Implementierung einer automatischen DDC-Klassifikation für die Suchmaschine BASE auf Basis von Annif

Vorstellung der gleichnamigen Masterarbeit

Christoph Broschinski

Hintergrund: Klassifikation

- Bestandteil der Sacherschließung
- Thematische Einordnung von Publikationen in ein (hierarchisches) Klassifikationssystem
- Ziele: Unterstützung der Recherche, Schaffung einer Aufstellungssystematik
- Intellektuelle, üblicherweise von Menschen durchgeführte Aufgabe
- In Bielefeld: Erschließung von (Print-)Neuerwerbungen nach der eigenen Haussystematik durch die Fachreferate

Hintergrund: DDC

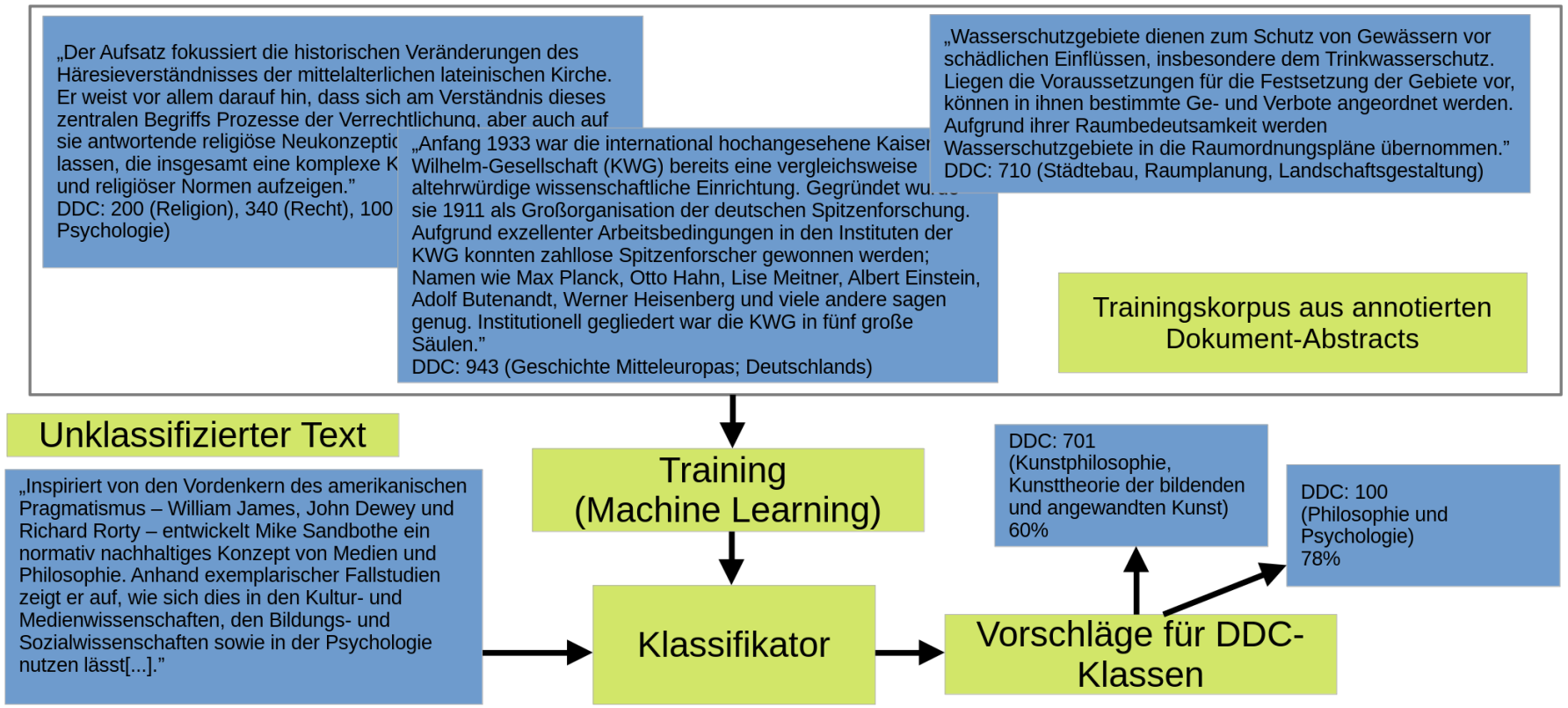
- DDC: Dewey Decimal Classification
- Entwickelt im 19. Jahrhundert vom US-amerikanischen Bibliothekar Melvil Dewey
- Dezimal: Jede Klasse besteht aus (maximal) zehn Unterklassen, Hierarchietiefe prinzipiell unbegrenzt. Beispiel:
-
    - *600 Technik, Technologie*
        - *660 Chemische Verfahrenstechnik*
            - *662 Explosivstoffe, Brennstoffe und verwandte Produkte*
                - *662.1 Feuerwerk (Pyrotechnik)*
- Internationaler Anspruch, weltweit verbreitet - aber weniger im deutschsprachigen Raum
- Wichtige Ausnahme: Erschließung von Publikationsreihen durch die DNB seit 2016 ("DDC-Sachgruppen")
- In BASE: DDC-Browsing

