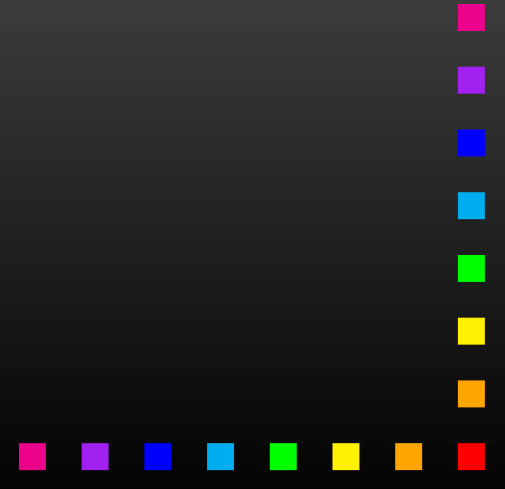
Das Erkennen bibliographischer Dubletten

Dr. Harald Jele

harald.jele@uni-klu.ac.at

Universität Klagenfurt



Kontext des Projekts

- Phase 2: Integration der Datensätze des ISIS-Bibliothekssystems der IFF-Fakultät der Uni Klagenfurt (Schottenfeldgasse, 1090 Wien) in das Verbundsystem. Ca. 10.000 Titeldatensätze

Kontext des Projekts

- Phase 2: Integration der Datensätze des ISIS-Bibliothekssystems der IFF-Fakultät der Uni Klagenfurt (Schottenfeldgasse, 1090 Wien) in das Verbundsystem. Ca. 10.000 Titeldatensätze
- Phase 1: Titeldublette = gleiche ISBN + gleiches Jahr + gleiche Auflage (abgeschlossen 2005, ca. 6.300 Titeldatensätze + Items + Holdings)

Prinzipien, die dabei verfolgt werden (1):

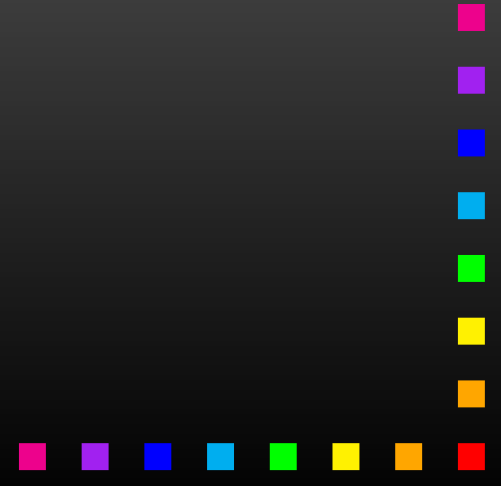
- Datensätze sind sich ähnlich oder gleich, wenn auch deren „Bestandteile“ sich ähnlich oder gleich sind (Kategorien, Begriffe, Silben, Zeichen)

Prinzipien, die dabei verfolgt werden (1):

- Datensätze sind sich ähnlich oder gleich, wenn auch deren „Bestandteile“ sich ähnlich oder gleich sind (Kategorien, Begriffe, Silben, Zeichen)
- unterschiedliche Regeln zur Katalogisierung bzw. die Auslegung dieser kann durch eine Reduzierung von Inhalten ausgeglichen werden (Normalisierung)

Prinzipien, die dabei verfolgt werden (2):

- nicht alle Datensätze einer beliebigen Menge sind gleich behandelbar (Vorsortierung ist sinnvoll bzw. notwendig)

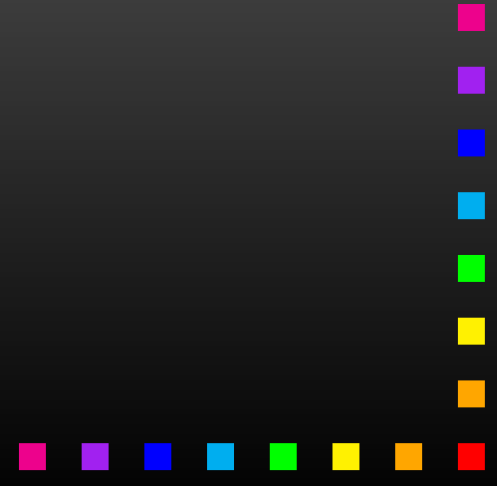


Prinzipien, die dabei verfolgt werden (2):

