technische universität
dortmund

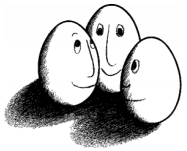
Fakultät für Informatik
LS 8

RapidMiner als Werkzeug für die textorientierten Geisteswissenschaften

Katharina Morik

technische universität
dortmund

Fakultät für Informatik
LS 8



Informatik:

Methoden +
Verfahren

Linguistik:

dortmund

Forschungsfragen,
Anforderungen + Evaluation

GEFÖRDERT VOM

Bundesministerium
für Bildung
und Forschung

BBAW
berlin-brandenburgische
AKADEMIE DER WISSENSCHAFTEN
DWDS

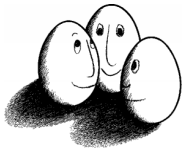
CLARIN-D
WEBLICHT
WEB-BASED
LINGUISTIC CHAINING TOOL

SfS Uni Tübingen

INSTITUT FÜR
DEUTSCHE SPRACHE
Mitglied der Leibniz-Gemeinschaft

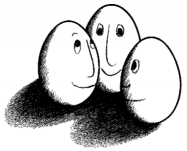
IdS

Sprachtechnologie-Partner aus CLARIN:
Daten, Werkzeuge, Infrastrukturen



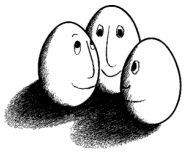
Zielsetzung

- Routineaufgaben durch maschinelles Lernen unterstützen:
  - Klassifikation
  - Ausreißerentdeckung
  - Clustering
  - Exploration der Daten
- Korpus-basierte Sprachforschung für einen breiten Anwenderkreis:
  - ETL: Extract Transform Load
  - Repräsentation von Texten
  - Leicht bedienbares Werkzeug
  - Theorie-basierte Methoden



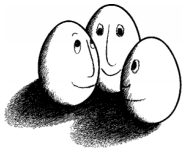
KobRA Studien

- Repräsentation von Texten
  - Bag of Words
  - Meta-Daten
- Klassifikation
  - Stützverb vs. Vollverb
- Topic Models
  - Latent Dirichlet Allocation



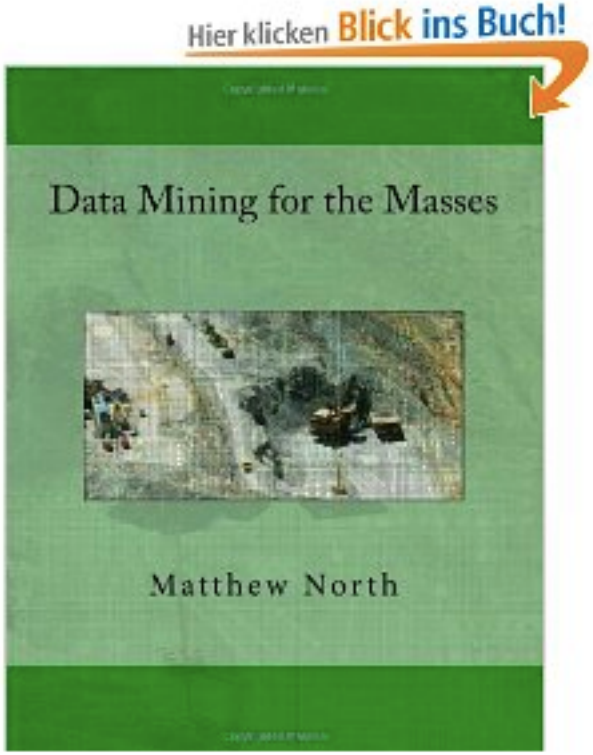
Repräsentationen

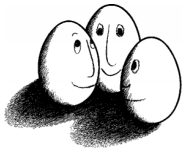
- **Bag of words** Repräsentation der Dokumente als Vektor, wobei jede Komponente ein Wort des Korpus ist –
  etwa 10 000 bis 100 000 Dimensionen!
  - Jedes Wort ist wichtig: Lässt man auch nur ein Wort weg, das bzgl. der Korrelation mit der Klasse auf Rang 9947 ist, sinkt die Klassifikationsgüte.
  - Zwei Texte über das selbe Thema müssen kein einziges gemeinsames Wort enthalten (außer Stoppwörter).
  - Wortvektoren sind dünn besetzt– die meisten Wörter kommen in einem Dokument nicht vor.
  - T. Joachims 2002 *Learning to Classify Text using Support Vector Machines*
- **Baum** Repräsentation für morphologische and grammatische Modelle -- anspruchsvoll.



RapidMiner

- Datenintegration, Analytical „Extract Transform Load“, Datenanalyse und Reporting in einer einzigen Suite
- Mächtige und zugleich intuitive graphische Benutzeroberfläche zum Design von Analyseprozessen
- Repositories zur Prozess-, Daten- und Metadatenverwaltung
- Einzige Lösung, welche eine On-the-Fly Fehlererkennung und Quick Fixes unterstützt
- Vollständig und flexibel: Hunderte Methoden für Datenintegration, Datentransformation, Modellierung und Visualisierung





Texte - Daten – Beispiele bringenW-N_K2

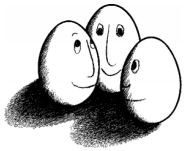
Result Overview | ExampleSet (Filter Examples (2))

Data View | Meta Data View | Plot View | Advanced Charts | Annotations

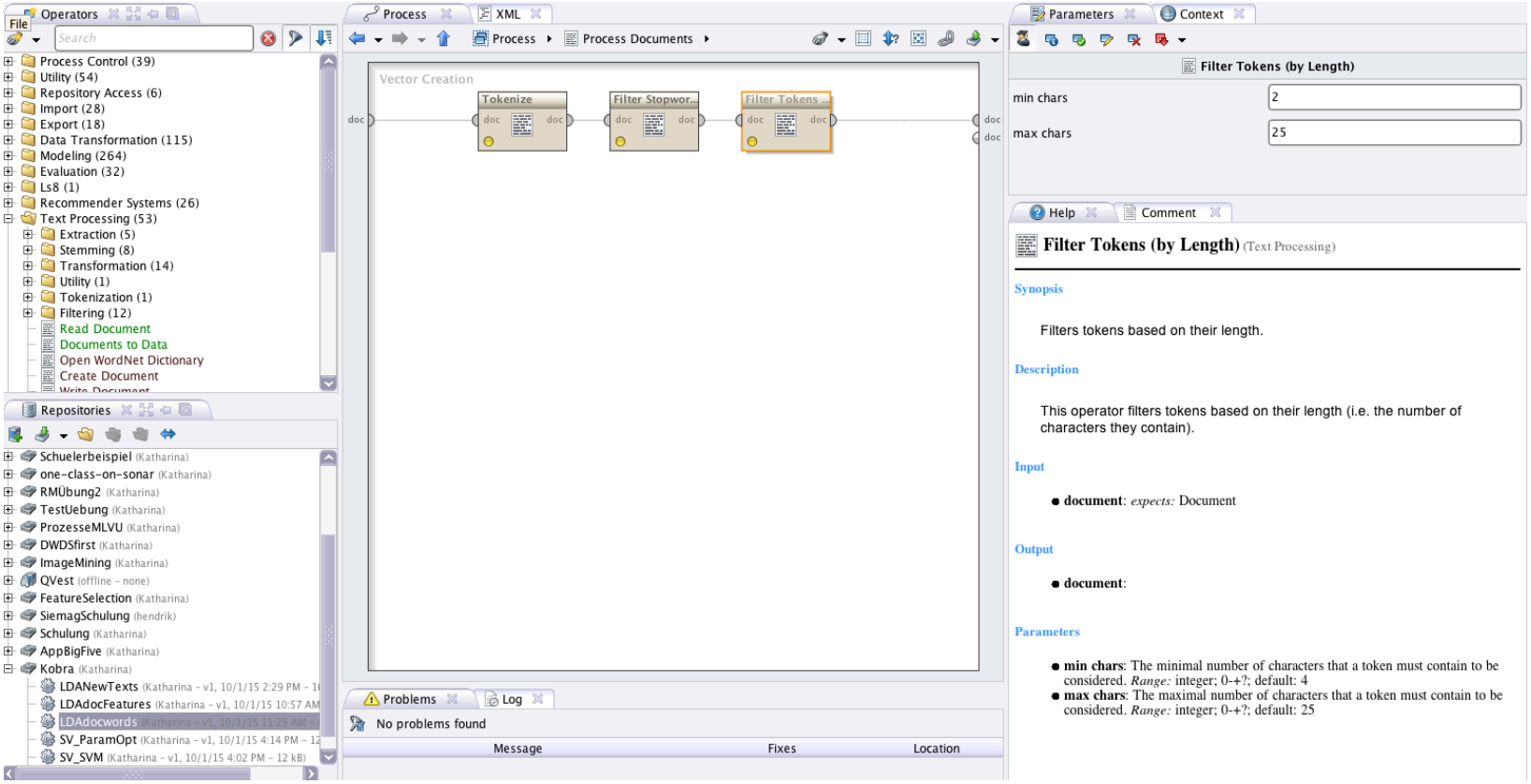
ExampleSet (181 examples, 1 special attribute, 1 regular attribute)

