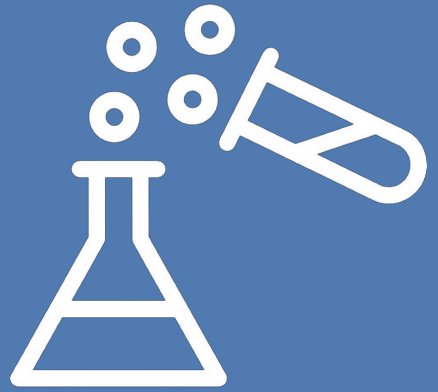
Att digitalisera vår historia

Utmaningar och möjligheter

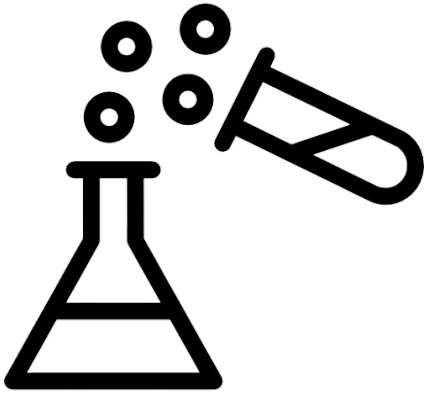


Erik Lenas, Viktoria Löfgren

2024-05-16

Riksarkivet

Agenda



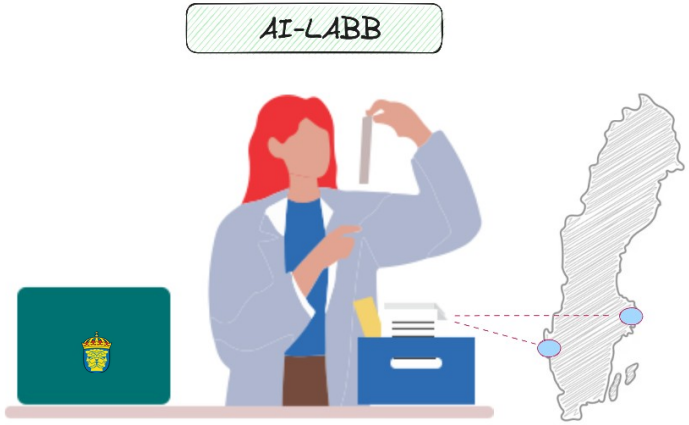
- Lite kort om Riksarkivet
- AI-labbet
- Digitalisering av historisk text - vad menar vi?
- HTR/OCR - mer än bara textigenkänning
- HTRFLOW - en mångsidig kodbas
- OCR av Kubhist2 - vad innebär det?
- Historiska språkmodeller - varför?

Riksarkivet

- 800,000 hyllmeter arkiv, från medeltid till nutid
- 270 miljoner skannade dokument, vilket utgör ca 5% av vårt arkiv
- Myndighetsdata, taget brett

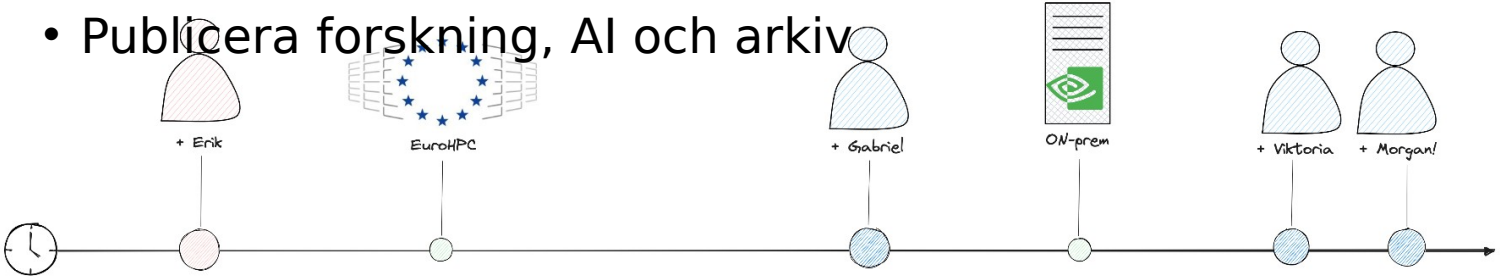
Riksarkivet

AI-labbet



Varför ett AI-labb på Riksarkivet?

