Predicting LCSH

Comparing the output of predicted Library of Congress subject headings for politics, medicine and physics

Christian Wartena Michael Franke-Maier

Hochschule Hannover, Abteilung Information und Kommunikation

Freie Universität Berlin, Universitätsibliothek

May 12, 2017

Introduction

B3KAT

- Library Union Catalogue of Bavaria, Berlin and Brandenburg
- Linked open data Representation `http://lod.b3kat.de`
- ca. 26 Mio bibliographic entities
- Freie Universität Berlin is shared cataloguing partner library

Library of congress Subject Headings (LCSH)

- Used since 1898 for cataloging materials at the LoC
- Linked open data Representation `http://id.loc.gov/authorities/subjects.html`
- Widely adopted in the anglophone world
- cross-referenced to GND and Rameau (MACS-Project)

Data

Source

- SPARQL-Endpoint of the B3-Catalogue (B3Kat)
- 4 Datasets
- Different from the one used in the 2016 TPDL Paper

Criteria

- Title and Abstract
- Abstract of at least 200 characters
- Abstract in English, according to metadata **and** language detection
- LCSH in metadata
- Restricted by subject area (see next slide)
- Query results might differ due to time-outs

4 Sets

Restrictions and Sizes

Politics	• DDC: 32 … • 1769 records
Medicine	• DDC: 61 … • 1427 records
Physics	• DDC: 53 …; RVK: U…; LCC: QC • 371 records
General	• Anything not in the previous sets • 20 618 records

LCSH in Text

Labels of LCSH

- LCSH have ids and labels
- Labels could occur in text
- LCSH are not classes!
  - Training a classifier for *each* LCSH is impossible

Types of Labels

- Preferred labels
- Labels including scope notes
- Labels of precombined headings containing "- -"
- Inverted labels

LCSH

Many highly specific headings

sh00000172 Halle 13 (Expo, International Exhibitions Bureau, 2000, Hannover, Germany)

sh2005002460 Brown versus Board of Education of Topeka

sh85120114 Septets (Piano, flute, zither, percussion)

Ambiguity

- Many LCSH with identical variants!
- Especially, after removing scope notes from the label

LCSH

Selection of LCSH

414 355	subject headings
162 569	pre-combined headings (like *Voyages and travel - Mythology*)
497 427	terms after removing pre-combined labels with dashes, subdivisions, inverted labels and Children's headings
572 697	terms after adding (non-ambiguous) singular forms
15 661	ambiguous terms after removing scope notes (like in *Taxis (Biology)* and *Taxis (Vehicles)*)