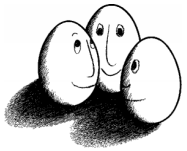
Row No.	SV	Beleg
1	-1	fragte Doktor Langhals, indem er dem kleinen Johann mit seinen eitlen Augen ins Gesicht blickte... Hanno verfärbte sich. W
2	-1	Dieser Transport hatte den noch nicht ganz trockenen Bildern nicht wohl getan. Der junge Maler, den man so wenig aufmu
3	-1	Dabei hörte ich aus dem Innern der schönen Kirche den Klang der Orgel und fühlte die heftige Versuchung, wieder einmal
4	-1	Sie wußte jetzt, daß im Augenblicke des Erwachens vorhin, als sie, die Quittung schreibend, begriff, daß es Abend war, de
5	-1	Herr Piesch sieht seinen Sohn an, spürt dessen Mißtrauen und spürt auch dessen Eifersucht. Das bringt ihn in Laune, er lacl
6	-1	Als der Wagen hielt, wälzte sich Pakulat aus der Tür, drückte den Daumen auf die Klingel, drückte und rüttelte am Gartent
7	1	Und schließlich sah ich den Mann, den Professor, auf einem Kinderhocker sitzen. Es war ihm ganz gleichgültig, daß ich kam
8	-1	Am Vorabend von Illes Geburtstag, sie war am 14. Juli 1912 geboren, war ich noch einmal an den Frauenzaun gegangen, s
9	1	Ich hangele mich wieder in die Zentrale zurück und klettere nach oben. Über den Achtersteven wird das Boot von seinen E
10	-1	Man wußte ja gar nicht, ob man da eigentlich lachen dürfe, und dann war wieder nicht klar, ob da eigentlich diese schaude
11	-1	Sie machen die großen Blätter ab und tun die Kolben heraus. Die werden dann ins Stall gebracht und geschichtet... « » Un
12	-1	In der Fahrt zwang etwas ihn immer wieder, sich selber zu bespiegeln: Ich hätte nicht zittern sollen, als die Nachricht mich
13	1	Und dann, am Fuß der Treppe, traf mich ein Stück des brennenden Geländers, fast an der gleichen Stelle, an der mich dar
14	-1	Wenn du nicht weißt, was du mit deinen Gästen anfangen sollst, dann hol dein Album und deine Tafeln. Das bringt immer S
15	1	In diesem Falle möchte ich nicht gerade behaupten, daß der neue Bezirksrichter von Andruss drauf aus gewesen wäre, sei
16	-1	Und wie er so länger sitzt und am liebsten möcht er den Dluga erschlagen, da hängt er sich an einen Menschen, einen Einb
17	-1	Nicht nötig? Hugo, Hugo, bring die Herrschaften auf Zimmer 212 und nimm eine Weinkarte mit ich werde die jungen Herrs
18	1	Gerade diese Ihre Wünsche zeigen mir, daß wir uns bisher mißverstanden haben. Wenn ich einem Fremden gestatte, den A
19	1	Sogleich nach dem Tode Stephans hatte der Arzt seine Frau von all ihren sonstigen Pflichten befreit und in das Krankenzelt
20	-1	Freilich, die Geschäfte waren nicht mehr gut, seit die Hofgesellschaft sich nur noch auf Schloß Monbijou beschränkte. Aber
21	1	Ich beginne, den weg zu ihm zurück zu legen. zögernd bringe ich mich in bewegung, setze einen fuss vor den andern. mei
22	-1	Der der? War der es nicht, der ihm dieses Geschenk ins Haus brachte, dies seltsame Wesen, an dem er nun zugrunde gir
23	1	Wieder auch ministrierte das Zwillingspaar Gisbert und Vincenz. Sie sangen gemeinsam mit Maximilian die Sequenz, und d
24	-1	Komisch, wie unfeierlich sie wirken, ich kann das Gefühl nicht loswerden, die da oben spielen Richter, wie wir als Kinder Pf
25	-1	So eine Dusche ist eine Wohltat. Nun bring mir heißen Tee, aber schnell, Boy! Ich klopfte dem schwarzen Jungen auf die S
26	1	Und was sie auch sagten und von ihren Präparaten herbeitrugen, es geschah mit so hingebungsvollem Eifer und einer fast
27	-1	Und sobald der Krieg es zuläßt, mußt du schreiben. Denn was aus deinem Briefe geworden ist, den wir an jenem Tage, als
28	-1	Verknusen konnte der Alte auch überhaupt nicht, daß der pensionierte Pastor der Sankt Marien Gemeinde, Karl Roggenkar

The data for the attribute 'Beleg'.

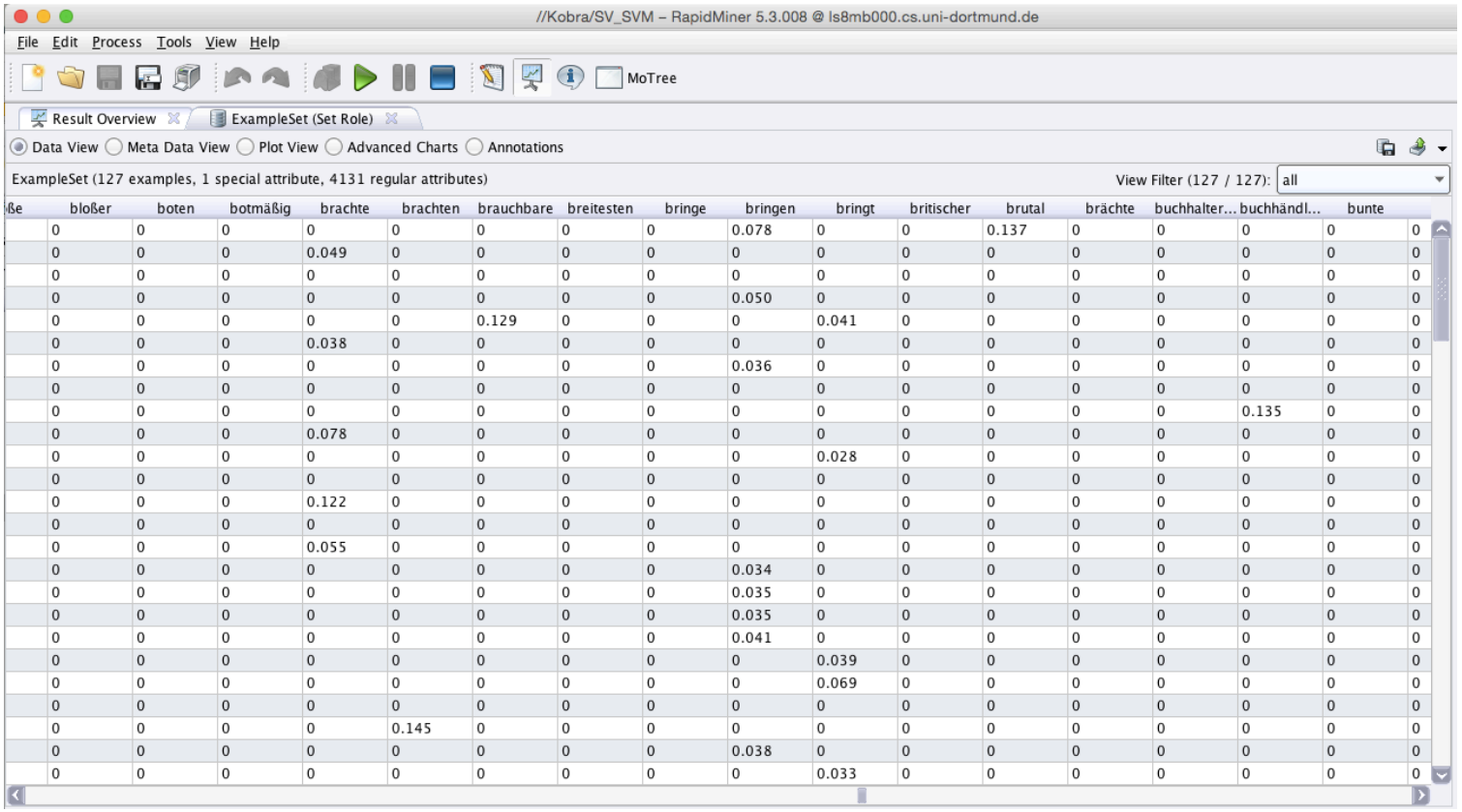


Texte vorverarbeiten

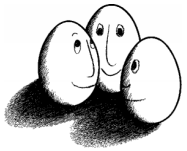




Bag of Words

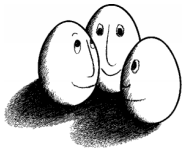


ße	bloßer	boten	botmäßig	brachte	brachten	brauchbare	breitesten	bringe	bringen	bringt	britischer	brutal	brächte	buchhalter...	buchhändl...	bunte	
	0	0	0	0	0	0	0	0	0.078	0	0	0.137	0	0	0	0	0
	0	0	0	0.049	0	0	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0.050	0	0	0	0	0	0	0	0
	0	0	0	0	0	0.129	0	0	0	0.041	0	0	0	0	0	0	0
	0	0	0	0.038	0	0	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0.036	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.135	0	0
	0	0	0	0.078	0	0	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0.028	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0.122	0	0	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0.055	0	0	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0.034	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0.035	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0.035	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0.041	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0.039	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0.069	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0.145	0	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0.038	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0.033	0	0	0	0	0	0	0



KobRA Studien

- Repräsentation von Texten
  - Bag of Words
  - Meta-Daten
- Klassifikation
  - Stützverb vs. Vollverb
- Topic Models
  - Latent Dirichlet Allocation



Klassifikation Stützverb vs. Vollverb

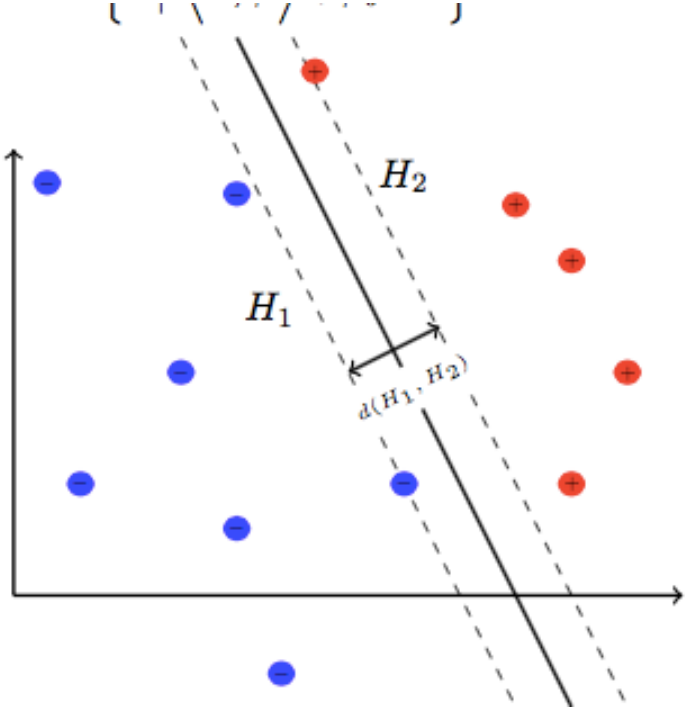
- Stützvektormethode für Stützverbkonstruktionen
- SVM lernt Ebene, die die Klassen trennt, durch Gewichtung der Attribute:

