	749.313	fragte
8	195	746.421	fuhr
9	153	713.390	medard
10	268	601.185	ging
11	134	596.550	baron
12	210	560.351	wagen
13	291	531.383	sah

Distinktiv?

Burrow’s Zeta

Burrow’s Zeta

Dispersionsmaß zum Auffinden distinktiver Wörter für Textgruppen



Heimat	Kriegs	Krimi	Liebes	SciFi	Hochlit
madl	leutnant	streifenwagen	dienerschaft	galaxis	klo
bissel	mg	ford	gnädigen	raumschiff	it
brotzeit	munition	field	diwan	planet	me
tonis	russen	officer	gnädiges	universums	hitler
bös	russischen	schalldämpfer	anerbieten	schleuse	texte
obstler	deutscher	handschellen	unbeschreiblichen	weltraum	for
förster	einschläge	dienstwaffe	vornehmer	hangar	weltkrieg
ausschaut	flugzeuge	detective	gottlob	schutzschirm	on
gell	oberst	brooklyn	destille	raumschiffs	andauernd
leut	funker	notebook	teetisch	raumfahrer	wischt
trenker	meldet	inspektoren	liebenswürdigkeit	jahrtausenden	juden
bergwald	flanke	ermittlung	mancherlei	terra	russland
bursch	lastwagen	mafia	reizendes	lichtjahre	cola
bergführer	feindes	plaza	umzukleiden	humanoiden	wörter
feschen	pistole	ganoven	umkleiden	geortet	präsidenten
gerad	ne	bewußte	namenlos	unsterblichen	what
bergtour	flieger	datenbanken	frohen	projektion	christus

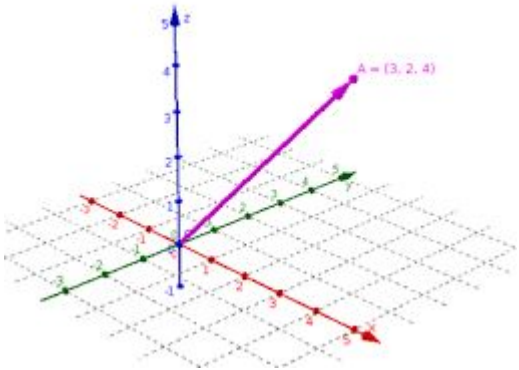
Textvektoren

- Repräsentation eines Textes als Vektor zur besseren maschinellen Verarbeitung
- Grundlage: Worthäufigkeiten
- Mehrere Dokumente werden als Matrix dargestellt und erlauben Vergleichbarkeit

$m \times n$-Matrix

n Spalten

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2n} \\ a_{31} & a_{32} & a_{33} & & a_{3n} \\ \vdots & \vdots & & \ddots & \vdots \\ a_{m1} & a_{m2} & a_{m3} & \cdots & a_{mn} \end{bmatrix}$$ m Zeilen



Textvektoren

Gegeben sind zwei Dokumente D und deren Sätze:

D 1: Wir bauen unser Haus. Es wird ein kleines Haus.

D 2: Jetzt haben wir ein kleines Haus.

Aus diesen Sätzen lässt sich eine sog. Document-Term-Matrix erzeugen. Diese enthält die Information darüber, wie oft ein Wort in einem Text enthalten ist.

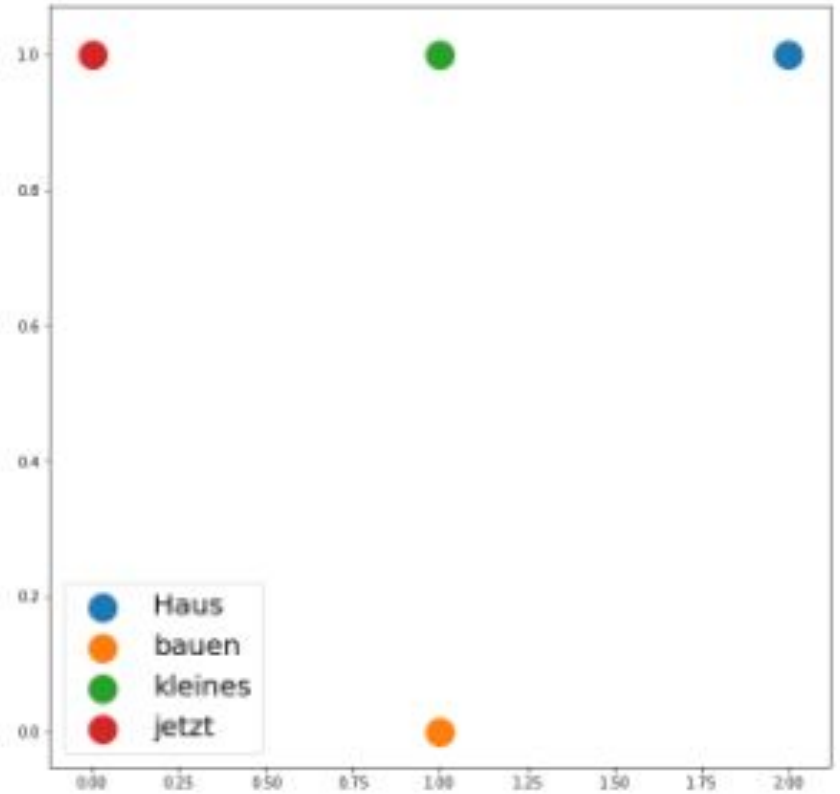
	wir	ein	bauen	Haus	kleines	unser	jetzt	wird
$D1$	1	1	1	2	1	1	0	1
$D2$	1	1	0	1	1	0	1	0

Dokumentvektor

Wortvektor

Textvektoren

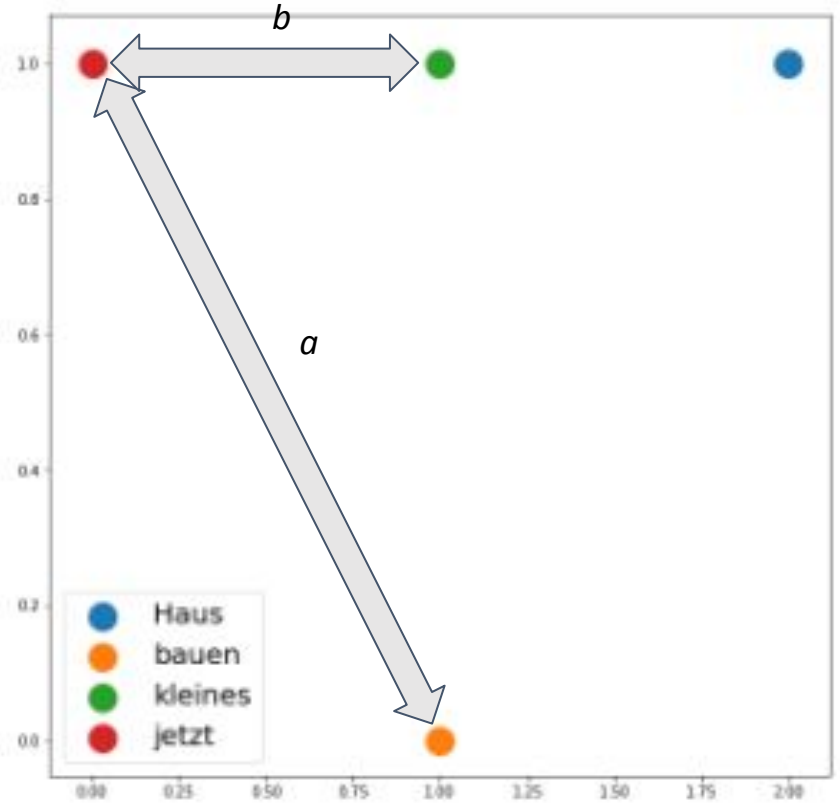
	wir	ein	bauen	Haus	kleines	unser	jetzt	wird
$D1$	1	1	1	2	1	1	0	1
$D2$	1	1	0	1	1	0	1	0



Textvektoren

	wir	ein	bauen	Haus	kleines	unser	jetzt	wird
$D1$	1	1	1	2	1	1	0	1
$D2$	1	1	0	1	1	0	1	0

Distanzen zwischen Worten



Textvektoren

Distanzen zwischen Texten?

	wir	ein	bauen	Haus	kleines	unser	jetzt	wird
$D1$	1	1	1	2	1	1	0	1
$D2$	1	1	0	1	1	0	1	0

Textvektoren

Distanzen zwischen Texten?

	wir	ein	bauen	Haus	kleines	unser	jetzt	wird
$D1$	1	1	1	2	1	1	0	1
$D2$	1	1	0	1	1	0	1	0