- nicht alle Datensätze einer beliebigen Menge sind gleich behandelbar (Vorsortierung ist sinnvoll bzw. notwendig)
- Datenelemente, die nicht vorhanden sind, können nicht berechnet – aber als gleich angenommen – werden.
  (Claudia: Die Leere Menge ist immer wahr)

Ansätze, die konkret realisiert wurden:

- Berechnung des Jaccard-Koeffizienten (häufig auch bei Plagiate-Software anzutreffen)



Ansätze, die konkret realisiert wurden:

- Berechnung des Jaccard-Koeffizienten (häufig auch bei Plagiate-Software anzutreffen)
- euklidische Distanz von Zeichenketten (math. Standard-Methode seit den 70er-Jahren, gut dokumentiert)

Ansätze, die konkret realisiert wurden:

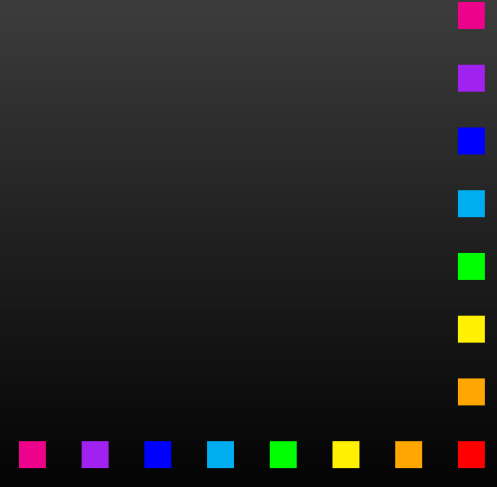
- Berechnung des Jaccard-Koeffizienten (häufig auch bei Plagiate-Software anzutreffen)
- euklidische Distanz von Zeichenketten (math. Standard-Methode seit den 70er-Jahren, gut dokumentiert)
- Methode des KOBV: Euklidische Distanz gewichtet (Ergebnisse überprüfbar)

Vorgehensweise, bei allen drei Ansätzen gleich (1):

- Entladen der Datensätze aus Aleph und ISIS sowie Selektion der zu vergleichenden Kategorien: 100, 200, 331, 403 (Ausgabebez.), 410 (Erscheinungsorte), 412 (Verleger), 425 (Erscheinungsjahr), 433 (Umfangsangabe), 540 (ISBN)

Vorgehensweise, bei allen drei Ansätzen gleich (2):

- Laden der Daten in eine externe Datenbank



Vorgehensweise, bei allen drei Ansätzen gleich (2):

- Laden der Daten in eine externe Datenbank
- Normalisierung der Daten nach den Vorschlägen von Beate Rusch (KOBV)

Vorgehensweise, bei allen drei Ansätzen gleich (2):

- Laden der Daten in eine externe Datenbank
- Normalisierung der Daten nach den Vorschlägen von Beate Rusch (KOBV)
- Berechnung

Vorgehensweise, bei allen drei Ansätzen gleich (2):

- Laden der Daten in eine externe Datenbank
- Normalisierung der Daten nach den Vorschlägen von Beate Rusch (KOBV)
- Berechnung
- Ergebnisse in eine Ergebnistabelle zur weiteren Interpretation

Die wesentlichsten Schritte bei der Normalisierung (1):

- Umsetzen in `US-ASCII-7-bit` Großbuchstaben

Die wesentlichsten Schritte bei der Normalisierung (1):

- Umsetzen in `US-ASCII-7-bit` Großbuchstaben
- Löschen von Diakritika, Akzenten, Sonderzeichen, Steuerzeichen (wie z.B. die entsprechenden Stoppwortzeichen) sowie sämtlicher Zeichensatzzeichen, die außerhalb des definierten Zeichensatzes von `ASCII-7-bit` liegen

Die wesentlichsten Schritte bei der Normalisierung (1):

