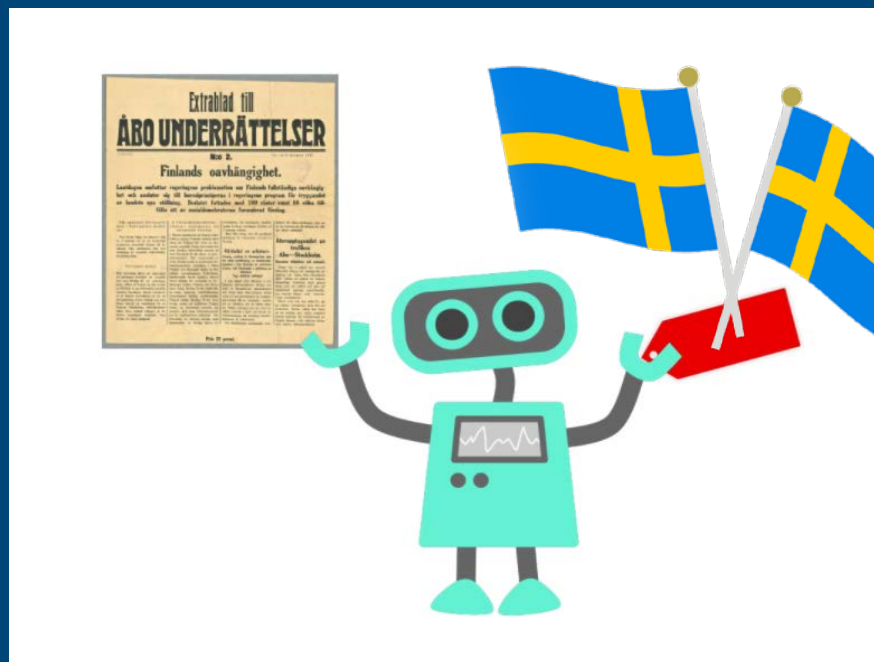


Annif & Swedish DDC

– AI in Swedish



EDUG 2021
Harriet Agaard, National Library of Sweden

Annif + Dewey + Libris = true!!!

annif

Annif.org

TRY THE DEMO!

INPUT TEXT

About EDUG

This is the website of the European DDC Users Group (EDUG). Over the past fifteen years, extensive translation projects have been completed in Europe to produce the French, Italian, German, Swedish and Norwegian editions of the DDC21, DDC22 and DDC23. Experience has shown that the translation and adaptation of the DDC in the European context can be difficult. It is in this context that EDUG was established in 2007. EDUG works in partnership with OCLC to foster cooperation in the development of DDC in Europe. In particular, EDUG aims:

To promote professional interests of all users of the DDC in Europe by the exchange of experience in the use of the DDC;

To coordinate proposals for the development of the DDC according to the bibliographic needs of European libraries and users in collaboration with the Dewey Classification Editorial Policy Committee (EPC) and OCLC;

To encourage the development and dissemination of techniques, applications, software, documentation and procedures in the areas of translation and access;

To encourage and promote co-operation in the translation of the DDC into European languages. More information can be found in the EDUG cooperation agreement (updated 2019-09-02).

PROJECT (VOCABULARY AND LANGUAGE)

YSO NN ensemble English

MAX # OF SUGGESTIONS

10 15 20

Get suggestions →

annif

SUGGESTED SUBJECTS

- Europe
- software development
- language policy
- libraries
- cooperation (general)
- applications (documents)
- translating
- professional development
- use of language
- library policy

Why a project? 1)

KB tillgängliggör kraftfulla modeller för språkförståelse

4 februari 2020

Forskning

I dag publicerar KB tre svenska språkmodeller baserade på Googles "BERT" (Bidirectional Encoder Representations from Transformers). De första testerna visar att KB:s modeller överträffar Googles flerspråkiga modell.