0	0	1	1	0	1	-1	1

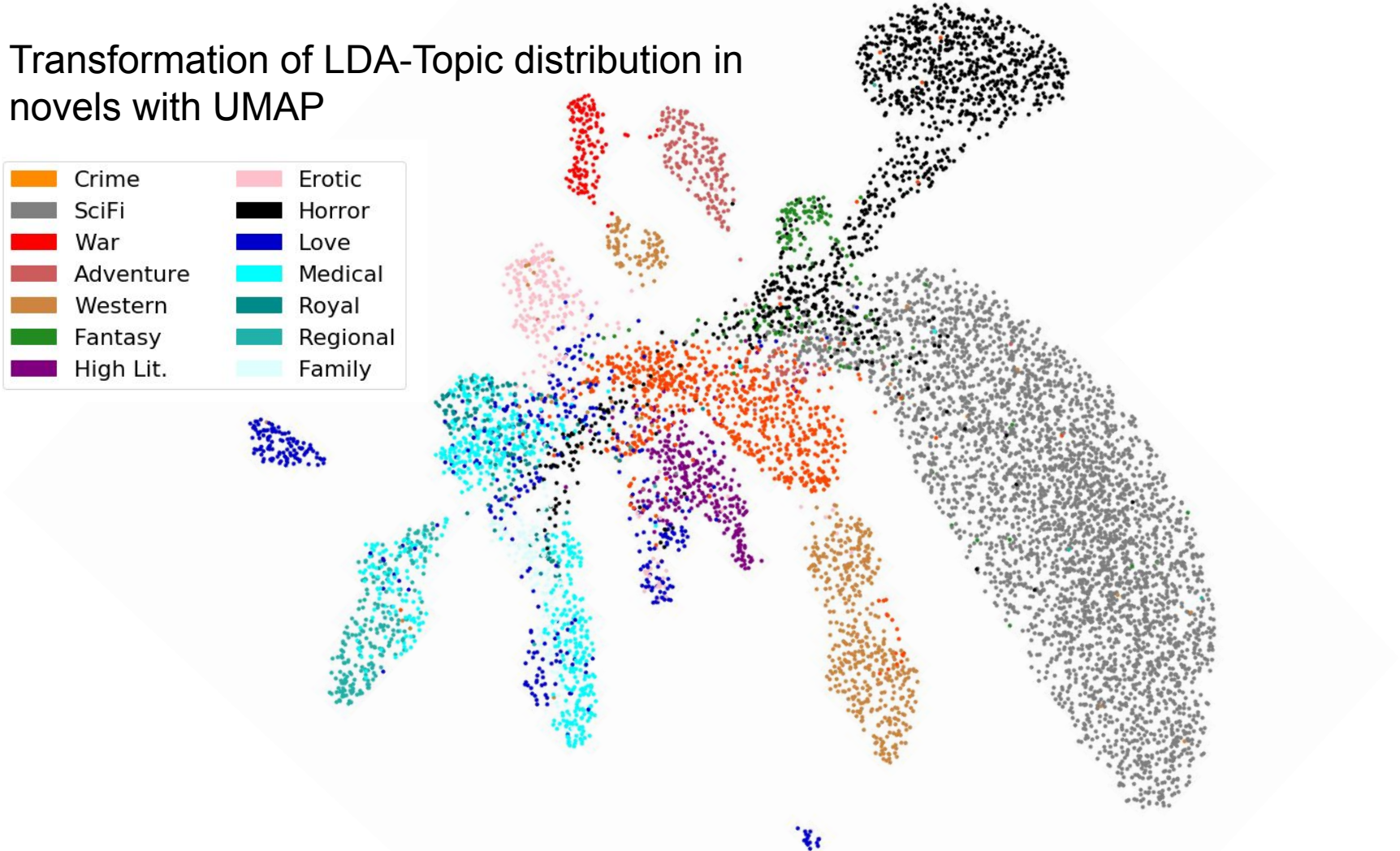
Summe des Betrags der Werte (Abstand der Texte):
5

Textvektoren

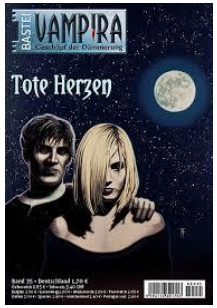
Anwendung in der Dokumentensuche:

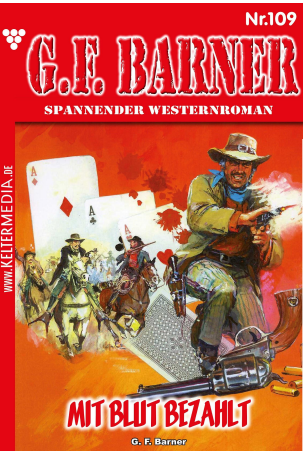
- Suche nach Texten mit geringer Distanz zu Dokument X
- Erweiterung einer Suche um ähnliche Worte zu Suchwort X

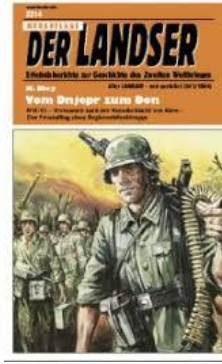
Transformation of LDA-Topic distribution in novels with UMAP



2D Darstellung der Vektoren von Heftromanen







Western/War

Horror/Love

War

Western

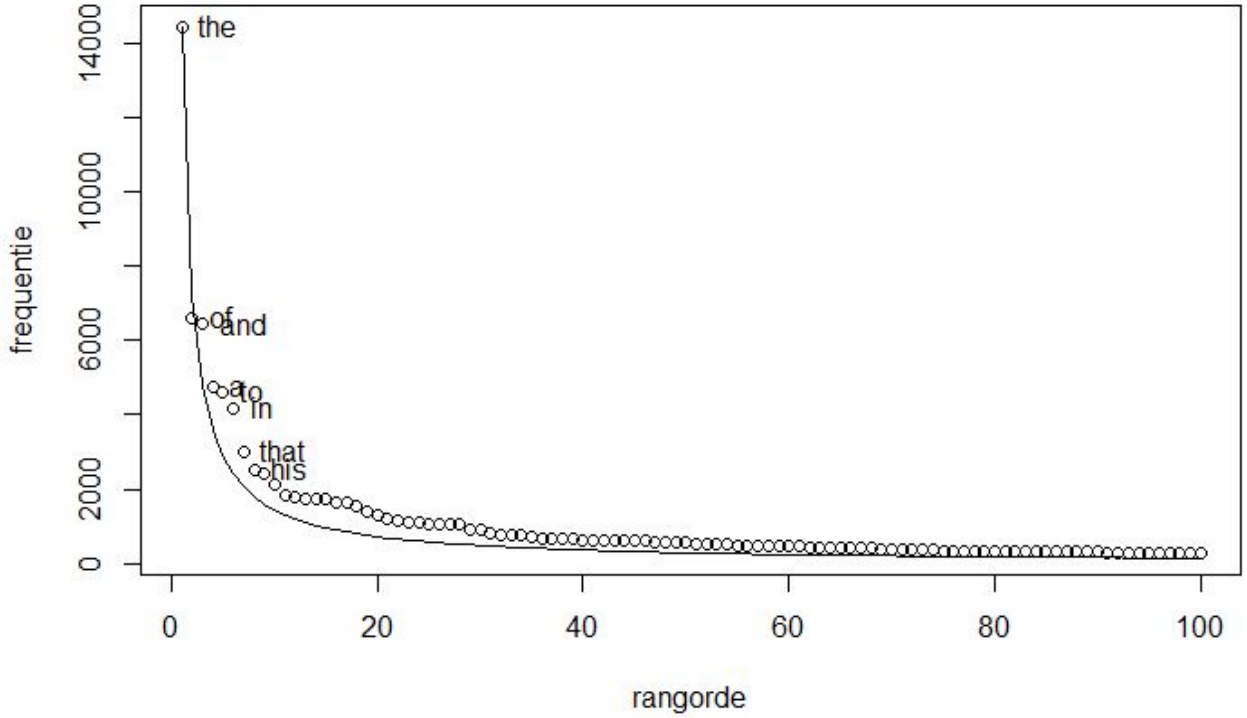
Crime	Erotic
SciFi	Horror
War	Love
Adventure	Medical
Western	Royal
Fantasy	Regional
High Lit.	Family

Burrow’s Delta

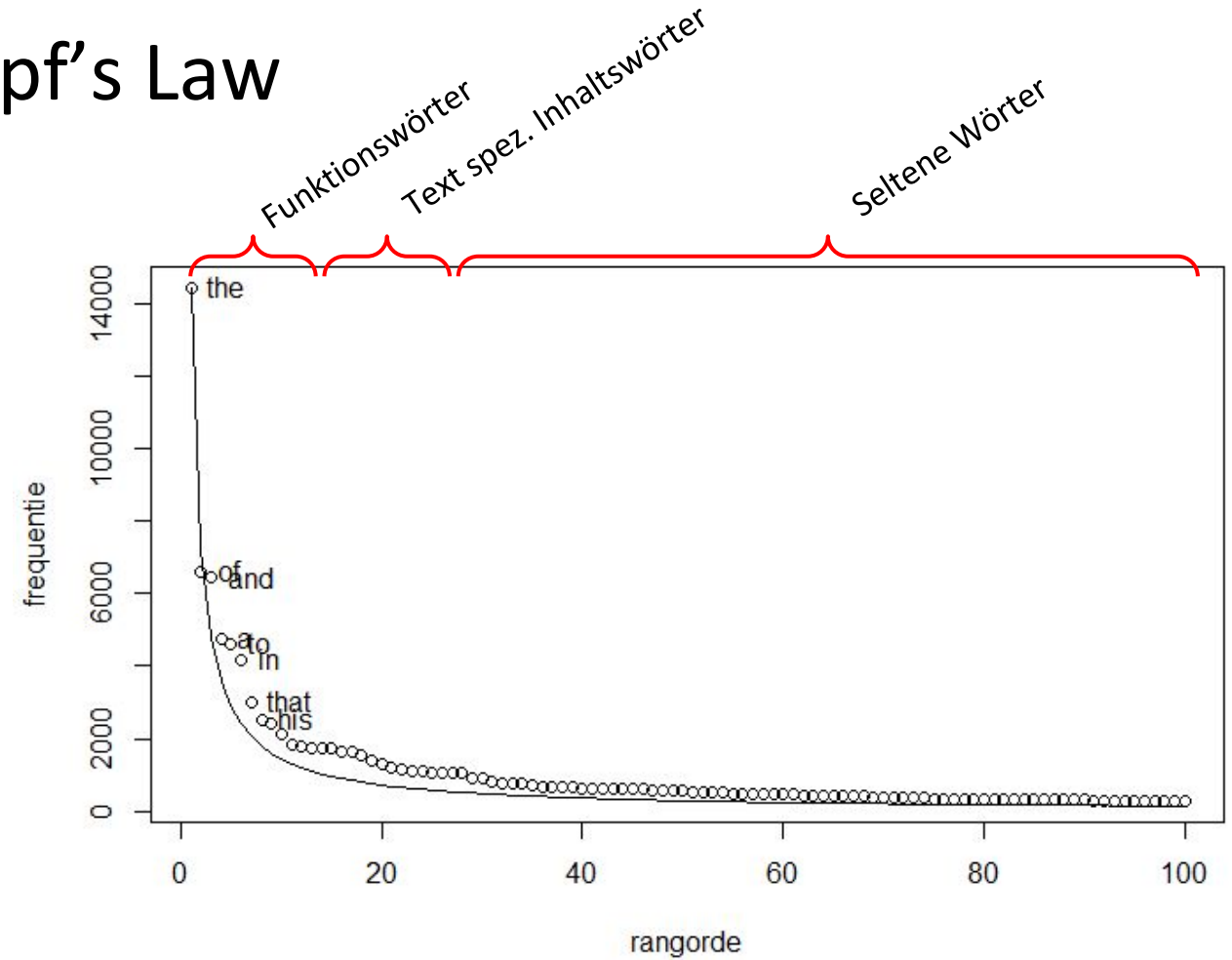
Stilometrie

Zipf's Law

Die absolute Häufigkeit eines Wortes verhält sich antiproportional zu seinem Rang nach Häufigkeiten