- Umsetzen in `US-ASCII-7-bit` Großbuchstaben
- Löschen von Diakritika, Akzenten, Sonderzeichen, Steuerzeichen (wie z.B. die entsprechenden Stoppwortzeichen) sowie sämtlicher Zeichensatzzeichen, die außerhalb des definierten Zeichensatzes von `ASCII-7-bit` liegen
- Umsetzen der deutschen Umlaute sowie des ß nach AE, OE, UE, SS

Die wesentlichsten Schritte bei der Normalisierung (2):

- Entfernen doppelter Leerzeichen (Blanks)

Die wesentlichsten Schritte bei der Normalisierung (2):

- Entfernen doppelter Leerzeichen (Blanks)
- Löschen der führenden sowie der abschließenden Leerzeichen

Die wesentlichsten Schritte bei der Normalisierung (2):

- Entfernen doppelter Leerzeichen (Blanks)
- Löschen der führenden sowie der abschließenden Leerzeichen
- Trunkierung der Feldeinträge an der feldspezifisch sinnvollen Stelle (z.B. bei Personennamen hinter dem ersten Buchstaben des zweiten Vornamens und bei bei Erscheinungsorten nach dem fünften Zeichen des ersten Wortes)

Die wesentlichsten Schritte bei der Normalisierung (3):

- Entfernen recherchierter Informationen, die Daten zur katalogisierenden Vorlage ergänzt und dazu entsprechend den angewandten Katalogisierungsregeln in eckigen Klammern hinzugefügt wurden

Beispiele zur Normalisierung (1):

- Kat 100:
  `Berendt, Hans A.` → `BERENDT HANS A`
  `Huizinger, Franz Ernst »von«` → `HUIZINGER FRANZ E`

Beispiele zur Normalisierung (1):

- Kat 100:
  `Berendt, Hans A.` → `BERENDT HANS A`
  `Huizinger, Franz Ernst »von«` → `HUIZINGER FRANZ E`
- Kat 200:
  `Zentrum für Umfragen, Methoden und Analysen <Mannheim>` → `ZENTRUM FUER UMFRAGEN METHODEN UND ANALYSEN MANNHEIM`

Beispiele zur Normalisierung (2):

- Kat 403:
  `Ausgabe 2001` $\rightarrow$ `2001`
  `2., vollst. durchgesehende und überarb. Aufl.` $\rightarrow$ `2`

Beispiele zur Normalisierung (2):

- Kat 403:
  `Ausgabe 2001` → `2001`
  `2., vollst. durchgesehende und überarb. Aufl.` → `2`
- Kat 410:
  `Frankfurt am Main [u.a.]` → `FRANK`
  `München` → `MUENC`

Beispiele zur Normalisierung (3):

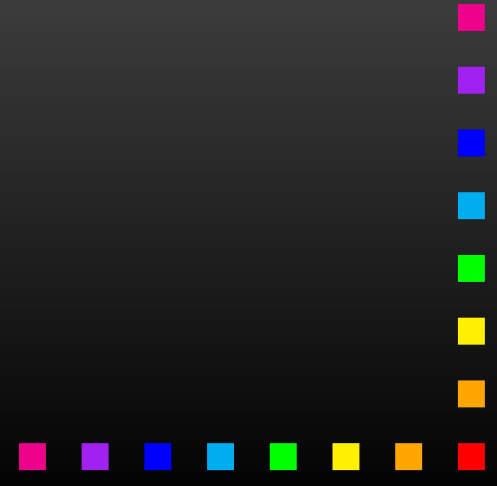
- Kat 425:
  `2000 [erschienen] 1999` → `2000`
  `[19]56` → `56`

Beispiele zur Normalisierung (3):

- Kat 425:
  `2000 [erschienen] 1999` → `2000`
  `[19]56` → `56`
- Kat 433:
  `211 S., [21] Bl.` → `211`
  `XVI, 123 S. : Ill., graph. Darst., Kt.` → `123`

Ngramme als kleinste Einheit zur Berechnung der Ähnlich-keit:

- Trigramme werden in den meisten Verfahren bevorzugt gegenüber 2- oder 4-Grammen