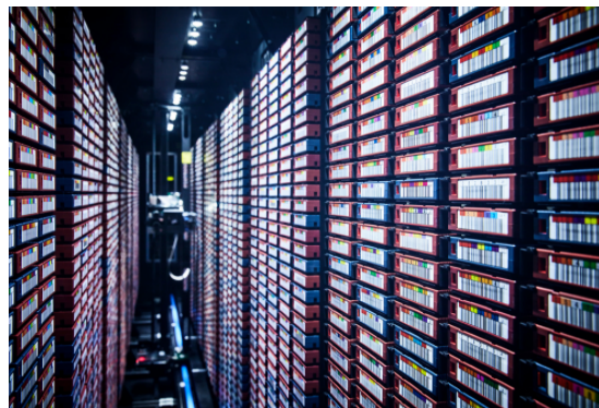
En språkförståelsemodell är ett omfattande artificiellt neuralt nätverk som tränats på stora mängder text för att få en flexibel och djup språkförståelse. KB:s BERT har lärt sig mekanismerna i svenska språket och kan analysera och strukturera text i stora underlag.

SpaCy – ny svensk modell för storskalig textanalys

23 november 2020

Forskning

KB-labb har utvecklat en ny modell till verktyget SpaCy. Modellen gör det betydligt smidigare att utföra storskaliga textanalyser på svenska. Det här är en viktig del av den infrastruktur för datadriven forskning som labbet etablerar.



SpaCy bygger på natural language processing (NLP) – eller språkteknologi – på svenska – som fokuserar på interaktioner mellan datorer och mänskligt språk. Genom att lära sig hur människor använder språk kan en datamodell analysera

Why a project? 2)

- Svein Arne Brygfjeld, National library of Norway
 - Libris user meeting 2019
 - Nancy: a digital AI lady



Koralka Golub, EDUG 2019



DE GRUYTER
G
OPEN

Research Paper

Automatic Classification of Swedish Metadata Using Dewey Decimal Classification: A Comparison of Approaches

Koralka Golub^{1*}, Johan Hagelbäck², Anders Ardo³ (emeritus)

¹Department of Cultural Sciences, Faculty of Arts and Humanities, Linnæus University, Växjö, Sweden
²Department of Computer Science and Media Technology, Faculty of Technology, Linnæus University, Kalmar, Sweden
³Department of Electrical and Information Technology, Lund University, Lund, Sweden

Citation: Golub, Koralka, Johan Hagelbäck, and Anders Ardo. "Automatic Classification of Swedish Metadata Using Dewey Decimal Classification: A Comparison of Approaches." *Journal of Data and Information Science*, vol. 5, no. 1, 2020, pp. 18–38.
DOI: 10.2478/jdis-2020-0003
Received: Feb. 4, 2020
Revised: Mar. 20, 2020
Accepted: Mar. 25, 2020

Abstract
Purpose: With more and more digital collections of various information resources becoming available, also increasing is the challenge of assigning subject index terms and classes from quality knowledge organization systems. While the ultimate purpose is to understand the value of automatically produced Dewey Decimal Classification (DDC) classes for Swedish digital collections, the paper aims to evaluate the performance of six machine learning algorithms as well as a string-matching algorithm based on characteristics of DDC.

Design/methodology/approach: State-of-the-art machine learning algorithms require at least 1,000 training examples per class. The complete data set at the time of research involved 143,838 records which had to be reduced to top three hierarchical levels of DDC in order to provide sufficient training data (totaling 802 classes in the training and testing sample, out of 14,413 classes at all levels).

Findings: Evaluation shows that Support Vector Machine with linear kernel outperforms other machine learning algorithms as well as the string-matching algorithm on average; the string-matching algorithm outperforms machine learning for specific classes when characteristics of DDC are most suitable for the task. Word embeddings combined with different types of neural networks (simple linear network, standard neural network, 1D convolutional neural network, and recurrent neural network) produced worse results than Support Vector Machine, but reach close results, with the benefit of a smaller representation size. Impact of features in machine learning shows that using keywords or combining titles and keywords gives better results than using only titles as input. Stemming only marginally improves the results. Removed stop-words reduced accuracy in most cases, while removing less frequent words increased it marginally. The greatest impact is produced by the number of training examples: 81.90% accuracy on the training set is achieved when at least 1,000 records per class are available in the training set, and 66.13% when too few records (often less than

JDIS
Journal of Data and Information Science

* Corresponding author: Koralka Golub (E-mail: koralka.golub@lnu.se).

