OCR @ KBLab

Robin Kurtz

National Library of Sweden, KBLab


KONGL. BIBLIOTEKET

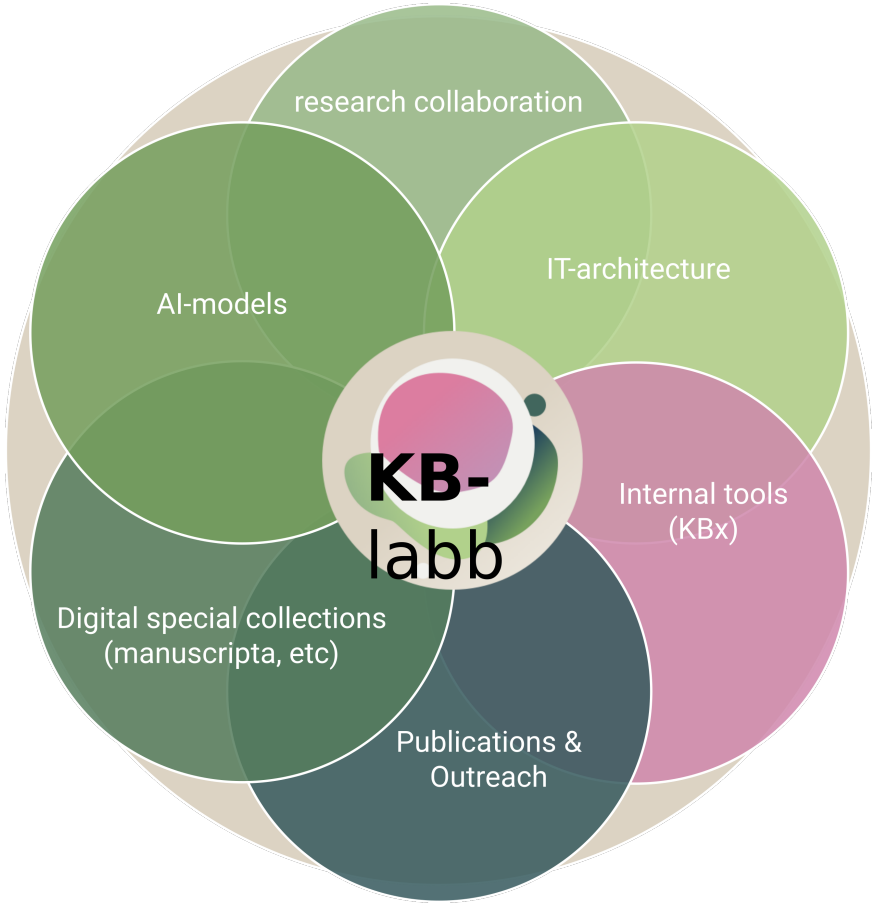
The National Library of Sweden

Kungliga Biblioteket (KB)

- collects, preserves and gives access to almost everything that is published in Sweden
- legal deposit act from 1661 required all printers to deliver one copy to KB
- a censorship law that now helps preserve Sweden's cultural heritage
- expanded in 1900 to include sound, moving images and video games
- collections currently hold over 18 million items
- ongoing digitization process

KBLab

- started in 2019 to give researchers the possibility to do large-scale quantitative research
- curate data maintained by the National Library
- train models on data to be used by academia, governmental organizations and industry



Legal deposit

- The Swedish Act (1993:1392) on legal deposit copies of documents applies to printed matter and to radio and TV broadcasts, music, film and multimedia that are intended for distribution to the general public.
- The Swedish Act (2012:492) on legal deposit copies of electronic material applies to, for example, music recordings, film, web-specific magazine and daily newspaper articles, and e-books and government agency publications.

OCR Data at KB

- newspapers
- books
- magazines
- Swedish Governmental Official Reports (SOU)
- Swedish Parliament
- ...

What is OCR?


ledare

Gamla hjulspår

DAGENS CITAT

Det kommer drabba alla våra medlemmar

Lördag

Vad sa Ebba Busch när svarta dog i covid-19?

