Sachliche Einordnung von Dokumenten in Bibliotheken: praktische Erfahrungen mit maschinellen Lernverfahren

Monika Lösse (Deutsche Nationalbibliothek)
Mathias Lösch (Universitätsbibliothek Bielefeld)

Workshop on Classification and Subject Indexing in Library and Information Science (LIS'2012)

Universität Bielefeld
INFORMATION.plus! Universitätsbibliothek
1912 100 JAHRE 2012 DEUTSCHE NATIONAL BIBLIOTHEK

Motivation

- Zahl der elektronischen Publikationen steigt massiv
- Mehrzahl ist nicht klassifikatorisch erschlossen
- Nachträgliche intellektuelle Klassifikation bei den Aggregatoren aufgrund der Menge nicht möglich/finanzierbar

Automatische Dokumentenklassifizierung in Bibliotheken

- Idee nicht neu
- Ergebnisse jedoch bisher nicht zufriedenstellend, kein Produktiveinsatz
- Seit den 1990er Jahren jedoch zunehmend erfolgreiche Anwendung von maschinellen Lernverfahren zur Textkategorisierung (z.B. Spam-Filter)

Zwei Praxisbeispiele

1. Projekt PETRUS (Deutsche Nationalbibliothek)
   - Automatische Klassifikation von Netzpublikationen nach DDC-Sachgruppen
2. Projekt „Automatische Anreicherung von OAI-Metadaten“ (UB Bielefeld)
   - Automatische Klassifizierung von Metadaten nach DDC-Sachgruppen für die *Bielefeld Academic Search Engine* (BASE)
3. Gemeinsame Erkenntnisse

DDC-Sachgruppen

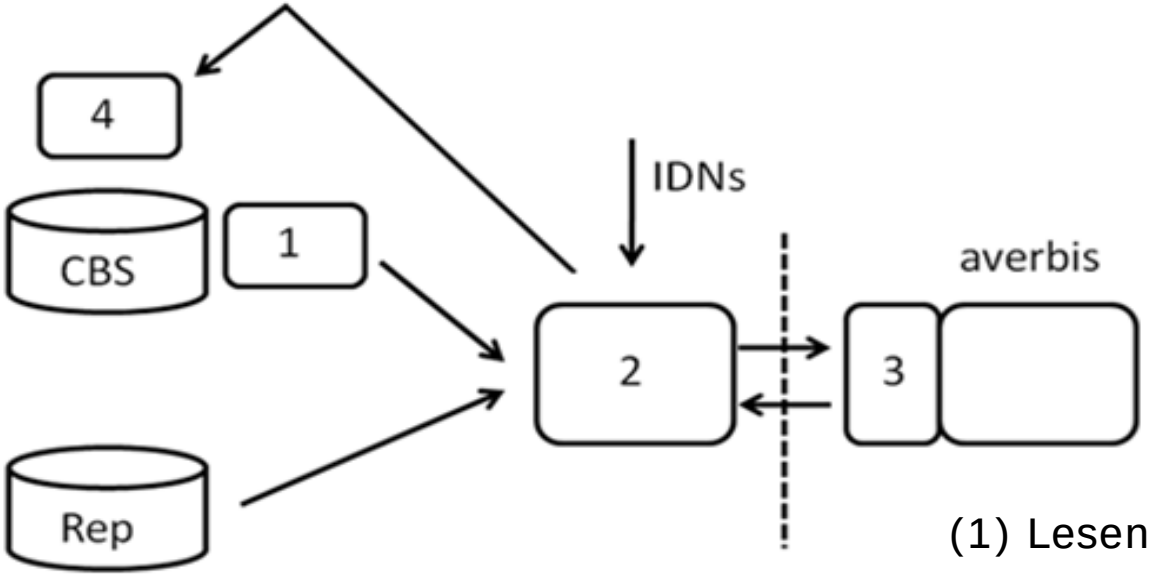
1912 100 JAHRE 2012 DEUTSCHE NATIONAL BIBLIOTHEK

- Seit Bibliografie-Jahrgang 2004: Gliederung der Bibliografie-Reihen nach DDC-Sachgruppen
- System zur thematischen Ordnung von Titeldaten unabhängig von der Sprache der Publikationen
- System beruht auf der Dewey-Dezimalklassifikation (DDC) und gliedert sich zurzeit in 104 DDC-Sachgruppen

000	Allgemeines, Wissenschaft
004	Informatik
010	Bibliografien
...	
500	Naturwissenschaften
510	Mathematik
520	Astronomie, Kartographie
530	Physik
...	
600	Technik
610	Medizin, Gesundheit
620	Ingenieurwissenschaften und Maschinenbau
621.3	Elektrotechnik, Elektronik
...	

Abläufe im Produktivbetrieb bei der DDC-Sachgruppen-Vergabe

realisiert mit der Averbis Extraction Platform seit Januar 2012
für deutsch- und englischsprachige Netzpublikationen



(1) Lesende Schnittstelle CBS
(2) DnBPetrus-Service
(3) Averbis-Webservice
(4) Schreibende Schnittstelle CBS
(Rep) Repository