- Vi har data, mycket, och den är värdefull!
- Effektivisera vår verksamhet
- Tillgängliggöra arkiv
- Publicera forskning, AI och arkiv

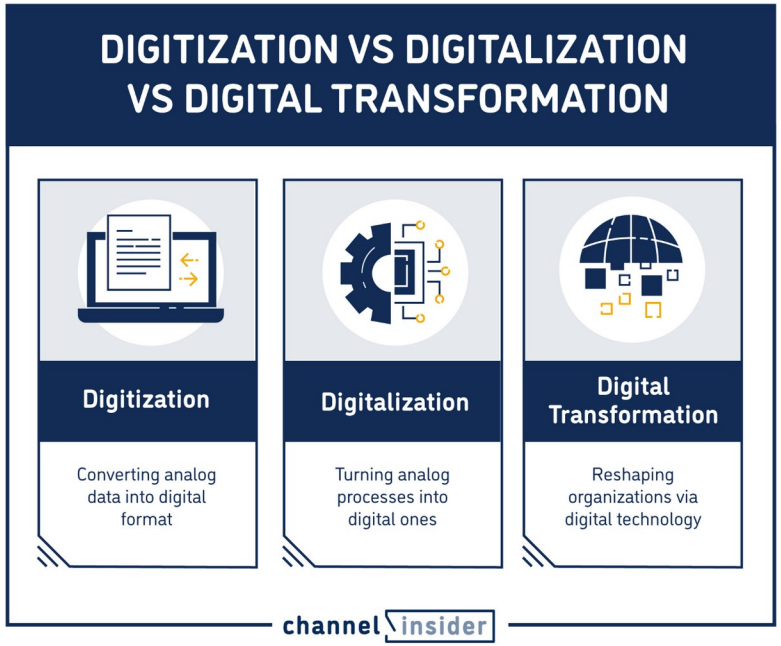


Vad menar vi med digitalisering?

Och varför kommer den att revolutionera den historiska forskningen?

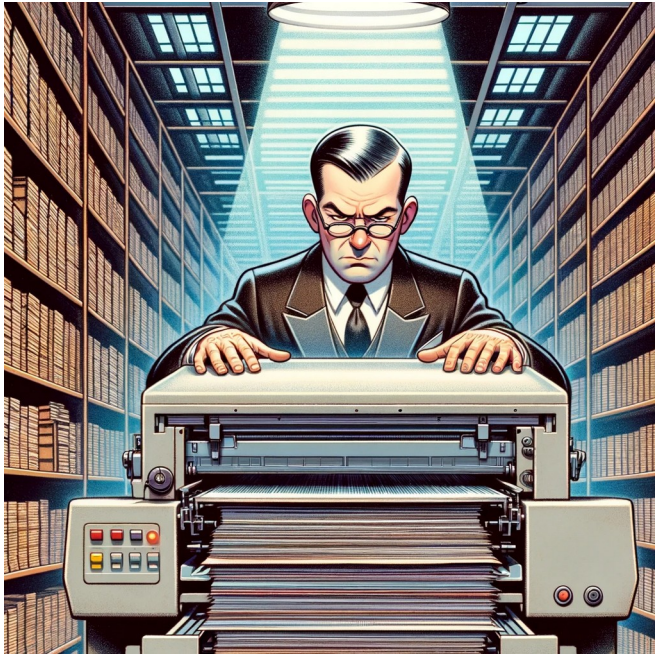
”Digitalisering” - i ordets fulla bemärkelse

- Skanning
- Utveckling av state-of-the-art OCR/HTR modeller
- Köra dessa modeller på stora delar av våra arkiv
- Finetuna LLM:s på högkvalitativ historisk text
- AI-baserad sökfunktionalitet för historisk text
- NLP/NLU för historisk text
  - tematisk uppmärkning
  - summering
  - text-modernisering
  - entitets och relationsextraktion
  - kunskapsbaser



Skanning (digitisering)