Zipf's Law

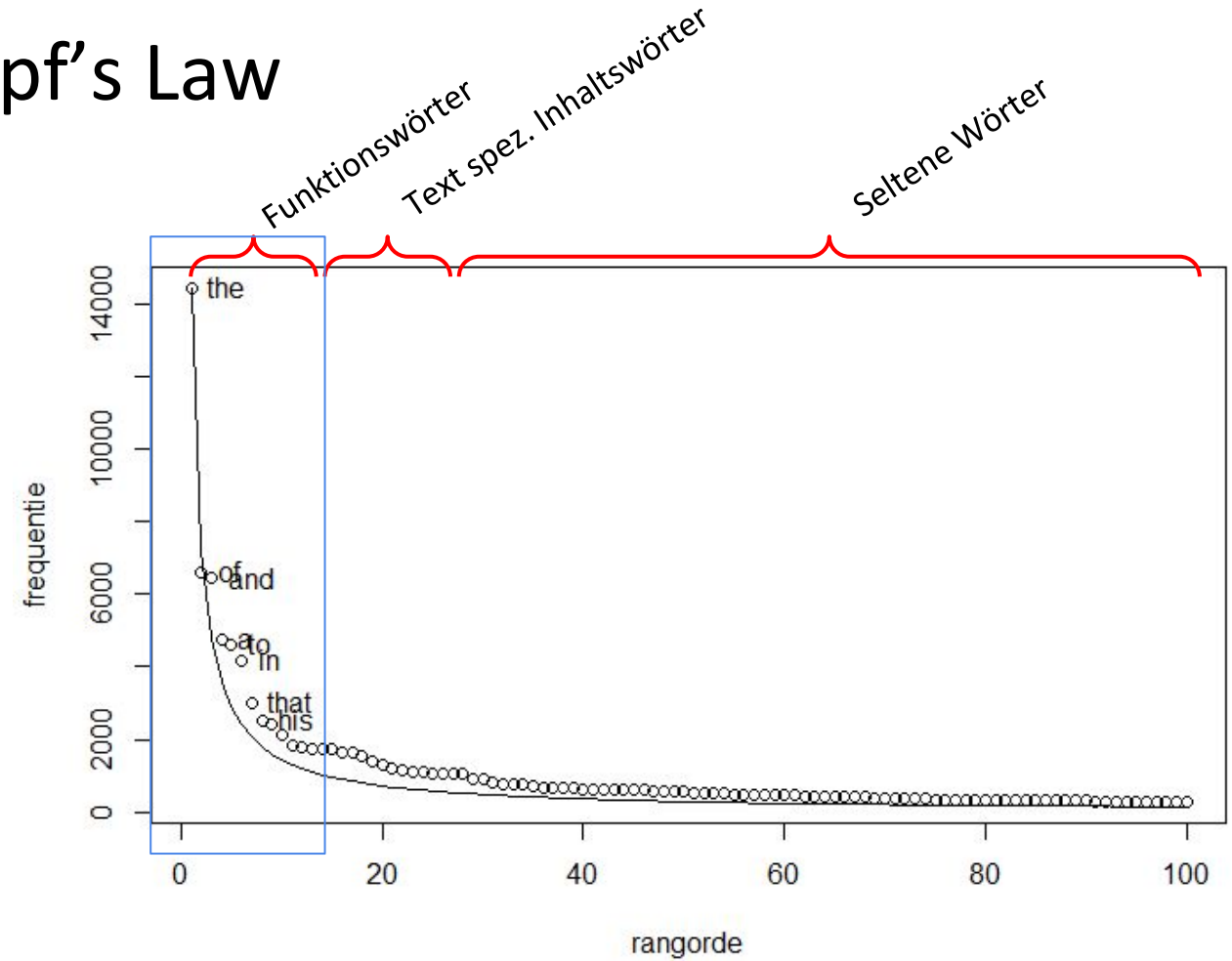

Funktionswörter
Text spez. Inhaltswörter
Seltene Wörter
the
of
and
a
to
in
that
his
frequentie
0
2000
6000
10000
14000
0
20
40
60
80
100
rangorde

Burrow’s Delta

Problemstellung:

In einer Sammlung an Texten mit bekannten Autoren befindet sich ein Text ohne Verfasser. Welcher der bekannten Autoren (Kandidaten) hat diesen Text geschrieben?

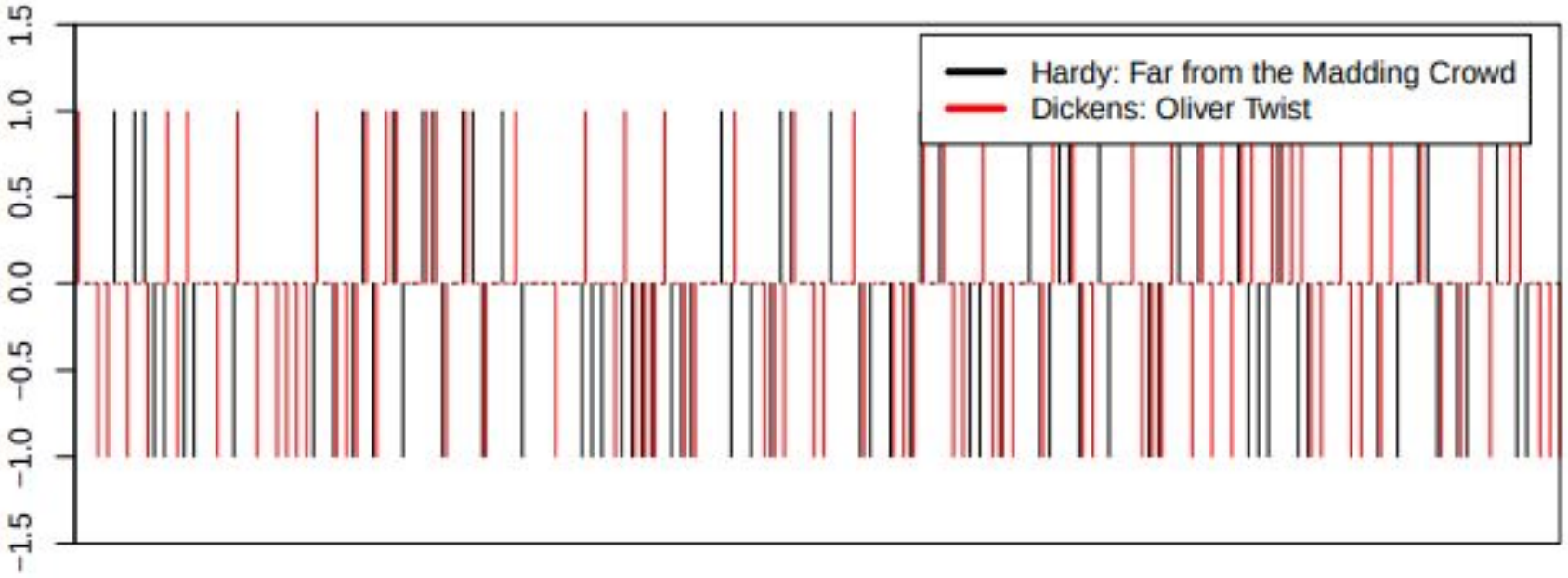
Zipf's Law


Funktionswörter
Text spez. Inhaltswörter
Seltene Wörter
the
of
and
a
to
in
that
his
frequentie
0
2000
6000
10000
14000
0
20
40
60
80
100
rangorde

Algorithmus

1. Ermitteln der häufigsten Wörter in der Sammlung
2. Relative Häufigkeiten für alle Texte und für jeden Einzeltexte berechnen
3. Abweichung der Häufigkeit für jeden Text von der Sammlung bestimmen
4. Tertiarisierung der Werte in:
    a. 0: keine/geringe Abweichung
    b. -1: unterdurchschnittliche Verwendung
    c. 1: überdurchschnittliche Verwendung