Modellbildung
für die maschinelle Klassifikation

1912 100 JAHRE 2012 DEUTSCHE NATIONAL BIBLIOTHEK

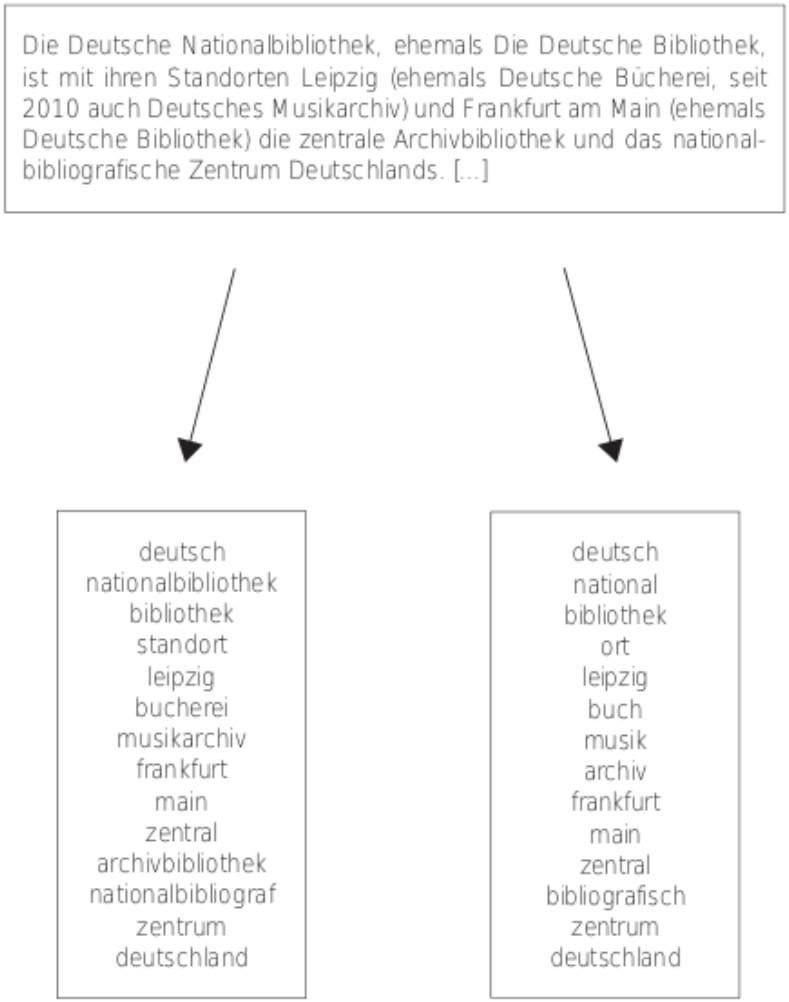
- Modellbildung mit statistischen Lernverfahren (z.B. SVM): Das System lernt anhand von Trainingsbeispielen und leitet daraus mathematische Gesetzmäßigkeiten ab.

- Trainings- und Testbeispiele sind Netzpublikationen und gescannte Inhaltsverzeichnisse mit intellektuell vergebenen Sachgruppen.

Gewinnung der Merkmalsvektoren für Training und Klassifikation

- Extraktion der Merkmale einer Publikation mit linguistischen Verfahren
  (optional: Stemming-, Segment- oder Konzept-Modus)

- Reduktion und Gewichtung der Merkmale
  (Entfernen von Stoppwörtern; Gewichtung nach Position, Textgröße, Häufigkeit im Text vs. Häufigkeit in der Kollektion etc.)

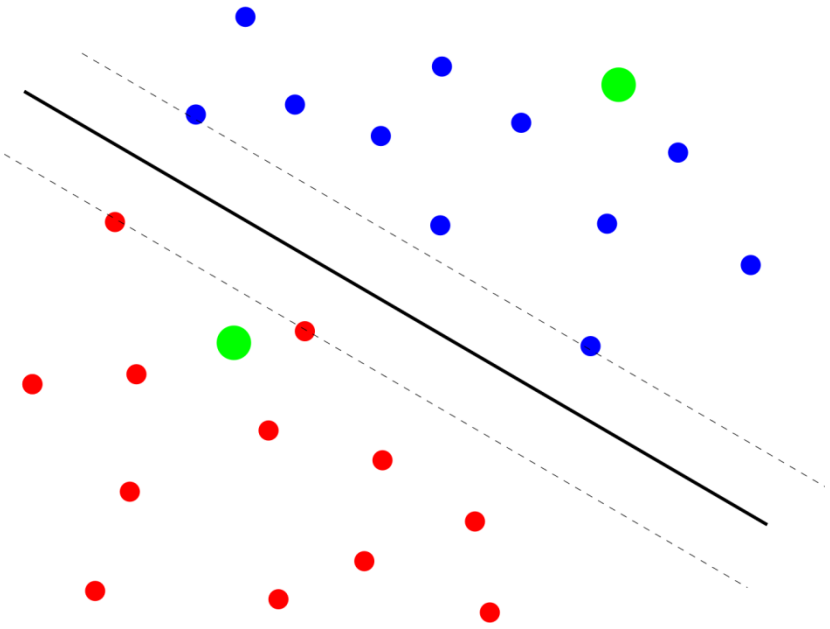


Quelle: Averbis

Durchführung der maschinellen Klassifikation

- Vergleich der Merkmalsvektoren mit den beim Training für jede der DDC-Sachgruppen ermittelten Klassifikatoren

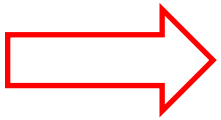
- Berechnung der Konfidenzwerte für jede Klasse (d.h. Bestimmung der Wahrscheinlichkeit der Klassenzugehörigkeit) über den Abstand zu einer Trennebene im mehrdimensionalen Vektorraum



Quelle: Averbis

Qualitätssicherung im Geschäftsprozess

- Falls der Konfidenzwert der übernommenen Sachgruppe unterhalb eines Schwellenwerts liegt, wird der Statusindikator für die intellektuelle Nachbearbeitung gesetzt.
- Einmal pro Monat werden die maschinell vergebenen Sachgruppen mit intellektuell zugewiesenen Sachgruppen verglichen. Das sind z.B.:
  - aus parallelen Ausgaben übernommene Sachgruppen,
  - intellektuell vergebene Sachgruppen aus der Qualitätssicherung.
- Der berechnete F-Measure-Wert berücksichtigt gleichermaßen Recall (= Ausbeute) und Precision (= Genauigkeit).



Zurzeit werden pro Monat 5.000 – 10.000 Netzpublikationen maschinell klassifiziert. Dabei wird ein F-Measure-Wert von 0.7 erreicht.

