Quantitative Analyse

Einführung in die Digital Humanities
2. Februar 2017

Christof Schöch
(Computerphilologie / CLiGS, Universität Würzburg)

Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

CLiGS

GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

Überblick

1. Statistische Grundlagen
2. Kontrastive Analyse
3. Maschinelles Lernen: Klassifikation
4. Maschinelles Lernen: Clustering

1. Statistische Grundlagen

Merkmalserhebung

Beispieltext (Doyle)

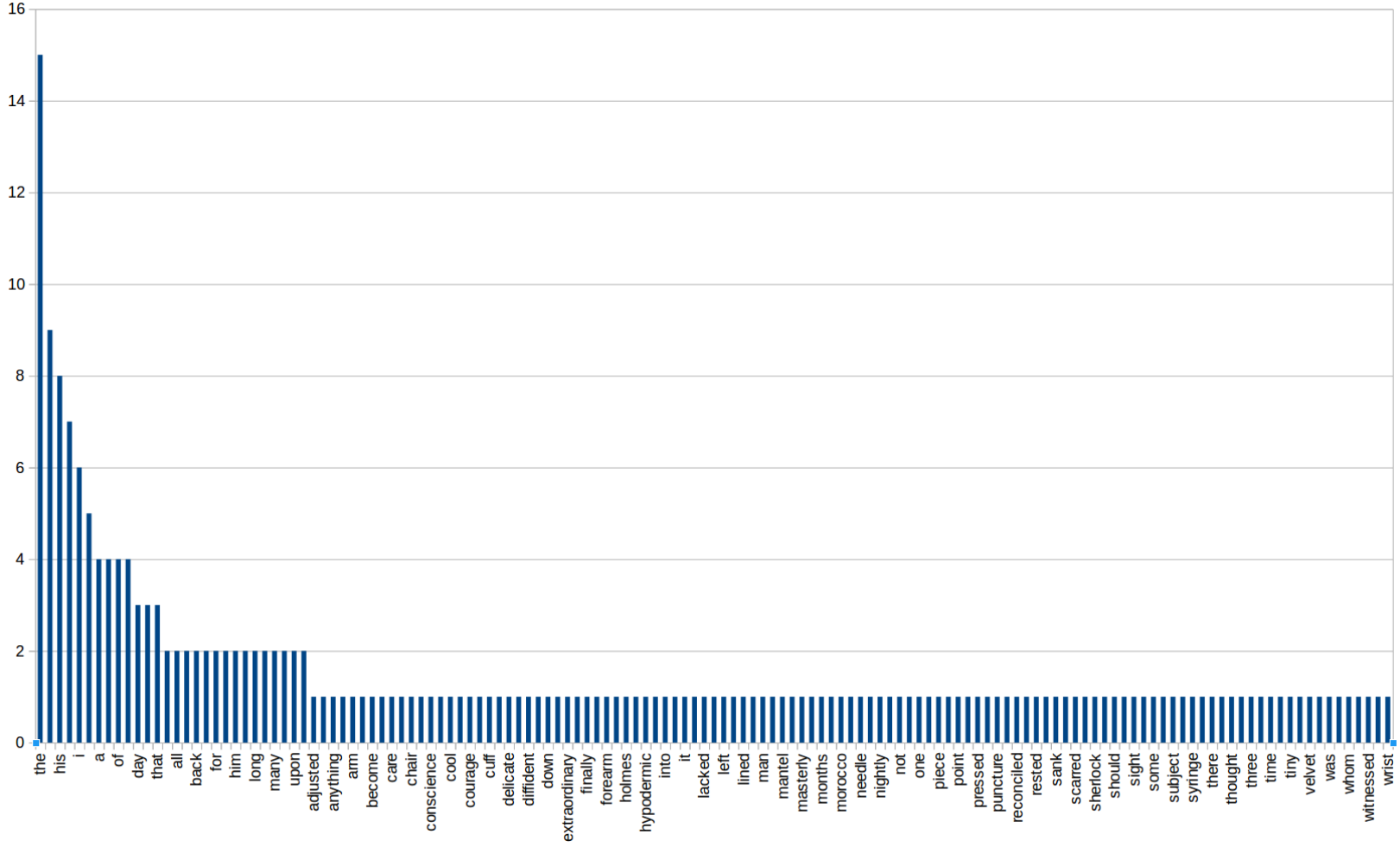
Sherlock Holmes took his bottle from the corner of the mantel-piece and his hypodermic syringe from its neat morocco case. With his long, white, nervous fingers he adjusted the delicate needle, and rolled back his left shirt-cuff. For some little time his eyes rested thoughtfully upon the sinewy forearm and wrist all dotted and scarred with innumerable puncture-marks. Finally he thrust the sharp point home, pressed down the tiny piston, and sank back into the velvet-lined arm-chair with a long sigh of satisfaction.

Three times a day for many months I had witnessed this performance, but custom had not reconciled my mind to it. On the contrary, from day to day I had become more irritable at the sight, and my conscience swelled nightly within me at the thought that I had lacked the courage to protest. Again and again I had registered a vow that I should deliver my soul upon the subject, but there was that in the cool, nonchalant air of my companion which made him the last man with whom one would care to take anything approaching to a liberty. His great powers, his masterly manner, and the experience which I had had of his many extraordinary qualities, all made me diffident and backward in crossing him.