LCSH

Ambiguity

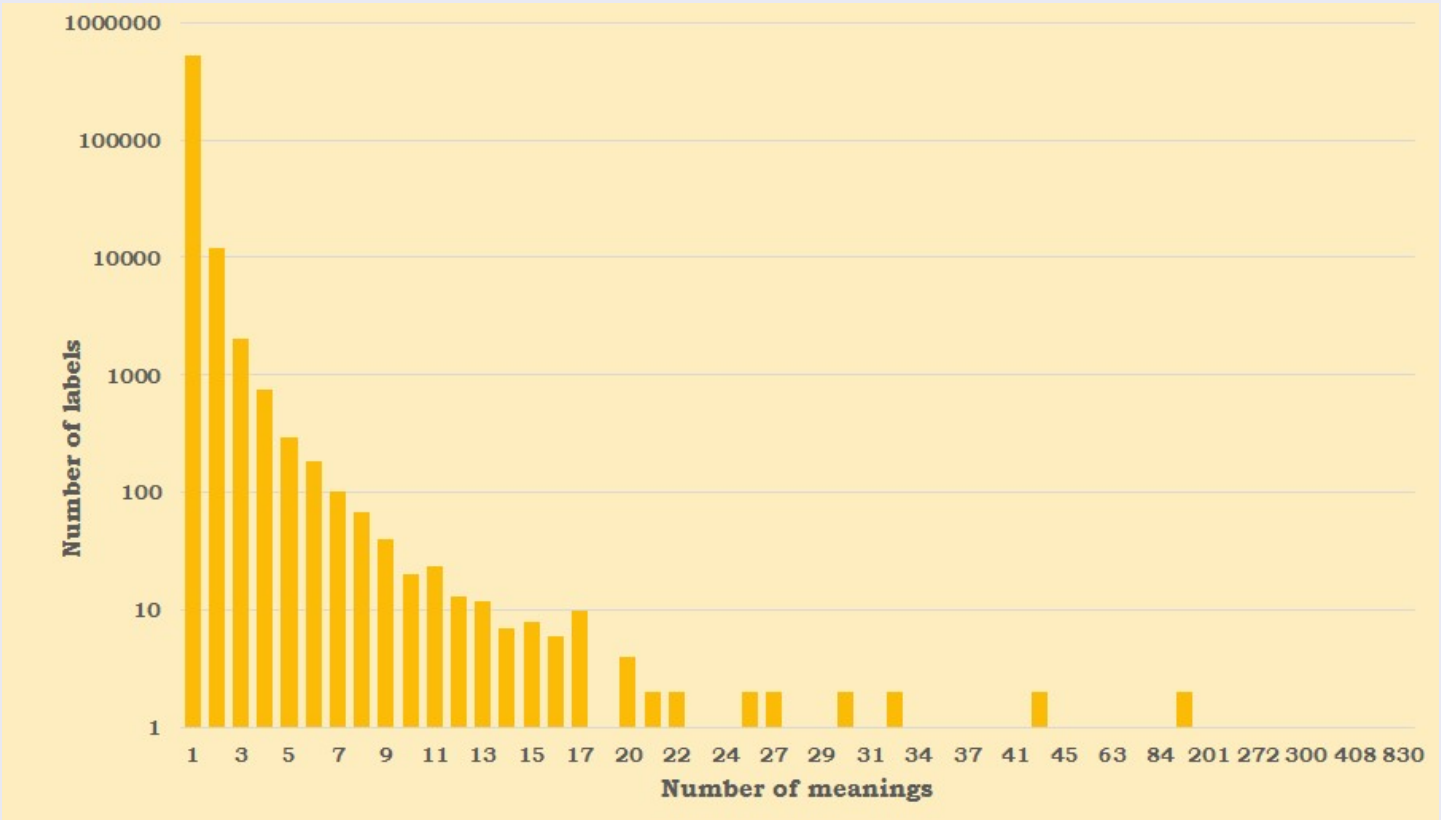