Hintergrund: Automatische Erschließung

- Grundidee: Klassifikation wird maschinell durchgeführt
- Ansätze: Halbautomatisch (Computer macht Vorschläge) oder vollautomatisch (Klassenzuweisung ohne menschliche Eingriffe)
- Methoden des (überwachten) maschinellen Lernens
- Eingesetzt etwa von der DNB: Seit 2010 zunehmende automatische Klassifikation/Verschlagwortung (vor allem für elektronische Publikationen)

- Klassifikationssystem **baseclf**: Im Rahmen des DFG-Projekts "Automatische Anreicherung von OAI-Metadaten" ab 2009 für BASE implementiert
- Seit 2013 im Regelbetrieb
- Größter Teil der DDC-Daten in BASE ist heute automatisch erschlossen

Maschinelle Klassifikation: Grundprinzip



Gewinnung von Trainingsdaten aus BASE

- Feld **description**: Enthält mögliche Trainingstexte (Abstracts/Beschreibungen)
- Feld **classcode**: Enthält normalisierte DDC-Klassen (im Dokument selbst enthalten)
- Feld **autoclasscode**: Durch baseclf automatisch zugewiesene DDC-Klassen

```
<record>
  <header xmlns="http://www.openarchives.org/OAI/2.0/">
    <identifier>ftbastrassen:oai:opus4-bast:2</identifier>
    <datestamp>2022-05-23T14:04:48Z</datestamp>
  </header>
  <metadata xmlns="http://www.openarchives.org/OAI/2.0/" xmlns:base_dc="http://oai.base-search.net/base_dc/" xmlns:xsi=
  "http://www.w3.org/2001/XMLSchema-instance" xmlns:dc="http://purl.org/dc/elements/1.1/">
    <base_dc:dc xsi:schemaLocation="http://oai.base-search.net/base_dc/ http://oai.base-search.net/base_dc/base_dc.xsd">
      <base_dc:global_id>ftbastrassen:oai:opus4-bast:2</base_dc:global_id>
      <base_dc:continent>ceu</base_dc:continent>
      <base_dc:country>de</base_dc:country>
      <base_dc:collection>ftbastrassen</base_dc:collection>
      <base_dc:collname>German Federal Highway Research Institute (BASt): Electronic BASt Archive (ELBA)</base_dc:collname>
      <dc:title>EEVC approach to the improvement of crash compatibility between passenger cars</dc:title>
      <dc:creator>Faerber, Eberhard</dc:creator>
      <dc:subject>Kompatibilität</dc:subject>
      [...]
      <dc:subject>ddc:600</dc:subject>
      <dc:description>The objective of European Enhanced Vehicle-safety Committee (EEVC) Working Group (WG) 15 Car Crash
      Compatibility and Frontal Impact is to develop a test procedure(s) with associated performance criteria and limits for
      car frontal impact compatibility. This work should lead [...]</dc:description>
      <dc:date>2009-04-06</dc:date>
      <base_dc:year>2009</base_dc:year>
      <dc:type>conferenceobject</dc:type>
      <dc:type>doc-type:conferenceobject</dc:type>
      <base_dc:typenorm>13</base_dc:typenorm>
      <dc:format>application/pdf</dc:format>
      <dc:identifier>https://bast.opus.hbz-nrw.de/frontdoor/index/index/docId/2</dc:identifier>
      [...]
      <base_dc:link>https://bast.opus.hbz-nrw.de/frontdoor/index/index/docId/2</base_dc:link>
      <dc:language>eng</dc:language>
      <dc:relation>https://bast.opus.hbz-nrw.de/frontdoor/index/index/docId/2</dc:relation>
      [...]
      <dc:rights>info:eu-repo/semantics/openAccess</dc:rights>
      <base_dc:autoclasscode type="ddc">380</base_dc:autoclasscode>
      <base_dc:classcode type="ddc">600</base_dc:classcode>
      <base_dc:oa>1</base_dc:oa>
      <base_dc:lang>eng</base_dc:lang>
    </base_dc:dc>
  </metadata>
</record>
```