Dokumentation aller Erfassungsarten im bibliografischen Datensatz

```
...
3010   !130464511!Hartinger, Anselm [Hrsg.]
4000   Vergnügte Pleißenstadt : Bach in Leipzig /
       Anselm Hartinger
...

5050 780$Ei$D2011-02-10

5050 780$Ep$D2011-03-05

5050 780$Em$K0,8$D2012-01-19

5050 330$Ea$D2010-03-04
```

Die dargestellte Reihenfolge entspricht der Rangfolge der Sachgruppen:

$Ei	– intellektuell erstellt
$Ep	– aus paralleler Ausgabe
$Em	– maschinell gewonnen
$Ea	– aus Fremddaten bei der Ablieferung

Produktivbetrieb 2012

1912 100 JAHRE 2012 DEUTSCHE NATIONAL BIBLIOTHEK

Monat	Anzahl Abgelieferte NP	mSG	pSG + mSG	SG qs
Januar	**13.410**	**9.970**	4.110	439
Februar	**14.140**	**8.995**	4.117	250
März	**13.195**	**6.248**	2.399	231
April	**12.347**	**3.742**	1.446	159
Mai	**11.596**	**11.294**	3.618	577

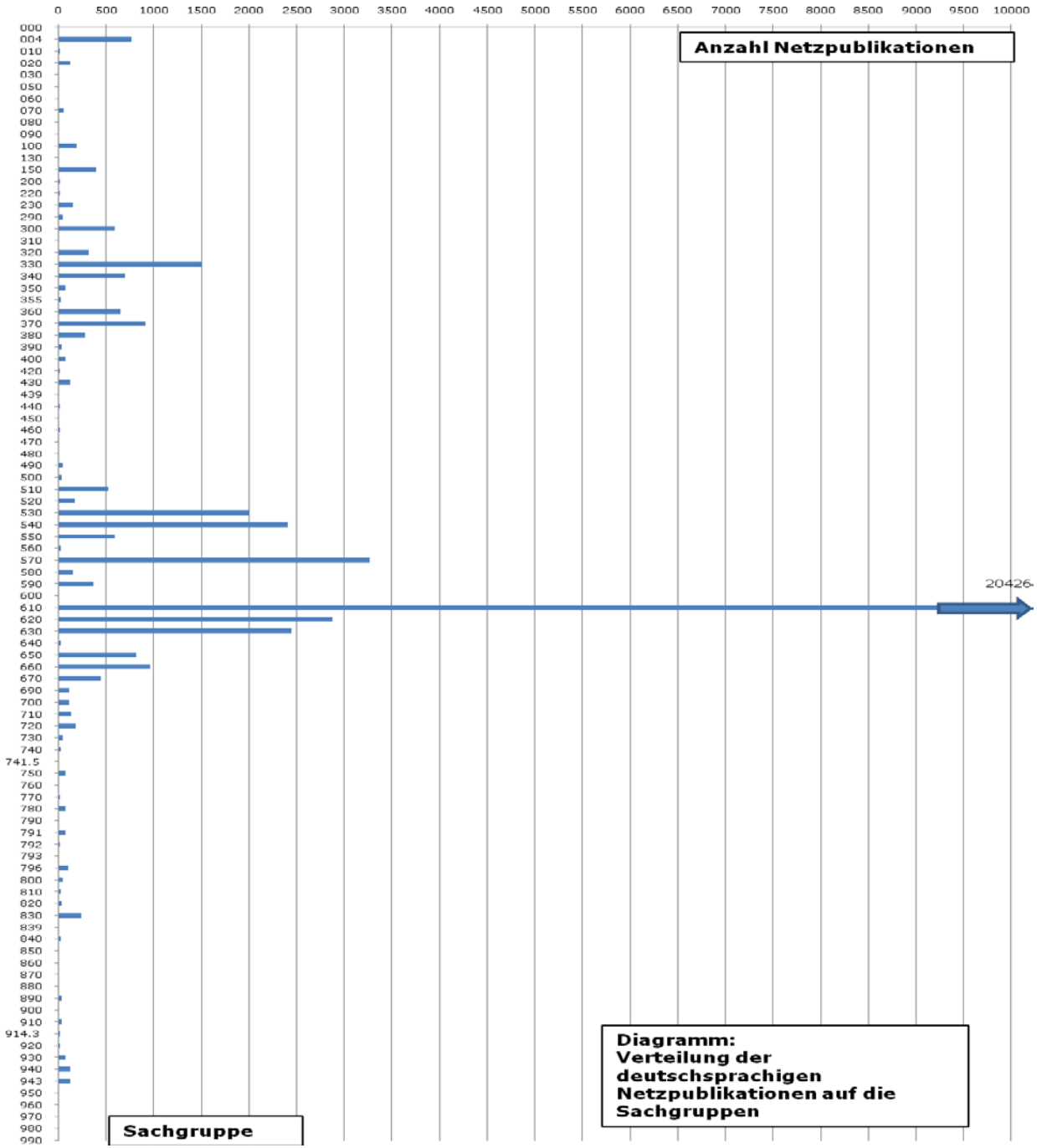
Monatlicher Abgleich
aus parallelen Ausgaben übernommene Sachgruppen mit
maschinell vergebenen Sachgruppen

Monat	F-Measure micro-average
Januar	0,69
Februar	0,76
März	0,61
April	0,70

Probleme: Training

- Gemischtes Trainingsmaterial: Netzpublikationen und gescannte Inhaltsverzeichnisse!
- Ungleichmäßige Verteilung der Netzpublikationen und der gescannten Inhaltsverzeichnisse auf die Sachgruppen!
- Kein Trainingsmaterial für alle Sachgruppen!
- 2004 bis 2006 Vergabe der Sachgruppen vor Einführung der DDC-Notationsvergabe; damit Fehler im Trainingsmaterial.
- Neue Sachgruppe seit 2010:

333.7	Natürliche Ressourcen, Energie und Umwelt
620	Ingenieurwissenschaften und Maschinenbau
621.3	Elektrotechnik, Elektronik
624	Ingenieurbau und Umwelttechnik
491.8	Slawische Sprachen
891.8	Slawische Literatur