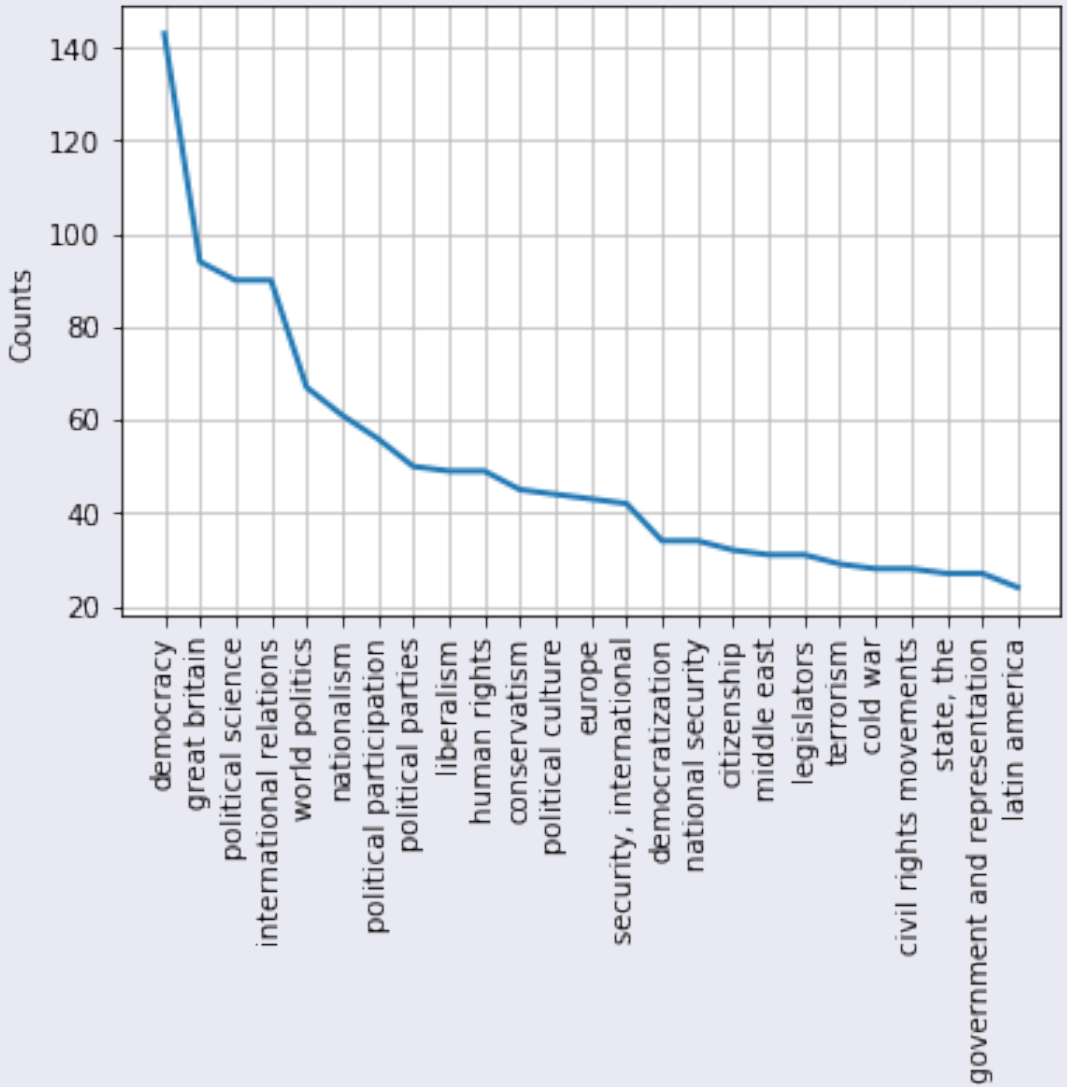