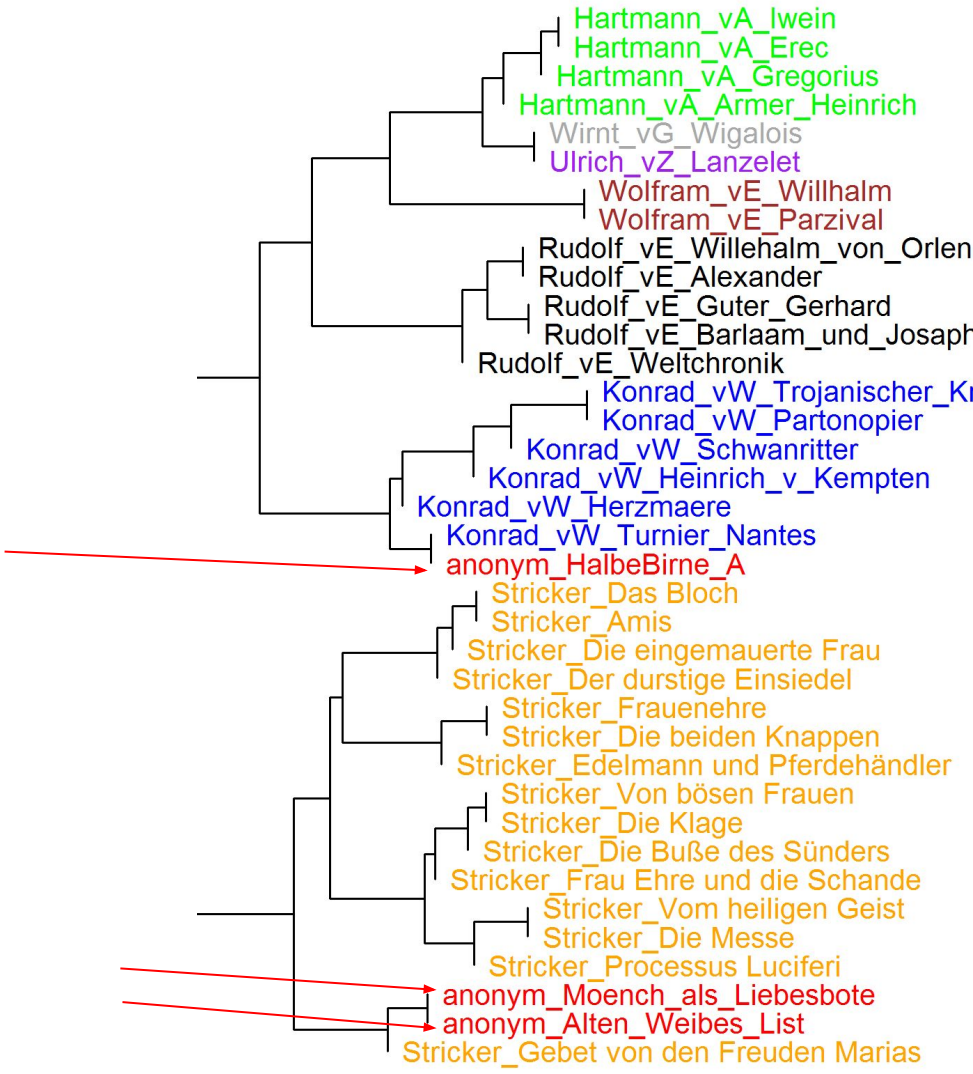
Ergebnis


Hardy: Far from the Madding Crowd
Dickens: Oliver Twist
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5

Auswertung: Dendrogram

Erstellung:
Berechnung der Distanzen aller Texte zueinander. Gruppen von Texten geringer Distanz werden zusammengehalten.

Tutorial:
https://fortext.net/routinen/lerneinheiten/stilometrie-mit-stylo



Key Profile Hypothese

Bei einer ausreichenden Anzahl an Texten eines Verfassers kann man durch mitteln der Delta Vektoren ein Schlüsselprofil erzeugen.

Dieser Schlüssel ermöglicht es neue Texte Autoren zuzuordnen.

Außerdem enthält er biographische Informationen wie Bildung, Herkunft und Alter einer Person.

DH as Data Science

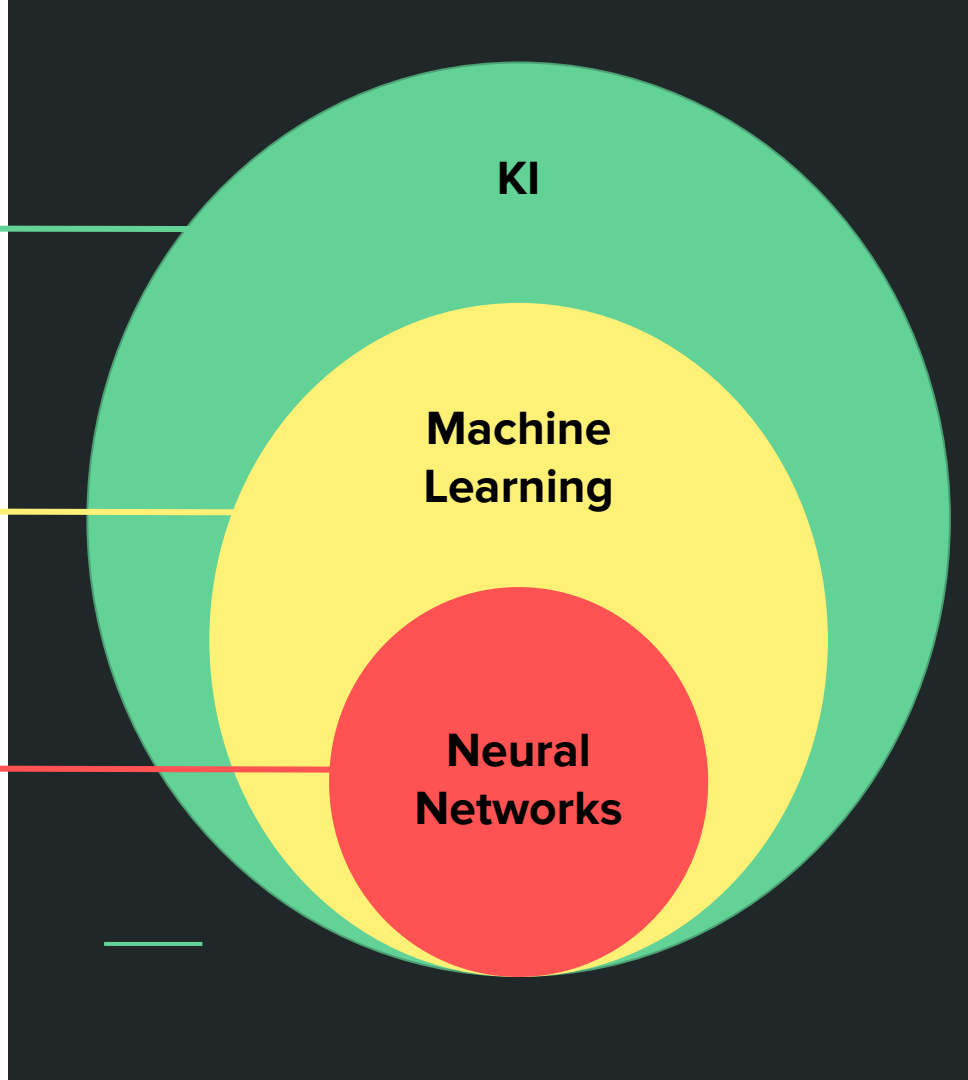
Neuronale Netze, Word Embeddings und Sprachmodelle

Machine Learning, Deep Learning & KI

Simulation of human decision structures by algorithms in order to solve problems as autonomously as possible.

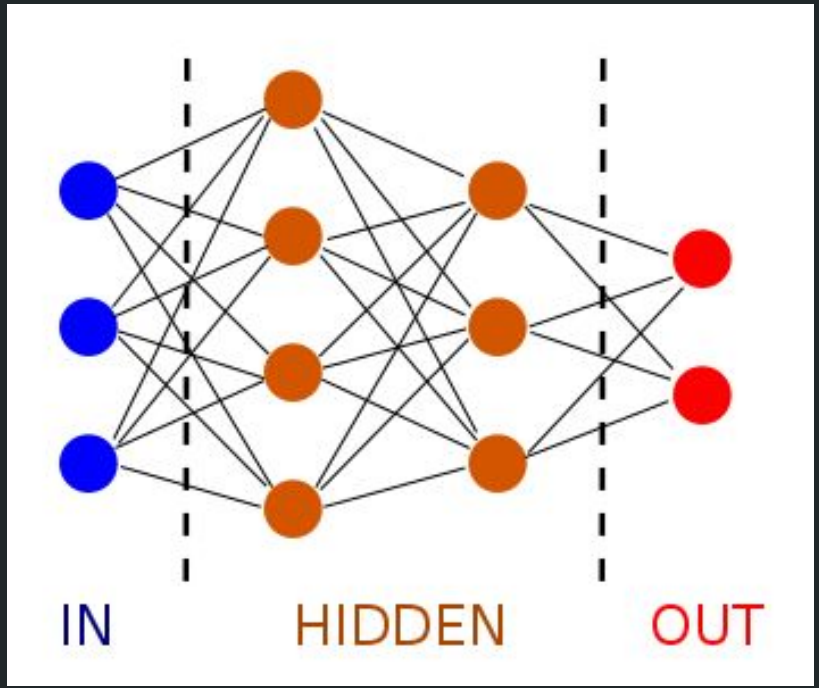
Implicit replication of these structures by adaptation of algorithms using examples

Distribution of the learning process to a net structure



Neural Networks

Ziel:
Gegeben ein Set von Paaren
{(x1,x2,x3,...x4),(y1,y2,y3,...y4)}
soll eine Funktion f(x)=y gefunden werden, die auch für neue Werte x zuverlässige Ergebnisse produziert.



Fully-Connected Feedforward Network

Neural Nets - Neurons

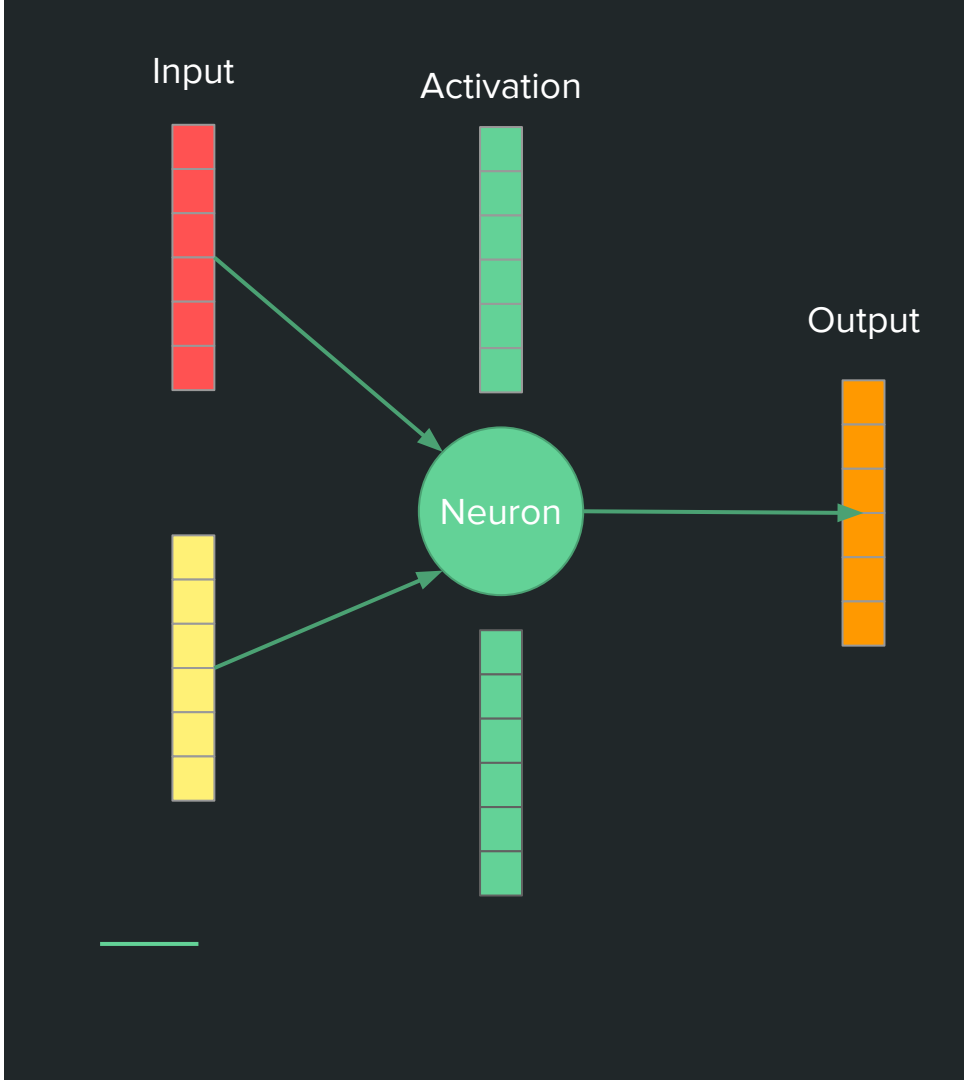
The output of a neuron is determined by its activation function:

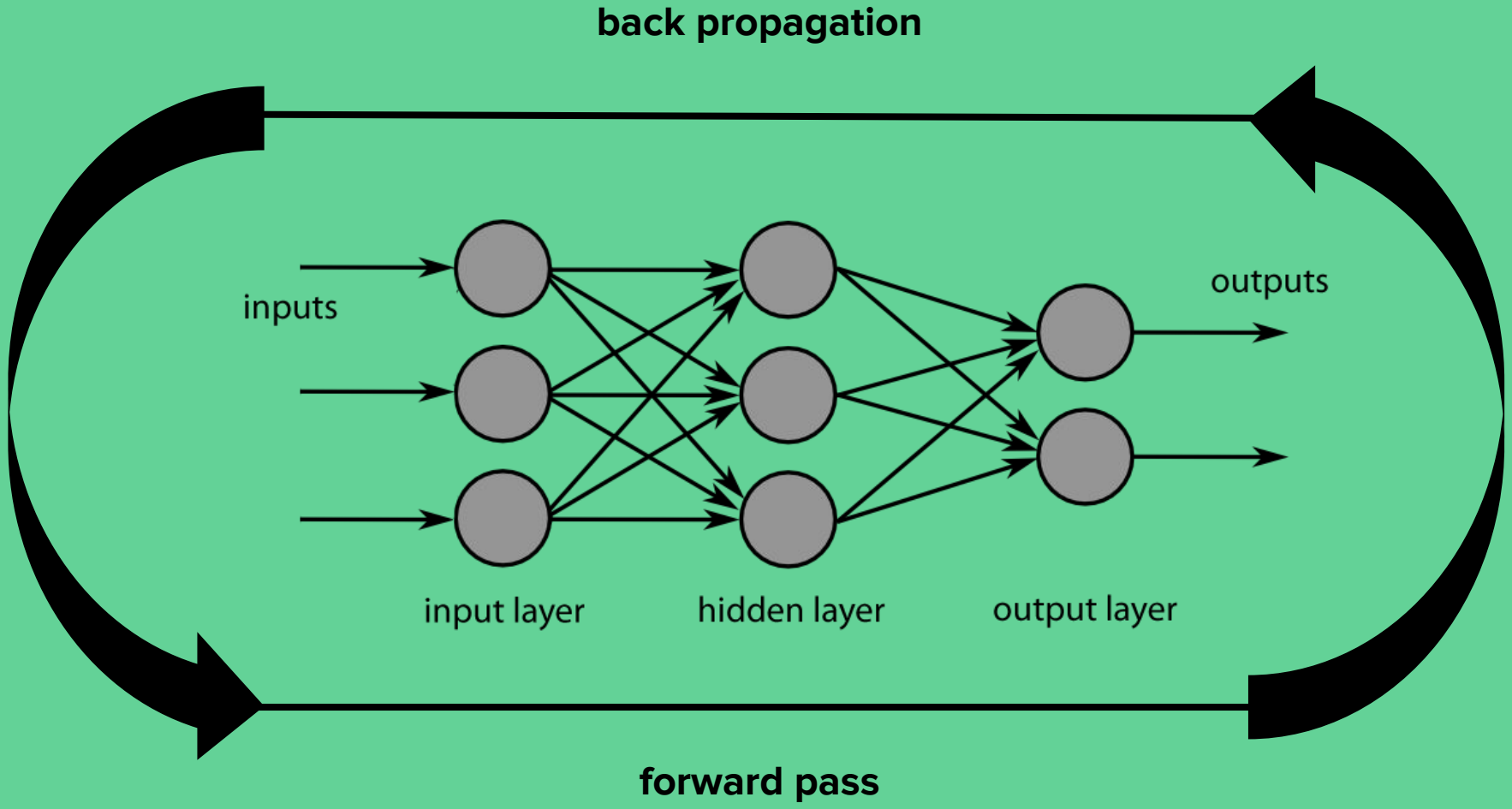
$$y = wx + b$$

x: Input

w: weights (random)

b: bias (random)




back propagation
inputs
outputs
input layer
hidden layer
output layer
forward pass

Distributionelle Semantik

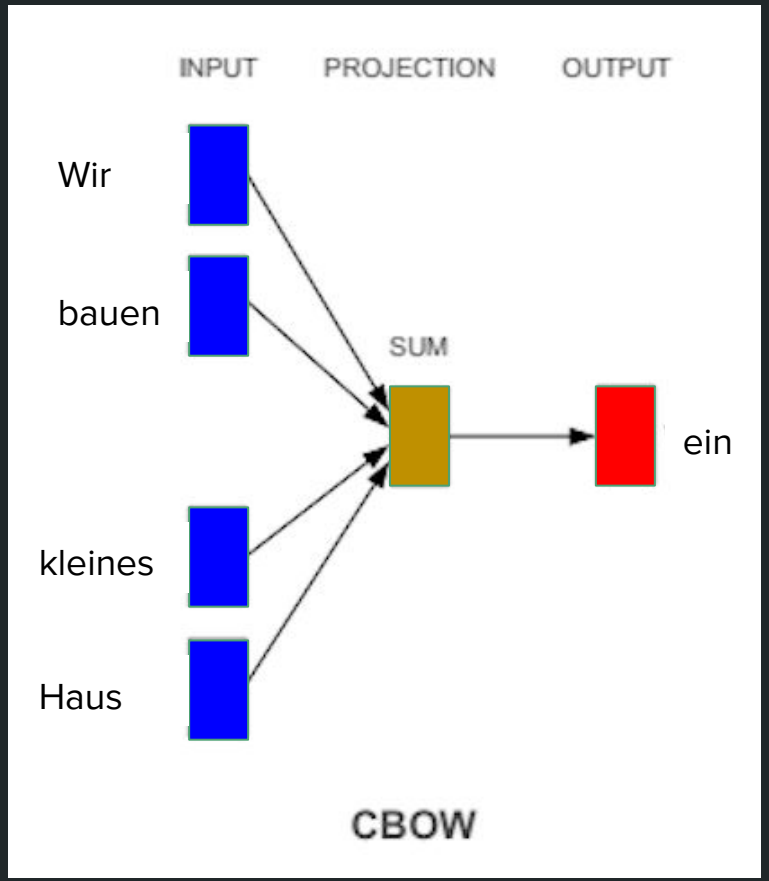
„Language can be described in terms of a distributional structure, i.e., in terms of the occurrence of parts relative to other parts.“ Harris 1954

„You shall know a word by the company it keeps.“ Firth 1957

word2vec

Continuous Bag-of-Words (CBOW)

- Training the prediction of a word given its context
- Order of context not relevant