1912 100 JAHRE 2012 DEUTSCHE NATIONAL BIBLIOTHEK

Trainingsmaterial

Netzpublikationen

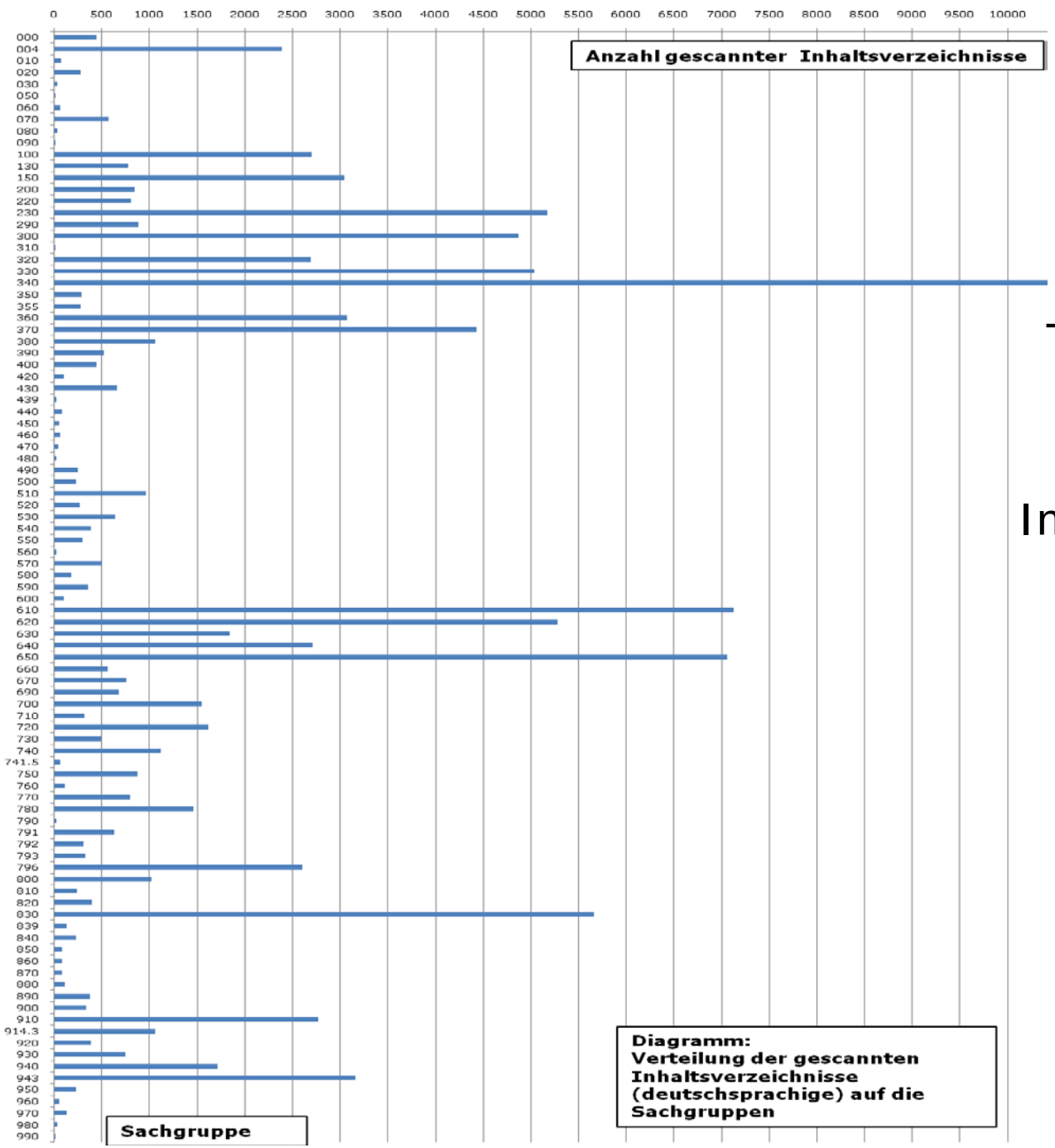


Diagramm:
Verteilung der gescannten Inhaltsverzeichnisse (deutschsprachige) auf die Sachgruppen

1912 100 JAHRE 2012 DEUTSCHE NATIONAL BIBLIOTHEK

Trainingsmaterial gescannte Inhaltsverzeichnisse

Probleme:
Heterogene Netzpublikationstypen

– Sehr heterogene Dokumententypen:
  - Belletristik
  - Hochschulschriften
  - BoD

– Unregelmäßige Ablieferung über die Jahre.

Verschiedene Dokumenttypen Netzpublikationen	2010 Anzahl NP	2011 Anzahl NP
NP BoD	87.885	27.783
NP Springer	6.932	53.777
NP Hochschulschriften	12.984	12.265
NP Reihe A ohne Springer	998	928
Alle abgelieferten NP	112.765	101.181

Probleme:
Häufig vertauschte Sachgruppen

004 Informatik	⟷	620 Ingenieurwissenschaften
150 Psychologie	⟷	610 Medizin, Gesundheit
330 Wirtschaft	⟷	650 Management
360 Soziale Probleme, Sozialdienste, Versicherungen	⟷	610 Medizin, Gesundheit
570 Biowissenschaften, Biologie	⟷	610 Medizin, Gesundheit

Weitere Probleme:

- Nur 40.000 kB des Volltextes stehen zur Verfügung!
- Fehler bei der Spracherkennung! Z.B. russische Netzpublikationen werden oft nicht erkannt und bekommt fälschlicherweise eine Sachgruppe.
- Nur Netzpublikationen im PDF-Format bekommen maschinell eine Sachgruppe zugewiesen. EPUB und andere Formate können nicht bearbeitet werden.
- Laden der Trainingskorpora sehr langsam, deshalb wurde bisher nur mit 22% des zur Verfügung stehenden Trainingsmaterials Modelle erstellt.