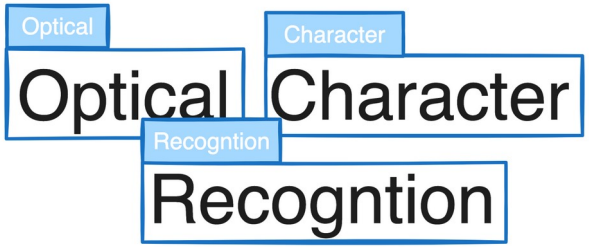
- Digitiseringsenhet i Fränsta.
- **Bevarande**: Säkerställer att information från ömtåliga dokument bevaras
- **Global tillgänglighet**: Arkiven globalt tillgängliga genom vår bildvisare
- **Globalt samarbete**: Möjliggör samarbete över nationsgränserna



HTR/OCR

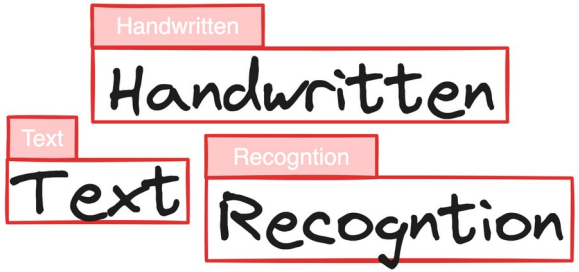
Mer än bara textigenkänning

Textigenkänning → HTR/OCR

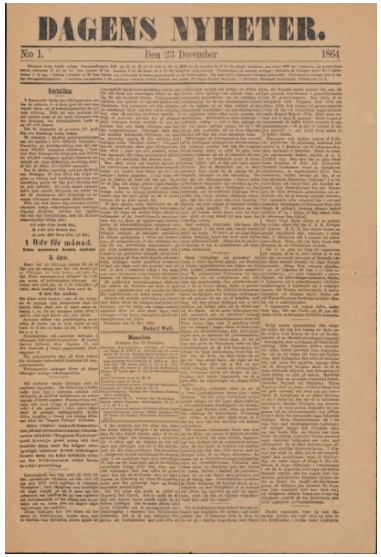


Layoutanalys och läsordning ofta svårare, men linjesegmentering och textigenkänning enklare

Segmentering och textigenkänning kan vara utmanande



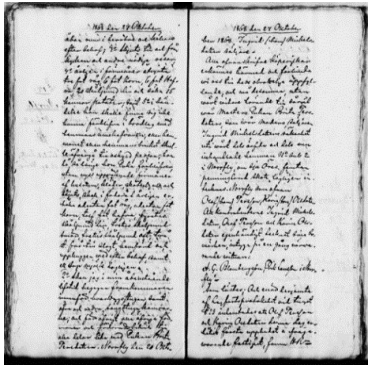
Olika typer av material - Olika utmaningar


DAGENS NYHETER.