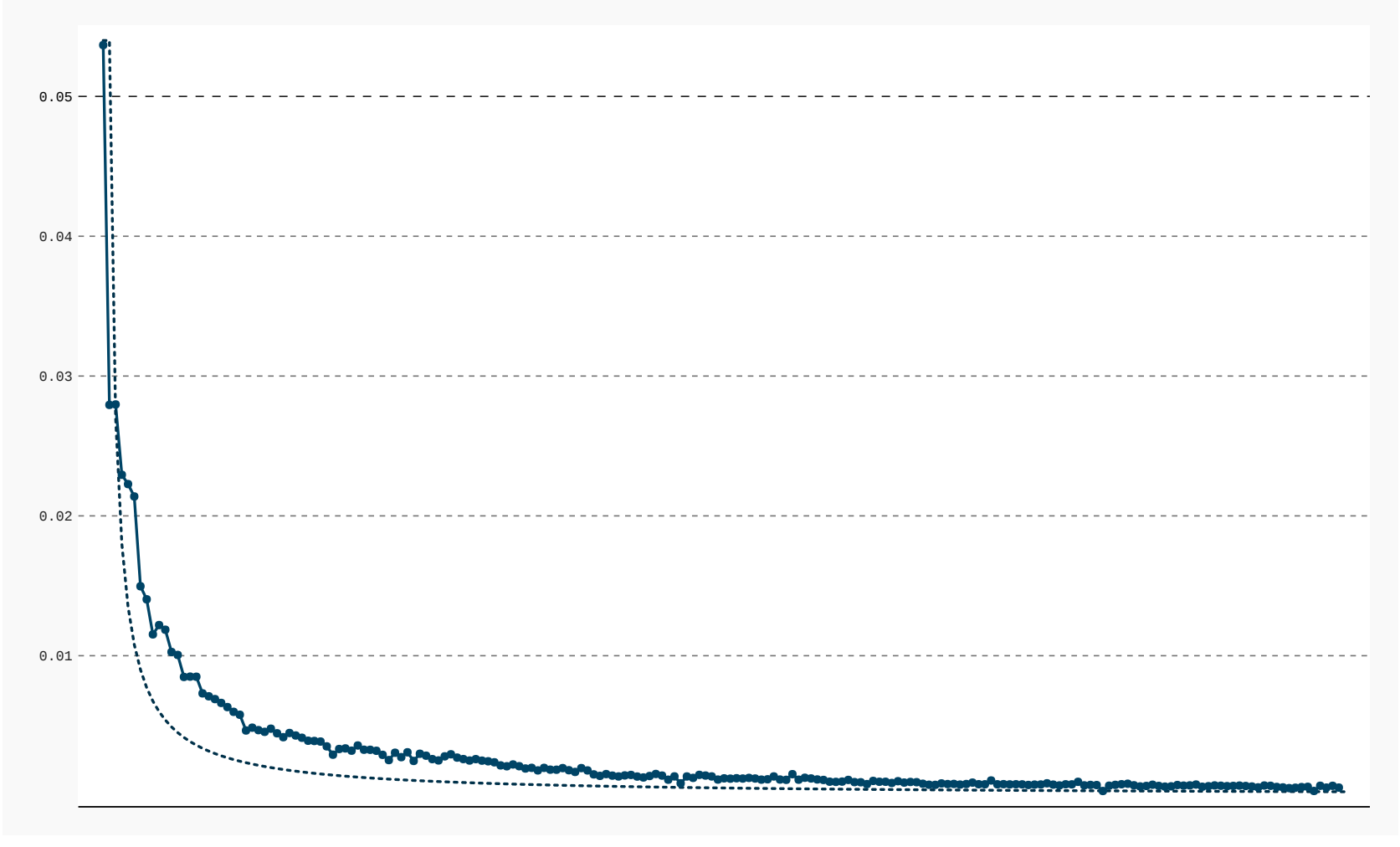
Types und Token

- Begriffe
  - Token: jede einzelne Wortform ist ein Token
  - Type: jede unterschiedliche Wortform ist ein Type
- Beispiel
  - *A rose is a rose is a rose.*
  - 9 Token (A, rose, is, a, rose, is, a, rose, .)
  - 5 Types (A, rose, is, a, .)
- Allgemeiner
  - Token = Untersuchungseinheit: Wort, Satz, Objekt in Bild, Einstellung im Film, etc.
  - Type = Klassen von Token: bestimmte Eigenschaften von bestimmten Tokens

Texte: Typische Häufigkeitsverteilung



Doyle vs. Zipf



Merkmalsmatrix

Rang	Type	The Hound of the Baskervilles	The Sign of Four	A Study in Scarlet	The Valley of Fear	Lost World	The Mystery of Cloomber	Poison Belt	Raffles Haw	The Refugees	The White Company	Summe
1	the	3056	2147	2319	1994	4156	2632	1618	1842	7264	9124	37152
2	and	1505	1154	1313	1343	2000	1442	732	1042	3899	5226	19656
3	of	1580	1105	1195	1424	2502	1362	864	1018	3362	4781	19121
4	to	1381	1070	1070	1256	1678	1218	646	936	2606	3131	14992
...	...											...
21	which	417	235	315	307	639	341	203	280	751	844	4332
22	my	420	323	274	221	542	466	175	213	541	918	4093
23	at	335	311	289	341	430	288	172	196	813	821	3996
24	be	328	260	248	302	413	310	189	262	650	763	3725
...	...											...
101	know	115	69	49	104	74	55	32	73	146	142	859
102	eyes	69	36	59	54	77	34	26	49	219	225	848
103	like	51	58	33	59	127	56	47	47	164	204	846
...	...											...
1001	pocket	7	8	14	14	4	6	1	3	7	7	71
1002	extraordi-nary	8	5	7	5	16	7	3	8	2	8	69
1003	long	9	4	1	10	6	6	0	2	16	9	63
...	...											...
Summe		69802	51583	51218	69676	88221	56534	34544	44910	146303	178233	791024