Nu går botten ur Kristersson

6 JUNI 2020

Moderaterna

Svårt att vifta bort

Ingenting att ge

Zina Al-Dewany

AFTONBLADET

TIPSA

Motivation for better OCR

- better search
- better data for researchers
- better (Large) Language Models

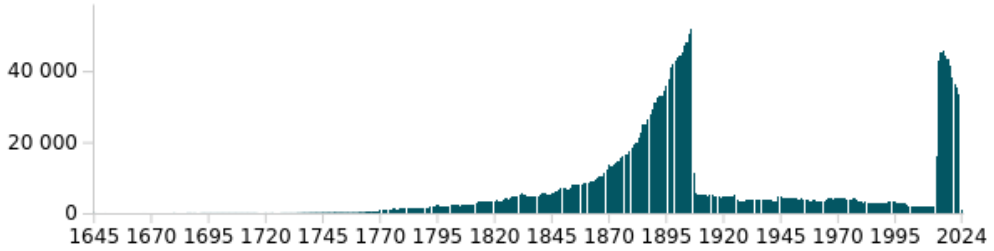
Motivation for better OCR

Tidsperiod

Från ÅÅÅÅ-MM-DD

Till ÅÅÅÅ-MM-DD

Applicera datum



Titel

Dagens nyheter	67 802
Aftonbladet	65 871
Svenska dagbladet	49 692
Göteborgs handels- och sjöfartsti...	41 674
Expressen	40 202
Göteborgsposten	36 988
Arbetet (1887)	34 854
Stockholms dagblad	28 076
Post- och inrikes tidningar	26 144

OCR so far

- Abbyy FineReader
- Zissor
- Tesseract
- A Two-OCR Engine Method for Digitized Swedish Newspapers

Challenges

- better OCR systems
- post-OCR error correction
- layout analysis
- reading order
- editorial content vs. advertisement


Nr 437. SYDSVENSKA DAGBLADET 1897
Morgonnummer
Onsdagen SNÄLLPOSTEN. 22 September
Kungsparken. Konsert
Filmurarelogens Restaurant
Wienerkafé
Browns Conditoriträdgårds Restaurant
Med anledning af h. m:t konungens
festlig middag
Kapplöpningarna
Ladugårdsgärdet
Lindarängen
Dr Axel Bjerkén,
Dr P. Rosenius,
Hilda Erikson,
gymnastikdirektör,
Gustaf M. Nyströms
N. Gust. Åberg.
Patentagentur
Konstruktionsbyrå.
H. Sjögrens Eftr.
Fotograf
Axel Sjöberg,
Inkasseringar
Fördelaktig försäkring
Brandstodskomitén.
Gas-förbrukare
Amicitia Loge N:o III
Malmö Tjenstemannaklubb.
Bolagsstämma.
Delegarne
Helsingborgs Sockerfabriks
Aktiebolag
Styrelsen.
Kullans Chokolad och Cacao-Maxelli
Guldmedalj
10 öre Jubileumscigarren 10 öre
HAVANA STJERNAN
C. G. Kockum & Co, Malmö.
10 öre Nytt! Rök! Nytt! 10 öre
ROYAL
VELOCIPEDER
Parasiticide.
Fodervaror.
Carl Engström.
20, 18 och 10 % Superfosfat,
Gust. A. Hedman & Co.
Malmö.
Aktiebolaget Skånska Handelsbanken,
Malmö.
Auktion
å bättre möbler.
Till salu:
Briggs Silkstone
Hushållskol
"Frankfort"
A. Scholander & Co.
M. ZADIG.
neutrala Lanolin Crème-Tvål
Lanolin Crème
Fjäder,
Fredrik Nyström.
Prima Stettiner Ragskakor
Använ Patenterad Svafvel-Lanolin-tvål
KÖP 4711:s VIOL-PARFYM
Försäljes i flaskor
Äkta Eau de Cologne
Fort Wilhem No. 4711, Köln
Ur sista utl. posterna.