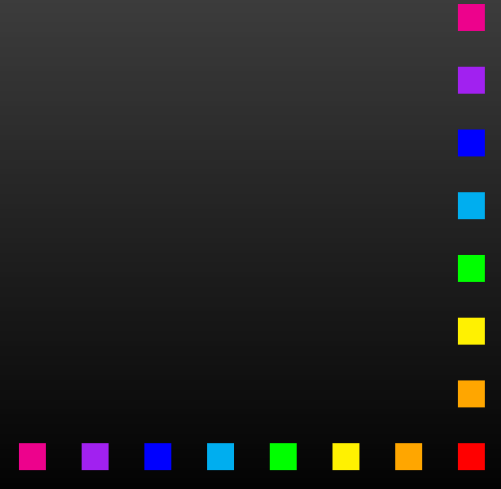


Ngramme als kleinste Einheit zur Berechnung der Ähnlich-keit:

- Trigramme werden in den meisten Verfahren bevorzugt gegenüber 2- oder 4-Grammen
- Ähnlichkeit von Trigrammen $\longrightarrow$ Ähnlichkeit der repräsentierten Zeichenketten

Bildung der Trigramme (1):

- Begriffe werden sequentiell zu Einheiten von jeweils drei Schriftzeichen zerlegt, wobei das jeweils letzte Zeichen eines Trigramms am Beginn des nächsten wiederholt wird.



Bildung der Trigramme (2):

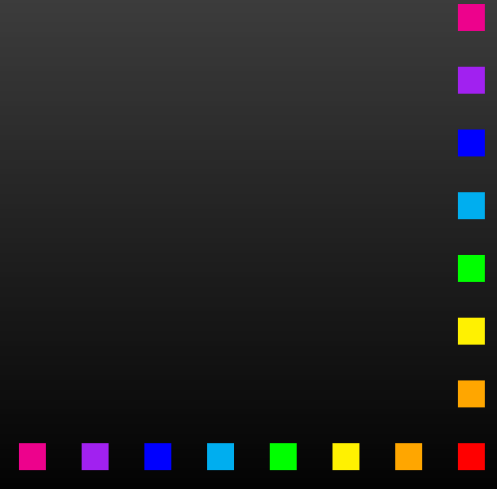
- Bauernmarkt:
  $\vec{A} = \{\_ba, bau, aue, uer, ern, rnm, nma,$
  $mar, ark, rkt, kt\_\}$

Bildung der Trigramme (2):

- Bauernmarkt:
  $\vec{A} = \{\_ba, bau, aue, uer, ern, rnm, nma,$
  $mar, ark, rkt, kt\_\}$
- Marktbauern:
  $\vec{B} = \{\_ma, mar, ark, rkt, ktb, tba, bau,$
  $aue, uer, ern, rn\_\}$

Das Jaccard-Maß (1):

- ist das Verhältnis der Skalarprodukte der beiden Zeichenvektoren



Das Jaccard-Maß (1):

- ist das Verhältnis der Skalarprodukte der beiden Zeichenvektoren
- das Skalarprodukt wird gebildet als

$$w_i \cdot q = \sum_{k=1}^{n} w_{i,k} q_k$$

in unserem Fall ist $w = \vec{A}$ und $q = \vec{B}$ zu verstehen

Das Jaccard-Maß (2):

- das Verhältnis der Skalarprodukte ist demnach

$$s_j(w_i, q) = \frac{\sum_{k=1}^{n} w_{i,k} q_k}{\sum_{k=1}^{n} w_{i,k} + \sum_{k=1}^{n} q_k - \sum_{k=1}^{n} w_{i,k} q_k}$$

Das Jaccard-Maß (3):

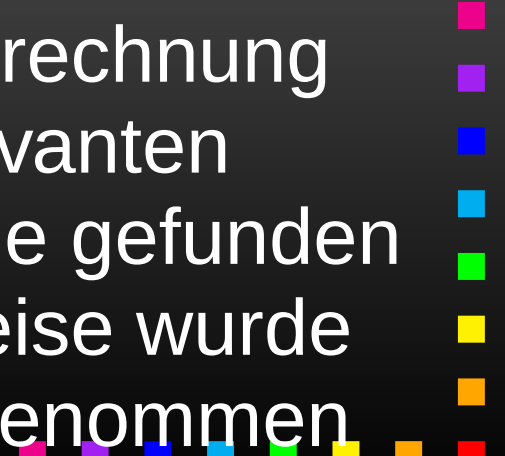
- für Vektoren, die mit den Werten `0` und `1` gewichtet werden (ein Trigramm kommt in beiden Zeichenketten vor – oder eben nicht) vereinfacht sich diese Darstellung auf folgende Gleichung:
  $Jaccard - Maß =$
  $Schnittmenge \div Vereinigungsmenge$
  (=Wert zwischen `0` und `1`)