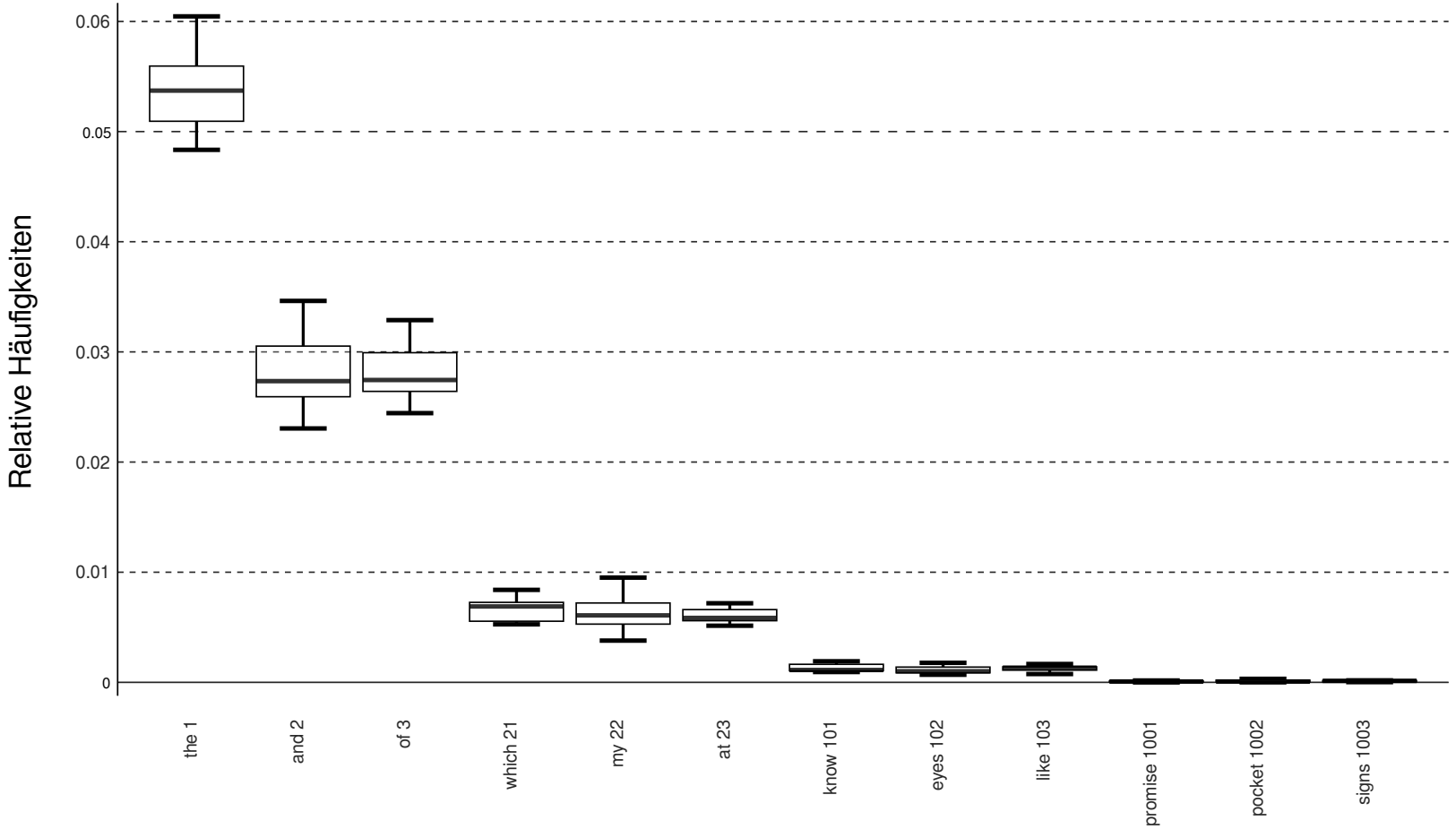
Basisoperationen

- Summen über Spalten oder Zeilen
- Absolute und relative Häufigkeiten
- Maße der zentralen Tendenz
  - Mittelwert
  - Median
- Streuungsmaße
  - Spannweite
  - Standardabweichung

Standardabweichung: Berechnung

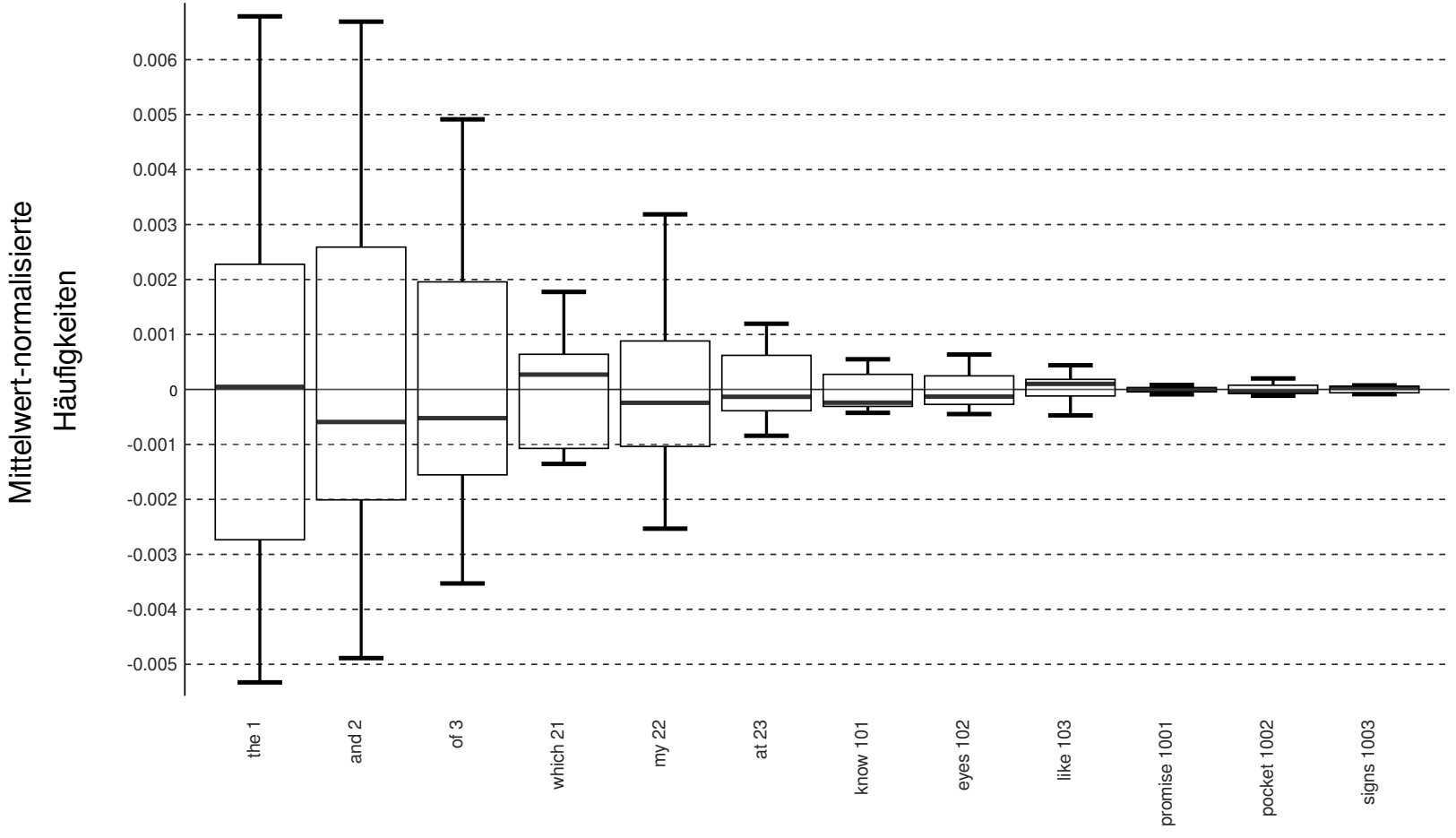
$$\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{n}}$$