Grundidee der Masterarbeit

- baseclf wird nicht mehr aktiv gepflegt / involvierte Personen nicht mehr an der UB tätig
- Codebasis völlig veraltet
- "Black Box": Performance des Systems seit Jahren völlig unklar!

Idee: Erstellung eines Nachfolgesystems auf Basis von **Annif**

annif

- Entwickelt an der finnischen Nationalbibliothek
- Toolkit zur automatischen Inhaltserschließung
- Unter anderem eingesetzt von der DNB zur automatischen Inhaltserschließung (EMa)

Aufbau der Arbeit

Ausgangslage: Snapshot der BASE-Daten ("Dump"), bestehend aus ~ 220 Mio. Dokumenten

Empirischer Teil: Analyse / Exploration des Dumps (Data Science) im Hinblick auf Nutzbarkeit für einen Trainingskorpus.

Wichtige Fragestellungen:

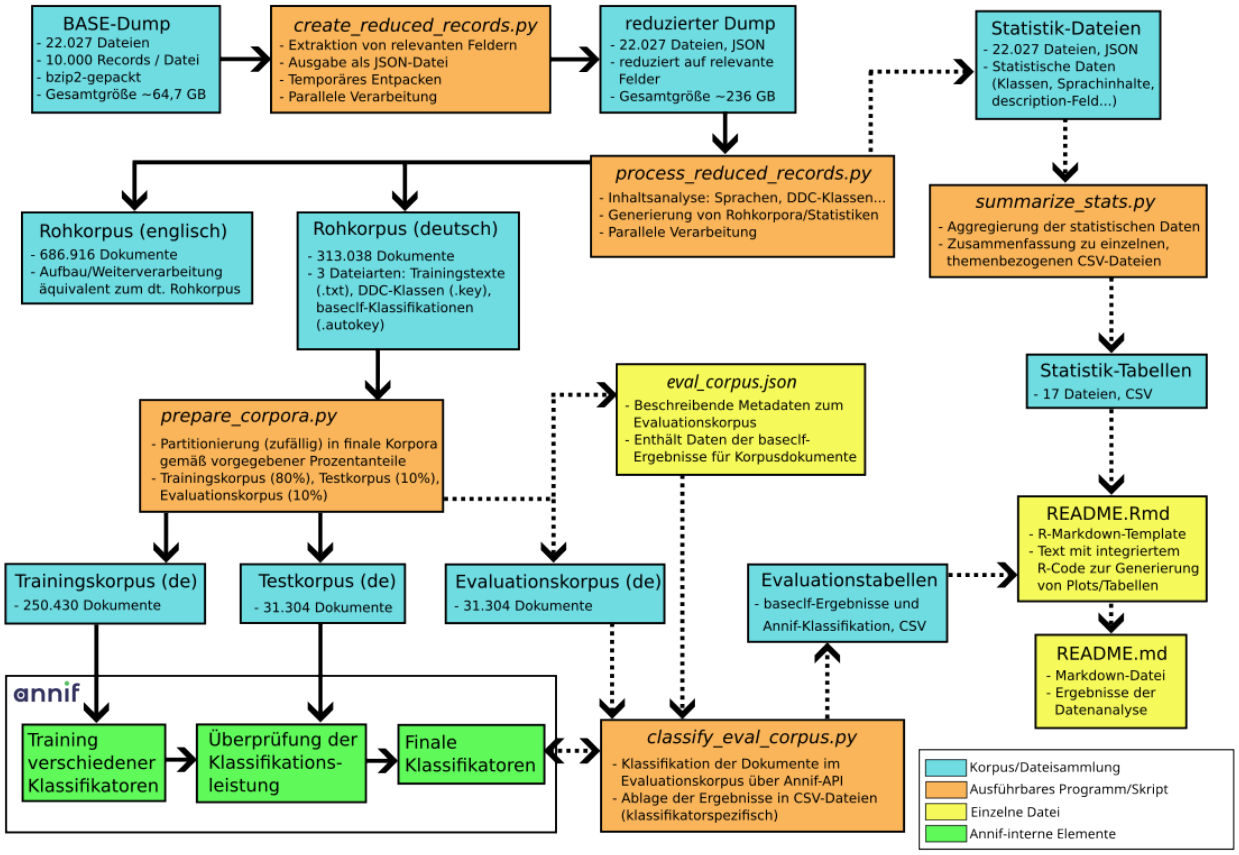
- Wieviele Dokumente sind mit Abstracts versehen? Wie verteilt sich die Länge der Abstracts?
- In welchen Sprachen sind die Abstracts verfasst? (automatische Spracherkennung)
- Wieviele Dokumente sind mit DDC-Informationen versehen? Wie verteilen sich die DDC-Klassen über den Gesamtbestand?
- Wie gut funktioniert die Klassifikationen des bestehenden Systems baseclf?