Das Jaccard-Maß (4):

- nachdem das Jaccard-Maß Werte zwischen 0 und 1 annimmt, muss festgelegt werden, ab welchem Schwellenwert zwei miteinander verglichene Datensätze als ähnlich (oder ident) angesehen werden

Das Jaccard-Maß (4):

- nachdem das Jaccard-Maß Werte zwischen 0 und 1 annimmt, muss festgelegt werden, ab welchem Schwellenwert zwei miteinander verglichene Datensätze als ähnlich (oder ident) angesehen werden
- dafür liegen in der Literatur zur Berechnung bibliographischer Daten keine relevanten Ergebnisse vor (bzw. konnten keine gefunden werden). Für unsere Vorgehensweise wurde ein Wert zwischen 0.7 und 0.8 angenommen bzw. empirisch ermittelt



Der euklidische Abstand (1):

- wird gebildet als die Vektordifferenz zweier Vektoren, die das Vorkommen jener Trigramme abbilden, die in den beiden zu vergleichenden Zeichenketten vorhanden sind

Der euklidische Abstand (2):

- $\vec{A}_{S1} = \{bau, aue, uer, ern, rnm, nma, mar, ark, rkt\}$

Der euklidische Abstand (2):

- $\vec{A}_{S1} = \{bau, aue, uer, ern, rnm, nma, mar, ark, rkt\}$
- $\vec{A}_{S2} = \{mar, ark, rkt, ktb, tba, bau, aue, uer, ern\}$

Der euklidische Abstand (2):

- $\vec{A}_{S1} = \{bau, aue, uer, ern, rnm, nma, mar, ark, rkt\}$
- $\vec{A}_{S2} = \{mar, ark, rkt, ktb, tba, bau, aue, uer, ern\}$
- steht für `Bauernmarkt` und `Marktbauern`

Der euklidische Abstand (3):

- die Differenz ist

$$\begin{aligned}\|\vec{D}\| &= \vec{A}_{S1} - \vec{A}_{S2} \\ &= \{rnm, nma, ktb, tba\} \\ &= \sqrt{1^2 + 1^2 + 1^2 + 1^2} = 2\end{aligned}$$

Der euklidische Abstand (4):

- zum Abstand hinzu kommt die empirische Ermittlung eines Schwellenwerts (T), über dem zwei miteinander zu vergleichende Zeichenketten nicht mehr als ähnlich angesehen werden

$$T = 2.486 + 0.025 * n$$

Der euklidische Abstand (4):

- zum Abstand hinzu kommt die empirische Ermittlung eines Schwellenwerts (T), über dem zwei miteinander zu vergleichende Zeichenketten nicht mehr als ähnlich angesehen werden

$$T = 2.486 + 0.025 * n$$

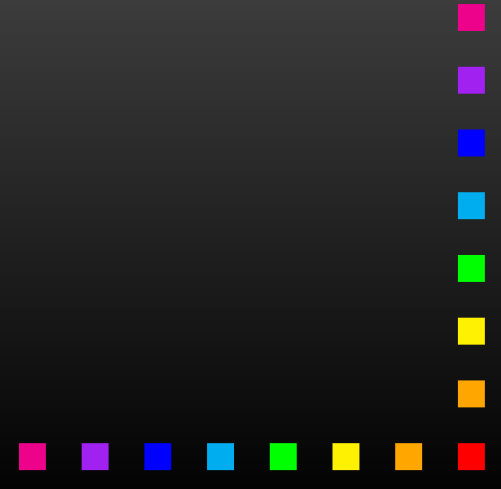
- wobei n die Anzahl der zu vergleichenden Teilstrings in der Vereinigungsmenge beider Vektoren darstellt Hylton (1996, S.47)

Die Methode des KOBV (1):