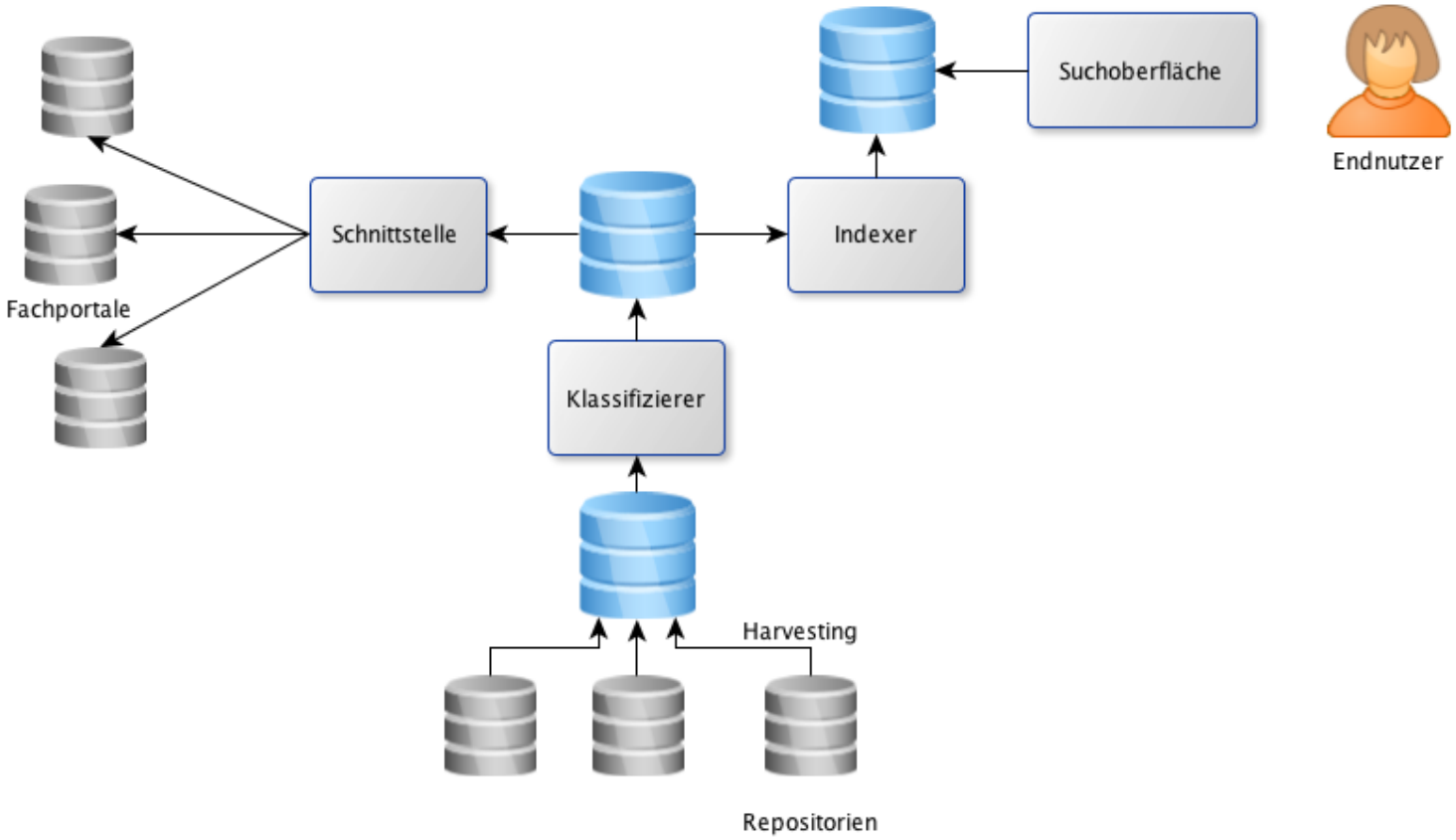
Lösung der Probleme?

- Durch schrittweise Optimierung der Trainingskorpora sollen die Modelle laufend weiter verbessert werden, insbesondere hinsichtlich der „schwierigen“ Sachgruppen:
  - Sachgruppen, für die nicht genügend Trainingsbeispiele vorhanden sind (z.B. seltene oder neue Sachgruppen),
  - Sachgruppen, die inhaltlich schwer trennbar sind (z.B. 330 Wirtschaft vs. 650 Management).
- Verbesserung der Klassifizierungs-Software in einem Folgeprojekt durch die Firma Averbis.
- Verbesserung des Produktivprozesses:
  - durch Ausweitung auf weitere Formate, insbesondere EPUB
  - Verbesserung der Spracherkennung
  - Strukturanalyse
  - Nutzung der abgelieferten Warengruppen in der Belletristik

DFG-Projekt „Automatische Anreicherung von OAI-Metadaten“

- Automatische Klassifikation wissenschaftlicher Dokumente im Index der *Bielefeld Academic Search Engine* (BASE) nach DDC
- Ziele:
  - DDC-Browsinginterface für BASE
  - BASE als Content-Provider für virtuelle Fachportale

Workflow der automatischen Anreicherung



Trainingsmaterial

- Nutzung von Metadaten (Abstracts) aus der BASE-Datenbasis als Trainingsdokumente
- DDC-Sachgruppen in deutschen Repositorien verbreitet (z.B. Bedingung für das DINI-Zertifikat)
- Derzeit sind rund 500.000 intellektuell nach DDC-Sachgruppen klassifizierte Dokumente in BASE indexiert

Konstruktion des Klassifikators

- Vorfilterung der Metadatensätze:
  - Enthält DDC-Sachgruppe
  - Mindestlänge des Abstracts von 500 Zeichen
  - Abstract in Deutsch oder Englisch
- Konstruktion eines „balancierten“ Trainingskorpus (gleiche Anzahl von Dokumenten pro Klasse)
- Stoppworteliminierung und Merkmalsextraktion
- Lernen des Modells (zur Zeit SVMs)