Figure: Number of labels for each degree of ambiguity.

Labels in the dataset

Politics



Label in the dataset

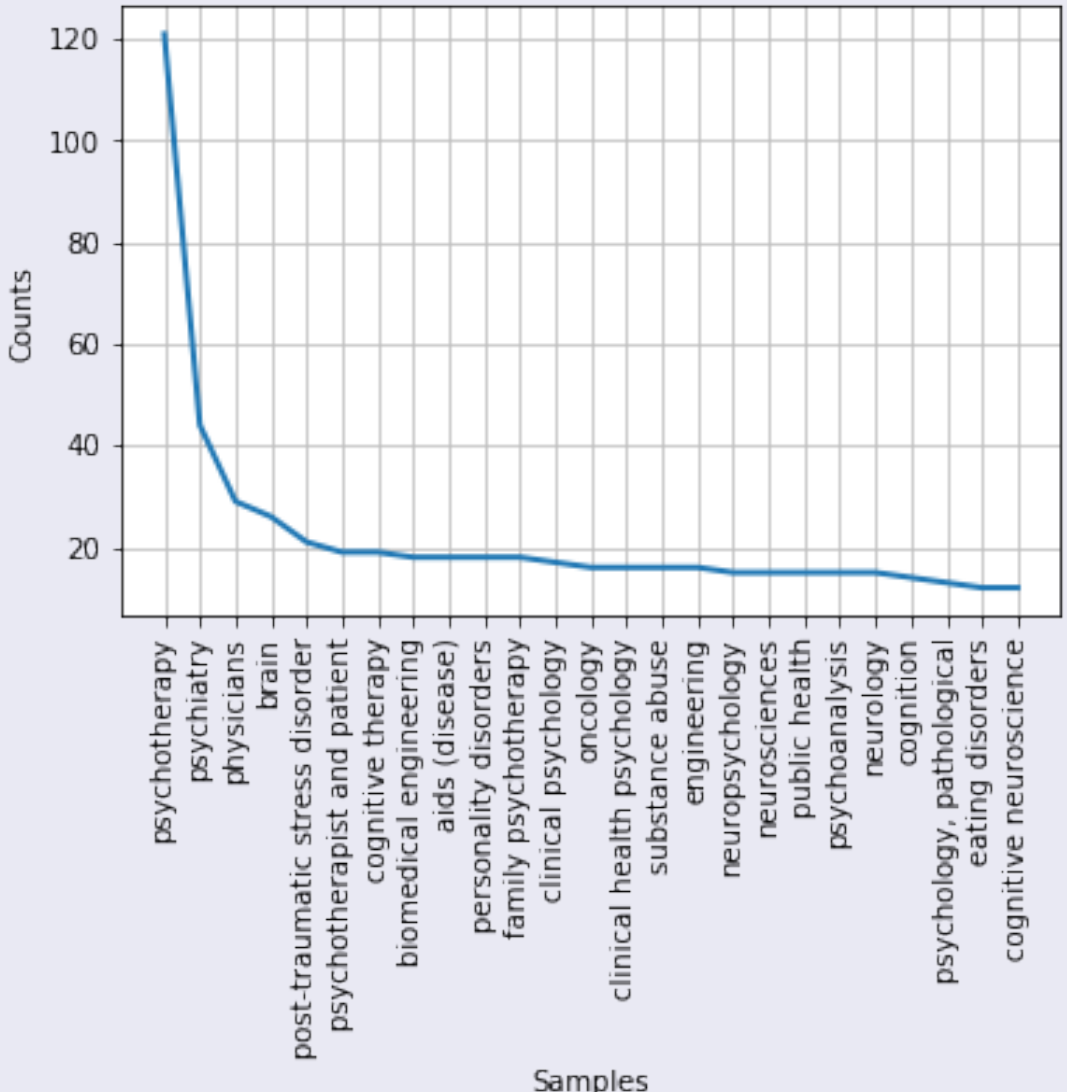
Politics

ID	Frequency	Concept
sh85036647	143 / 72	democracy
sh85056605	94 / 324	great britain
sh85104440	90 / 54	political science
sh85067435	90 / 49	international relations
sh85148216	67 / 80	world politics
sh85090150	61 / 55	nationalism
sh85104370	56 / 38	political participation
sh85104371	50 / 10	political parties
sh85076443	49 / 36	liberalism
sh85026379	49 / 62	human rights