- bei diesem Verfahren wird die euklidische Distanz berechnet

Die Methode des KOBV (1):

- bei diesem Verfahren wird die euklidische Distanz berechnet
- diese wird auf eine Skala von 0 bis 1 transformiert



Die Methode des KOBV (1):

- bei diesem Verfahren wird die euklidische Distanz berechnet
- diese wird auf eine Skala von 0 bis 1 transformiert
- jede Kategorie wird gesondert verrechnet und gewichtet – und zwar positiv als auch negativ

Die Methode des KOBV (1):

- bei diesem Verfahren wird die euklidische Distanz berechnet
- diese wird auf eine Skala von 0 bis 1 transformiert
- jede Kategorie wird gesondert verrechnet und gewichtet – und zwar positiv als auch negativ
- zur Entscheidung, wie eine Kategorie gewichtet wird, gilt der Wert 0.8 als Schwellenwert

Die Methode des KOBV (2):

- wird bei der Berechnung der Schwellenwert erreicht oder überschritten, wird die Kategorie pos. gewichtet (anderenfalls negativ)

Die Methode des KOBV (2):

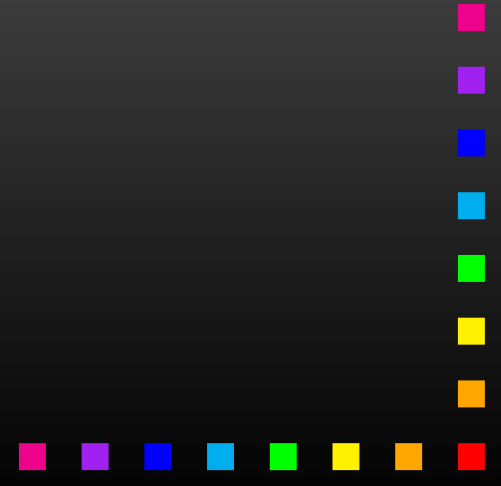
- wird bei der Berechnung der Schwellenwert erreicht oder überschritten, wird die Kategorie pos. gewichtet (anderenfalls negativ)
- zwei Datensätze gelten als gleich, wenn die Summe aller pos. Werte ≥ 0.5 und zugleich die Summe aller neg. Werte ≤ 0.3 ist

Gewichtungen im KOBV

Kategorie	pos. Gewichtung	neg. Gewichtung
Personenname	30	50
Körperschaftsname	30	50
Sachtitel	70	70
Standardnummer (z.B. ISBN)	70	60
Erscheinungsjahr	30	60
Erscheinungsorte	20	30
Verleger/in	20	20
Ausgabe-bezeichnung	30	60
Umfangsangabe	20	40

Prinzip der Datensatzabfrage (1)

- Ausgangspunkt einer Abfrage ist, dass nicht jeder Datensatz mit jedem anderen verglichen werden muss. Das dauert zulange und bringt ein zu großes Übermaß an Berechnungen, die eigentlich nicht notwendig sind



Prinzip der Datensatzabfrage (2)

- daher wird die kleinere Menge an Datensätzen als Anfragemenge genommen und mit dem ersten Begriff aus dem Autorenfeld (=meist wohl ein Teil des Familiennamens) eine Query in der Abfragemenge (=die größere Menge) getätigt

Prinzip der Datensatzabfrage (2)

- daher wird die kleinere Menge an Datensätzen als Anfragemenge genommen und mit dem ersten Begriff aus dem Autorenfeld (=meist wohl ein Teil des Familiennamens) eine Query in der Abfragemenge (=die größere Menge) getätigt
- alle Datensätze, die sich durch diese Query qualifizieren werden mit dem Anfragedatensatz verglichen

Prinzip der Datensatzabfrage (3)