Produktivbetrieb: BASE

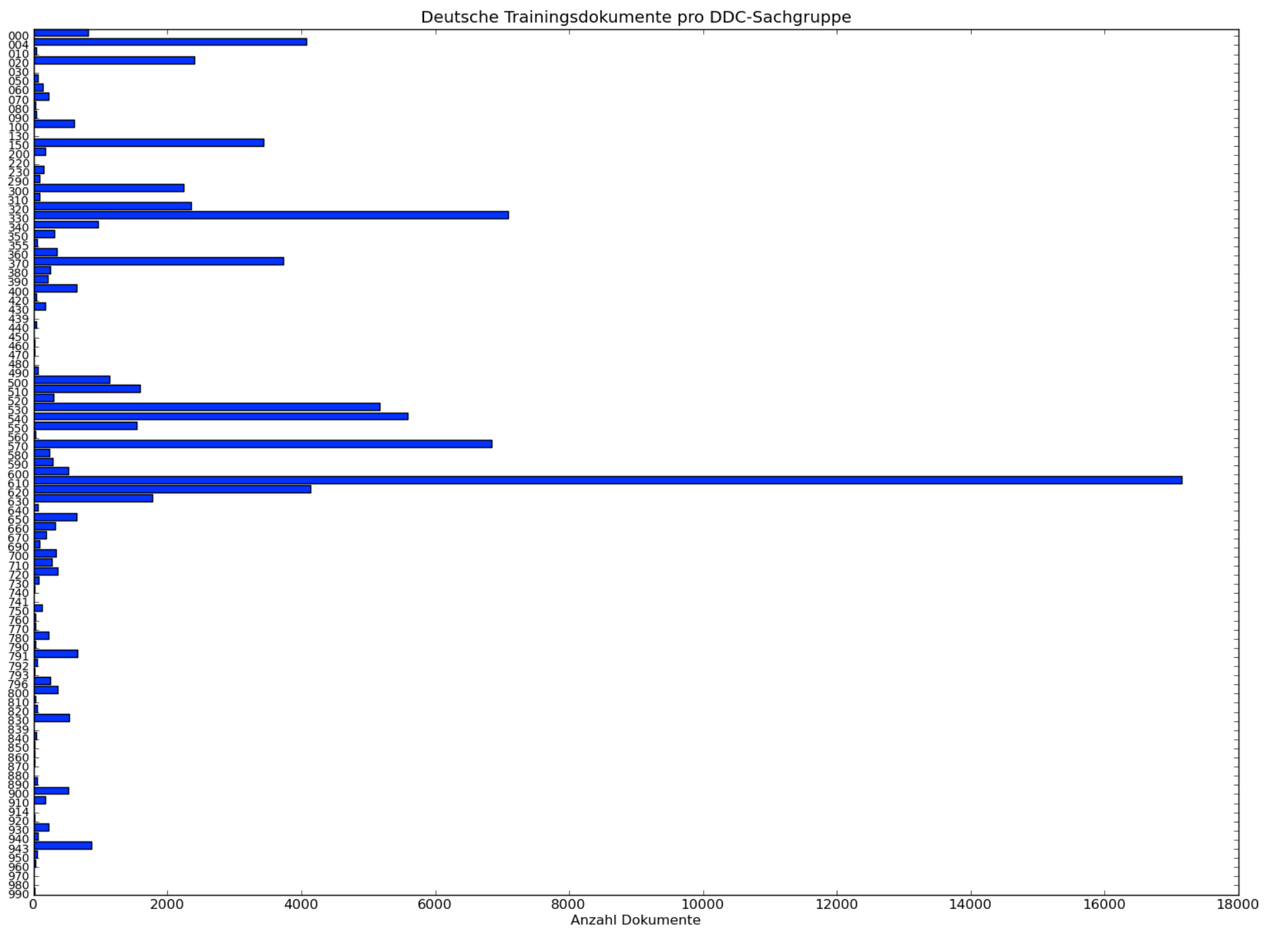
- Schrittweise Klassifizierung der BASE-Datenbasis
- Aktuell ~ 3 Mio. Dokumente nach DDC klassifiziert, davon ~ 2,5 Mio. automatisch
- DDC-Browsinginterface wurde freigeschaltet