- One geographic concept
- The rest: More or less of general nature

Labels in the dataset

Medicine



Label in the dataset

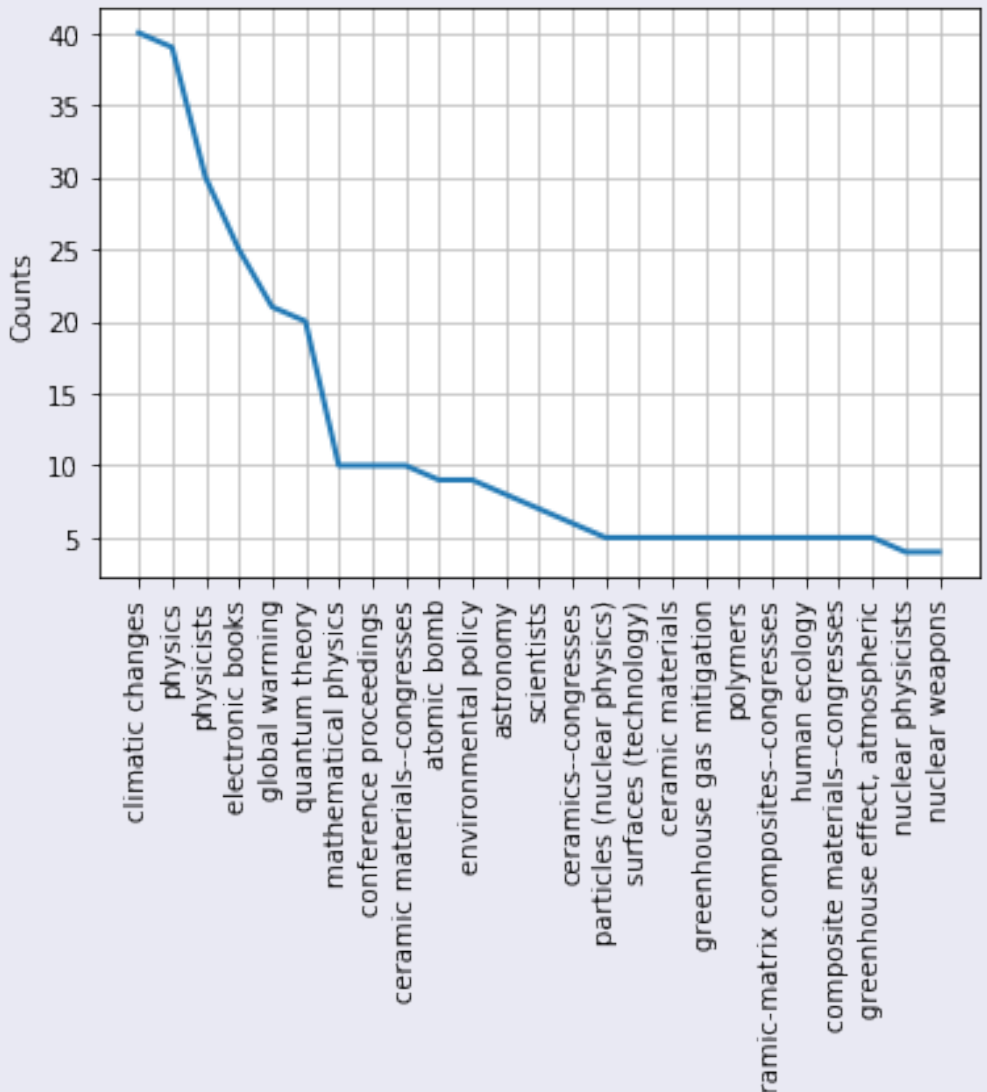
Medicine

ID	Frequency	Concept
sh85108516	121 / 20	psychotherapy
sh85108381	44 / 24	psychiatry
sh85101610	29 / 24	physicians
sh85016319	26 / 23	brain
sh85105424	21 / 10	post-traumatic stress disorder
sh85108509	19 / 4	psychotherapist and patient
sh85027756	19 / 7	cognitive therapy
sh85014237	18 / 3	biomedical engineering
sh85002541	18 / 26	aids (disease)

- Most frequent terms from Psychology and Psychiatry
- The rest: More or less of general nature