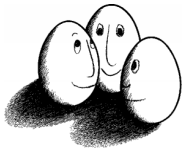
$$f(x) = wx + b$$

- "Margin" wird maximiert:

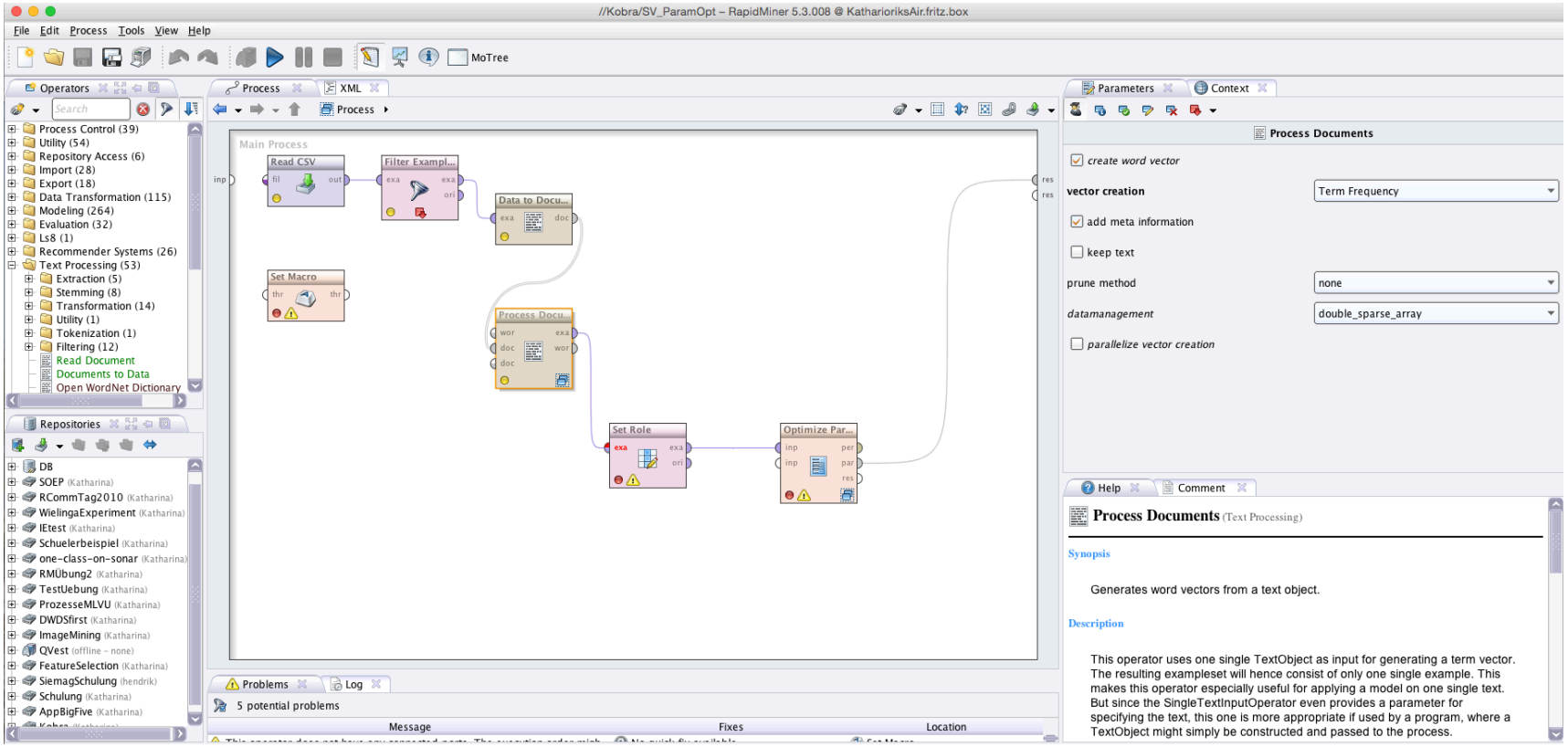
$$w = \sum \alpha_i y_i x_i$$

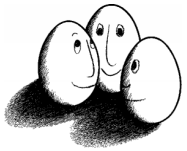
- Lernen gewichtet auch die Beispiele: $\alpha_l > 0$, x_i ist Stützvektor Linearkombination der Stützvektoren.



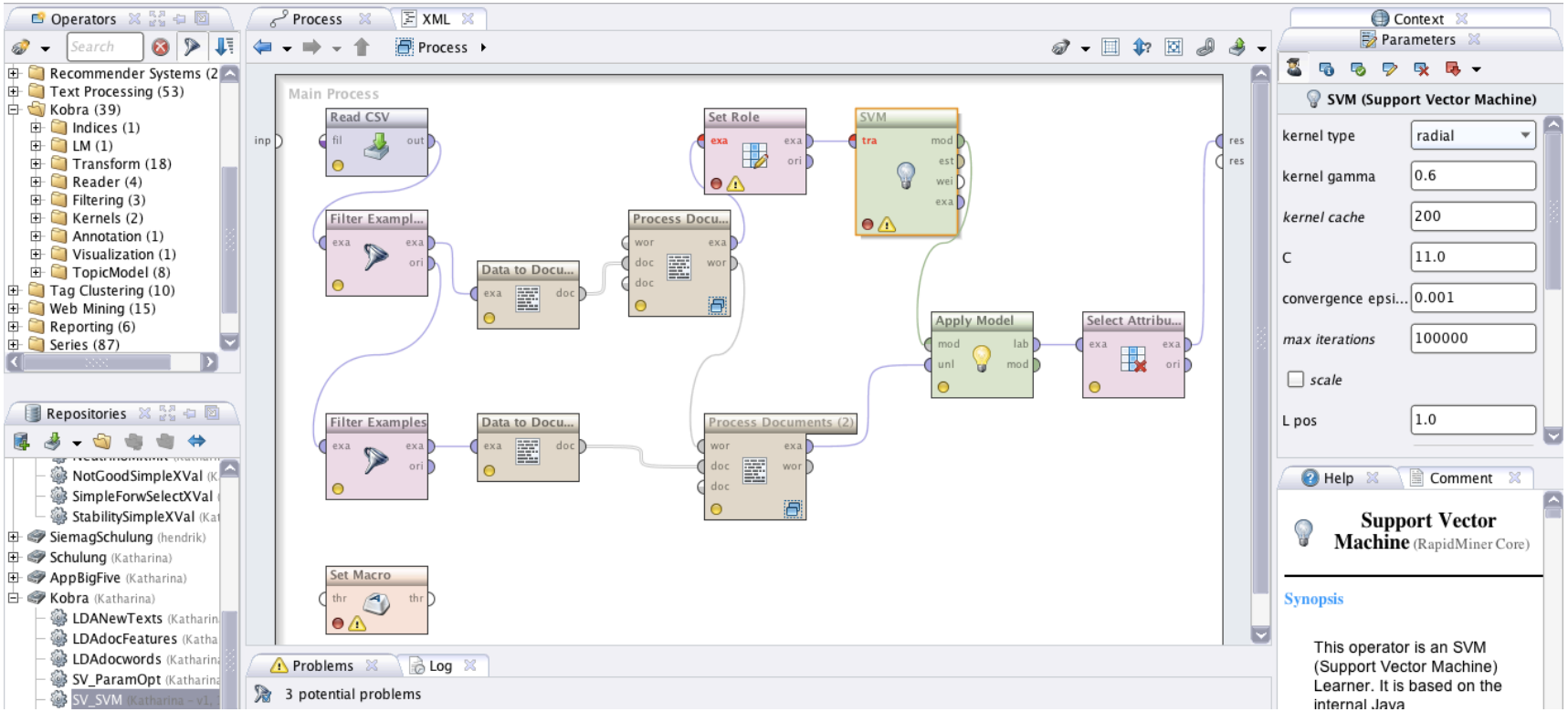


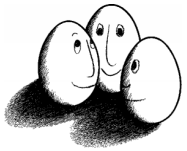
Parameter Optimierung





SVM mit optimierten Parametern lernen lassen





Ergebnis

Result Overview | ExampleSet (Select Attributes)

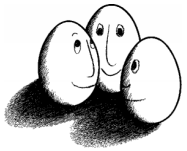
Data View | Meta Data View | Plot View | Advanced Charts | Annotations

ExampleSet (17825 examples, 2 special attributes, 0 regular attributes)