Aufbau der Arbeit (2)

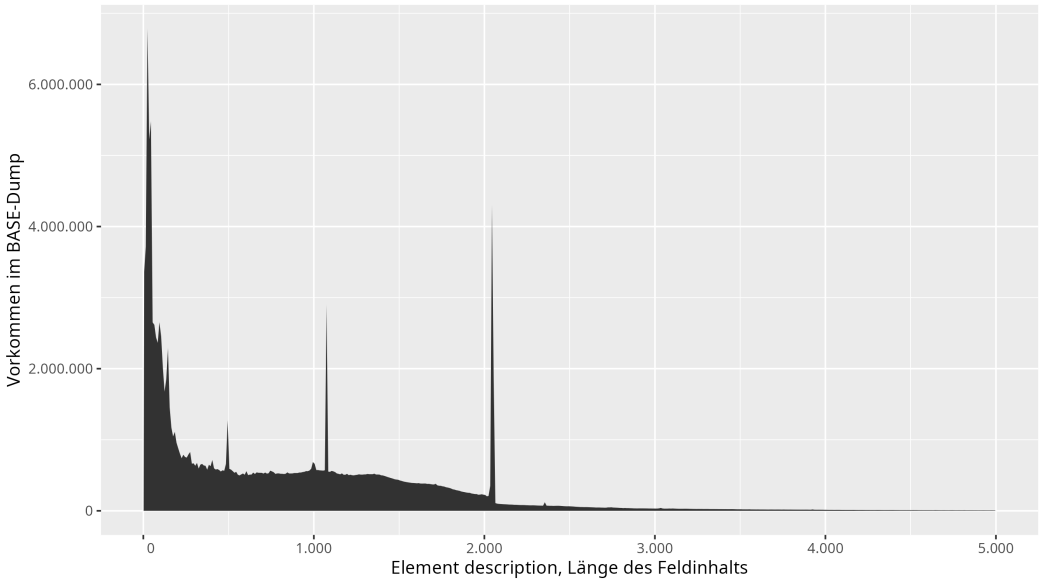
Projektbezogener Teil: Erstellung der Korpora, Training und Auswahl eines Annif-Klassifikators

- Erstellung von zwei neuen Trainingskorpora (deutsch/englisch) aus dem BASE-Dump auf Basis der empirischen Erkenntnisse
- Training unterschiedlicher Annif-Backends, Performance-Vergleich, Auswahl des besten Kandidaten
- Vergleich der Ergebnisse des neuen Annif-Klassifikators mit dem alten baseclf auf einem dezidierten Evaluationskorpus
- Für alle Schritte (auch für den empirischen Teil) muss eine eigene Software-Toolchain programmiert werden