No 1. Den 23 December 1864

1 Rdr för månad.

Tidningar

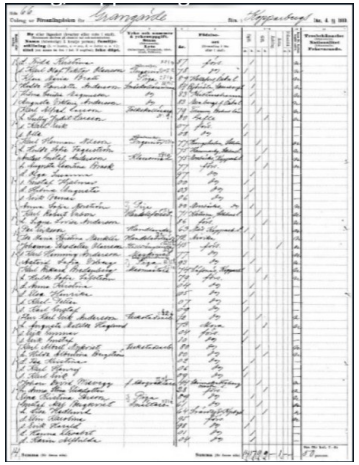


Löptext

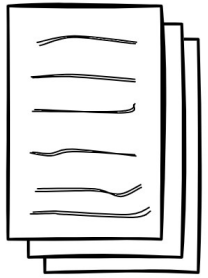
Löptext + tabeller


Anstag i bevis

Formulär



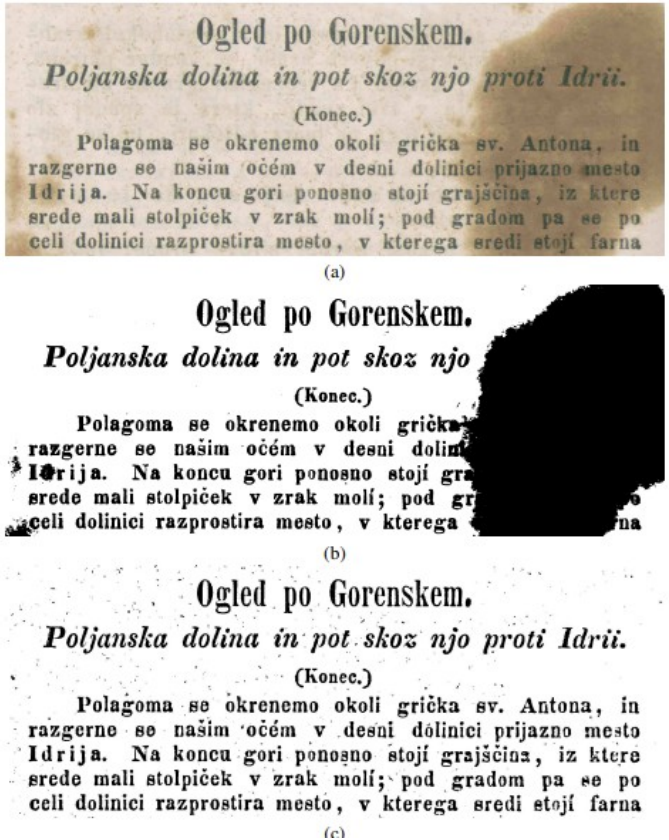
Tabeller



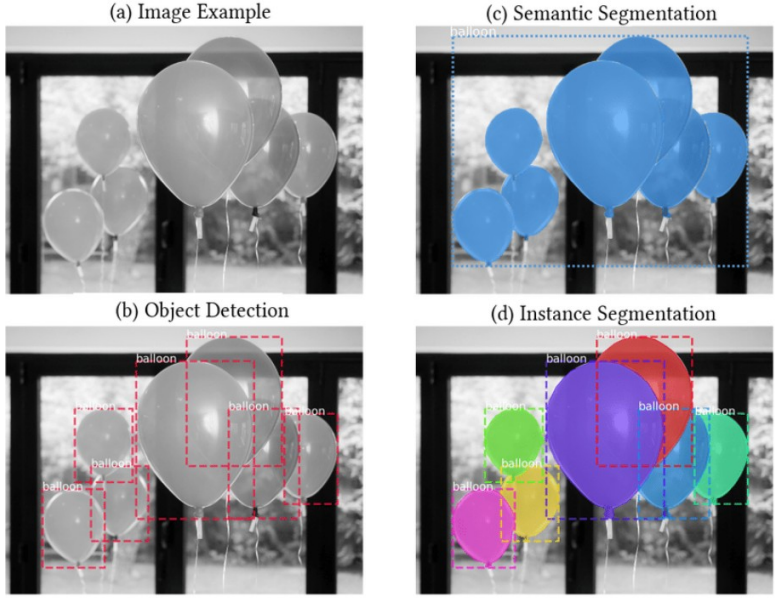
+ mycker mer!

Bildförbehandling

- Målet är inte ett “rent” dokument, utan en bättre OCR/HTR
- Allt från enkel binarisering till avancerade generativa modeller
- Olika OCR/HTR arkitekturer olika bra på att hantera “noise”
- Transformerbaserade arkitekturer mindre känsliga för noise
- Strategi beror på materialet