Normalisierung und Standardisierung



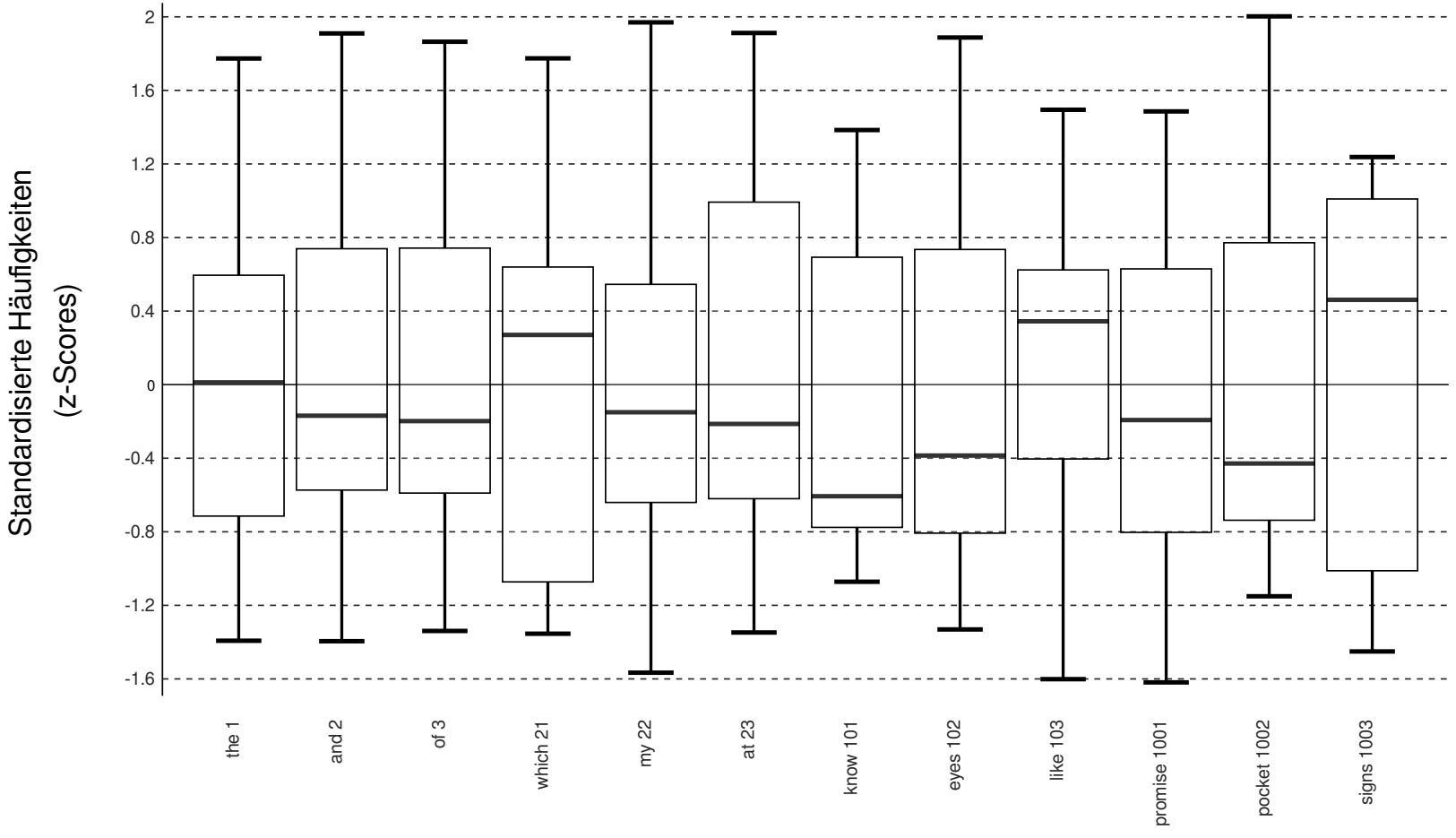
(Relative Häufigkeiten)

Mittelwert-Normalisierung



(Jeder Wert minus Mittelwert des Types)

Z-Scores



(z-score der Werte)

Z-Score: Berechnung

$$z = \frac{x - \mu}{\sigma}$$

Zwischenbilanz

- Deskriptive (univariate) Statistik (Kennwerte)
- Multivariate Statistik (Korrelation, Regression)
- Inferenz-Statistik (Hypothesen-Prüfung)

2. Kontrastive Analyse

Grundidee: Vergleich

- Gleichrangige Partitionen
  - Zwei Autoren: Shakespeare vs. Marlowe
  - Zwei Gattungen: Tragödien vs. Komödien
  - Zwei Epochen: Klassik vs. Romantik
- Unterschiedliche Partitionen
  - Tragikomödien vs. alle übrigen Dramengattungen
  - Ein Roman vs. das gesamte Romanwerk des Autors
  - Autor vs. Referenzkorpus