18

Osma Suominen, Annif



2021-05-06

Senior Cataloguing Experts



2021-05-06



Mixed translation

- Translated into Swedish
- Automated updates
 - not translated numbers stay in English

The screenshot shows a hierarchical menu for '618.1 *Gynekologi'. The top level is '618.1 *Gynekologi'. Below it are several sub-levels, some of which are expanded. The sub-levels include:

- 600 ▾ Teknik
- 610 ▾ Medicin & hälsa
- 618 ▾ Gynekologi, obstetrik, pediatrik, geriatrik
- 618.1-618.8 ▾ Gynekologi och obstetrik
- 618.1 *Gynekologi**
- 618.10083 ▾ Young people
- 618.10088796 Sports gynecology
- 618.10092 Gynekologer
- 618.10231 Female genital diseases--humans--nursing, . . .
- 618.10232 Gynecologists--role and function
- 618.1059 ▾ Female genital diseases--humans--surgery, . . .
- 618.106 Female genital diseases--humans--therapy
- 618.107 ▾ Gynecologic pathology
- 618.11 ▾ *Diseases of ovaries
- 618.12 ▾ *Diseases of fallopian tubes
- 618[.13] [Unassigned]
- 618.14 ▾ *Diseases of uterus
- 618.15 ▾ *Diseases of vagina
- 618.16 ▾ *Diseases of vulva
- 618.17 ▾ *Functional and systemic disorders
- 618.18 ▾ Födelsekontroll
- 618.19 ▾ *Bröstsjukdomar

The bottom of the screenshot shows a 'Historik' (History) section with a right-pointing arrow.

Project plan - goals

- Evaluating indexing quality directly
 - through assessment by an evaluator
 - by comparison with a gold standard
 - (in the context of an indexing workflow - new project)

- Accuracy of the algorithmic suggestions from Annif for DDC numbers
 - as deep as possible
 - 3 top DDC numbers

Project plan

- 20 documents in 3 subject areas, 60 in total
 - random choice from records from the National bibliography in Libris 2019-2020
 - at least 2 users who classify them – from students at Växjö university
 - ~~classification by subject experts~~
 - evaluated by 2 senior cataloging experts

What will Annif do? 1)