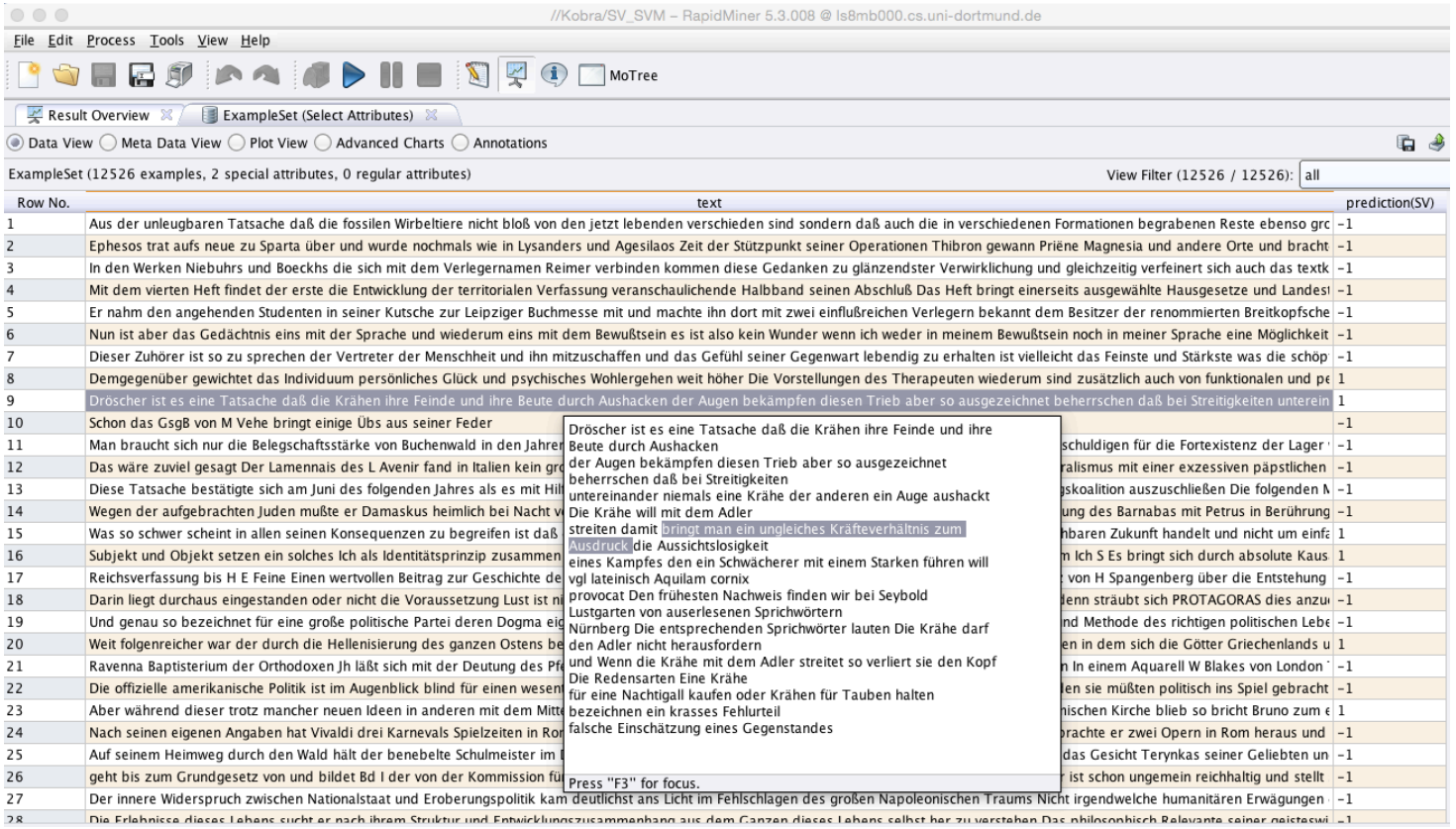
Row No.	text	predicti...
2281	Du hast mich ja nicht gelassen sagt sie und fängt endlich an zu weinen Er sagt daß er sie auch nicht gesehen hat	1
8426	Sein Blick lief über Hugo Bandilla hinweg zur schiefen Schmiede zu den ausgedie	
13934	Es ist wirklich zum Lachen Im Anfang ja damals hat es mich fast zum Wahnsinn g	
1	Und dann flammten plötzlich Bogenlampen auf so blendend hell und violett daß i	
2	Und du hast es gewußt verdammt noch mal Hawa fühlte Hitze vom Magen hochst	
3	Es tickt die Uhr Da kommts aus West und Ost Was bringt mir morgens so die Pos	
4	Nervös griff sie in ihre Handtasche und holte den Veilchentraum hervor Der ande	
5	Ich rief den Kellner und bat ihn um einen Atlas Er brachte mir so ein schmieriges	
6	Was hatte Guébriant seitdem ausgestanden stets hatte er vermitteln einlenken vo	
7	Wenn der merkt daß ich etwas weiß Rose ließ den Blick nicht von dem Bullen Also wohin habt ihr das arme Wurm	-1
8	Der Vater schien durch die Worte der Schwester auf bestimmtere Gedanken gebracht zu sein hatte sich aufrecht	-1
9	Der Sohn einer Mutter ich sage es biblisch wie es sich zugetragen war gestorben und man trug ihn hinaus auf de	-1
10	Ja log Lisa Er hob die Schultern ging in die Küche und brachte einen kleinen Fisch den er auf die Terrasse warf Ir	-1
11	Darum war es von größter Wichtigkeit das Feuer auf seinen Herd zu beschränken denn sobald es weiter um sich	-1
12	Alle unsere Dienstleistungen lassen wir uns vorher bezahlen Da hatte er Leporis sich durchgesetzt bis Henri den	-1
13	Bloß gut daß man so was auch aus eigenem Interesse tut Ich flitzte zwischen den Hauptquartieren hin und her m	-1
14	Er hat in Firgun Pintg ausgerichtet daß du über den Schmalberg nach Pro de Peadra absteigen wirst Du bringst n	-1
15	Dann schüttelte er das ab kniete sich über den Leblosen und machte die vorgeschriebenen Bewegungen mit seir	-1
16	Und Sklarz nicht zu vergessen Der Katz Kahn Vertrag brachte uns Millionen Goldmark Verluste Das Hanauer Pio	-1
17	Fritz schob sich den einzigen Stuhl zu ihr und liess sich nieder Ich bring dir ein Buch sagte er wichtig Hättest du r	-1

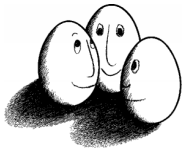
Du hast mich ja nicht gelassen sagt sie und fängt endlich an zu weinen Er sagt daß er sie
auch nicht gesehen hat bis zum Ende des Zuges nicht vielleicht
haben sie die Gefahr rechtzeitig gespürt
und sich in Sicherheit gebracht Er weiß wie lächerlich das ist nach
drei Worten merkt er wie
nutzlos er lügt aber er bringt die Sätze zu Ende wie aufgezogen

Press "F3" for focus.



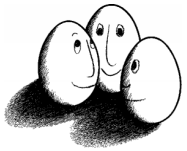
Ergebnis


//Kobra/SV_SVM – RapidMiner 5.3.008 @ ls8mb000.cs.uni-dortmund.de
File Edit Process Tools View Help
MoTree
Result Overview
ExampleSet (Select Attributes)
Data View Meta Data View Plot View Advanced Charts Annotations
ExampleSet (12526 examples, 2 special attributes, 0 regular attributes)
View Filter (12526 / 12526): all
Row No. text prediction(SV)
1 Aus der unleugbaren Tatsache daß die fossilen Wirbeltiere nicht bloß von den jetzt lebenden verschieden sind sondern daß auch die in verschiedenen Formationen begrabenen Reste ebenso grc -1
2 Ephesos trat aufs neue zu Sparta über und wurde nochmals wie in Lysanders und Agesilaos Zeit der Stützpunkt seiner Operationen Thibron gewann Priëne Magnesia und andere Orte und bracht -1
3 In den Werken Niebuhrs und Boeckhs die sich mit dem Verlegernamen Reimer verbinden kommen diese Gedanken zu glänzendster Verwirklichung und gleichzeitig verfeinert sich auch das textk -1
4 Mit dem vierten Heft findet der erste die Entwicklung der territorialen Verfassung veranschaulichende Halbband seinen Abschluß Das Heft bringt einerseits ausgewählte Hausgesetze und Landest -1
5 Er nahm den angehenden Studenten in seiner Kutsche zur Leipziger Buchmesse mit und machte ihn dort mit zwei einflußreichen Verlegern bekannt dem Besitzer der renommierten Breitkopfsche -1
6 Nun ist aber das Gedächtnis eins mit der Sprache und wiederum eins mit dem Bewußtsein es ist also kein Wunder wenn ich weder in meinem Bewußtsein noch in meiner Sprache eine Möglichkeit -1
7 Dieser Zuhörer ist so zu sprechen der Vertreter der Menschheit und ihn mitzuschaffen und das Gefühl seiner Gegenwart lebendig zu erhalten ist vielleicht das Feinste und Stärkste was die schöp -1
8 Demgegenüber gewichtet das Individuum persönliches Glück und psychisches Wohlergehen weit höher Die Vorstellungen des Therapeuten wiederum sind zusätzlich auch von funktionalen und pe 1
9 Dröscher ist es eine Tatsache daß die Krähen ihre Feinde und ihre Beute durch Aushacken der Augen bekämpfen diesen Trieb aber so ausgezeichnet beherrschen daß bei Streitigkeiten unterein 1
10 Schon das GsgB von M Vehe bringt einige Übs aus seiner Feder -1
11 Man braucht sich nur die Belegschaftsstärke von Buchenwald in den Jahre schuldigen für die Fortexistenz der Lager -1
12 Das wäre zuviel gesagt Der Lamennais des L Avenir fand in Italien kein gr ralismus mit einer exzessiven päpstlichen -1
13 Diese Tatsache bestätigte sich am Juni des folgenden Jahres als es mit Hil skoalition auszuschließen Die folgenden M -1
14 Wegen der aufgebrachten Juden mußte er Damaskus heimlich bei Nacht v ung des Barnabas mit Petrus in Berührung -1
15 Was so schwer scheint in allen seinen Konsequenzen zu begreifen ist daß nbaren Zukunft handelt und nicht um einfa 1
16 Subjekt und Objekt setzen ein solches Ich als Identitätsprinzip zusammen n Ich S Es bringt sich durch absolute Kaus 1
17 Reichsverfassung bis H E Feine Einen wertvollen Beitrag zur Geschichte de von H Spangenberg über die Entstehung -1
18 Darin liegt durchaus eingestanden oder nicht die Voraussetzung Lust ist n denn sträubt sich PROTAGORAS dies anzu -1
19 Und genau so bezeichnet für eine große politische Partei deren Dogma eig nd Methode des richtigen politischen Lebe -1
20 Weit folgenreicher war der durch die Hellenisierung des ganzen Ostens be en in dem sich die Götter Griechenlands u 1
21 Ravenna Baptisterium der Orthodoxen Jh läßt sich mit der Deutung des Pf n In einem Aquarell W Blakes von London -1
22 Die offizielle amerikanische Politik ist im Augenblick blind für einen wesen en sie müßten politisch ins Spiel gebracht -1
23 Aber während dieser trotz mancher neuen Ideen in anderen mit dem Mitt nischen Kirche blieb so bricht Bruno zum e 1
24 Nach seinen eigenen Angaben hat Vivaldi drei Karnevals Spielzeiten in Ro brachte er zwei Opern in Rom heraus und -1
25 Auf seinem Heimweg durch den Wald hält der benebelte Schulmeister im das Gesicht Terynkas seiner Geliebten un -1
26 geht bis zum Grundgesetz von und bildet Bd I der von der Kommission fü ist schon ungemein reichhaltig und stellt -1
27 Der innere Widerspruch zwischen Nationalstaat und Eroberungspolitik kam deutlichst ans Licht im Fehlschlagen des großen Napoleonischen Traums Nicht irgendwelche humanitären Erwägungen -1
28 Die Erlebnisse dieses Lebens sucht er nach ihrem Struktur und Entwicklungszusammenhang aus dem Ganzen dieses Lebens selbst her zu verstehen Das philosophisch Relevante seiner geisteswi -1
Dröscher ist es eine Tatsache daß die Krähen ihre Feinde und ihre
Beute durch Aushacken
der Augen bekämpfen diesen Trieb aber so ausgezeichnet
beherrschen daß bei Streitigkeiten
untereinander niemals eine Krähe der anderen ein Auge aushackt
Die Krähe will mit dem Adler
streiten damit bringt man ein ungleiches Kräfteverhältnis zum
Ausdruck die Aussichtslosigkeit
eines Kampfes den ein Schwächerer mit einem Starken führen will
vgl lateinisch Aquilam cornix
provocat Den frühesten Nachweis finden wir bei Seybold
Lustgarten von auserlesenen Sprichwörtern
Nürnberg Die entsprechenden Sprichwörter lauten Die Krähe darf
den Adler nicht herausfordern
und Wenn die Krähe mit dem Adler streitet so verliert sie den Kopf
Die Redensarten Eine Krähe
für eine Nachtigall kaufen oder Krähen für Tauben halten
bezeichnen ein krasses Fehlurteil
falsche Einschätzung eines Gegenstandes
Press "F3" for focus.