Verhältnis der relativen Häufigkeiten

$$rrf_i = \frac{rf_i(Z)}{rf_i(V)}$$

- Z = Zielpartition
- V = Vergleichspartition
- rfi = relative frequency von Wort i

Zeta-Maß (1: document proportions)

$$dp_i(Z) = \frac{df_i(Z)}{n(Z)}$$

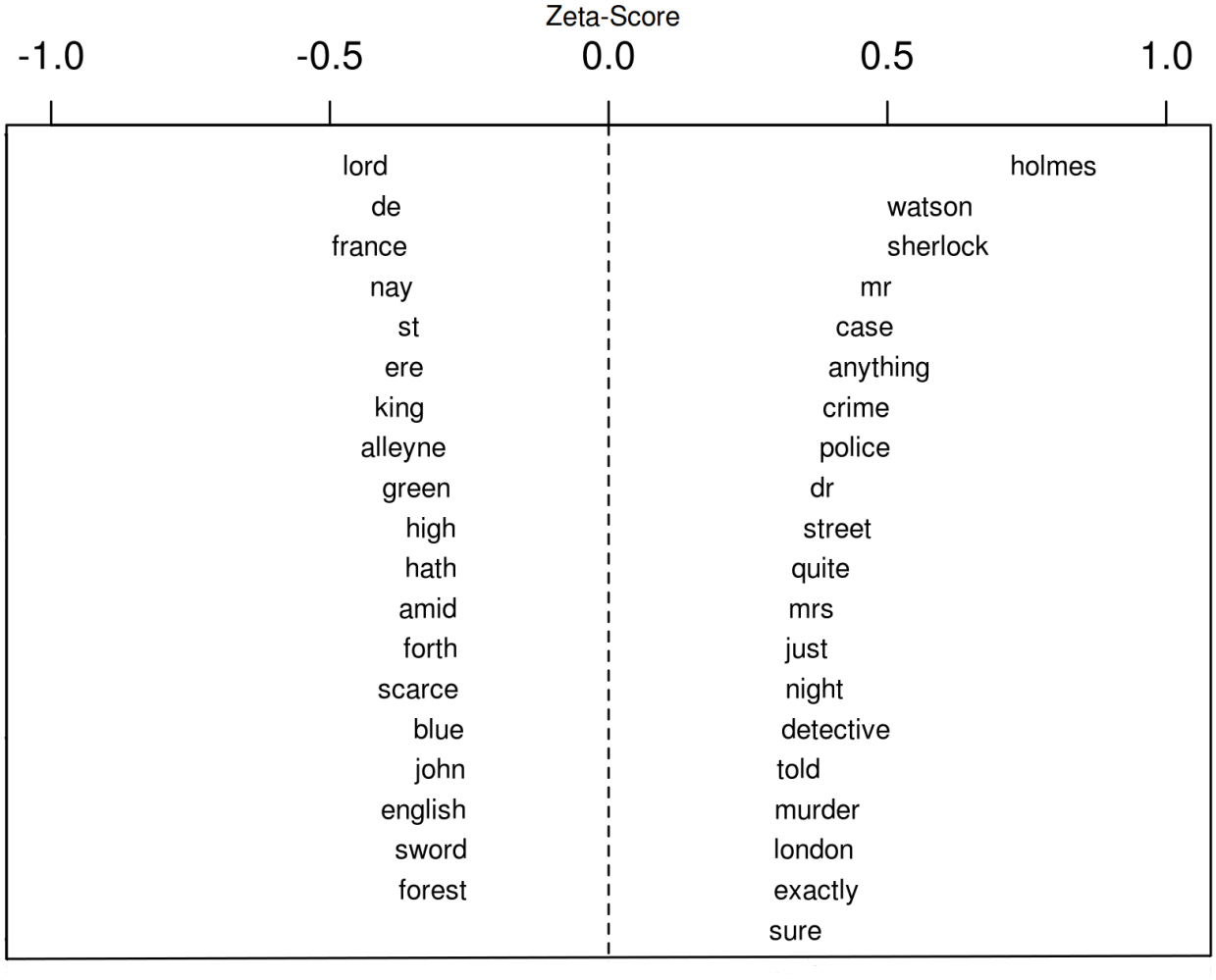
- dfi = document frequency von Wort i
- n = Anzahl der Dokumente

Zeta-Maß (2)

$$Zeta_i = dp_i(Z) - dp_i(V)$$

Zeta: einfache Subtraktion

Zeta: grafische Darstellung



(Doyle: andere Romane vs. Detektivromane)