Toolchain


BASE-Dump
- 22.027 Dateien
- 10.000 Records / Datei
- bzip2-gepackt
- Gesamtgröße ~64,7 GB
create_reduced_records.py
- Extraktion von relevanten Feldern
- Ausgabe als JSON-Datei
- Temporäres Entpacken
- Parallele Verarbeitung
reduzierter Dump
- 22.027 Dateien, JSON
- reduziert auf relevante Felder
- Gesamtgröße ~236 GB
Statistik-Dateien
- 22.027 Dateien, JSON
- Statistische Daten (Klassen, Sprachinhalte, description-Feld...)
process_reduced_records.py
- Inhaltsanalyse: Sprachen, DDC-Klassen...
- Generierung von Rohkorpora/Statistiken
- Parallele Verarbeitung
summarize_stats.py
- Aggregierung der statistischen Daten
- Zusammenfassung zu einzelnen, themenbezogenen CSV-Dateien
Rohkorpus (englisch)
- 686.916 Dokumente
- Aufbau/Weiterverarbeitung äquivalent zum dt. Rohkorpus
Rohkorpus (deutsch)
- 313.038 Dokumente
- 3 Dateiarten: Trainingstexte (.txt), DDC-Klassen (.key), baseclf-Klassifikationen (.autokey)
Statistik-Tabellen
- 17 Dateien, CSV
eval_corpus.json
- Beschreibende Metadaten zum Evaluationskorpus
- Enthält Daten der baseclf-Ergebnisse für Korpusdokumente
prepare_corpora.py
- Partitionierung (zufällig) in finale Korpora gemäß vorgegebener Prozentanteile
- Trainingskorpus (80%), Testkorpus (10%), Evaluationskorpus (10%)
README.Rmd
- R-Markdown-Template
- Text mit integriertem R-Code zur Generierung von Plots/Tabellen
Trainingskorpus (de)
- 250.430 Dokumente
Testkorpus (de)
- 31.304 Dokumente
Evaluationskorpus (de)
- 31.304 Dokumente
Evaluationstabellen
- baseclf-Ergebnisse und Annif-Klassifikation, CSV
README.md
- Markdown-Datei
- Ergebnisse der Datenanalyse
annif
Training verschiedener Klassifikatoren
Überprüfung der Klassifikations-leistung
Finale Klassifikatoren
classify_eval_corpus.py
- Klassifikation der Dokumente im Evaluationskorpus über Annif-API
- Ablage der Ergebnisse in CSV-Dateien (klassifikatorspezifisch)
Korpus/Dateisammlung
Ausführbares Programm/Skript
Einzelne Datei
Annif-interne Elemente

Datenanalyse 1: Länge des Description-Feldes



Längenbereich	Anzahl Records	Anteil (%)
1 - 100	37.290.995	22.87
101 - 200	16.019.709	9.83
201 - 300	7.668.779	4.70
301 - 400	6.336.854	3.89
401 - 500	6.672.317	4.09
501 - 600	5.360.613	3.29
601 - 700	5.278.667	3.24
701 - 800	5.347.301	3.28
801 - 900	5.260.935	3.23
901 - 1000	5.664.255	3.47

- Minimale Feldlänge: Abwägung zwischen Korpusgröße und Datenqualität
- Letzlich festgesetzt auf 100 Zeichen -> Verlust von ~ 37 Mio. Records

Datenanalyse 2: Sprachinhalte des Description-Feldes

- Die Inhalte der Abstracts müssen sprachspezifisch sein (deutsch/englisch), um als Trainingsdaten nutzbar zu sein.
- Sprachauswertung erfolgt automatisiert mithilfe des Python-Moduls polyglot

Kategorie	Anzahl
Erkennung fehlgeschlagen	21.682
Status "unreliable"	1.560.491
Konfidenz zu niedrig	17.159.180
Sprache erkannt	106.997.582

Sprache	Anzahl	Anteil (in %)
English	87.695.083	81.96
Spanish	3.591.911	3.36
Other Languages	3.578.950	3.34
French	3.057.203	2.86
German	2.496.552	2.33
Portuguese	1.925.485	1.80
Indonesian	1.839.472	1.72
Polish	1.243.505	1.16
Italian	941.502	0.88
Russian	627.919	0.59

