STAATS- UND UNIVERSITÄTS-
BIBLIOTHEK
HAMBURG
CARL VON OSSIETZKY

Herzlich Willkommen

Relevanzsortierung bei beluga

Imke Rulik, BIS Oldenburg / Hajo Seng, SUB Hamburg

Hamburg, 14.10.2015

Programm

- Herausforderung Relevanzsortierung
- Vorbereitende Überlegungen
- Statistische Analysen und heuristische Betrachtungen
- Ausblicke
- Die Known-Item-Suchanfrage
- Identifizierung von Known-Item-Anfragen
- Retrievaleffektivität von Known-Item-Anfragen
- Fazit

Konsortialsystem beluga

Bibliothekssystem Universität Hamburg	8.601.229 Datensätze
Zentralbibliothek Wirtschaftswissenschaften	4.852.720 Datensätze
Helmut-Schmidt-Universität	1.837.414 Datensätze
Technische Universität Hamburg-Harburg	855.485 Datensätze
Hochschule für angewandte Wissenschaften	404.864 Datensätze
Commerzbibliothek	348.096 Datensätze
Staatsarchiv	234.842 Datensätze
Hamburger Lehrerbibliothek	221.517 Datensätze
Hochschule für Musik und Theater	188.142 Datensätze
Hafencity Universität	179.632 Datensätze
Hamburg Media School	16.573 Datensätze

Herausforderung Relevanzsortierung

- **schwieriger Relevanzbegriff**
- **unterschiedliche Metadatenqualitäten**
- **Abhängigkeiten der Metadaten untereinander**
- **solr-interne Datenverarbeitung (tokenizing, stemming)**
- **nicht-lineare Bewertungsalgorithmen (tf-idf)**
- **viele Parameter, die z.T. voneinander abhängen**

Herausforderung Relevanzsortierung

Beispiel: Erste zehn Treffer mit einem E-Zeitschriften-Boosting von 28,95

Rang:1 PPN:788889680 Handbook of Human Resources Management (Book, 2016)

Rang:2 PPN:802100163 Operations management : Processes and supply chains (Book, 2016)

Rang:3 PPN:816726515 Teufelswerk oder Himmelsgeschenk? ... (Book, 2016)

Rang:4 PPN:823734730 Investing in China Through Shanghai Free Trade Zone (Book, 2016)

Rang:5 PPN:797284907 Project management achieving competitive advantage (Book, 2016)

Rang:6 PPN:799528900 Financial management core concepts (Book, 2016)

Rang:7 PPN:722134363 Wissenschaftstheorie und wissenschaftliches Arbeiten ... (Book, 2016)

Rang:8 PPN:79696727X Business a changing world (Book, 2016)

Rang:9 PPN:798452102 Understanding financial statements (Book, 2016)

Rang:10 PPN:819481149 Slavery in the circuit of sugar Martinique and ... (Book, 2016)

Herausforderung Relevanzsortierung

Beispiel: Erste zehn Treffer mit einem E-Zeitschriften-Boosting von 29,00

Rang:1 PPN:816694591 Journal of service theory and practice (eJournal, 2015)

Rang:2 PPN:819365017 Statistical yearbook Curaçao (eJournal, 2015)

Rang:3 PPN:818042699 Hohenheim discussion papers in business, ... (eJournal, 2015)

Rang:4 PPN:826235999 Prego das Kulturmagazin von Edel (eJournal, 2015)

Rang:5 PPN:826757251 Diskussionspapiere / Hannover Center of Finance e.V (eJournal, 2015)

Rang:6 PPN:815915756 Journal of Aquaculture Engineering and Fisheries ... (eJournal, 2015)

Rang:7 PPN:815917015 Ocular oncology and pathology (eJournal, 2015)

Rang:8 PPN:817363181 Nuclear materials and energy (eJournal, 2015)

Rang:9 PPN:818041668 Heidelberger Zeitschrift für Iranistik (eJournal, 2015)

Rang:10 PPN:818041994 BMC nutrition (eJournal, 2015)

Herausforderung Relevanzsortierung

Aber: **Die Sortierung bestimmt, was Nutzer/innen mit ihren Suchen in erster Linie finden.**

Ziele:

Transparente, nachvollziehbare Sortierung

Sichtbarmachung *relevanter* Medien

Felder clustern

Ziel: Reduzierung von Komplexität

Zu berücksichtigen:

Feldabbildungen: pica → marc → Index

Gegenseitige Überschneidungen der Felder

Feldabdeckungen

Qualität der Feldinhalte

Felder clustern

Beispiel: Titelcluster

Abhängigkeiten der Felder untereinander:

- ***title = series + title_auth*** *(disjunkte Vereinigung)*
- ***title_full > title_auth > title_sub*** *(Enthaltensein)*
- ***title_full > title_short***

Feldabdeckungen:

- ***title 94% Feldabdeckung***
- ***title_auth 92% Feldabdeckung***
- ***title_short 92% Feldabdeckung***
- ***title_full 92% Feldabdeckung***
- ***title_sub 38% Feldabdeckung***
- ***series 10% Feldabdeckung***

Felder clustern

Identifizierte Cluster:

- *Autorencluster:* *author, author2*
- *Titelcluster (1):* *series, title_auth, title_short, series2, journal*
- *Titelcluster (2):* *title_alt, title_old, title_new*
- *Erschließungscluster:* *topic, class, (class_local), geographic, dewey-xyz, bklname*
- *Volltextcluster:* *contents, abstract, fulltext, (allfields)*
- *gering bewertete Felder:* *isbn, issn, publisher, publishPlace, ctrlnum*

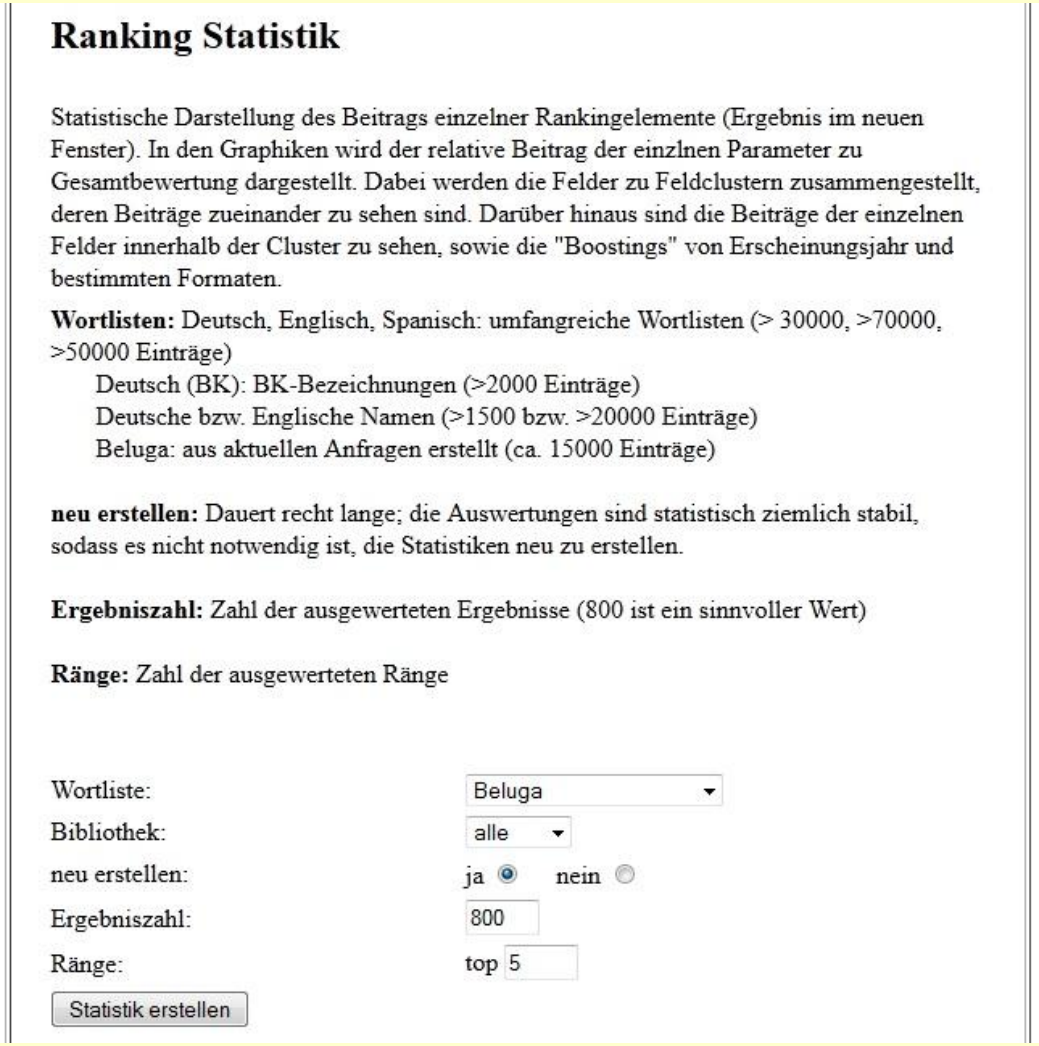
Felder (grob) vorbewerten

Feldbewertungen (Abdeckung, Überschneidung, Qualität):

- **Autorencluster:** - ***author*** sollte höher bewertet sein als ***author2***.
- **Titelcluster (1):** - ***title_auth*** und ***title_short*** sollten höher bewertet sein als ***series*** und ***series2***, aber nicht zu sehr, da sie sich überlappen.
  - ***series*** und ***series2*** sollten höher bewertet sein als ***journal*** (da eher die Zeitschriften gefunden werden sollen, als enthaltene Artikel).
- **Titelcluster (2):** - ***series***, ***title_auth*** und ***title_short*** sollten höher bewertet sein als ***title_alt***, ***title_old***, ***title_new***.
- **Erschließung:** - ***class*** sollte höher bewertet werden als ***topic***.
  - ***class_local*** wird nicht verwendet (Qualität).
  - ***geographic*** sollte wegen der geringen Feldabdeckung nicht zu hoch bewer-tet werden.
  - ***bklname*** *sollte wegen der geringen Feldabdeckung ebenfalls nicht zu hoch bewertet werden.*
- **Volltextcluster:** - ***contents*** und ***abstract*** können ähnlich hoch bewertet werden.
  - ***fulltext*** wird nur gering bewertet
  - ***allfields*** wird nicht verwendet.