3. Maschinelles Lernen: Klassifikation

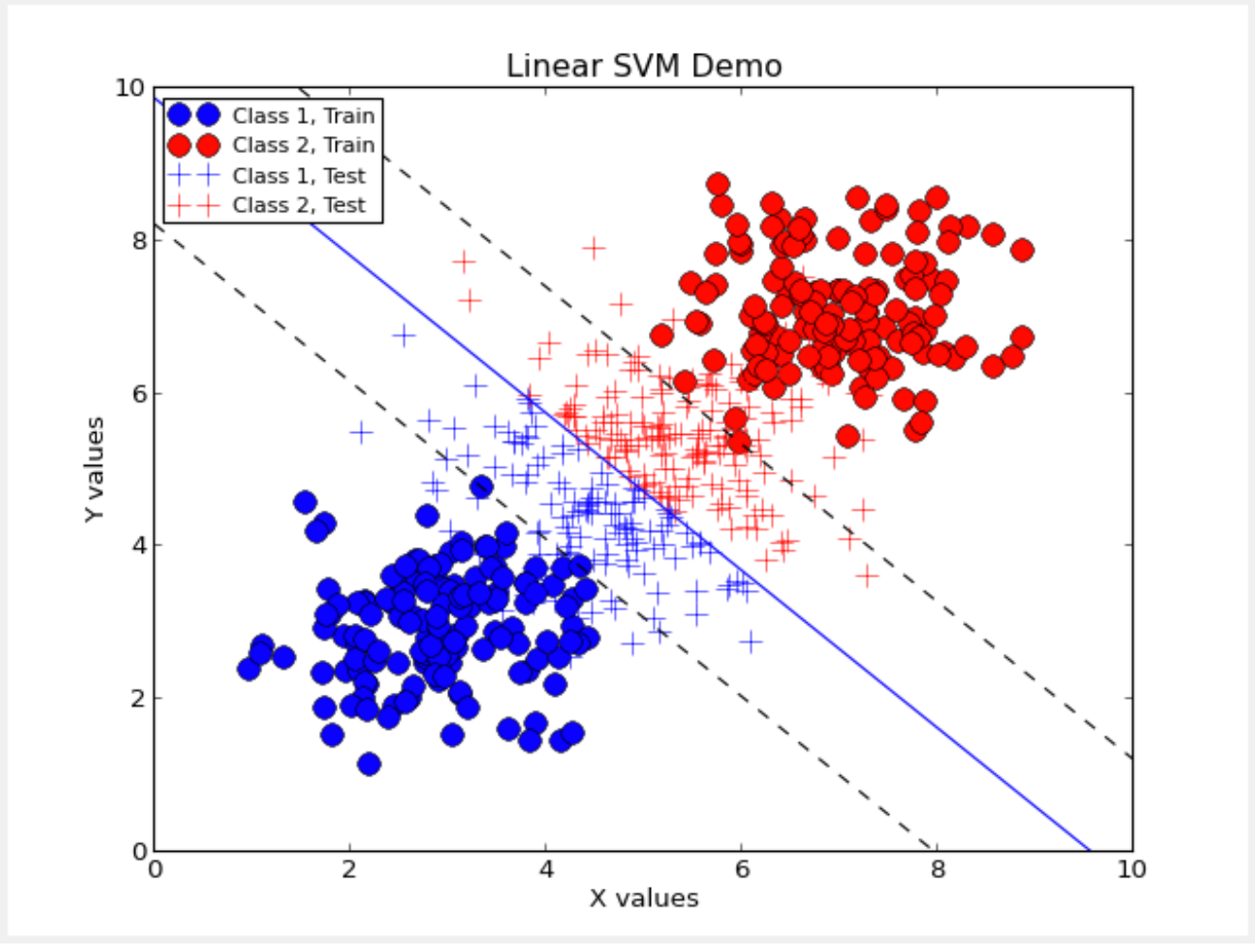
Klassifikation vs. Clustering

- Klassifikation
  - Klassen sind vorher bekannt
  - Drei Phasen: Lernen, Testen, Anwenden
  - Klassifizierte Lern- und Testdaten
- Clustering
  - Keine Klassen vorgegeben
  - Nur eine Phase: Clusterbildung
  - Cluster: Gruppe von ähnlichen Items
  - Keine Annotation notwendig