Challenges

- better OCR systems
- post-OCR error correction
- layout analysis
- reading order
- editorial content vs. advertisement

ledare

Gamla hjulspår

DAGENS CITAT

Det kommer drabba alla våra medlemmar

Åsa Fahlén, ordförande för Lärarnas riksförbund, om förslagen i den nya tas-utredningen.

Lördag

Vad sa Ebba Busch när svarta dog i covid-19?

Nu går botten ur Kristersson

6 JUNI 2020

Moderaterna

De som protesterar vet vilka som drabbas värst av sjukdomen

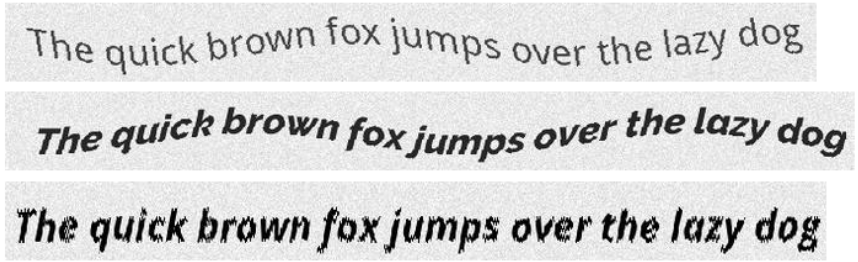
Zina Al-Dewany

AFTONBLADET

TIPSA!

Better OCR

- great open-source models available
- can be trained on synthetic data
- expensive to run on big datasets



---

https://github.com/Belval/TextRecognitionDataGenerator

Layout Analysis & Reading Order

- Zissor
- advertisement detection
- Donut 🍩, **Do**cument u**n**derstanding **t**ransformer
- LayoutLM
- LayoutReader
- SegFormer
- Surya

Layout Analysis & Reading Order

Problem

data...

Post-OCR Correction

KB-BERT

Error candidates checked and replaced with SALDO

- ö ⇆ o
- rn ⇆ m

Post-OCR Correction

- continue from *Post-OCR Correction of Digitized Swedish Newspapers with ByT5*
  - newspapers
  - literature bank
  - blackletter (*Swedish fraktur* & *Then swänska Argus*)
- add Project Runeberg
- SWERIK

Post-OCR Correction

Year	CER						WER				
	BL	M1	M2	M3	KB	KB2	BL	M1	M2	M3	KB
1817-1859	8.39	3.73	3.72	3.64	2.73	1.38	33.78	14.02	14.02	13.77	8.72
1859-1899	4.04	1.91	1.99	1.90	1.26	0.66	17.63	7.34	7.47	7.26	4.03
1899-1939	2.60	1.54	1.52	1.52	0.94	0.67	12.45	6.50	6.45	6.52	3.32
1939-1979	1.46	0.96	1.04	0.97	0.63	0.42	7.38	4.07	4.00	4.02	2.25
1979-2018	0.83	0.49	0.49	0.48	0.41	0.21	4.45	2.37	2.35	2.39	1.51
1817-2018	3.2	1.64		1.62	1.1	0.64	14.52	6.65		6.59	3.73

Post-OCR Correction

Small

Model	CER	WER
Baseline	3.2	14.52
Model 1	1.64	6.65
Model 2	-	-
Model 3	1.62	6.59
KB	1.1	3.73
KB2	0.64	2.44

Post-OCR Correction

Base

Model	CER	WER
Baseline	3.2	14.52
Model 1	1.54	6.3
Model 2	1.6	6.41
Model 3	1.57	6.31
KB	0.98	3.37
KB-large	0.97	3.22

Post-OCR

- more training steps
- longer sequences
- bigger models
- other models
- more data
- synthetic data

Where to find us

- https://huggingface.co/KBLab
- https://kb-labb.github.io/
- https://www.kb.se/in-english/research-collaboration/kblab.html
- https://lab.kb.se/bildsok/