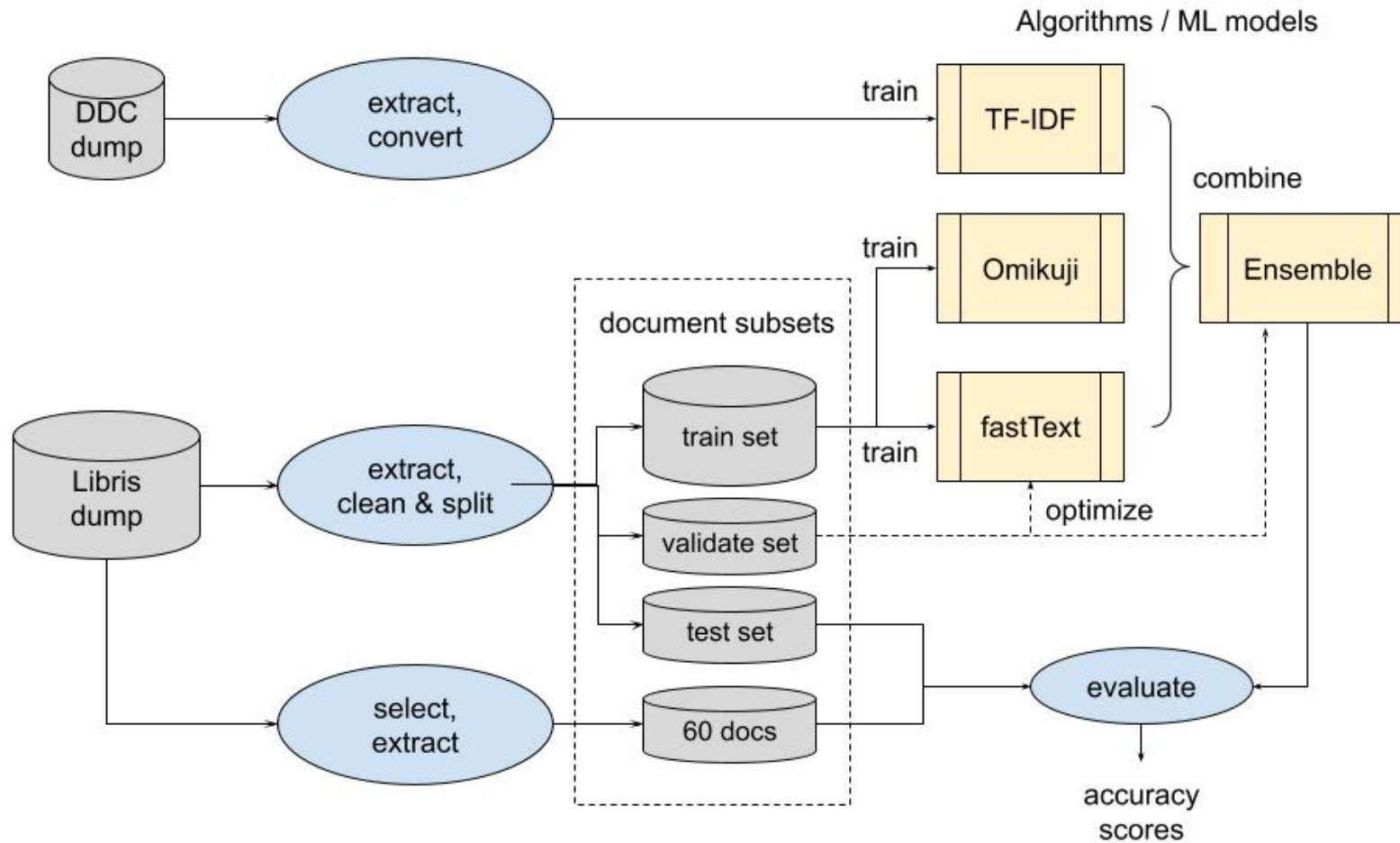
- Access to Libris cataloguing data & DDC from the Swedish WebDewey
 - Records from Libris with DDC 233 000 records
 - Training records (90%) + validate records (5%) + test records (5%)
- Algorithms in Annif
 - TF-IDF (statistical method, term frequencies)
 - fastText (neural machine learning algorithm for text classification)
 - Omikuji (tree-based machine learning algorithm for classification),
 - Ensembles (combination of the three above algorithms)

Annif 2)

- Algorithms in Annif
 - Validate the models - training records
 - Decide on which of the algorithms to use



Annif 3)



Why DDC? 1)



- International system used all over the world
- Sharing classification metadata

Why DDC? 2)

553.7 Water

- [500](#) Science
- [550](#) Earth sciences & geology
- [553](#) Economic geology
- [553.2-553.9](#) Specific materials

553.7 Water

- [553.7/2](#) Saline water
- [553.7/3](#) Mineral waters
- [553.7/8](#) Surface water
- [553.7/9](#) Groundwater (Subsurface water)

History

Thermal waters relocated to [333.88](#)
1989-03-06, Edition 20

Notes

Including ice
Class interdisciplinary works on thermal waters in [333.88](#)
Class interdisciplinary works on ice in [551.31](#)
For geology of thermal waters, see [551.23](#)
See Manual at [363.61](#)

Comments

Create built number



[553.7](#) Water

SEARCH ADVANCED SEARCH BROWSE COMMENTS PRINT

Search Browse in Relative Index (en)