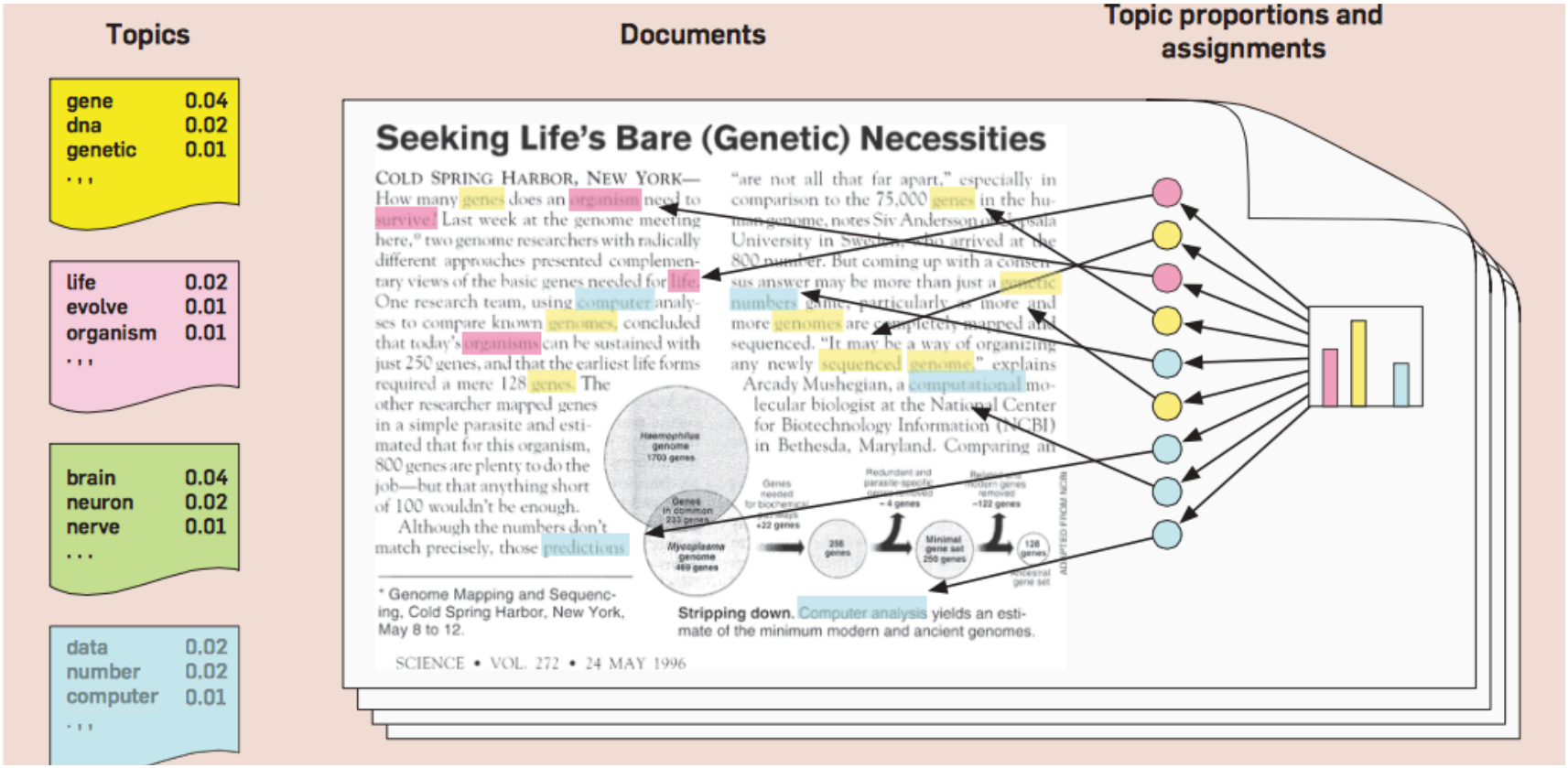


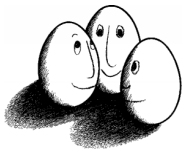
KobRA Studien

- Repräsentation von Texten
  - Bag of Words
  - Meta-Daten
- Klassifikation
  - Stützverb vs. Vollverb
- Topic Models
  - Latent Dirichlet Allocation



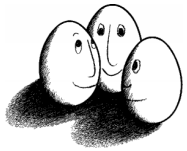
Topic Modelle





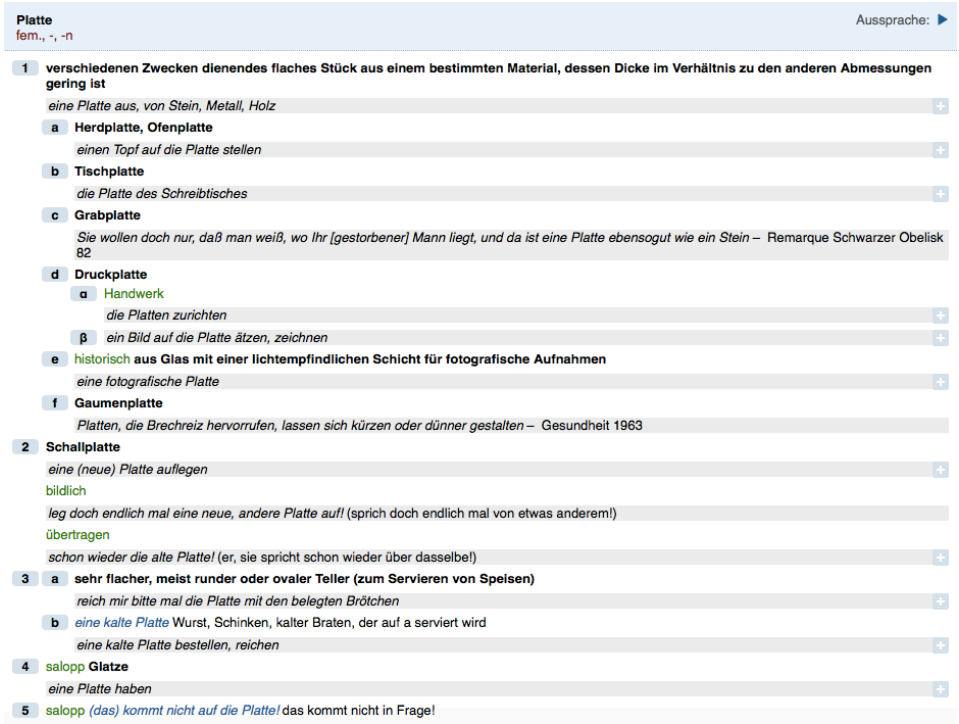
Latente Dirichlet Allocation

- Topics sind Verteilungen β über einem Vokabular V
  - Das Topic *Genetik* hat Wörter wie „Gen“, „Exon“, „Protein“ häufig
- Dokumente $\mathbf{w}=(w_1, .., w_N)$ zeigen die Worthäufigkeiten aus den Topic-Verteilungen
  1. Das Dokument hat viel *Genetik* und etwas *Datenanalyse.*
  2. Jedes Wort *w* wird gezogen:
     - gemäß der Topicverteilung θ
     - gemäß der Wortverteilung β im Topic.
- Ein Korpus ist eine Menge von Dokumenten $D=\{\mathbf{w}_1, ..., \mathbf{w}_M\}$
- Eine Topic-Zuweisung z_{dn} weist w_{dn}(dem n-ten Wort im Dokument d) einen Topic zu.