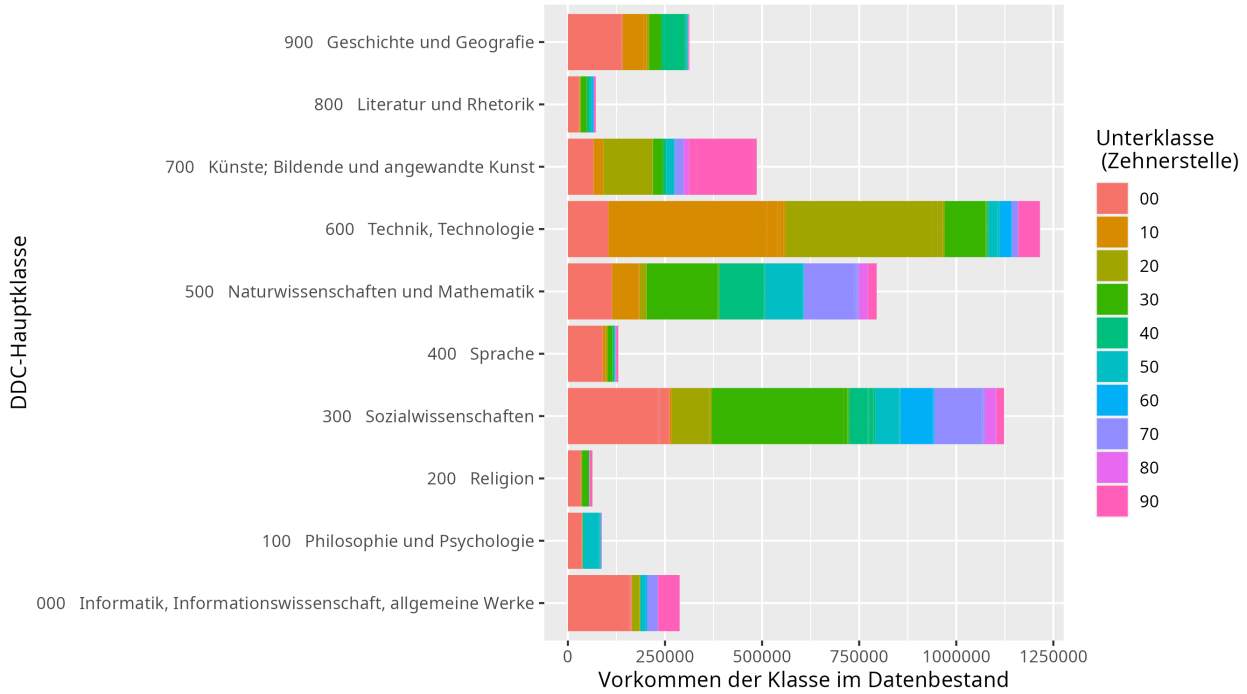
Datenanalyse 3: DDC-Klassen

DDC-Klassen pro Record	Anzahl Records
1	2.612.018
2	518.119
3	121.816
4	39.283
5	4.993
6	2.688
7	1.575
8	988
9	552
10	409
11	275
12	168
13	110
14	75
15	64

DDC-Klassen pro Record	Anzahl Records
16	32
17	21
18	14
19	23
20	8
21	12
22	2
23	3
24	1
25	1
26	1
30	1
38	1
46	1
Records mit DDC-Information aus base_dc:classcode	3.303.254

Erkenntnis: DDC-Informationen sind im BASE-Dump sehr selten, nur etwa 1,5% aller Records sind entsprechend klassifiziert

Datenanalyse 3: DDC-Klassen (Inhalte)



Erkenntnis: DDC-Klassen sind sehr ungleichmäßig (schief) verteilt

Processing / Korpuserstellung

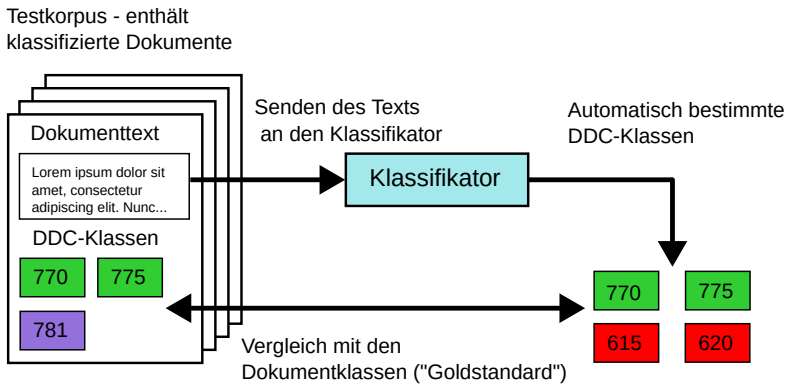
- Die Records im BASE-Dump durchlaufen während der Korpuserstellung ein Processing und werden ggf. ausgefiltert:

Testergebnis	Anzahl Records
Inhalt von `description` zu kurz oder nicht vorhanden	94.990.609
Keine DDC-Informationen im Record	124.375.833
Spracherkennung fehlgeschlagen	153
Spracherkenner meldet Status `unreliable`	10.209
Konfidenz der Spracherkennung zu niedrig	289.172
Andere erkannte Sprache als Deutsch oder Englisch	63.614
Alle Tests bestanden, Aufnahme in Korpus	999.954

- Zuletzt erfolgt noch eine Aufteilung gemäß der Sprachen:

Sprache	Größe des Rohkorpus
Deutsch	313.038
Englisch	686.916

Evaluation eines Klassifikators



3 mögliche Resultate:

- Klasse korrekt vorhergesagt: True Positive (TP)
- Klasse falsch vorhergesagt, die nicht im Dokument enthalten ist: False Positive (FP)
- Klasse falsch nicht vorhergesagt, die im Dokument enthalten ist: False Negative (FN)

Abgeleitete Metriken:

1. Precision: TP / (TP + FP) = 2 / (2 + 2) = 0,5 (Maß für **Genauigkeit**)
2. Recall: TP / (TP + FN) = 2 / (2 + 1) = 0,67 (Maß für **Vollständigkeit**)
3. F_1-Wert: (2*TP) / (2*TP + FP + FN) = 4 / (4 + 2 + 1) = 0,57 (**Harmonisches Mittel** aus Precision und Recall)

Trainierte Klassifikatoren in Annif

- Mithilfe der erstellten Trainingskorpora wurden in Annif 4 unterschiedliche Klassifikatoren trainiert (jeweils einer pro Sprache)
- Die Klassifikatoren beruhen jeweils auf unterschiedlichen ML-Algorithmen:
-
    - *TF-IDF (textstatistisch)*
    - *Omikuji (Entscheidungsbaum-basiert)*
    - *fastText (Neuronales Netz)*
    - *NN-Ensemble (Zusammenschluss der 3 erstgenannten)*
- Anschließend wurden alle Klassifikatoren mithilfe des Testkorpus evaluiert und auf einen möglichst hohen F1-Wert optimiert.
- Ergebnisse für die englischsprachigen Klassifikatoren:

	en-tfidf	en-omikuji	en-fasttext	en-nn_ensemble
Optimaler Threshold	0	0,1	0,15	0,2
Pre (doc avg)	0,1943	0,646	0,603	0,6474
Rec (doc avg)	0,3585	0,7531	0,7056	0,7417
F1 (doc avg)	0,2483	0,676	0,6305	0,6722
Pre (subj avg)	0,0288	0,088	0,0322	0,0765
Rec (subj avg)	0,0941	0,0562	0,0208	0,0371
F1 (subj avg)	0,0334	0,0622	0,0213	0,0427
Pre (weighted subj avg)	0,5218	0,5773	0,5096	0,5713
Rec (weighted subj avg)	0,342	0,7144	0,6607	0,7003
F1 (weighted subj avg)	0,3681	0,6309	0,5591	0,6101
Pre (microavg)	0,1943	0,5895	0,5412	0,5867
Rec (microavg)	0,342	0,7144	0,6607	0,7003
F1 (microavg)	0,2478	0,6459	0,595	0,6385
True positives	26697	55765	51577	54670
False positives	110687	38837	43716	38508
False negatives	51366	22298	26486	23393
Documents evaluated	68692	68692	68692	68692