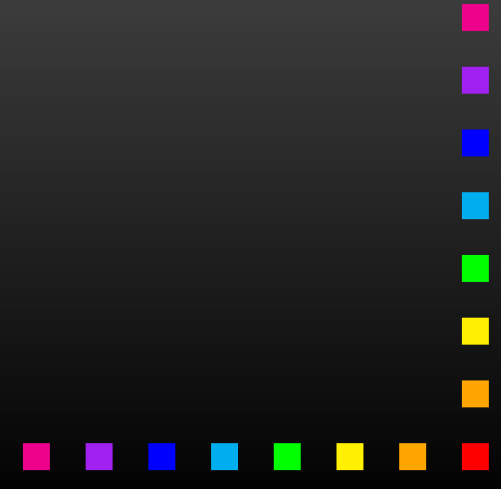
- sollte ein Datensatz keinen Inhalt in Kat. 100 haben, wird der erste Begriff aus Kat. 331, der kein Stoppwort ist, zur Abfrage herangezogen

Struktur/Inhalte der Ergebnistabelle (1)

- Begriff, der zur Query herangezogen wurde

Struktur/Inhalte der Ergebnistabelle (1)

- Begriff, der zur Query herangezogen wurde
- AC-Nummern beider Datensätze



Struktur/Inhalte der Ergebnistabelle (1)

- Begriff, der zur Query herangezogen wurde
- AC-Nummern beider Datensätze
- Jaccard-Maß, nur pos. gewichtet

Struktur/Inhalte der Ergebnistabelle (1)

- Begriff, der zur Query herangezogen wurde
- AC-Nummern beider Datensätze
- Jaccard-Maß, nur pos. gewichtet
- Euklidischer Abstand, nur pos.

Struktur/Inhalte der Ergebnistabelle (1)

- Begriff, der zur Query herangezogen wurde
- AC-Nummern beider Datensätze
- Jaccard-Maß, nur pos. gewichtet
- Euklidischer Abstand, nur pos.
- errechneter Schwellenwert zum euklid. Abstand

Struktur/Inhalte der Ergebnistabelle (2)

- KOBV-Wert, nur pos. gewichtet

Struktur/Inhalte der Ergebnistabelle (2)

- KOBV-Wert, nur pos. gewichtet
- Jaccard-Maß, pos. und neg. gewichtet

Struktur/Inhalte der Ergebnistabelle (2)

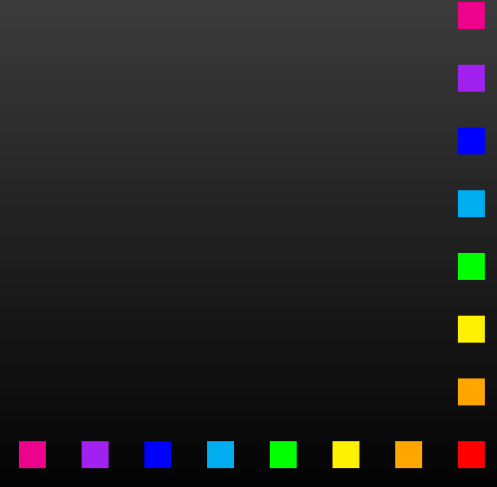
- KOBV-Wert, nur pos. gewichtet
- Jaccard-Maß, pos. und neg. gewichtet
- Euklidischer Abstand, pos. und neg. gewichtet

Struktur/Inhalte der Ergebnistabelle (2)

- KOBV-Wert, nur pos. gewichtet
- Jaccard-Maß, pos. und neg. gewichtet
- Euklidischer Abstand, pos. und neg. gewichtet
- KOBV-Wert, pos. und neg. gewichtet

Was bleibt??

- jede Menge Ergebniszahlen, die „richtig“ interpretiert werden müssen



Was bleibt??

- jede Menge Ergebniszahlen, die „richtig“ interpretiert werden müssen
- empirische Überprüfung der Ergebnisse (Stichproben)

Was bleibt??

- jede Menge Ergebniszahlen, die „richtig“ interpretiert werden müssen
- empirische Überprüfung der Ergebnisse (Stichproben)
- Justierung der Schwellenwerte beim Jaccard-Maß ($0.7 - 0.8$)

Was bleibt??

- jede Menge Ergebniszahlen, die „richtig“ interpretiert werden müssen
- empirische Überprüfung der Ergebnisse (Stichproben)
- Justierung der Schwellenwerte beim Jaccard-Maß ($0.7 - 0.8$)
- Vorteil dieses Ansatzes: Ergebnisse können einander direkt gegenübergestellt werden

Vielen Dank für die Aufmerksamkeit