Statistische Analysen: Analysewerkzeug

Eingabemaske:



Ranking Statistik

Statistische Darstellung des Beitrags einzelner Rankingelemente (Ergebnis im neuen Fenster). In den Graphiken wird der relative Beitrag der einzlnen Parameter zu Gesamtbewertung dargestellt. Dabei werden die Felder zu Feldclustern zusammengestellt, deren Beiträge zueinander zu sehen sind. Darüber hinaus sind die Beiträge der einzelnen Felder innerhalb der Cluster zu sehen, sowie die "Boostings" von Erscheinungsjahr und bestimmten Formaten.

Wortlisten: Deutsch, Englisch, Spanisch: umfangreiche Wortlisten (> 30000, >70000, >50000 Einträge)

Deutsch (BK): BK-Bezeichnungen (>2000 Einträge)
Deutsche bzw. Englische Namen (>1500 bzw. >20000 Einträge)
Beluga: aus aktuellen Anfragen erstellt (ca. 15000 Einträge)

neu erstellen: Dauert recht lange; die Auswertungen sind statistisch ziemlich stabil, sodass es nicht notwendig ist, die Statistiken neu zu erstellen.

Ergebniszahl: Zahl der ausgewerteten Ergebnisse (800 ist ein sinnvoller Wert)

Ränge: Zahl der ausgewerteten Ränge

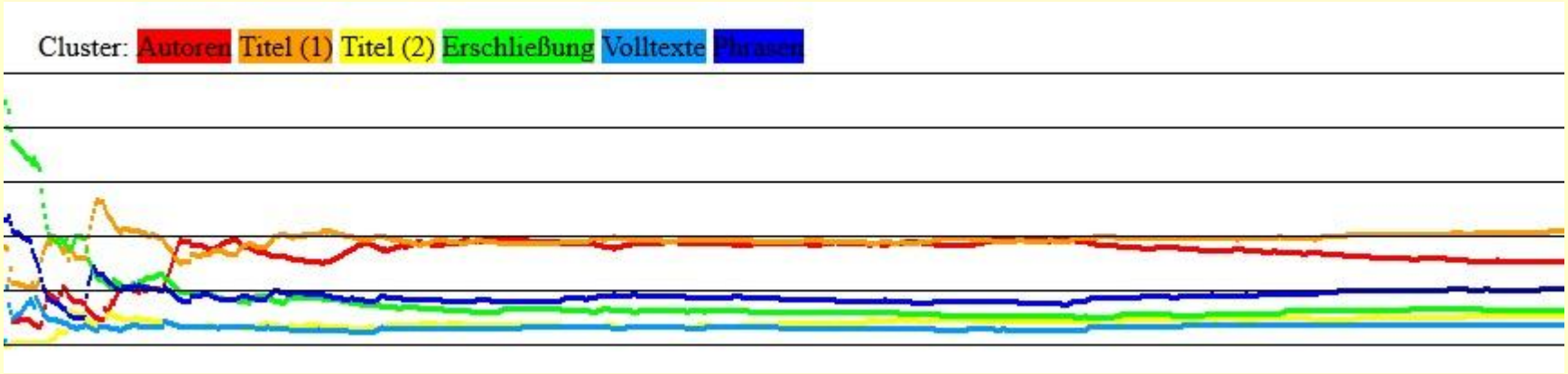
Wortliste: Beluga
Bibliothek: alle
neu erstellen: ja nein
Ergebniszahl: 800
Ränge: top 5
Statistik erstellen

Statistische Analysen: solr-Ausgabe

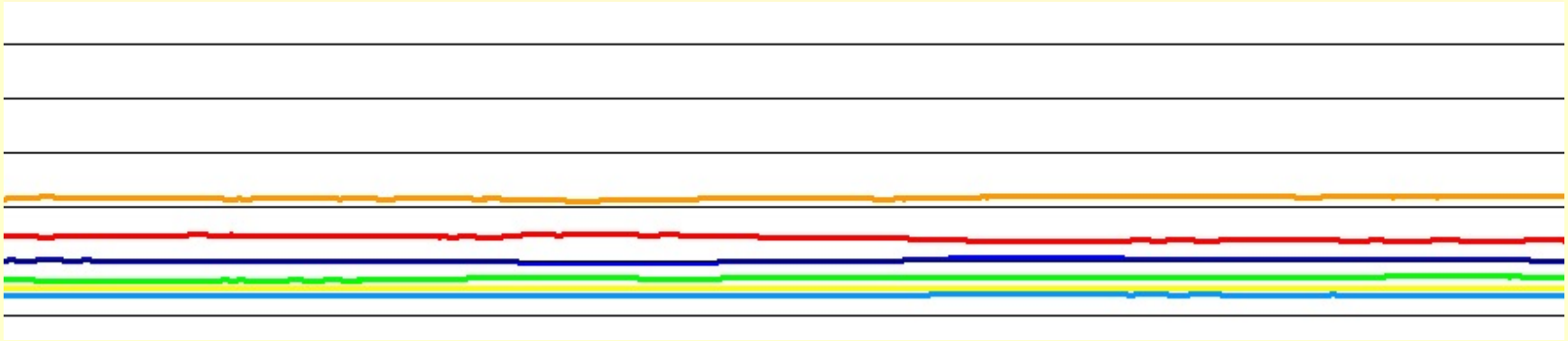
```
12.075157 = (MATCH) max of:
  4.346499 = (MATCH) weight(topic:"java script" in 248021) [DefaultSimilarity], result of:
    4.346499 = score(doc=248021,freq=2.0), product of:
      0.5640543 = queryWeight, product of:
        21.795341 = idf(), sum of:
          10.256164 = idf(docFreq=1442, maxDocs=15106620)
          11.539179 = idf(docFreq=399, maxDocs=15106620)
        0.02587958 = queryNorm
      7.7058167 = fieldWeight in 248021, product of:
        1.4142135 = tf(freq=2.0), with freq of:
          2.0 = phraseFreq=2.0
        21.795341 = idf(), sum of:
          10.256164 = idf(docFreq=1442, maxDocs=15106620)
          11.539179 = idf(docFreq=399, maxDocs=15106620)
        0.25 = fieldNorm(doc=248021)
    12.075157 = (MATCH) weight(title:"java script"^2.0 in 248021) [DefaultSimilarity], result of:
      12.075157 = fieldWeight in 248021, product of:
        1.0 = tf(freq=1.0), with freq of:
          1.0 = phraseFreq=1.0
        19.320251 = idf(), sum of:
          9.268343 = idf(docFreq=3874, maxDocs=15106620)
          10.0519085 = idf(docFreq=1769, maxDocs=15106620)
        0.625 = fieldNorm(doc=248021)
```