Topic Modeling zur Exploration

- Bedeutungen des Wortes “Platte” im Korpus.
- DWDS zeigt es pefekt.
- Was finden wir mit unüberwachtem Lernen?
  - Latent Dirichlet Allococation
  - Texte von 1588 bis1925



Platte
fem., -, -n

Aussprache: ▶

1 **verschiedenen Zwecken dienendes flaches Stück aus einem bestimmten Material, dessen Dicke im Verhältnis zu den anderen Abmessungen gering ist**

eine Platte aus, von Stein, Metall, Holz

a **Herdplatte, Ofenplatte**

einen Topf auf die Platte stellen

b **Tischplatte**

die Platte des Schreibtisches

c **Grabplatte**

Sie wollen doch nur, daß man weiß, wo Ihr [gestorbener] Mann liegt, und da ist eine Platte ebensogut wie ein Stein – Remarque Schwarzer Obelisk 82

d **Druckplatte**

α Handwerk

die Platten zurichten

β *ein Bild auf die Platte ätzen, zeichnen*

e historisch **aus Glas mit einer lichtempfindlichen Schicht für fotografische Aufnahmen**

eine fotografische Platte

f **Gaumenplatte**

Platten, die Brechreiz hervorrufen, lassen sich kürzen oder dünner gestalten – Gesundheit 1963

2 **Schallplatte**

eine (neue) Platte auflegen

bildlich

leg doch endlich mal eine neue, andere Platte auf! (sprich doch endlich mal von etwas anderem!)

übertragen

schon wieder die alte Platte! (er, sie spricht schon wieder über dasselbe!)

3 **a** **sehr flacher, meist runder oder ovaler Teller (zum Servieren von Speisen)**

reich mir bitte mal die Platte mit den belegten Brötchen

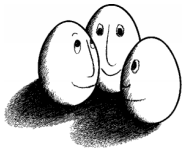
b *eine kalte Platte* Wurst, Schinken, kalter Braten, der auf a serviert wird

eine kalte Platte bestellen, reichen

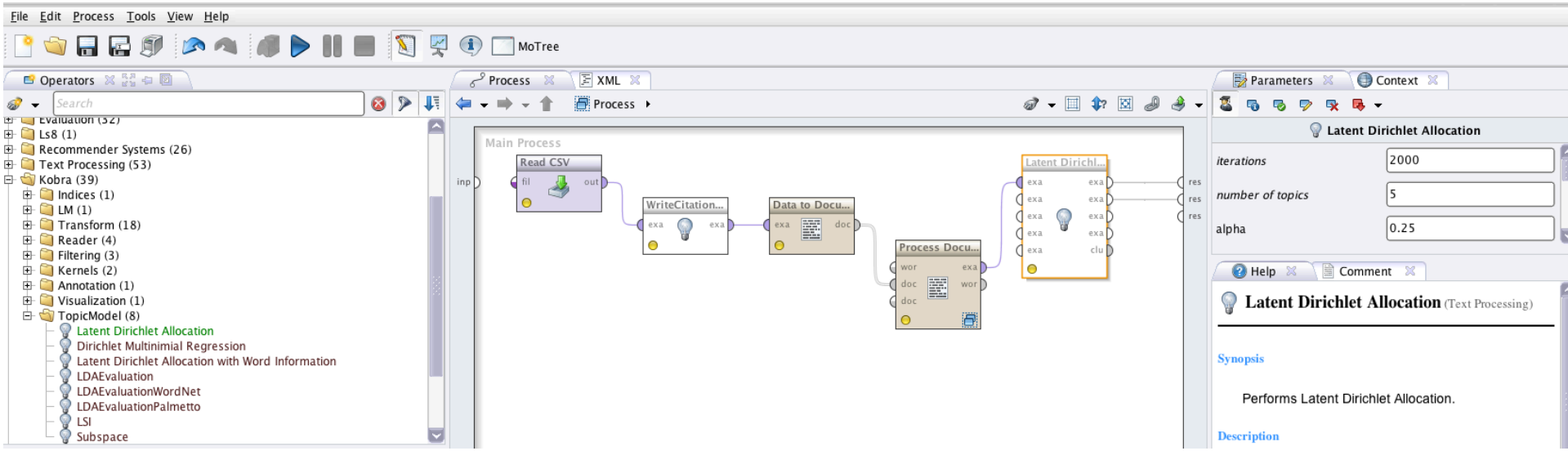
4 salopp **Glatze**

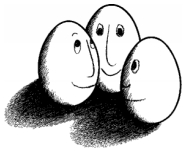
eine Platte haben

5 salopp *(das) kommt nicht auf die Platte!* das kommt nicht in Frage!



RapidMiner Prozess





Wahrscheinlichkeit der Topics für Dokumente und Wörter für Topics,

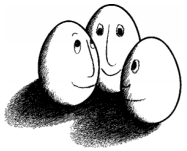
ExampleSet (4349 examples, 0 special attributes, 7 regular attributes)

Row No.	Doc	Topic	Topic_0	Topic_1	Topic_2	Topic_3	Topic_4
1	1	4	0.058	0.062	0.113	0.043	0.724
2	2	4	0.051	0.046	0.148	0.313	0.442
3	3	4	0.067	0.096	0.283	0.134	0.419
4	4	0	0.425	0.209	0.080	0.132	0.154
5	5	2	0.085	0.071	0.374	0.205	0.266
6	6	4	0.111	0.157	0.132	0.114	0.486
7	7	2	0.135	0.095	0.308	0.222	0.240
8	8	1	0.052	0.628	0.123	0.057	0.140
9	9	1	0.087	0.525	0.181	0.061	0.146
10	10	4	0.196	0.276	0.111	0.120	0.298
11	11	1	0.175	0.422	0.077	0.129	0.197
12	12	2	0.102	0.081	0.385	0.382	0.050
13	13	3	0.135	0.126	0.317	0.379	0.043
14	14	3	0.109	0.147	0.181	0.505	0.059
15	15	1	0.114	0.409	0.188	0.092	0.197
16	16	4	0.238	0.238	0.111	0.087	0.327
17	17	1	0.108	0.597	0.102	0.077	0.117
18	18	1	0.063	0.657	0.061	0.103	0.116
19	19	1	0.038	0.623	0.047	0.070	0.222
20	20	4	0.059	0.372	0.057	0.103	0.408
21	21	1	0.055	0.481	0.050	0.052	0.361
22	22	1	0.179	0.459	0.088	0.093	0.181
23	23	1	0.207	0.341	0.113	0.134	0.205
24	24	3	0.139	0.078	0.151	0.560	0.073
25	25	4	0.237	0.203	0.092	0.142	0.326
26	26	2	0.209	0.092	0.345	0.185	0.169
27	27	2	0.204	0.147	0.276	0.253	0.120
28	28	1	0.187	0.320	0.111	0.116	0.267
29	29	1	0.098	0.634	0.092	0.077	0.098
30	30	1	0.070	0.754	0.059	0.055	0.061
31	31	4	0.186	0.240	0.094	0.087	0.393

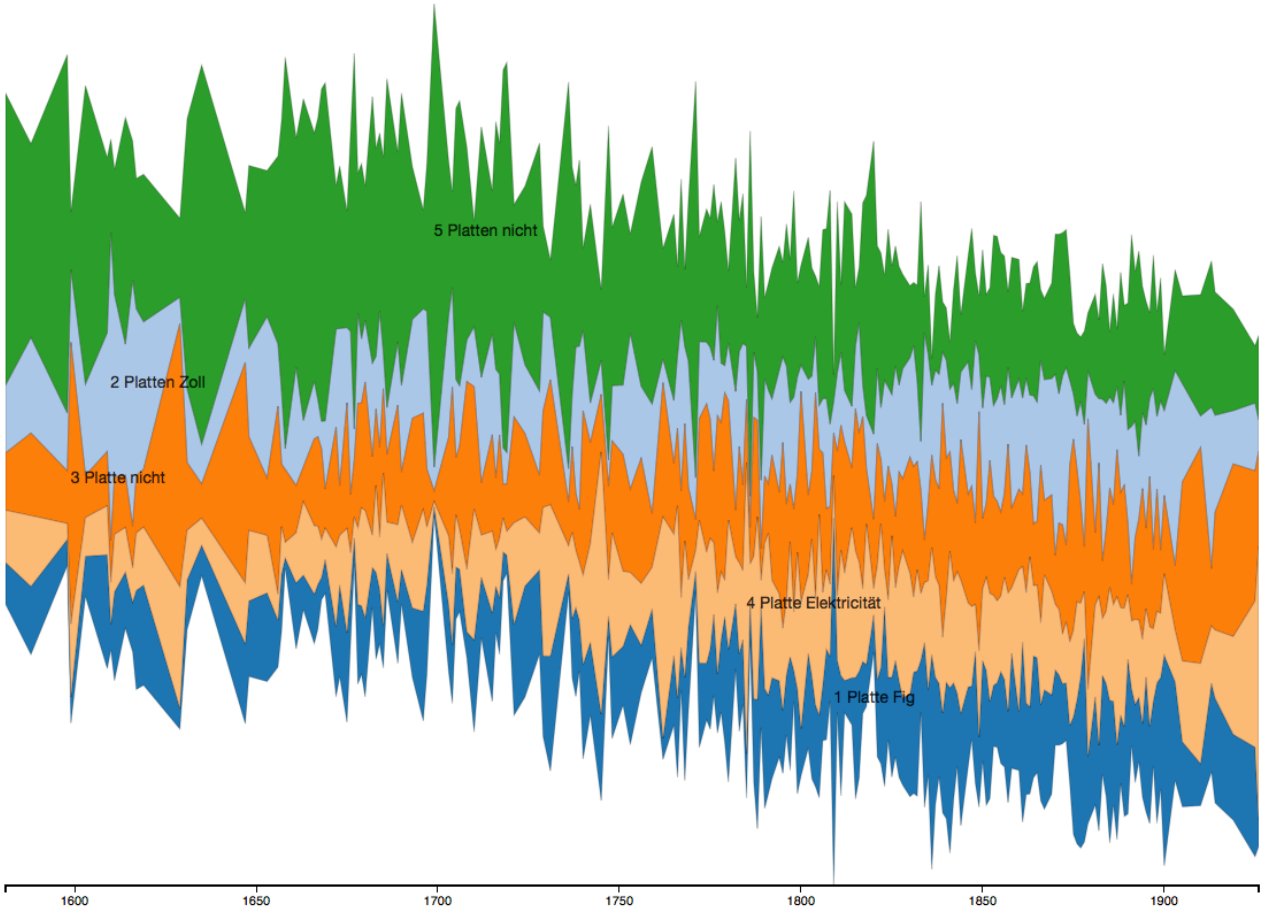
Data View | Meta Data View | Plot View | Advanced Charts | Annotations