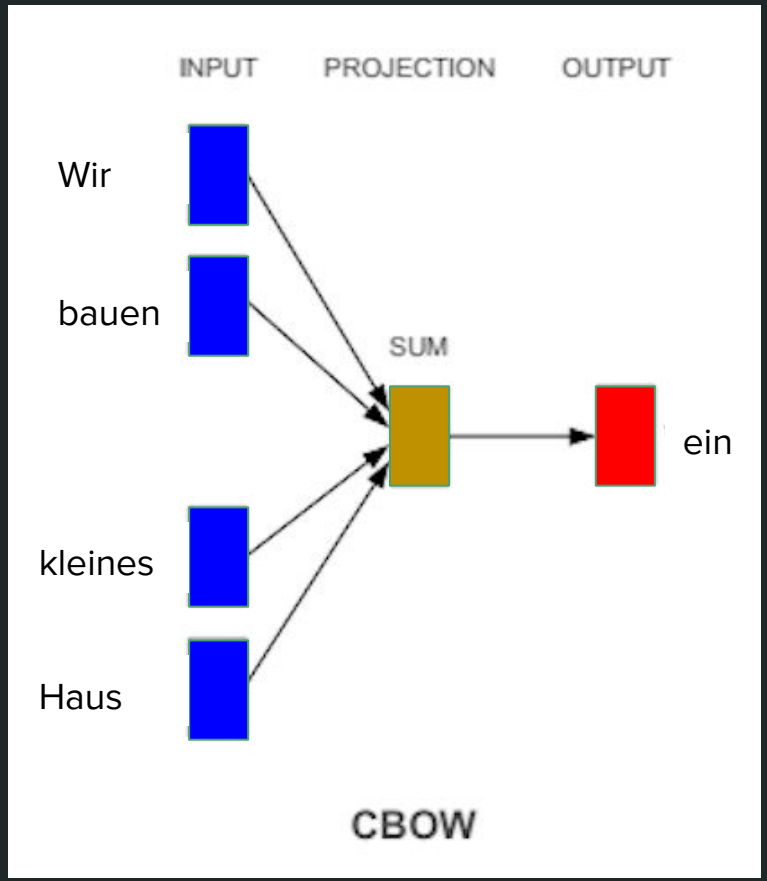
word2vec

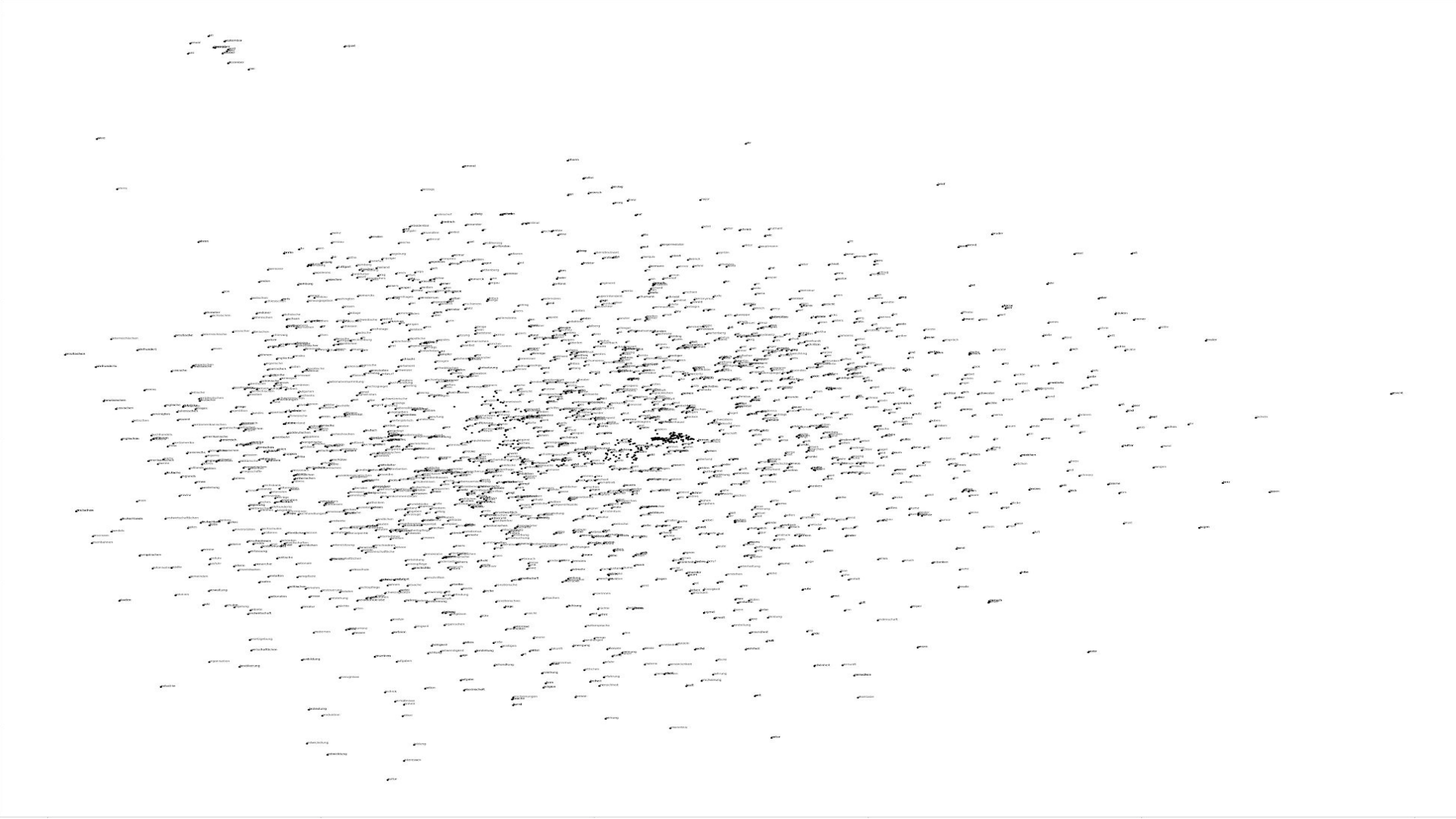
one-hot-encoding

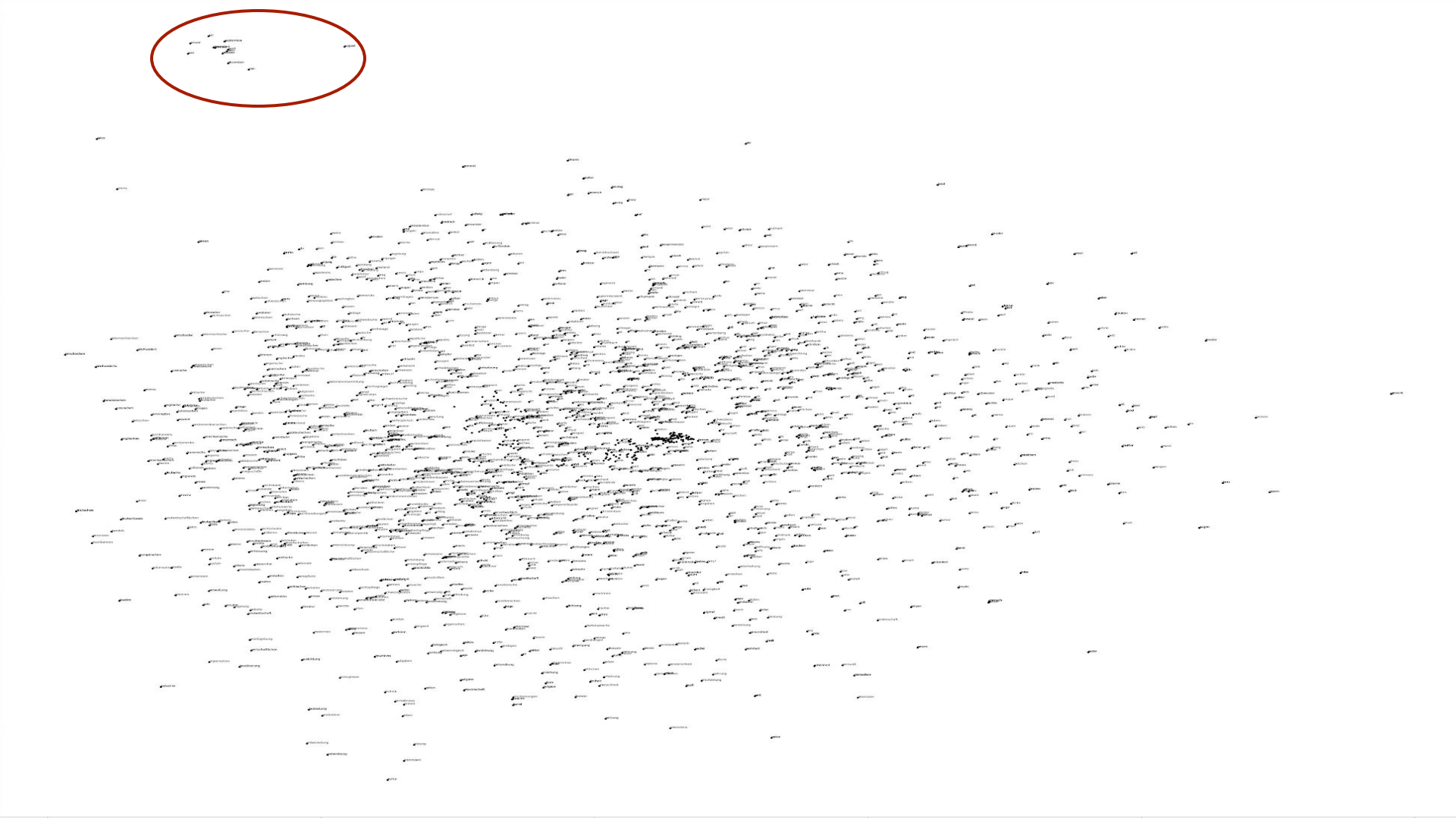
one-hot-encoding

continuous encoding

String	Wir	bauen
one-hot encoding	[1,0,0,0,0]	[0,1,0,0,0]
Continuous encoding	[0.1,1.3]	[4.55,1.8]