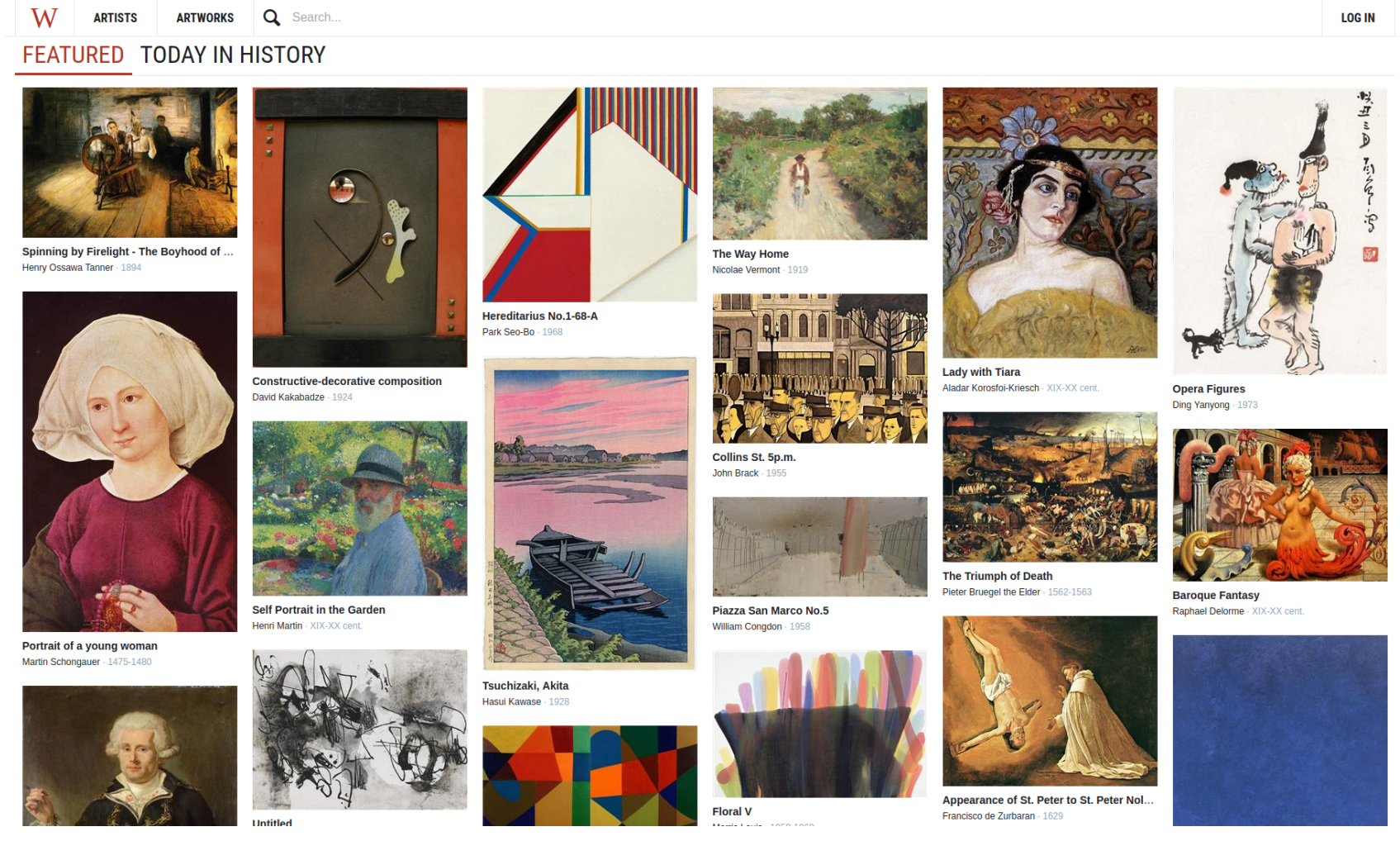
Vorgehen bei der Klassifikation

1. Vorbereitung
2. Annotation
3. Merkmalsgenerierung
4. Trainingsphase ("classifier")
5. Evaluationsphase
6. Anwendungsphase

Ein Classifier: Support Vector Machines



Beispiel: Gemälde und Epochen



https://www.wikiart.org/

Eckdaten

- Saleh & Elgammal: "Large-scale Classification of Fine-Art Paintings", 2016
- Bildsammlung: 63.691 Gemälde von Wikiart
- Klassen: 10 Epochenstile (bspw. Romantik, Neoklassik, Abstract Expressionism)
- Einfache Merkmale (bspw. Farbwerte von Pixeln) und komplexe Merkmale (e.g. repräsentierte Objekte)
- Performanz: F-Score 0.32 (vgl. die Baseline: 0.23)
- Fehleranalyse: Abstract Expressionism vs. Action Painting

4. Maschinelles Lernen: Clustering

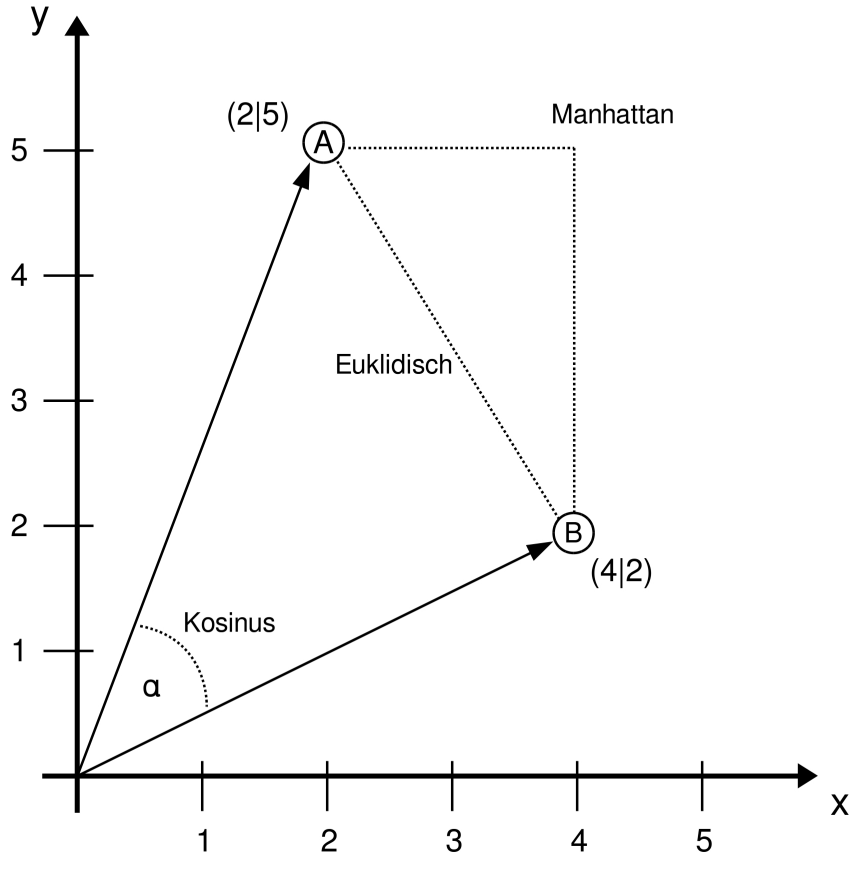
Vieldimensionaler Vektorraum

- vgl. Merkmals-Matrix
- Nachteile: Informationsverlust ("bag of words")
- Vorteile: bestimmte Berechnungen werden möglich

Vorgehen beim Clustering

1. Vorbereitung der Textsammlung
2. Merkmals-Matrix (Types x Texte)
3. Merkmals-Behandlung (Auswahl / Normalisierung)
4. Distanzmaß => Distanzmatrix
5. Clustering (=> "linkage matrix")
6. Visualisierung (bspw. als Dendrogramm)
7. Interpretation