PAGE UP PAGE DOWN

Browse Results

Watchworks--technology	681.112
Water	553.7
Water--biochemistry	572.539
Water--biochemistry--humans	612.01522
Water--chemical engineering	661.08
Water--chemistry	546.22
Water--disease transmission	614.43
Water--dowsing	133.3232
Water--economic geology	553.7
Water--folklore	398.26
Water--folklore--history and criticism	398.364
Water--geologic agent	551.35
Water--geologic agent <i>see Manual at 551.302-551.307 vs. 551.35</i>	
Water--health	613.287
Water--hydraulic engineering	627
Water--hydraulic-power technology	621.20422
Water--law	 346.04691
Water--materials science	620.198
Water--metabolism--human physiology	612.3923
Water--meteorology	551.57
Water--plant management	658.26
Water--prospecting	628.114
Water--public administration	354.36
Water--religious worship	202.12
Water--resource economics	333.91
Water--resource economics--development	 333.9115
Water--sanitary engineering	628.1
Water--supply services	363.61
Water--treatment engineering	628.162

Results?

Initial results:

- Test set accuracy approximate
 - 60% for 3-digit DDC
 - 40% for full DDC

First suggestions to evaluate

Libris-DDC example documents classified by Annif ☆ 🗨️ 📄

File Edit View Insert Format Data Tools Add-ons Help

100% Comment only

A1 Subset

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Subset	SubNr	Libris ID	Title	Keywords	DDC1_class	DDC1_heading	DDC1_score	DDC2_class	DDC2_heading	DDC2_score	DDC3_class	DDC3_heading	DDC3_score
27	Health	6	p2mm29m4mjgfbkv	Hälsa från bikupan	Honung; Hudvård; Hälsa	638	Insektsodling	0.1933	613	Hälsa och säkerhet	0.0757	641.38	Honung	0.0527
28	Health	7	z8z27xbjwmsfdgg	Lev livet längre : om evolution, hälsa ocl	Hälsa; Hälsobeteende; Livslängd; Åldrande	613	Hälsa och säkerhet	0.7177	612.67	Åldrande	0.0941	618.97	Geriatric	0.0474
29	Health	8	6k7g46jq44f1gp5b	Konsten att inte banta : den banbrytande	Bantningsmat; Viktminskning; Vegetarianis	613.25	Viktminskningskost (bantningskos	0.6018	613.2	Dietik	0.2352	641.563	Matlagning för hälsa, utsee	0.0705
30	Health	9	r2h4g2h9pg0k10kd	Rapport från ett slakteri : en veterinärs b	Slakterier; Svin; Veterinärer; Djurhälsa; Livs	338.4766492	Nötkött--kommersiell tillverknin--	0.1629	636.4	Svin	0.0365	338.76	Företag efter näringsgren, i	0.0261
31	Health	10	6gd7btv4zjfpkp5	Anteckningar från en orolig planet	Ängest; Psykisk hälsa; Stresshantering; Liv	616.8522	Ängeststörningar	0.1409	155.9042	Stress	0.0878	152.46	Rädsla, ångest, oro	0.0651
32	Health	11	22431516	Skärnhjärnan : hur en hjärna i osyn me	Hjärna; Stresshantering; Psykisk hälsa; Bra	612.82	Centrala nervsystemet	0.7416	155.9042	Stress	0.0601	612.8	Nervsystemet	0.0185
33	Health	12	r13qcm4xph4131pc	Vi borde vara lyckliga	Unga vuxna; Psykisk hälsa; Ängest; Persor	158.1	Personlig utveckling och analys	0.1621	362.2	Personer med psykisk störnin	0.0832	152.46	Rädsla, ångest, oro	0.0400