Statistische Analysen: Konvergenz

Clusterbewertung bei 2000 Top-5 Abfragen: erste Hälfte …

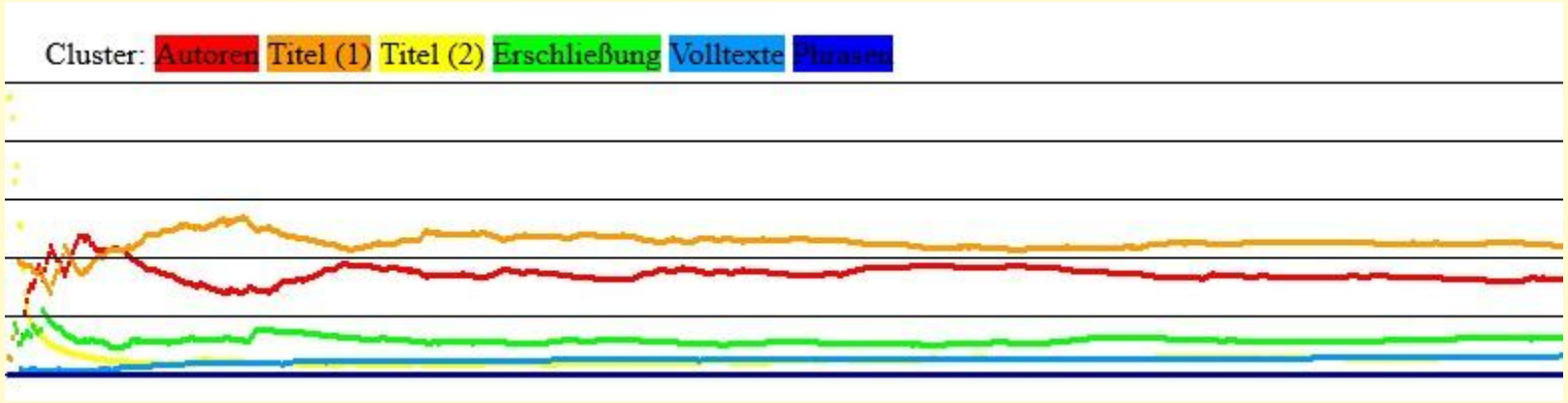


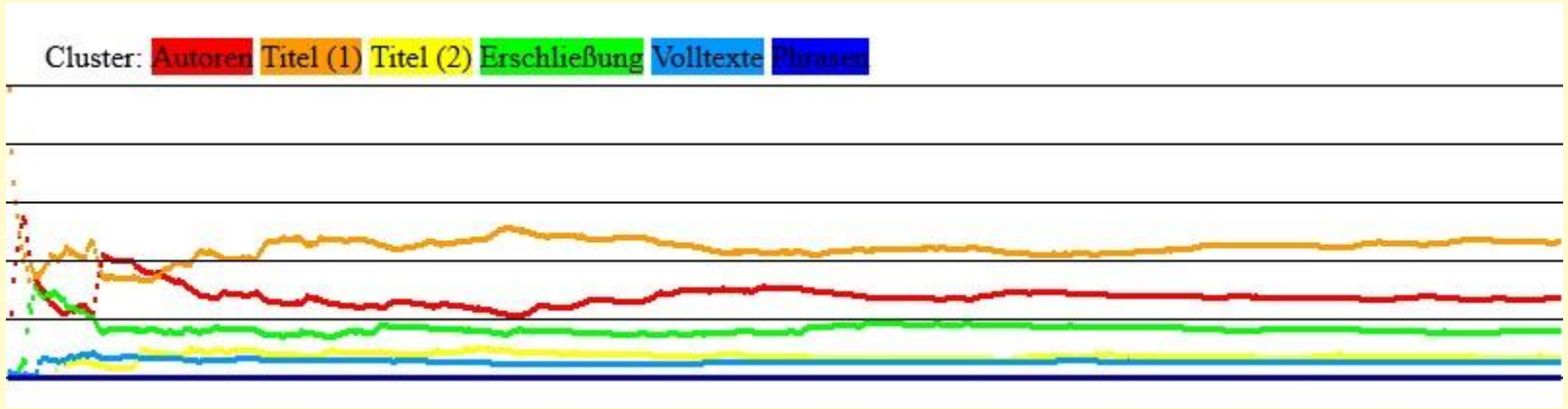
… und die zweite Hälfte



Statistische Analysen: Varianz

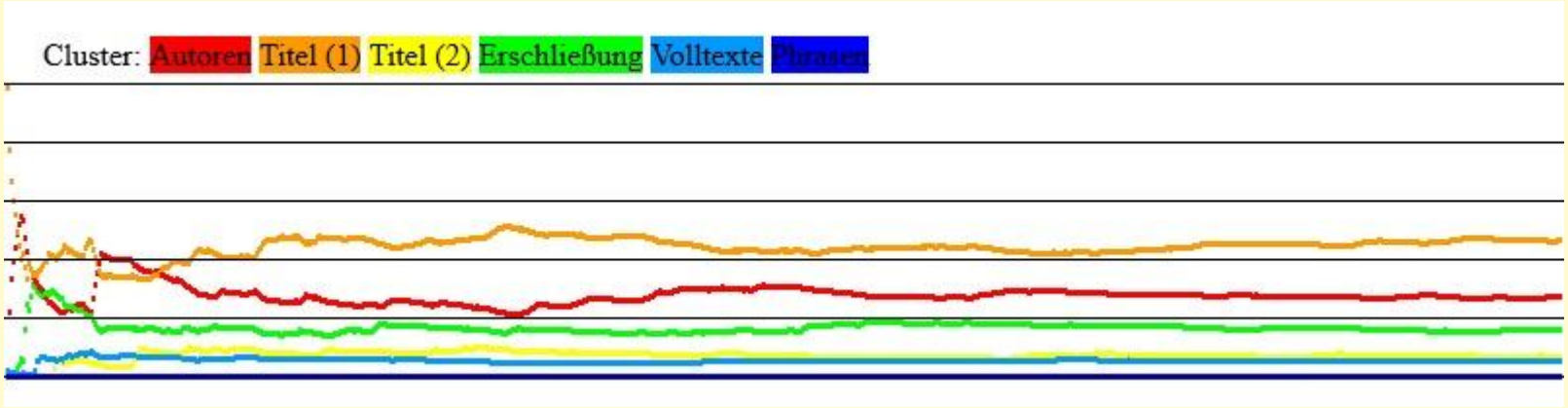
Zwei Analysen über beluga-Suchanfragen:

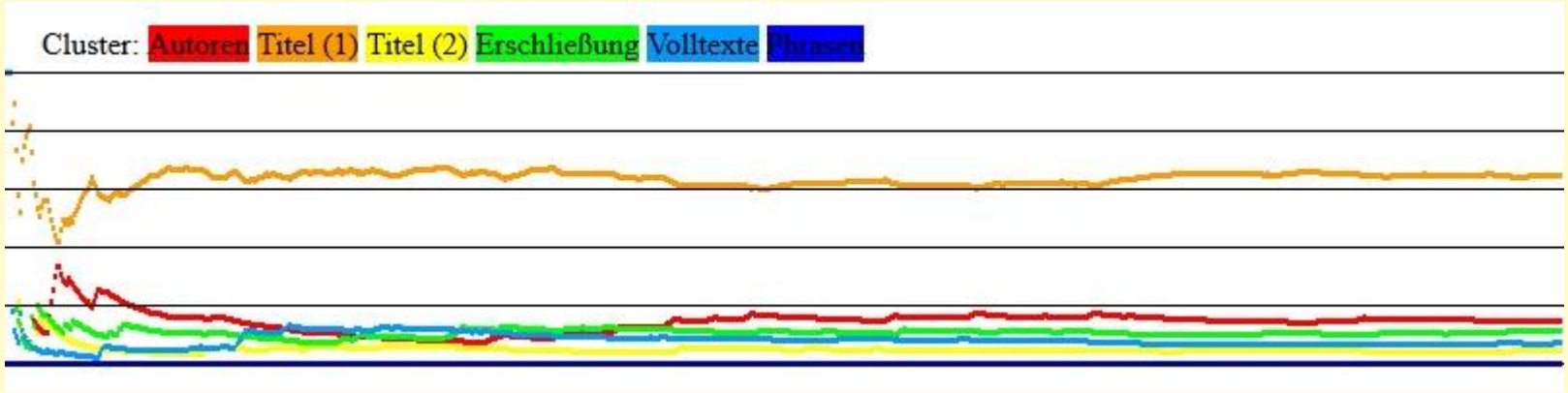




Statistische Analysen: Varianz

Zwei Analysen über verschiedene Wortmengen:





Statistische Analysen: Varianz

Zwei Analysen über verschiedene Wortmengen:

Wortliste	beluga-Suchanfragen
Entflut	*organon der*
Entfluten	*Artistic Citizenship A Public for Arts.*
Entflügel	*Vegetationsgeographie*
Entflügeln	*"Energiebewusstes Bauen"*
Entform	*fo fa gai lun*
Entformen	*trauma und Kindheit*
Entformung	*A1997/7913*
Entformungsschräge	*e-commerce* einzelandel**

Statistische Analysen: Zwischenergebnis

Titelcluster (1):	*title_auth*	*18*
	title_short	*18*
	series	*13*
	series2	*10*
	journal	*5*
Titelcluster (2):		
	title_alt	*5*
	title_new	*5*
	title_old	*5*
Autorencluster:		
	author	*17*
	author2	*12*
Erschließungscluster:		
	class	*30*
	topic	*20*
	geographic	*15*
	bklname	*15*
Volltextcluster:		
	abstract	*15*
	contents	*15*
	fulltext	*10*

...