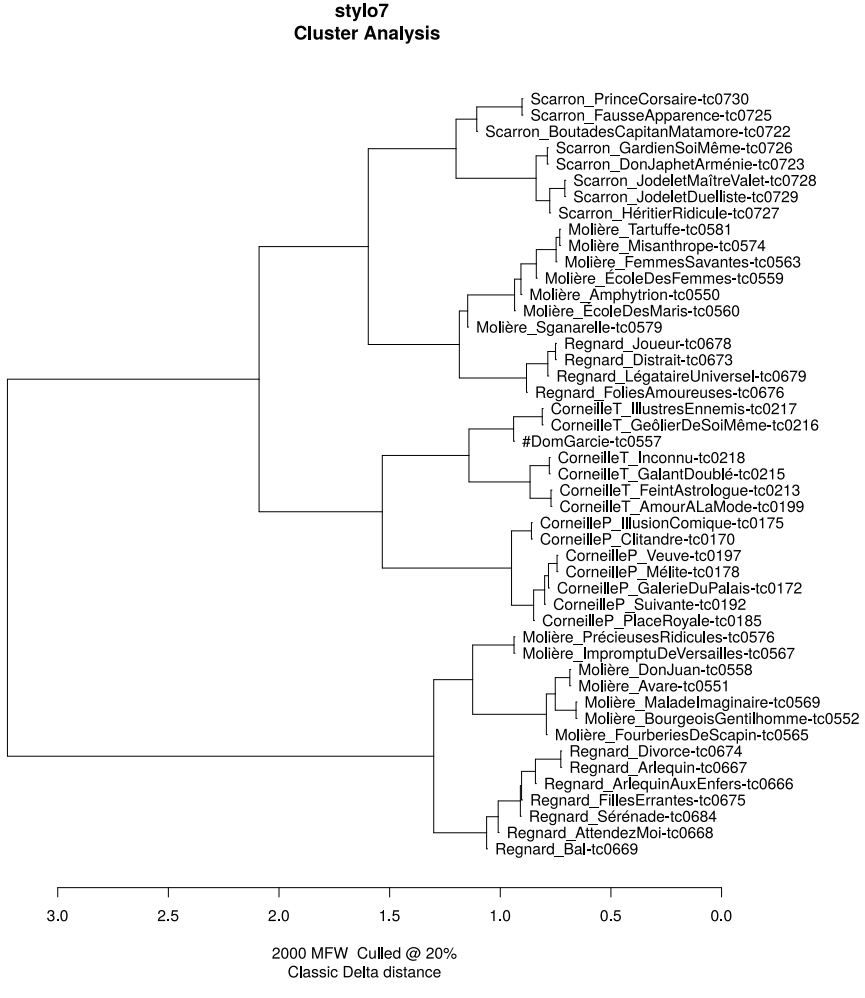
Verschiedene Distanzmaße



Beispiel: Autorschaftsattribution

- Französische Theaterstücke
- Mehrere Autoren
- Überwiegend Texte unstrittiger Autorschaft
- Ein strittiger Text: "Dom Garcie" (Molière, evtl. Pierre Corneille)
- Parameter: 2000 MFW, z-scores, Delta, Ward

Ergebnis


stylo7
Cluster Analysis
Scarron_PrinceCorsaire-tc0730
Scarron_FausseApparence-tc0725
Scarron_BoutadesCapitanMatamore-tc0722
Scarron_GardienSoiMême-tc0726
Scarron_DonJaphetArménie-tc0723
Scarron_JodeletMaîtreValet-tc0728
Scarron_JodeletDuelliste-tc0729
Scarron_HéritierRidicule-tc0727
Molière_Tartuffe-tc0581
Molière_Misanthrope-tc0574
Molière_FemmesSavantes-tc0563
Molière_ÉcoleDesFemmes-tc0559
Molière_Amphytrion-tc0550
Molière_ÉcoleDesMaris-tc0560
Molière_Sganarelle-tc0579
Regnard_Joueur-tc0678
Regnard_Distrait-tc0673
Regnard_LégataireUniversel-tc0679
Regnard_FoliesAmoureuses-tc0676
CorneilleT_IllustresEnnemis-tc0217
CorneilleT_GeôlierDeSoiMême-tc0216
#DomGarcie-tc0557
CorneilleT_Inconnu-tc0218
CorneilleT_GalantDoublé-tc0215
CorneilleT_FeintAstrologue-tc0213
CorneilleT_AmourALaMode-tc0199
CorneilleP_IllusionComique-tc0175
CorneilleP_Clitandre-tc0170
CorneilleP_Veuve-tc0197
CorneilleP_Mélite-tc0178
CorneilleP_GalerieDuPalais-tc0172
CorneilleP_Suivante-tc0192
CorneilleP_PlaceRoyale-tc0185
Molière_PrécieusesRidicules-tc0576
Molière_ImpromptuDeVersailles-tc0567
Molière_DonJuan-tc0558
Molière_Avare-tc0551
Molière_MaladeImaginaire-tc0569
Molière_BourgeoisGentilhomme-tc0552
Molière_FourberiesDeScapin-tc0565
Regnard_Divorce-tc0674
Regnard_Arlequin-tc0667
Regnard_ArlequinAuxEnfers-tc0666
Regnard_FillesErrantes-tc0675
Regnard_Sérénade-tc0684
Regnard_AttendezMoi-tc0668
Regnard_Bal-tc0669
3.0
2.5
2.0
1.5
1.0
0.5
0.0
2000 MFW Culled @ 20%
Classic Delta distance

Schluss

Weitere quantitative Methoden

- Topic Modeling
- Netzwerkanalyse
- "deep learning" / "word2vec" (neuronale Netze)
- Sentiment Analysis
- Allgemein: Natural Language Processing

Lektüreempfehlungen

Kapitel zur Sitzung

- "Quantitative Analyse", in: *Digital Humanities: Eine Einführung* , hg. von Fotis Jannidis, Hubertus Kohle und Malte Rehbein. Stuttgart: Metzler, 2017. http://link.springer.com/book/10.1007/978-3-476-05446-3

Weiterführende Lektüre

- Alpaydin, Ethem. Introduction to Machine Learning. 2nd ed. Cambridge MA: MIT Press, 2010.
- Burrows, John. "'Delta': A Measure of Stylistic Difference and a Guide to Likely Authorship." *Literary and Linguistic Computing* 17, no. 3 (2002): 267–87. doi:10.1093/llc/17.3.267.
- Jannidis, Fotis. "Methoden der computergestützten Textanalyse." In *Methoden der literatur- und kulturwissenschaftlichen Textanalyse*, hg. von Ansgar Nünning und Vera Nünning, 109–32. Stuttgart & Weimar: Metzler, 2010.
- Lijffijt, Jefrey et al. "Significance Testing of Word Frequencies in Corpora." *Digital Scholarship in the Humanities* 31, no. 2 (2014): 374–97. doi:10.1093/llc/ fqu064.
- Oakes, Michael P. Statistics for Corpus Linguistics. Edinburgh: Edinburgh Univ. Press, 1998.
- Stamatatos, Efstathios. " A Survey of Modern Authorship Attribution Methods." *Journal of the Association for Information Science and Technology* 60, no. 3 (2009): 538–556. doi:10.1002/ asi.v60:3.

Danke!

Christof Schöch, 2017

christofs.github.io