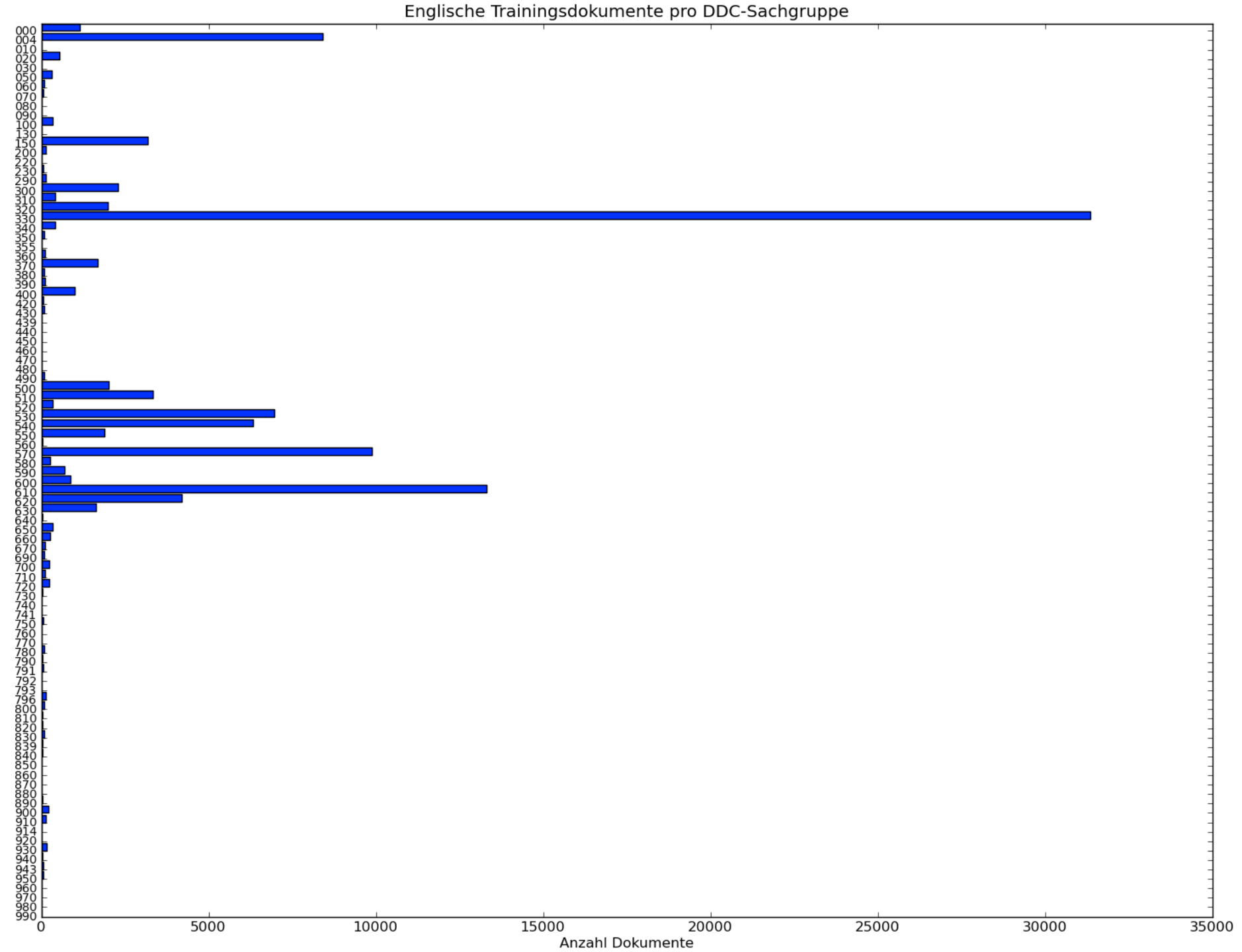
Produktivbetrieb: Nachnutzung

- Nachnutzung des Klassifikators
  - Möglich über offenen Webservice
  - Nachnutzung durch DFG-Projekt OA-Netzwerk wird evaluiert
- Pilotpartner für fachbezogenen Dokumentlieferungen: EconBiz.de (ZBW Kiel)

Probleme: Fehlende Trainingsdokumente

- Klassen unterschiedlich gut durch Beispieldokumente abgedeckt
- Für etliche Klassen existiert wenig bis kein Beispielmaterial
- Maschinelle Lernmodelle (in der Basiskonfiguration) verlangen gleichmäßige Klassenstärken


Deutsche Trainingsdokumente pro DDC-Sachgruppe
000
004
010
020
030
050
060
070
080
090
100
130
150
200
220
230
290
300
310
320
330
340
350
355
360
370
380
390
400
420
430
439
440
450
460
470
480
490
500
510
520
530
540
550
560
570
580
590
600
610
620
630
640
650
660
670
690
700
710
720
730
740
741
750
760
770
780
790
791
792
793
796
800
810
820
830
839
840
850
860
870
880
890
900
910
914
920
930
940
943
950
960
970
980
990
0
2000
4000
6000
8000
10000
12000
14000
16000
18000
Anzahl Dokumente


Englische Trainingsdokumente pro DDC-Sachgruppe
000
004
010
020
030
050
060
070
080
090
100
130
150
200
220
230
290
300
310
320
330
340
350
355
360
370
380
390
400
420
430
439
440
450
460
470
480
490
500
510
520
530
540
550
560
570
580
590
600
610
620
630
640
650
660
670
690
700
710
720
730
740
741
750
760
770
780
790
791
792
793
796
800
810
820
830
839
840
850
860
870
880
890
900
910
914
920
930
940
943
950
960
970
980
990
0
5000
10000
15000
20000
25000
30000
35000
Anzahl Dokumente

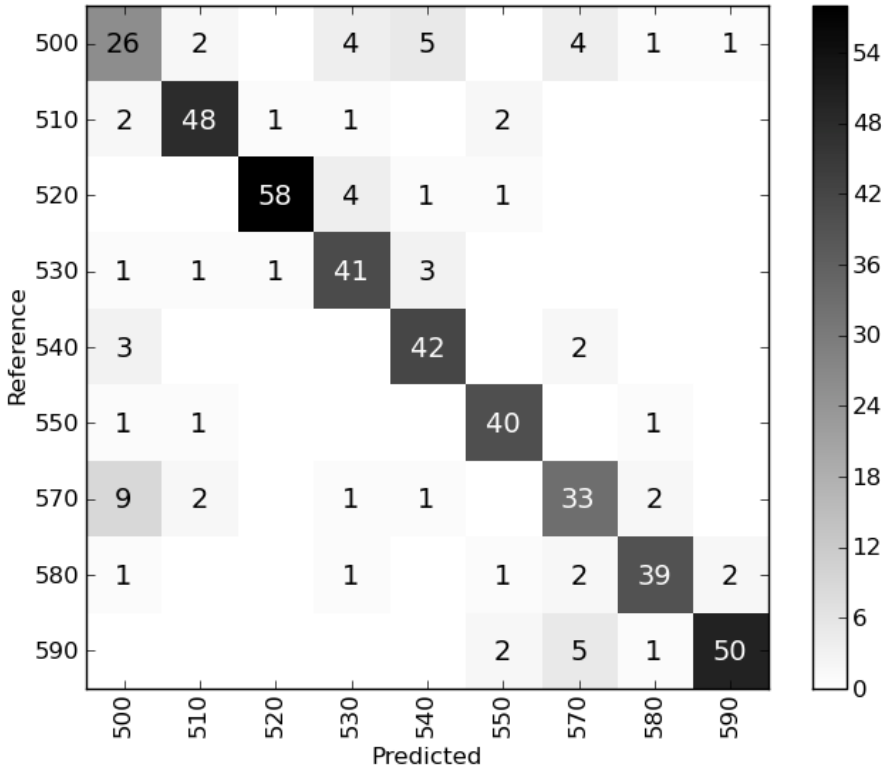
Probleme: Heterogene Dokumente

Dokumentlängen in Wörtern:

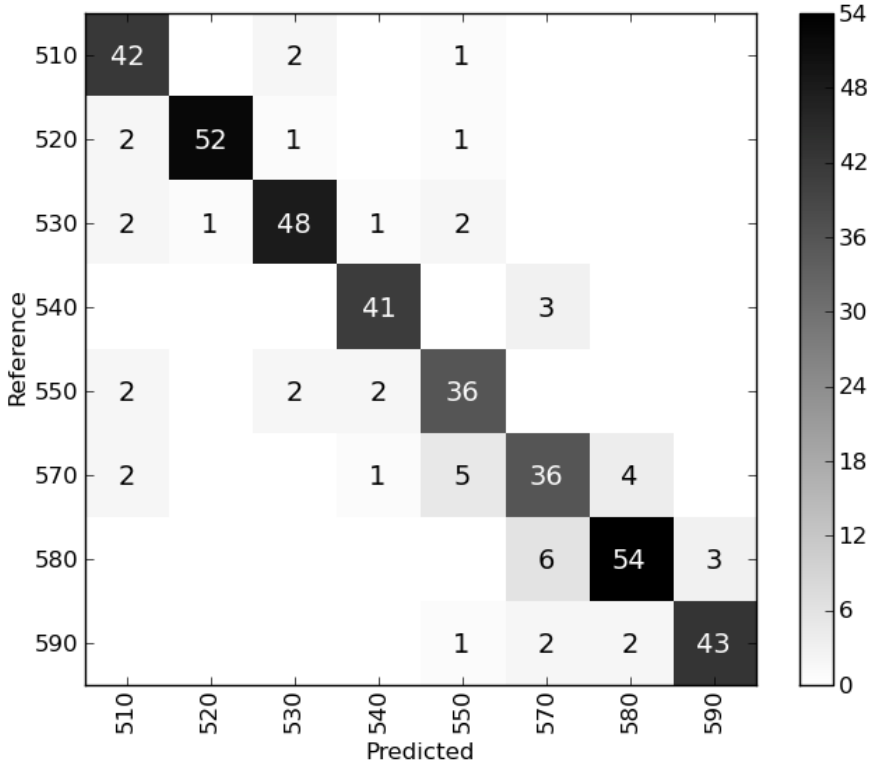
Englische Abstracts	Deutsche Abstracts
Minimale Länge: 59	Minimale Länge: 40
Maximale Länge: 10979	Maximale Länge: 8493
Durchschnittliche Länge: 271.34	Durchschnittliche Länge: 279.49
Standardabweichung: 204.34	Standardabweichung: 200.72

Probleme: Zu allgemeine Klassifizierung der Trainingsdaten in die X00-Klassen

F-Wert mit Klasse „500“: 0.85



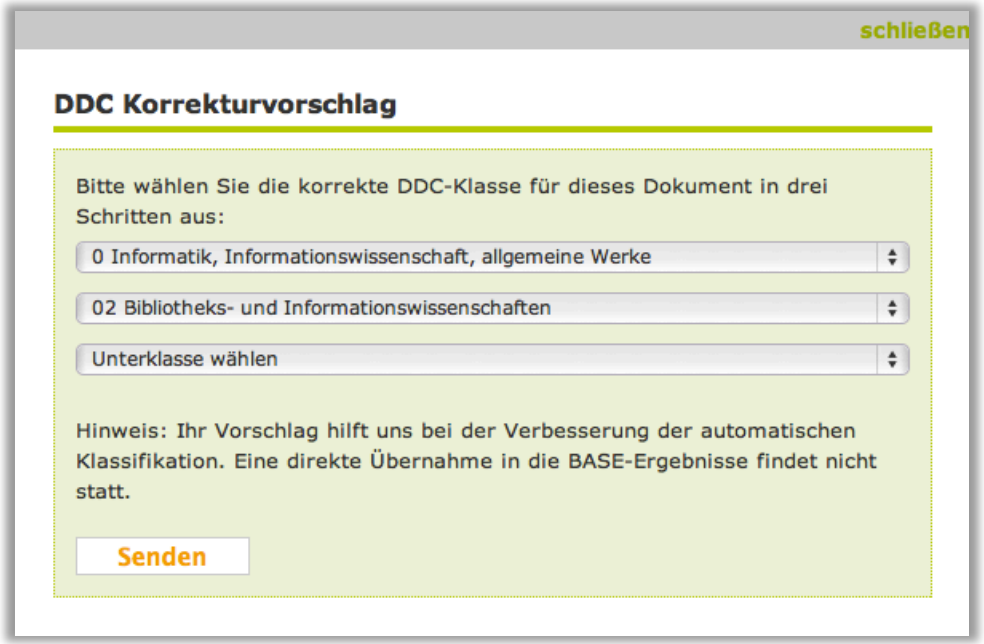
F-Wert ohne Klasse „500“: 0.88



Lösungsansätze: Vergrößerung des Trainingskorpus

- Die Zahl der Trainingsdokumente steigt durch die regelmäßige Aktualisierung der BASE-Datenbasis aus den Repositorien
- Regelmäßiges Neu-Training des Klassifikators mit der aktuellen Datenbasis

Lösungsansätze: Korpusvergrößerung durch Nutzerfeedback



- Wird jedoch noch wenig genutzt

Fazit

- Automatische Dokumentenklassifikation im Bibliotheksbereich erstmalig praxistauglich
- Verfahren haben dennoch noch Schwächen, an Verbesserung wird gearbeitet
- Qualität der Beispielkorpora entscheidend für den Erfolg

Vielen Dank für Ihre Aufmerksamkeit!

Dr. Monika Lösse
Deutsche Nationalbibliothek
Deutscher Platz 1
D-04103 Leipzig
+49-341-2271-582
M.Loesse@dnb.de

Mathias Lösch
Universitätsbibliothek Bielefeld
Universitätsstr. 25
D-33615 Bielefeld
+49 521-106-2546
Mathias.Loesch@uni-bielefeld.de

Universität Bielefeld
INFORMATION.plus! Universitätsbibliothek

1912 100 JAHRE 2012 DEUTSCHE NATIONAL BIBLIOTHEK