↓

Zwischenergebnis

↓

statistische Analyse

↓

heuristische Betrachtung

↓

Zwischenergebnis

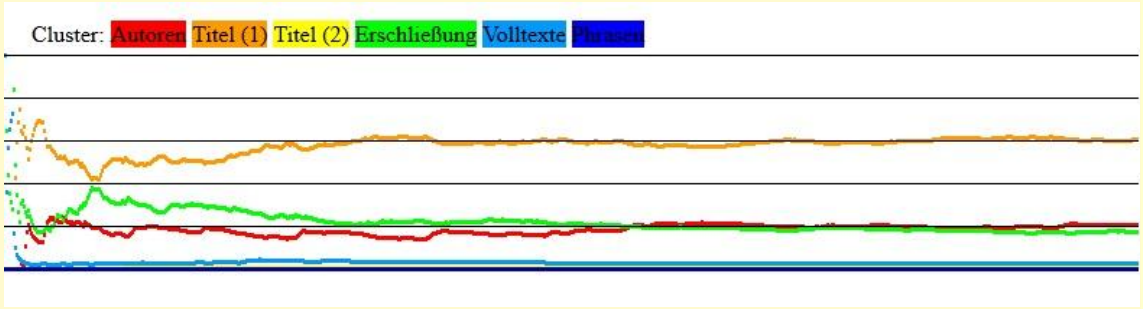
↓

...

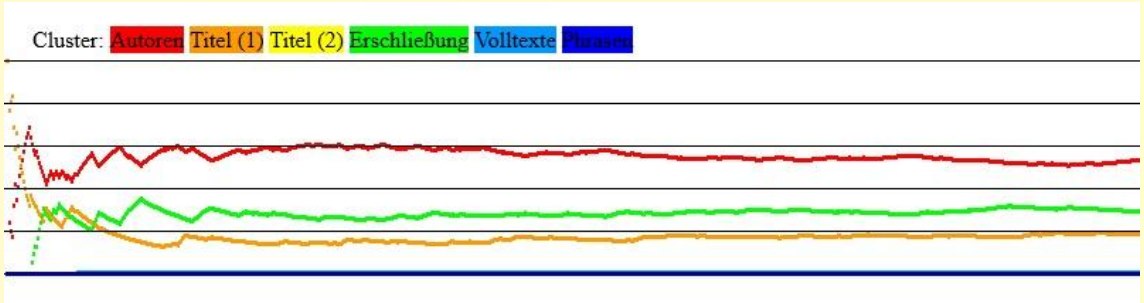
Statistische Analysen: Cluster zueinander

Cluster:

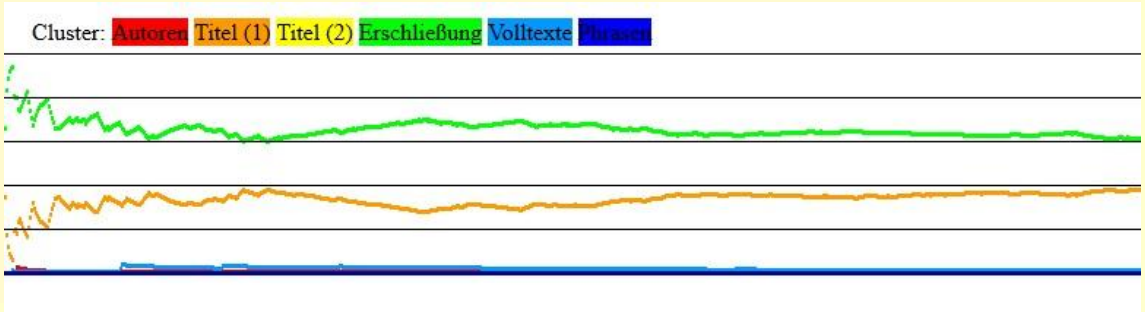
beluga-Abfragen



Deutsche Nachnamen

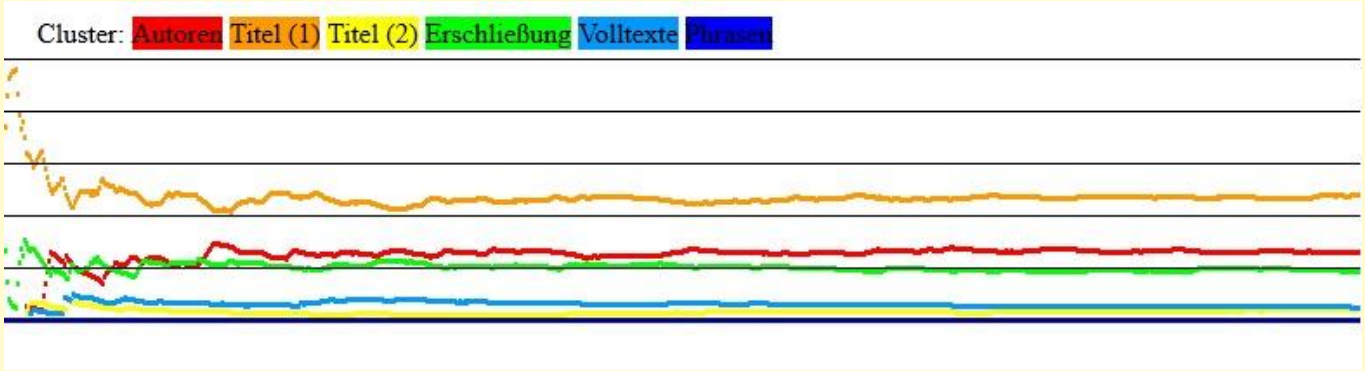


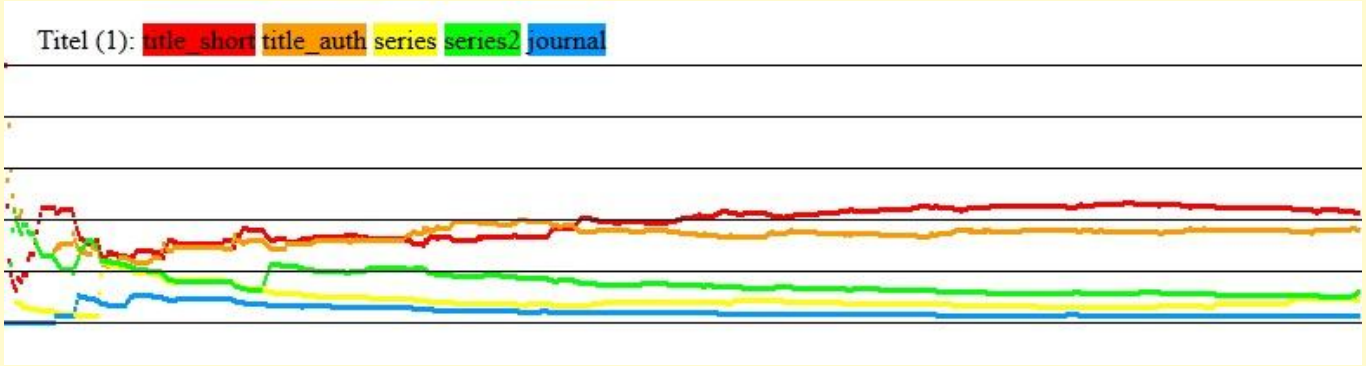
BK-Bezeichnungen



Statistische Analysen: Cluster intern

Cluster (oben) und Titelcluster-Details (unten)

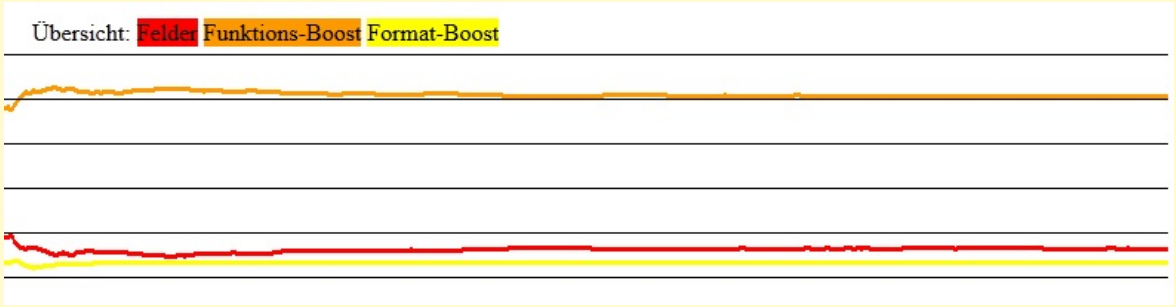




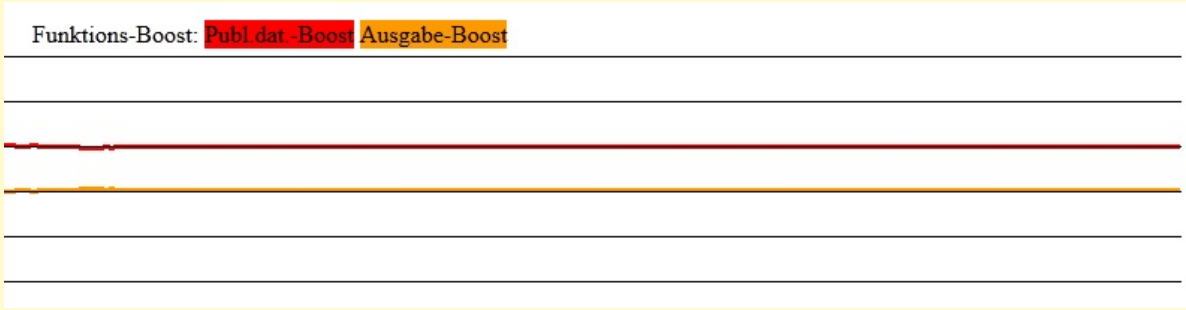
Statistische Analysen: Boosting

Boosting:

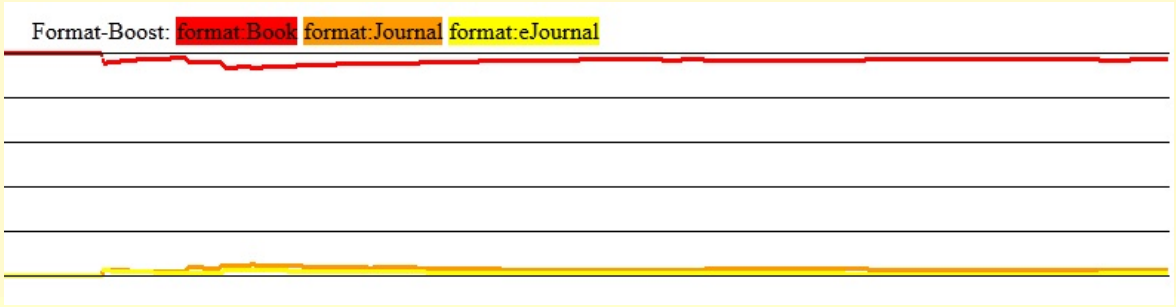
Übersicht



Funktionen



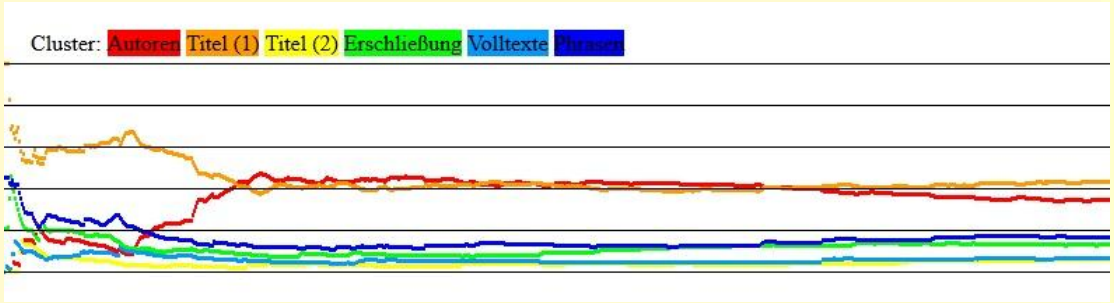
Formate



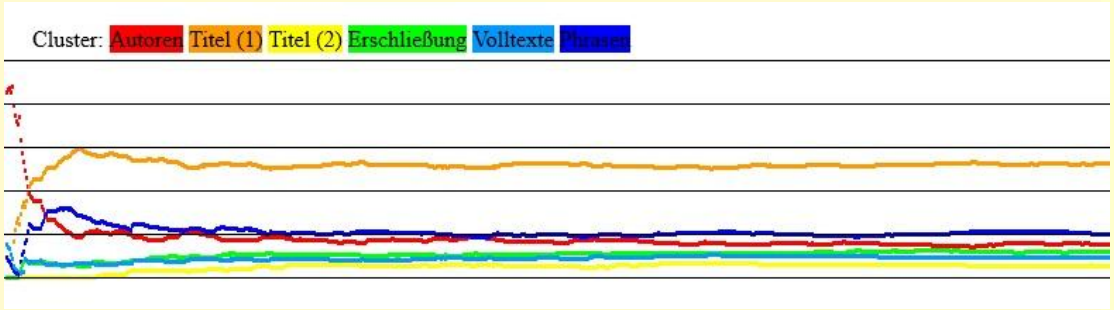
Statistische Analysen: Tiebreaker

tie-Parameter:

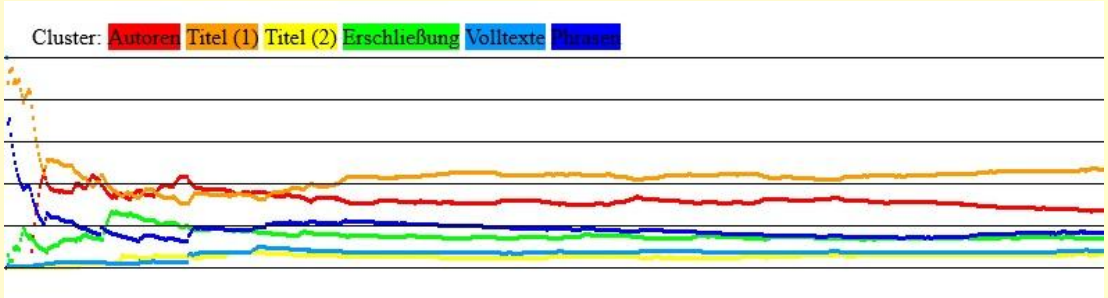
0



1



0,1



Parameter für die Relevanzsortierung

Bewertung der einzelnen Felder

Stärken zusammenhängender Felder („Tiebreaker“)

Anzahl der gefunden Teiltreffer bei mehreren Suchbegriffen

Bewertung von Phrasen

Zulässige „Entfernung“ in Phrasen / „gesplittete“ Phrasen

„Aufblasen“ („Boosting“) nach Feldeinträgen

Boosting nach anderen Merkmalen des Datensatzes

und weiter?

Nutzerverhalten weiter auswerten

Facettierungen betrachten

Known-entity-Suchen identifizieren

Mehr zum Thema im beluga-Blog:

http://beluga-blog.sub.uni-hamburg.de/blog/

Vielen Dank

Hajo Seng
Von-Melle-Park 3
20146 Hamburg

040 / 4 28 38-8336
hajo.seng@sub.uni-hamburg.de

www.sub.uni-hamburg.de
facebook.com/stabihh
twitter.com/stabihh