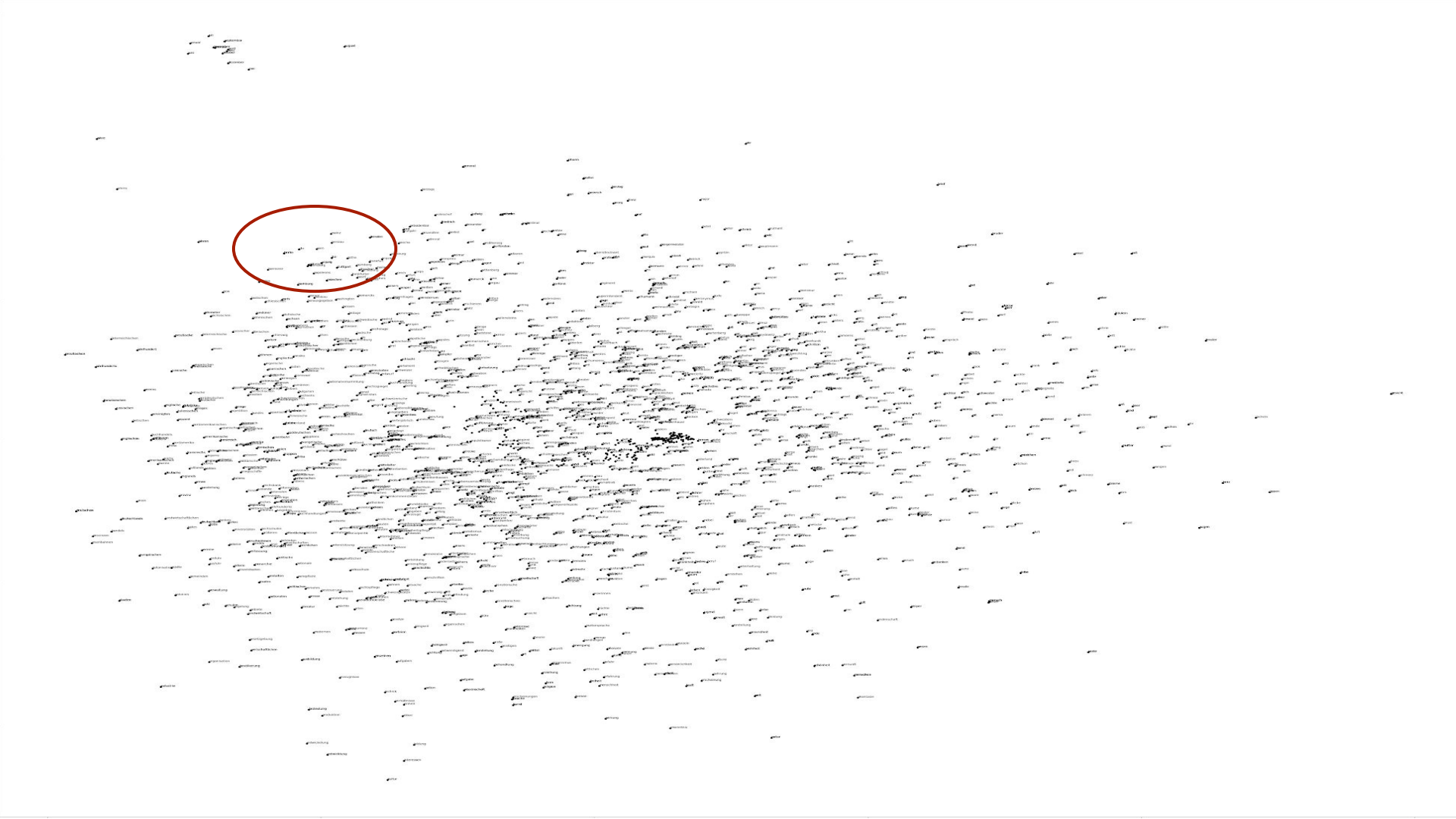




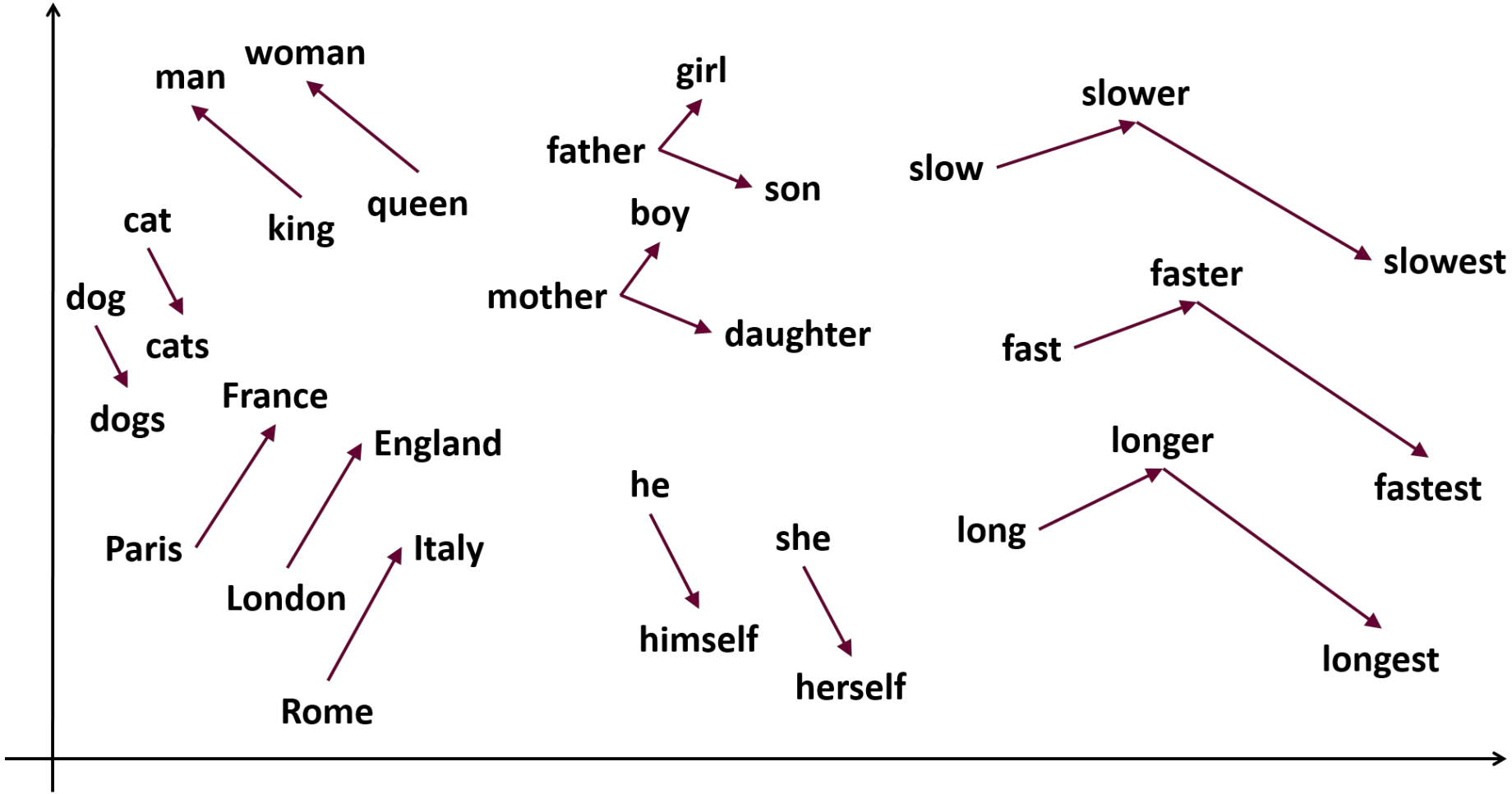


Word Embeddings

Kallimachos




präsiden
lord
frühjahr
mainz
dresden
breslau
pascha
chr
wien
berlin
augsburg
thlr
gotha
leipziger
frankfurt
leipzig
nürnberg
stuttgart
mailand
hannover
napoleons
zürich
corps
frankfurter
prag
erfurt
basel
königlichen
jahr
london
münchen
hamburg
kaisers
neapel
florenz
wiener
bismarcks
badischen
moskau
kopenhagen
paris
washington


man
woman
king
queen
cat
dog
cats
dogs
France
England
Paris
London
Italy
Rome
girl
father
son
boy
mother
daughter
he
himself
she
herself
slower
slow
slowest
faster
fast
fastest
longer
long
longest

Word Embeddings

https://projector.tensorflow.org/

Word2Vec shortcomings

- out-of-vocabulary words
- ambiguity

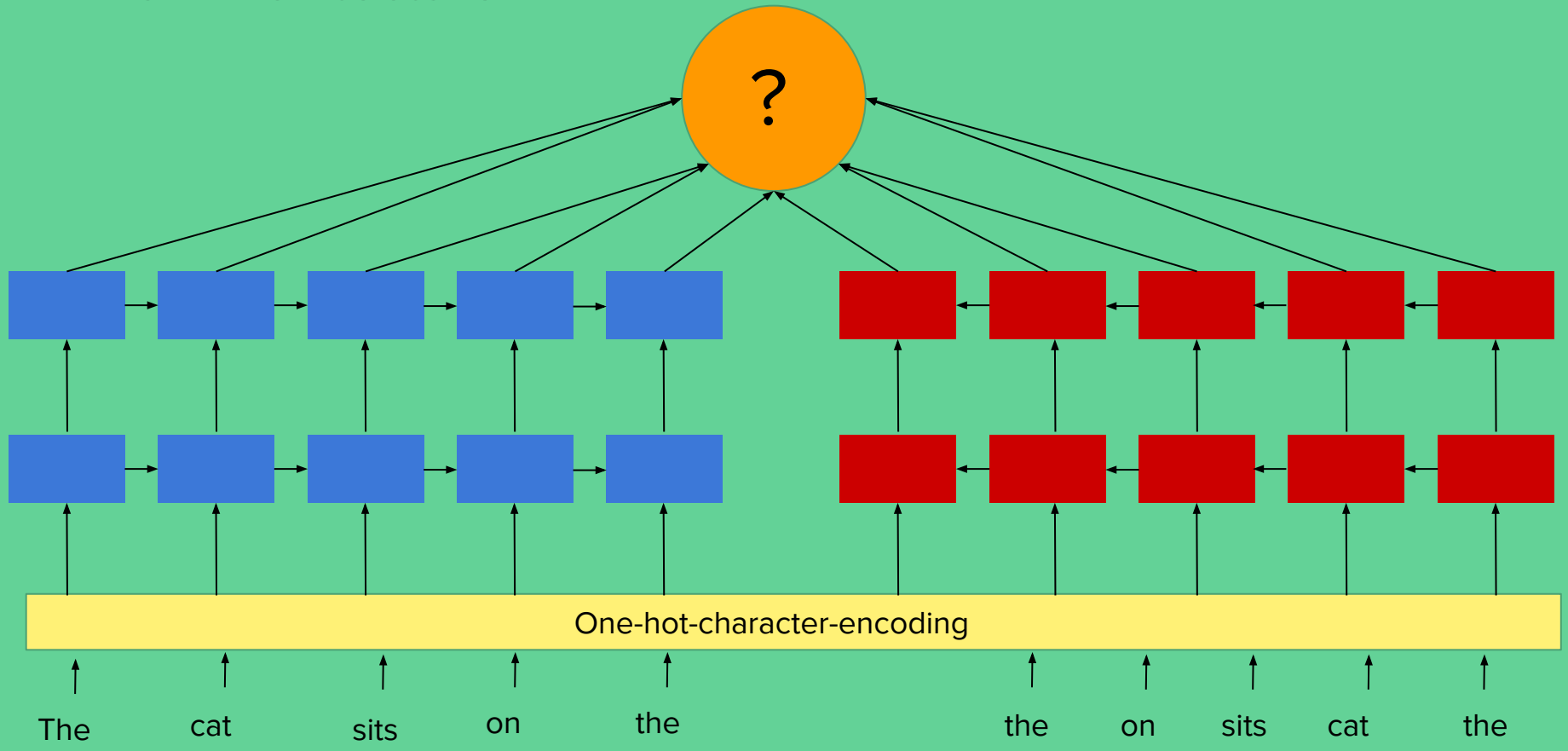
ELMo

Embeddings from Language Models

- context-sensitive
- character-based

Language Model: System cappable of predicting the next word given a sequence of previous words