Labels in the dataset

Physics



Label in the dataset

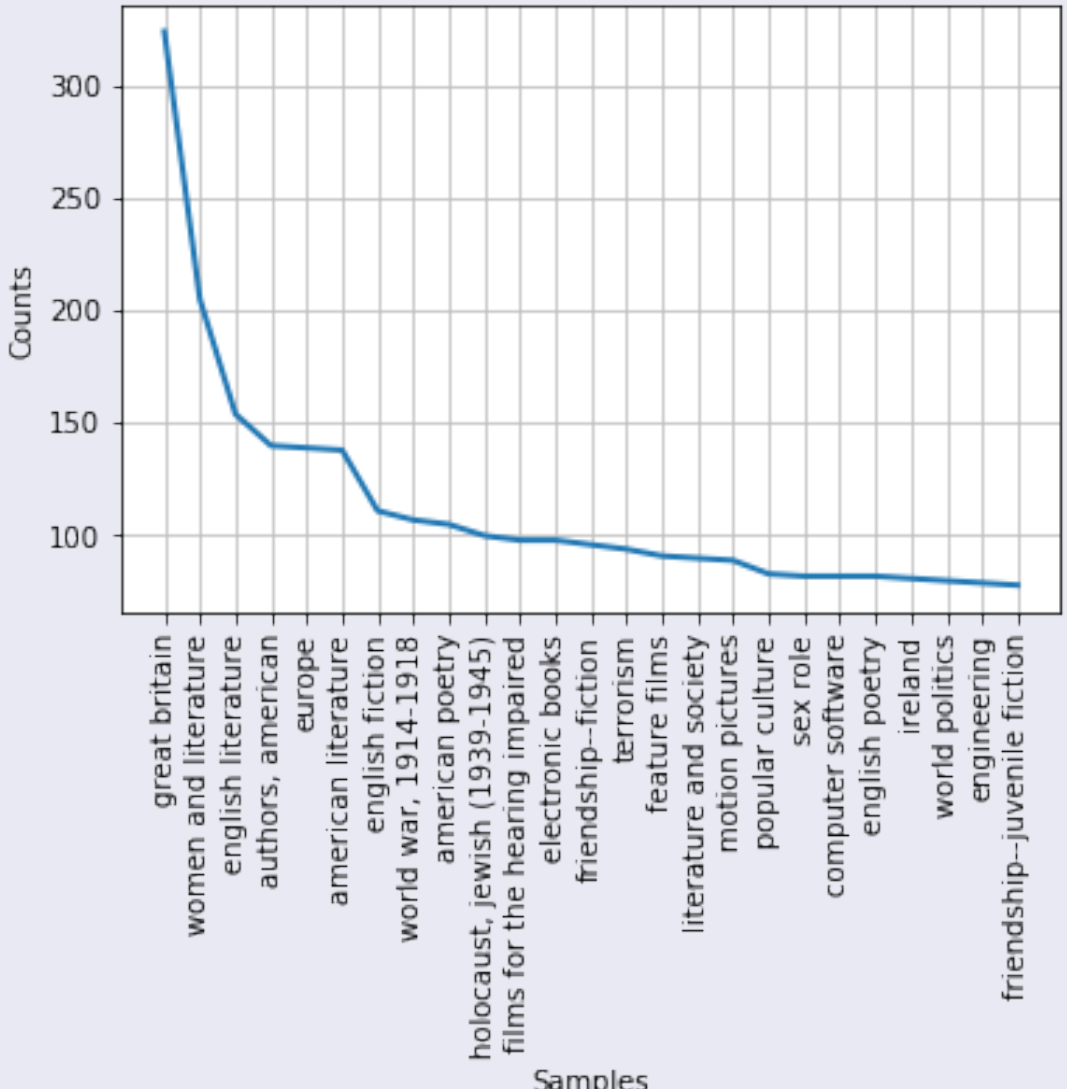
Physics

ID	Frequency	Concept
sh85027037	40 / 31	climatic changes
sh85101653	39 / 17	physics
sh85101651	30 / 5	physicists
sh93007047	25 / 98	electronic books
sh89000812	21 / 8	global warming
sh85109469	20 / 9	quantum theory
sh85030902	10 / 3	conference proceedings
sh2008100281	10 / 0	ceramic materials–Congresses
sh85082129	10 / 2	mathematical physics

- Three form headings, one precombined
- Two headings from environmental science
- The rest: More or less of general nature

Labels in the dataset

General



Matching labels

Results

	Prec	Recall	F1
Politics	0.078	0.38	0.12
Medicine	0.080	0.37	0.12
Physics	0.071	0.36	0.11
General	0.054	0.22	0.075

Training a classifier

Training

- We use the general set as training data
- We train classifiers for all 947 LCSH that occur > 10 times in training data (20 000 records!)
- Using mid-frequency words as features
- Logistic regression, one- vs. all classifiers
- We assign *n* most probable LCSH to each record

Training a classifier

Specific labels in data set *General*

- politics are well represented in data set *General*
- that's obvious, because LoC indexes for the US Congress
- the LoC focuses more upon a historical-political scope than on STEM
- this has also consequences for the availability of appropriate concepts for STEM

Training a classifier

Results for $n = 3$

	Prec	Recall	F1
Politics	0.20	0.30	0.22
Medicine	0.085	0.11	0.090
Physics	0.10	0.15	0.11

Example Politics : Method Extraction

The Irish constitutional tradition

- `http://lod.b3kat.de/page/title/BV009807040`
- existing LCSH: ireland, constitutional history
- …1782 to the present day and treats the **constitutional history** of **Ireland**, north and south, as an integrated whole…
- Correctly: ireland ; constitutional history
- Missing: –
- Wrongly: irish ; political science ; day ; north and south

Example Politics : Method Classifier

The Irish constitutional tradition

- `http://lod.b3kat.de/page/title/BV009807040`
- existing LCSH: ireland, constitutional history
- …1782 to the present day and treats the **constitutional history** of **Ireland**, north and south, as an integrated whole…
- Correctly: ireland ; constitutional history
- Missing: –
- Wrongly: constitutional law