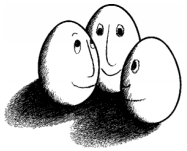
ExampleSet (52 examples, 0 special attributes, 8 regular attributes)

Row No.	Word	Word_id	Topic	Topic_0	Topic_1	Topic_2	Topic_3	Topic_4
1	Art	1	0	0.030	0.001	0.009	0.022	0.020
2	Dicke	2	1	0.000	0.049	0.008	0.000	0.000
3	Eisen	3	4	0.000	0.020	0.000	0.000	0.030
4	Elektricität	4	3	0.000	0.000	0.000	0.069	0.000
5	Fig	5	0	0.183	0.018	0.000	0.000	0.000
6	Fläche	6	3	0.002	0.002	0.001	0.056	0.000
7	Form	7	2	0.017	0.022	0.032	0.000	0.001
8	Höhe	8	1	0.000	0.023	0.011	0.003	0.009
9	Kupfer	9	2	0.000	0.003	0.030	0.002	0.019
10	Körper	10	3	0.001	0.002	0.001	0.054	0.002
11	Luft	11	3	0.000	0.001	0.001	0.044	0.001
12	Länge	12	1	0.001	0.044	0.000	0.001	0.000
13	Mitte	13	0	0.026	0.023	0.003	0.007	0.000
14	Oberfläche	14	2	0.002	0.014	0.025	0.002	0.000
15	Platte	15	3	0.243	0.055	0.346	0.364	0.006
16	Platten	16	1	0.109	0.301	0.000	0.012	0.251
17	Theil	17	3	0.002	0.010	0.004	0.046	0.015
18	Theile	18	0	0.018	0.002	0.001	0.016	0.006
19	Verbindung	19	0	0.028	0.006	0.000	0.011	0.000
20	Wasser	20	2	0.000	0.000	0.055	0.000	0.028
21	Weise	21	0	0.018	0.016	0.009	0.008	0.001
22	Zeit	22	2	0.001	0.002	0.039	0.001	0.028
23	Zoll	23	1	0.000	0.122	0.000	0.000	0.000
24	befestigt	24	0	0.048	0.004	0.000	0.000	0.000
25	befindet	25	0	0.023	0.000	0.000	0.013	0.000
26	bey	26	4	0.000	0.000	0.000	0.012	0.070
27	bildet	27	0	0.021	0.006	0.003	0.006	0.000



Visualisierung mit dem dfr-browser



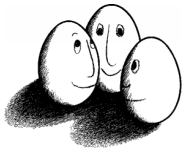


7 Topics auf neuen Texten

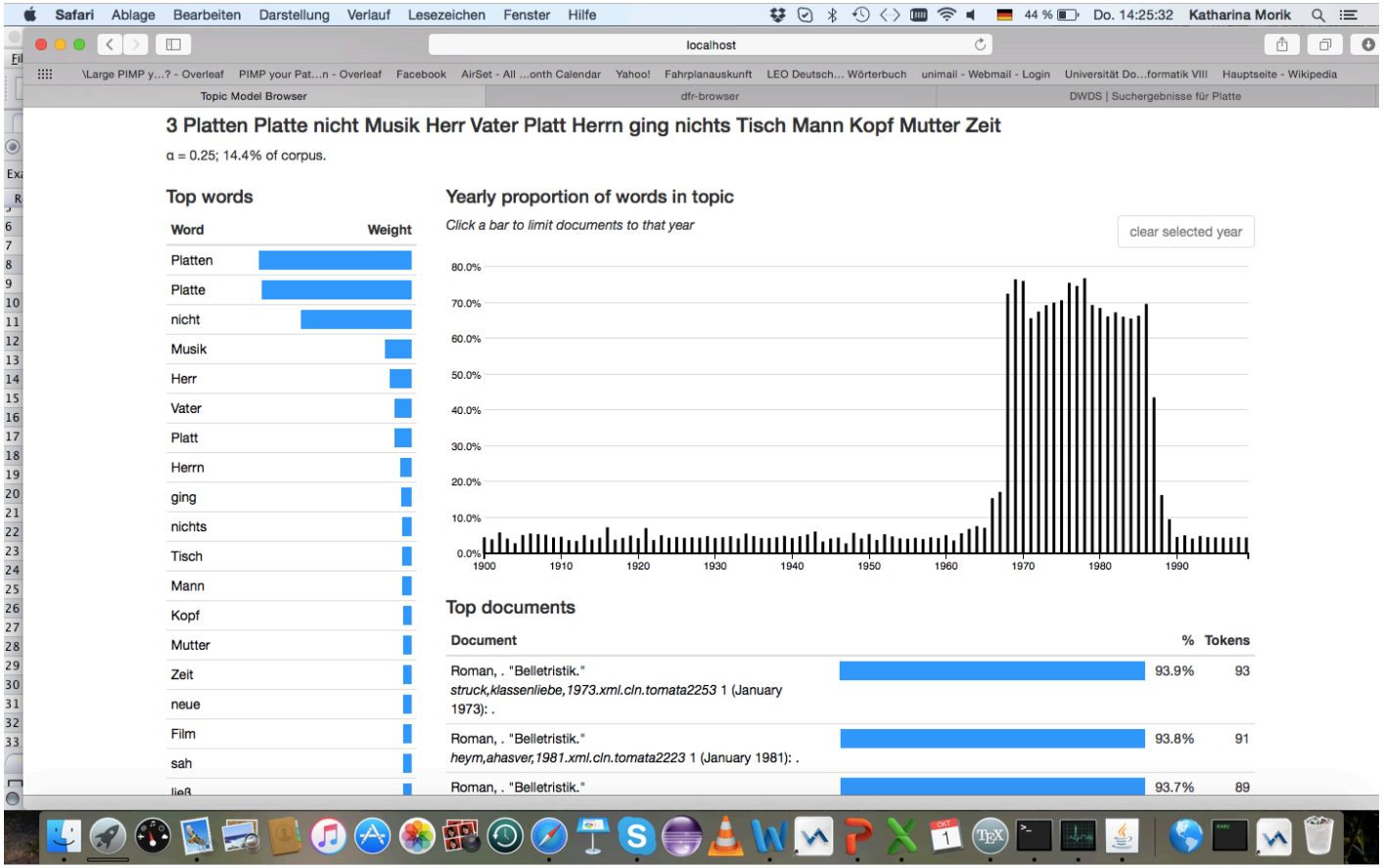
n *Kobra* | Overview | Topic | Document | Word | Bibliography | Word index | Settings | About

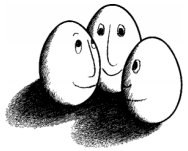
Scaled | List | Years — *click a column label to sort; click a row for more about a topic*

1900—1999	top words	proportion of corpus
	Platte Platten nicht Jh SCSI MByte Treiber Programm BIOS läßt Hersteller DOS Zeit Sektoren Festplatte	16.8
	Platte nicht Platten Judith Platter Abb Strom zeigt Licht oben großen Achse läßt senkrecht Strahlen	17.8
	Platten Platte nicht Musik Herr Vater Platt Herrn ging nichts Tisch Mann Kopf Mutter Zeit	14.4
	Platte nicht Platten BIOS Festplatte Daten SCSI Sektoren MByte Festplatten DOS verwendet Kopf neuen fast	7.8
	Platte Platten nicht Tisch große großen kleinen Formen Hand Abb Wasser Platt ging kleine Bild	10.8
	Platte Platten nicht Tisch sah Zeit großen kleinen Licht Augen große saß Hand stand kleine	15.0
	Platte Platten nicht Teller Tisch Kellner Hand Gast Bild Gäste kleine Art kleinen Fleisch Butter	17.3