34	Health	13	r2c4g6sgpglcsjl	Sjöstakoviij förändrade mitt liv	Musikpsykologi; Musikterapi; Psykisk hälsa	615.85154	Musikterapi	0.1582	780.92	Biografi	0.0654	781.66092	Rockmusiker	0.0412
35	Health	14	lx2jfmw0jm6rvcc	Dansa mjukt med tillvaron : om mening,	Tillämpad psykologi; Psykisk hälsa; Psycho	158.1	Personlig utveckling och analys	0.4560	616.89	Psykiska störningar	0.0997	362.2	Personer med psykisk störr	0.0892
36	Health	15	q1jc44xmn31q09sg	Livsbalans på 30 dagar	Personlig utveckling; Hälsa; Välbefinnande	158.1	Personlig utveckling och analys	0.8990	613	Hälsa och säkerhet	0.1039	158.16	Personlig utveckling och an	0.0243
37	Health	16	jsvswvbwj6wvnn29r	Ät dig frisk : revolutionerande forskning	Dietik; Nutrition; Funktionell mat	613.2	Dietik	0.8955	613.2833	Låg-kolhydrat diet (kolhydratf	0.0593	615.854	Näringssterapi	0.0347
38	Health	17	4dz9whf259713jn	Ersättningen och e-hälsan	Hälsoekonomi; E-hälsa; Economics; Medici	362.1028	Hälsa- och sjukvård--metoder	0.4013	338.433621	Sjukvård--ekonomi, ...	0.3392	362.10681	Hälsa- och sjukvård--ekono	0.0931
39	Health	18	8k99zkt627z36r8	Gejjerarvet : en släkthistoria om dikt och	Släktskap; Psykisk hälsa; Kreativitet; Kinshi	616.89	Psykiska störningar	0.1081	362.2	Personer med psykisk störnin	0.0560	929.209485	Familjehistorier, ...	0.0476
40	Health	19	nxzsq578lk8f95ws	Skogluft-effekten	Inomhusklimat; Hälsa	697	Uppvärmningsteknik, ventilationst	0.6802	613	Hälsa och säkerhet	0.0581	697.9	Ventilation och luftkondition	0.0308
41	Health	20	kv8chj2dh7b605e3	Ärr för livet	Självdestruktivt beteende; Självskadebete	616.8582	Antisocial personlighetsstörning, v	0.8847	616.89	Psykiska störningar	0.3078	616.8526	Åtstörningar	0.1965
42	NaturalSci	1	s3vmt2nkqcl5zcv7v	All dödlighets sång och skrian : studier i	Naturen i litteraturen	839.7	Svensk litteratur	0.0746	839.709	Svensk litteratur--historia och	0.0335	839.5	Nordiska litteraturer (nordg)	0.0165
43	NaturalSci	2	r2jq7dp0p15drbfj	Äventyrlig sommar	Pojkar; Kusiner; Sommarlov; Mysterier	839.738	Svenska romaner, ...	0.8171	823	Engelska romaner och novelle	0.0367	894.54134	Finsk litteratur, ...	0.0356
44	NaturalSci	3	lv9sb0b4jd26invhq	Haminguiden	Hamnar; Naturhamnar	623.8929	Lotsning i och seglingsbeskrivning	0.9283	387.1094234	Hamnar	0.0241	387.1094136	Hamnar	0.0241
45	NaturalSci	4	dpq4xcwbtkpsv3v	Stadsträdgården i Karlstad	Trädgårdskonst; Parker; Stadsträdgården i	712	Landskapsarkitektur	0.5875	712.5	Offentliga parker och marker	0.0700	839.738	Svenska romaner, ...	0.0695
46	NaturalSci	5	12453955	Berghandboken : Halle-Hunneberg	Naturreservat; Halle-Hunneberg (ekopark);	914.8	Skandinavien--geografi	0.0914	839.738	Svenska romaner, ...	0.0256	508.2	Årstider	0.0184
47	NaturalSci	6	mxxnrb9vk4srdq5q	Sydsvenska nationalparker : åtta pärlor	Nationalparker; Naturvård	914.8	Skandinavien--geografi	0.3005	333.72	Bevarande och skydd	0.1540	333.9516	Biologisk mångfald--bevara	0.1423