ELMo - Architecture



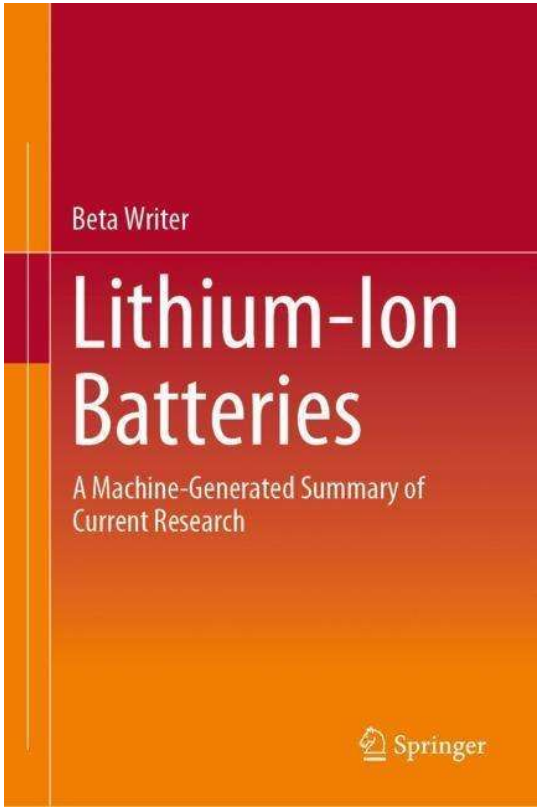
Bert

Bidirectional **E**ncoder **R**epresentations from **T**ransformer

- word pieces instead of token
- Task: Masked Language Model (MLM)
- Transformer instead of LSTMs

Für was werden Sprachmodelle verwendet?

- Translation
- Sentiment/Emotion Detection
- Argumentation Mining
- Text Generation
- Chatbots
- Coreference Resolution
- Question Answering
- Summarization



https://www.youtube.com/watch?v=3UwLhqcZqxc

Burrow’s Zeta

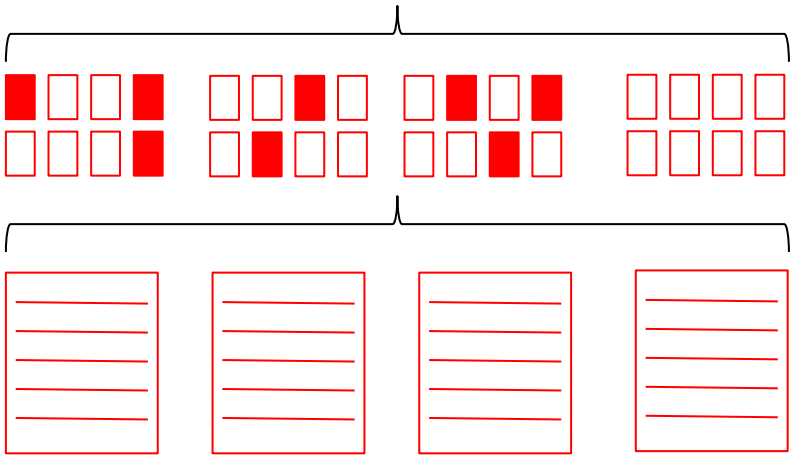
Berechnung von Zeta Wert für “Colt”

Wort in Segment

Wort in nicht Segment

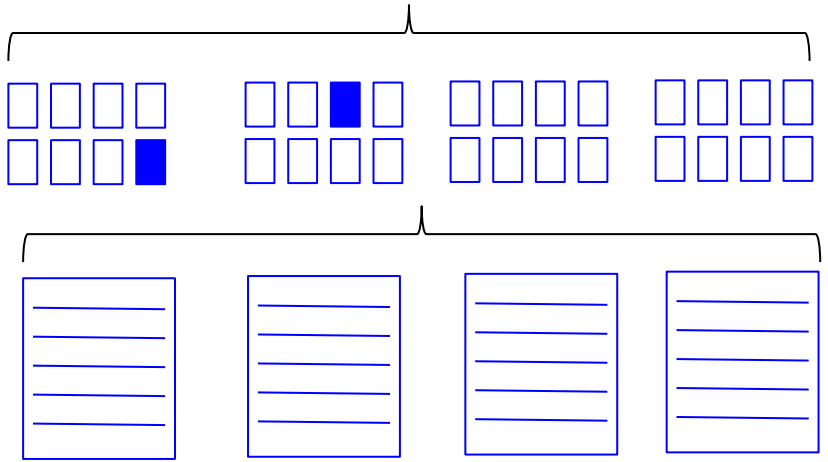
Zeta:
0.25 - 0.06 = 0.19

DP: 8/32 = 0.25



Western

DP: 2/32 = 0.06



Liebesroman

Problemstellung

Autor B verwendet statt “Colt” das Wort “Pistole”

- Das Auftauchen von Schusswaffen wird weniger distinktiv gewertet
- “Colt” nicht Teil der distinktiven Wörter
- Berechnung für Ähnlichkeit von Texten erhält einen Bias zu “Colt”-Western
- Neben Genre hat auch das Autorensignal starken Einfluss auf das Verfahren

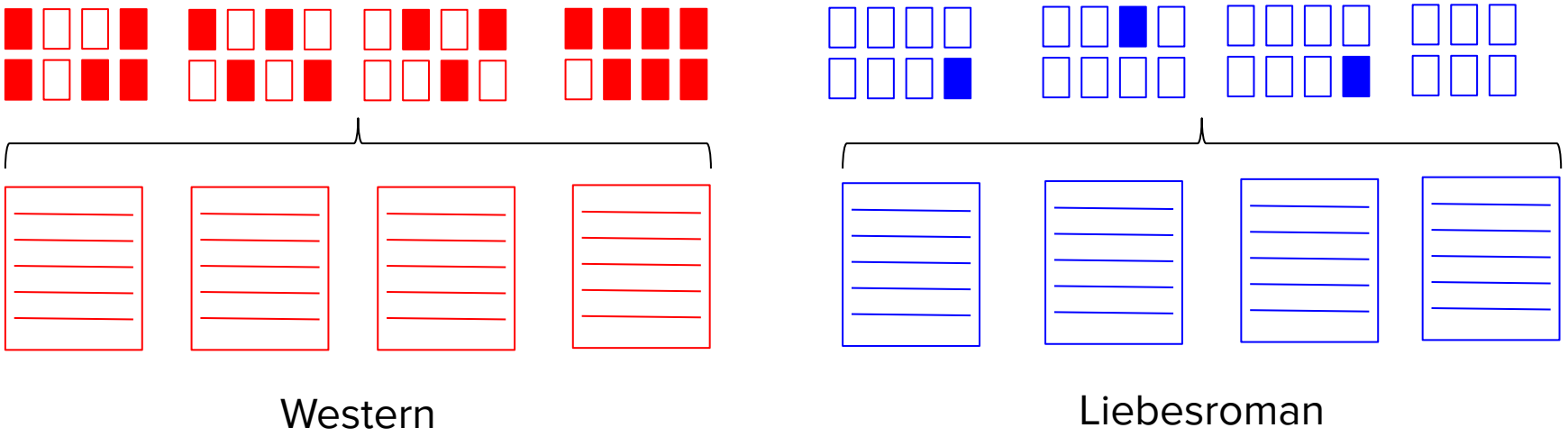
Cluster nach Zeta Wörtern

Semantic Zeta

Ähnliche Wörter werden zu abstrakten Klassen zusammengefasst

Berechnung von Zeta Wert für die Klasse

Colt/Pistole/Schießeisen...

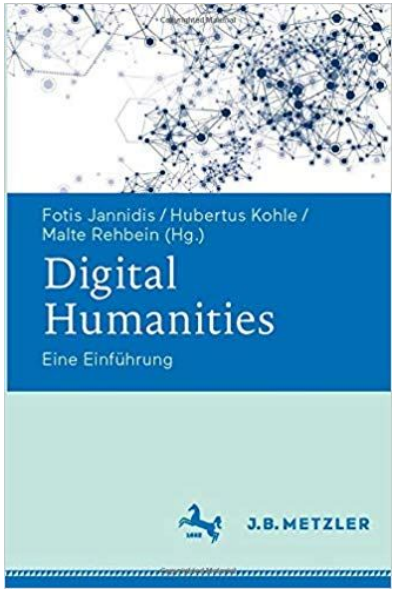


Beziehung der Klassen im Vektorraum

Aus dem Genre Adelsroman				
Tee	Tod	Baron	Reiten	Hauptportal
trinken	Trauer	Stallmeister	Pferde	Suite
Kamillentee	Unglück	Herr	Stall	Raum
Kuchen	Verlust	Fürst	Hengst	Ostflügel
Torte	Grab	Kriminalrat	Stute	Eingangshalle
Aus dem Genre Familienroman				
Haus	Wagen	Kinder	zärtlich	Spinat
Häuschen	Auto	Babys	sanft	Appetit
Nachbarn	Bus	Familien	strich	Eier
Villa	Bollerwagen	Mütter	behutsam	Brot
Hinterhof	Fahrrad	erziehen	küßte	Fleisch

Weiterführende Links

- DH für Historiker https://programminghistorian.org/
- DH für Literaturwissenschaftler https://fortext.net/
- DH für Kunsthistoriker http://www.digitale-kunstgeschichte.de/wiki
- Leipzig Summerschools http://esu.culintec.de
- DSH Journal https://academic.oup.com/dsh
- Cultural Analytics https://culturalanalytics.org/
- Zertifikat “Digitale Kompetenz”



Anprechpartner Uni Würzburg:

- Germantistik: Fotis Jannidis
- Geschichte: Jorit Wintjes, Markus Naser
- Philosophie/klass. Philologie: Dag Hasse
- Klass. Philologie: Holger Essler

Evaluation