Example Medicine : Method Extraction

Dictionary and handbook of nuclear medicine and clinical imaging

- `http://lod.b3kat.de/page/title/BV004061649`
- existing LCSH: diagnostic imaging, nuclear medicine, radioisotopes
- Dictionary that "bridges the gap between those highly sophisticated papers and . . . volumes dealing with basic sciences generally." Intended for generalists and specialists. Brief definitions. Handbook contains basic and reference data.
- Correctly: nuclear medicine ; diagnostic imaging
- Missing: radioisotopes
- Wrongly: paper ; science

Example Medicine : Method Classifier

Dictionary and handbook of nuclear medicine and clinical imaging

- `http://lod.b3kat.de/page/title/BV004061649`
- existing LCSH: diagnostic imaging, nuclear medicine, radioisotopes
- Dictionary that "bridges the gap between those highly sophisticated papers and . . . volumes dealing with basic sciences generally." Intended for generalists and specialists. Brief definitions. Handbook contains basic and reference data.
- Correctly: –
- Missing: nuclear medicine ; radioisotopes ; diagnostic imaging
- Wrongly: social sciences ; religion and science ; psychiatry

Example Physics : Method Extraction

Planck

- `http://lod.b3kat.de/page/title/BV042620297`
- existing LCSH: physicists, physics, national socialism and science
- Correctly: –
- Missing: physics ; physicists ; national socialism and science
- Wrongly: war ; vision ; process (law) ; law ; radiation ; universe ; quantum theory ; comprehension ; matter ; states ; twentieth century ; home ; science ; shorthand

Example Physics

Abstract: Planck

- **Planck's Law**, an equation used by physicists . . . **Max Planck** is credited with being the father of **quantum theory**, and his work laid the foundation for our modern understanding of **matter** and energetic processes. But Planck's story is not well known, especially in the United States. . . . What remains, . . . , are handwritten letters in German **shorthand**, . . . In Planck : Driven by **Vision**, Broken by **War**, Brandon R.
- Planck's Law not a concept in LCSH, but in GND: `http://d-nb.info/gnd/4174789-6`
- Max Planck as person not a concept in LCSH, but in LC Name Authority File `http://id.loc.gov/authorities/names/n80038130.html`
- vision and war very unspecific words from subtitle

Example Physics : Method Classifier

Planck

- `http://lod.b3kat.de/page/title/BV042620297`
- existing LCSH: physicists, physics, national socialism and science
- Correctly predicted: –
- Missing: physics ; physicists ; national socialism and science
- Wrongly predicted: modernism (art) ; cognition ; scientists

Conclusions I

Quality and errors

- Given the huge amount of possible LCSH: both methods have a high recall
- Low precision: many false positives

False positives

Extraction Very general terms, that occur in many texts

Classification Completely wrong (specific) areas, with some related terms and lack of training data.

Conclusions II

Comparison of disciplines

Extraction Absolutely no differences between disciplines!

Classification Differences can easiliy be explained by amount of training data.

Lessons learned

- LCSH is not very well suited for automatic assigment (ambiguity, pre-combined labels, etc.)
- For specific areas ot enough training data available.
- Extraction can help where classification is not possible due to data sparsity.

Discussion

Related work

- Aga R.T., Wartena C., Franke-Maier M. (2016) Automatic Recognition and Disambiguation of Library of Congress Subject Headings. In: Fuhr N., Kovács L., Risse T., Nejdl W. (eds) Research and Advanced Technology for Digital Libraries. TPDL 2016. Lecture Notes in Computer Science, vol 9819. Springer, Cham. `https://doi.org/10.1007/978-3-319-43997-6_40`

Future work

- Integrating extraction from text and classification
- Using hierarchical structure
- Results for other vocabularies (e.g. GND or MESH)

Kontakt

Vielen Dank für Ihre Aufmerksamkeit!

- Michael Franke-Maier: `franke@ub.fu-berlin.de`
- Christian Wartena: `christian.wartena@hs-hannover.de`