48	NaturalSci	7	jvv5x5mzgpsd0tww	Djur, natur och opretur : perspektiv på	Barnkonst; Bildanalys (konst); Djur och mår	809.933	Litteratur som behandlar särskilda	0.0480	800	Litteratur (skönlitteratur) och r	0.0412	839.7	Svensk litteratur	0.0306
49	NaturalSci	8	5fz2gvtl3fv024q8	Lär dig om sveriges vilda djur	Djur; Naturen	590	Djur	0.3020	591.9	Djur efter särskilda världsde	0.1308	599	Mammalia (daggdjur)	0.0249
50	NaturalSci	9	mwb612p3kv9sg2cc	Formåner : skattefritt & skattepliktigt	Anställningsformåner; Skatter	331.255	Tjänsteformåner	0.8893	336.2009	Historia, geografisk aspekt, bi	0.0291	336.2	Skatter	0.0233
51	NaturalSci	10	5g5hjk1303p20m7	Öar : världshavens unika utposter	Öar; Naturgeografi; Fotoböcker	914.8	Skandinavien--geografi	0.2330	839.738	Svenska romaner, ...	0.1201	823	Engelska romaner och novi	0.0170
52	NaturalSci	11	bmf08bdw8fr0603d	Människans spegel	Artificiell intelligens; Människans väsen; En	152.4	Kanslor	0.1001	6.3	Artificiell intelligens och natura	0.0986	153.9	Intelligens och anlag	0.0517
53	NaturalSci	12	cnx4dn3m916l2kqb	Kom igång med vetenskap	Naturvetenskap	500	Naturvetenskap och matematik	0.6829	501	Filosofi och teori	0.0564	505	Seriella resurser	0.0344
54	NaturalSci	13	lwtmxx0klctns58	Ölands natur : okända och ökända arter	Naturen; Växter; Fåglar; Daggdjur; Insekter	599	Mammalia (daggdjur)	0.0762	595.7	Insecta (insekter)	0.0654	590	Djur	0.0510
55	NaturalSci	14	5gn391z6324mgj57	Höga Kusten	Turism; Världsarv; Vandringsleder; Sverige	912.4	Europa	0.4193	338.4791	Turism--industri, ...	0.0447	914.8	Skandinavien--geografi	0.0389
56	NaturalSci	15	4fg0xrs12qjgbx7r	Antons anekdoter : experiment, kärlek o	Pojkar; Vardagsliv; Vänskap; Kärlek; Natur	839.738	Svenska romaner, ...	0.1689	507.8	Användning av apparatur och	0.1224	839.8238	Norska romaner och novelle	0.0367
57	NaturalSci	16	s3slqdhvq3r3kqsb	Skönhetsens evolution : hur Darwins bort	Människans utveckling; Sexuell selektion; F	576.8	Evolution	0.3188	591.56	Beteende i relation till livscyke	0.1833	591.3	Genetik, evolution, åldersk	0.0246
58	NaturalSci	17	lwg00n4j0l2z1tm	Identitet : ett socialpsykologiskt perspek	Identitet (psykologi); Människans väsen; Sc	302	Social interaktion	0.5376	155.2	Personlighetspsykologi	0.1272	305.8	Etniska och nationella grup	0.0668
59	NaturalSci	18	dp2wxxghb3xjs6ln	Storslagen fjällmiljö : underlagsrapport t	Fjällen--miljöaspekter--Sverige; highlands; i	363.7	Miljöfrågor	0.1512	333.7	Mark, friluftsområden och vild	0.0230	333.709	Miljögeografi	0.0225
60	NaturalSci	19	hsmh3g6fth8kdj9r	Vems rumpa?	Skogar; Vilda djur; Rumpor; Naturen	831	Tysk poesi	0.1389	839.738	Svenska romaner, ...	0.1052	839.718	Poeter--svensk litteratur, ...	0.0353
61	NaturalSci	20	mxxpqbxx8k70lc9z0	Överlevnadshandboken	Överlevnad i naturen; Överlevnadsteknik; C	613.69	Överlevnad	0.9740	796.5	Friluftsliv	0.0346	133.9013	Personlig överlevnad, den	0.0102

Thank you!

- harriet.aagaard@kb.se
- @haraag

2021-05-06