Topic 3

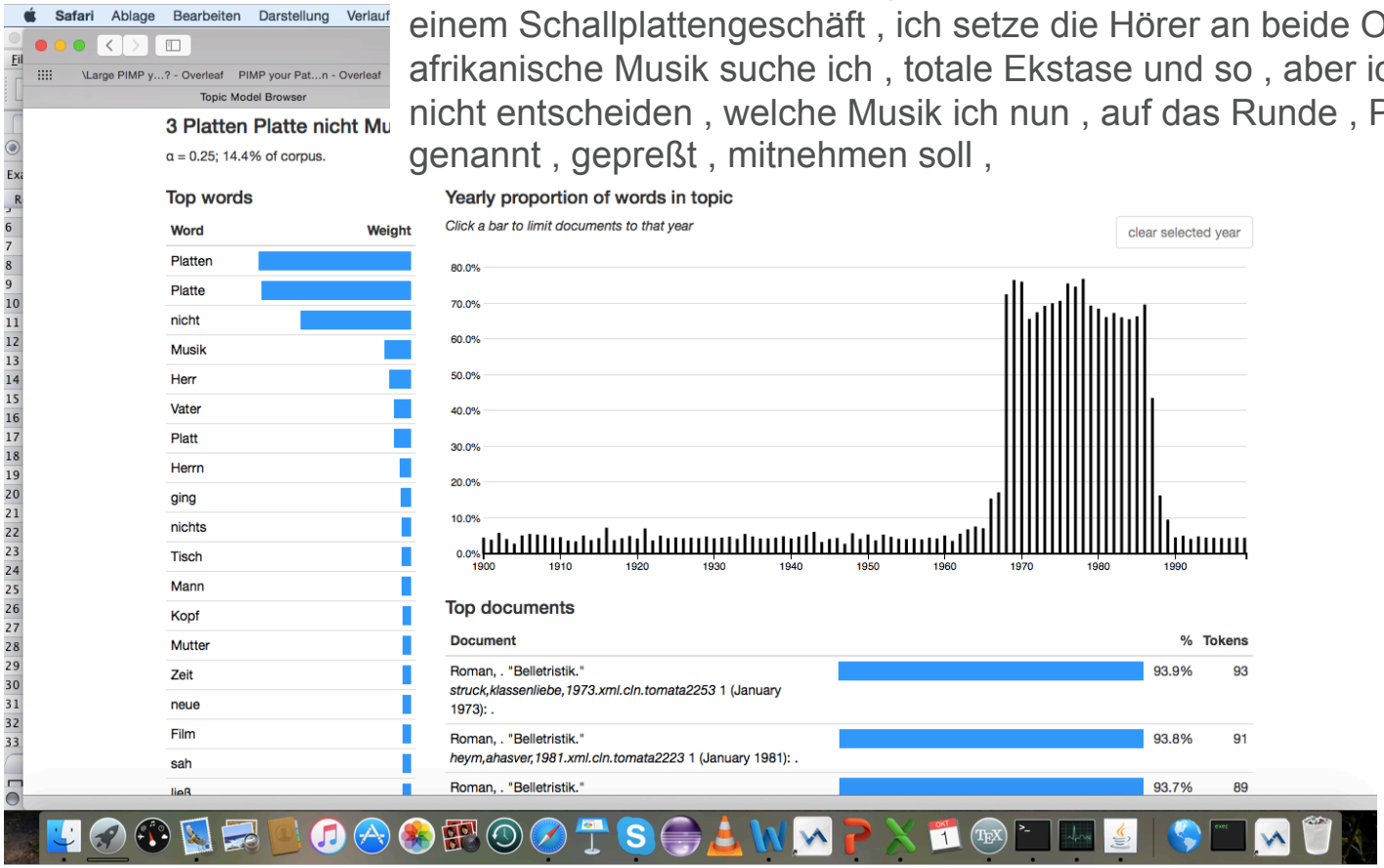

Safari Ablage Bearbeiten Darstellung Verlauf Lesezeichen Fenster Hilfe
Do. 14:25:32 Katharina Morik
localhost
Topic Model Browser
dfr-browser
DWDS | Suchergebnisse für Platte
3 Platten Platte nicht Musik Herr Vater Platt Herrn ging nichts Tisch Mann Kopf Mutter Zeit
α = 0.25; 14.4% of corpus.
Top words
Word
Weight
Platten
Platte
nicht
Musik
Herr
Vater
Platt
Herrn
ging
nichts
Tisch
Mann
Kopf
Mutter
Zeit
neue
Film
sah
Yearly proportion of words in topic
Click a bar to limit documents to that year
clear selected year
80.0%
70.0%
60.0%
50.0%
40.0%
30.0%
20.0%
10.0%
0.0%
1900
1910
1920
1930
1940
1950
1960
1970
1980
1990
Top documents
Document
%
Tokens
Roman, . "Belletristik." struck,klassenliebe,1973.xml.cln.tomata2253 1 (January 1973): .
93.9%
93
Roman, . "Belletristik." heym,ahasver,1981.xml.cln.tomata2223 1 (January 1981): .
93.8%
91
Roman, . "Belletristik."
93.7%
89

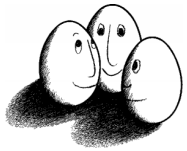


Topic 3 “record”

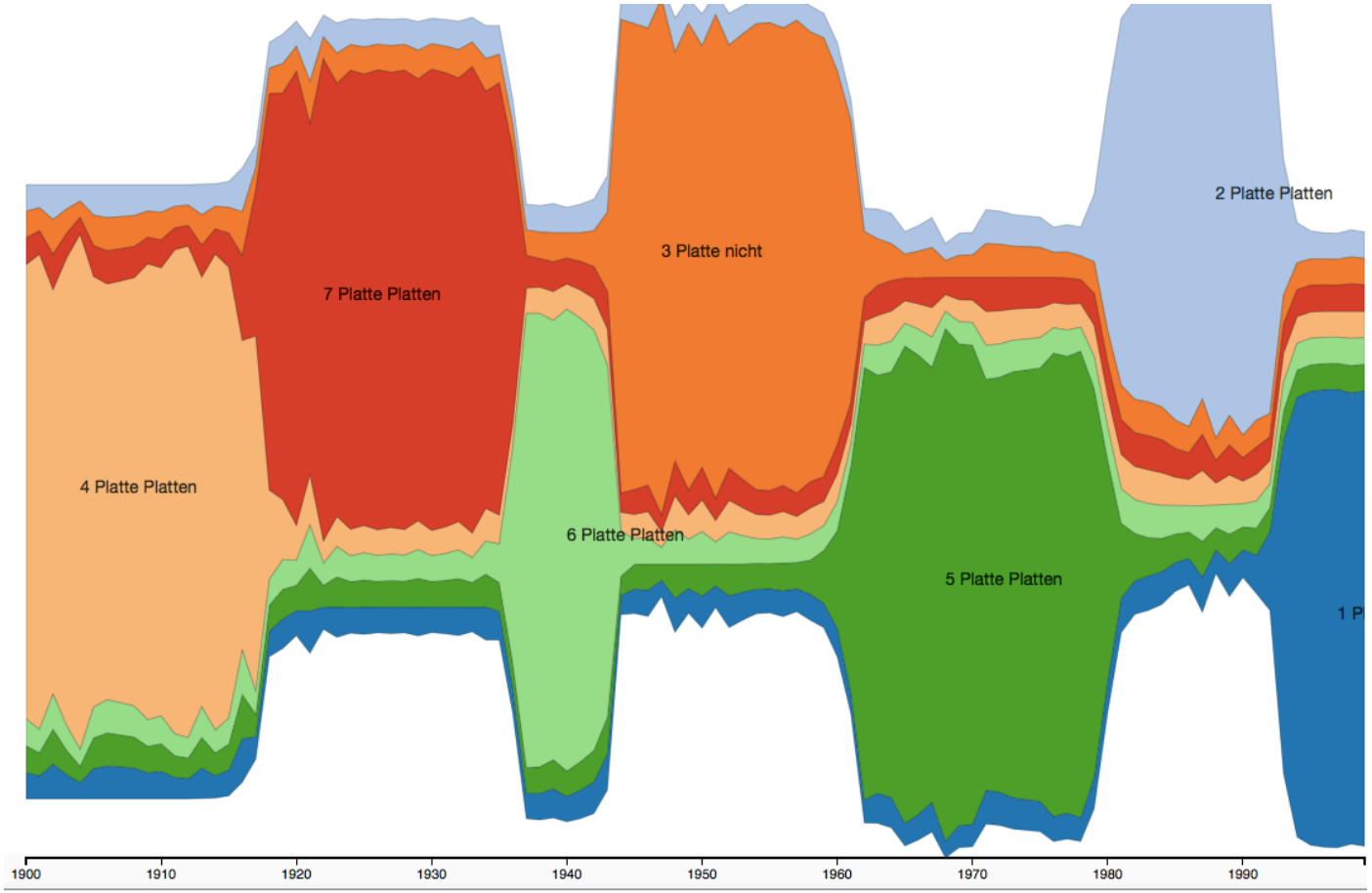
struck_klassenliebe_1973.xml.cln.tomata2:16

´ Etwa um fünf Uhr nachmittags setze ich mich auf einen Hocker in einem Schallplattengeschäft , ich setze die Hörer an beide Ohren , afrikanische Musik suche ich , totale Ekstase und so , aber ich kann nicht entscheiden , welche Musik ich nun , auf das Runde , Platte genannt , gepreßt , mitnehmen soll ,

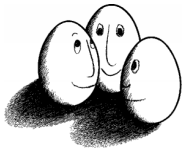




7 Topics im Zeitverlauf



hardware

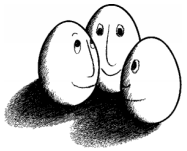


KobRA Studien

- Repräsentation von Texten
  - Bag of Words
  - Meta-Daten
- Klassifikation
  - Stützverb vs. Vollverb
- Topic Models
  - Latent Dirichlet Allocation
- Integration von RapidMiner und Weblicht, DWDS

technische universität
dortmund

Fakultät für Informatik
LS 8



Informatik:

Methoden +
Verfahren

Linguistik:

Forschungsfragen,
Anforderungen + Evaluation

GEFÖRDERT VOM

Bundesministerium
für Bildung
und Forschung

BBAW berlin-brandenburgische AKADEMIE DER WISSENSCHAFTEN
DWDS

CLARIN-D | WEBLICHT WEB-BASED LINGUISTIC CHAINING TOOL

SfS Uni Tübingen

INSTITUT FÜR DEUTSCHE SPRACHE
Mitglied der Leibniz-Gemeinschaft

IdS

Sprachtechnologie-Partner aus CLARIN:
Daten, Werkzeuge, Infrastrukturen