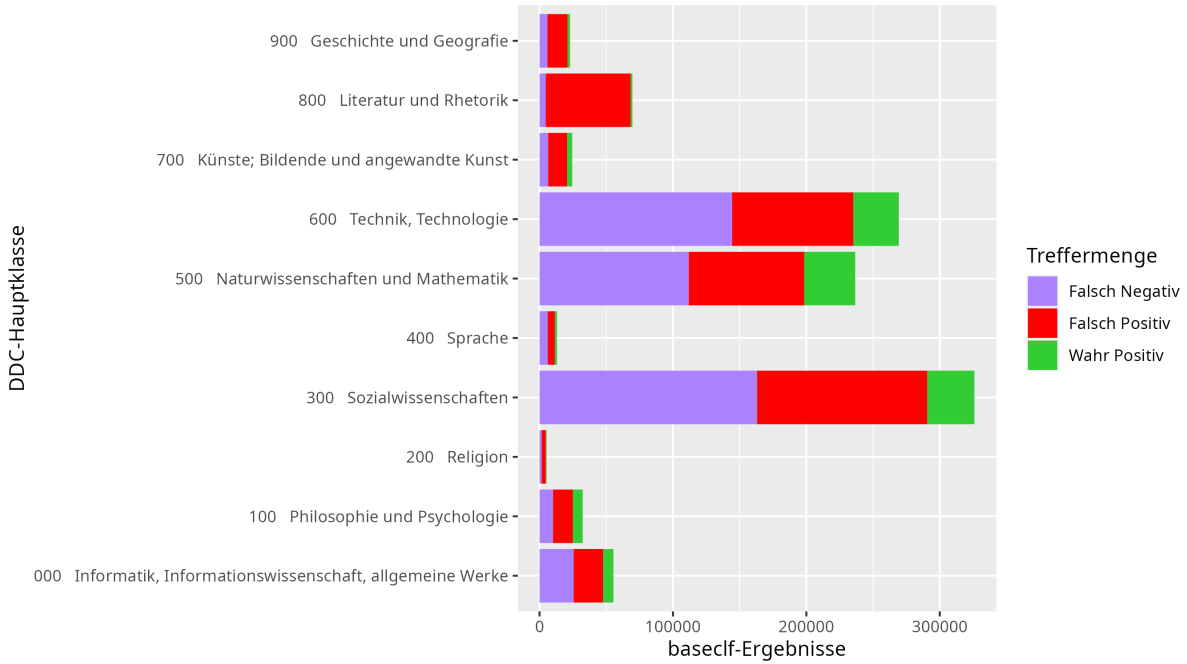
Vergleich mit baseclf

- Wichtige Frage: Wie schneiden die Annif-Klassifikatoren im Vergleich zum derzeit eingesetzten baseclf ab?
- Vergleich über Evaluationskorpus: Dokumente werden per Skript an Annif geschickt und die Ergebnisse verglichen
- baseclf-Vergleich erfolgt über Abgleich der Felder "classcode" und "autoclasscode"
- Ergebnisse:

	baseclf	en-tfidf	en-omikuji	en-fasttext	en-nn_ensemble
Korpusdokumente	68.692	68.692	68.692	68.692	68.692
Klassifiziert	35.827	68.692	67.547	67.883	67.768
Vollständig korrekt	6.226	258	32.478	29.514	33.343
Wahr-Positiv	8.532	26.751	55.829	51.596	54.850
Falsch-Positiv	30.026	110.633	38.996	43.909	38.265
Falsch-Negativ	69.615	51.396	22.318	26.551	23.297
Precision	0,22	0,19	0,59	0,54	0,59
Recall	0,11	0,34	0,71	0,66	0,70
F_1	0,15	0,25	0,65	0,59	0,64
Laufzeit (min)	–	19:00	16:11	5:06	36:49

Gesamtauswertung baseclf

- Für eine genauere Analyse von baseclf wurde dessen Performance auf dem gesamten Base-Dump analysiert
- Datengrundlage sind alle Records, die sowohl native DDC-Informationen haben und zugleich von baseclf klassifiziert wurden (classcode + autoclasscode)
- Dies trifft auf insgesamt 534.026 Records zu
- Ergebnisse:



Maß	Wert
Wahr Positiv	131.859
Falsch Positiv	443.148
Falsch Negativ	479.796
Precision (Microavg)	0,23
Recall (Microavg)	0,22
F1 (Microavg)	0,22

Ergebnisse

- Obwohl der Base-Datenbestand sehr umfangreich, ist nur ein sehr kleiner Bruchteil als Trainingsmenge für ein ML-Verfahren verwendbar
- Der baseclf-Klassifikator zeigt eine sehr schlechte Leistung und sollte ersetzt werden
- Als Ersatz bietet sich eine Annif-Installation mit einem Omikuji-Backend an
- Neu-Training in regelmäßigen Abständen ist empfehlenswert

Letzte Folie...

Vielen Dank für eure Aufmerksamkeit!