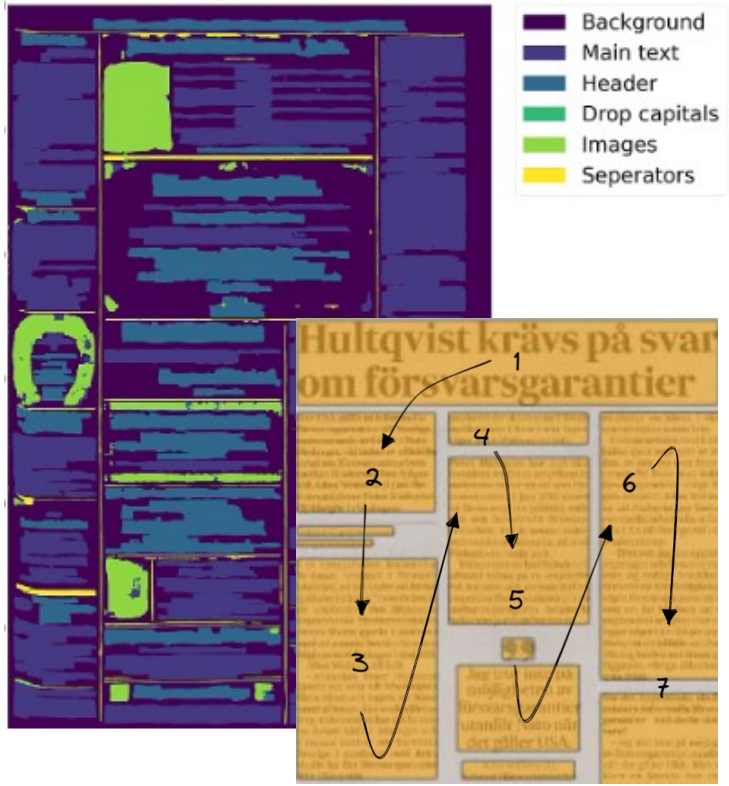


Att välja segmenteringsmetod



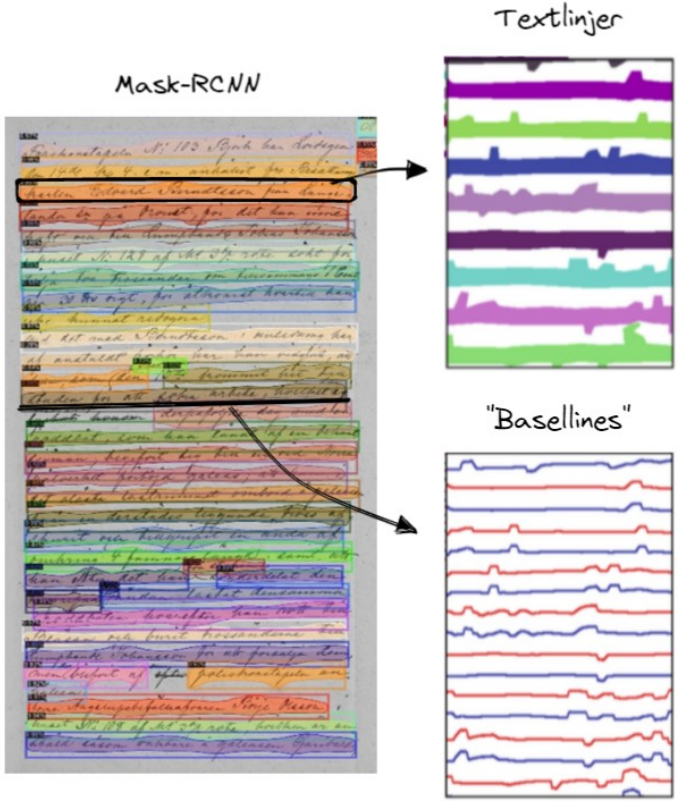
- Vilken typ av dokument?
  - ostrukturerad, semi-strukturerad eller strukturerad?
- Vad är målet med projektet, vilken information skall utvinnas?
- Vad är en acceptabel CER?
- Open-source?
  - Kodbaser
  - Modeller/arkitekturer
  - Publika dataset

Layout-analys



- Segmentering av dokument i meningsbärande regioner
  - Rubriker
  - Paragrafer
  - Separatorer
  - Tabeller
  - Listor
  - Bilder
- Reading-order: I vilken ordning läses regionerna?
  - Heurestik
  - Neurala nätverk

Textradssegmentering



- Fortfarande nödvändig, både för OCR och HTR (LLM:s ett undantag)
- Olika metoder:
  - base-line extraction + heuristics
  - instance segmentation models
  - oövervakade metoder
- Enklare för tryckt text
- Evaluering borde inkludera efterföljande OCR/HTR, inte endast IoU

Textigenkänning

- Omvandlar en textlinjebild till digital text

CER - metrik på teckennivå

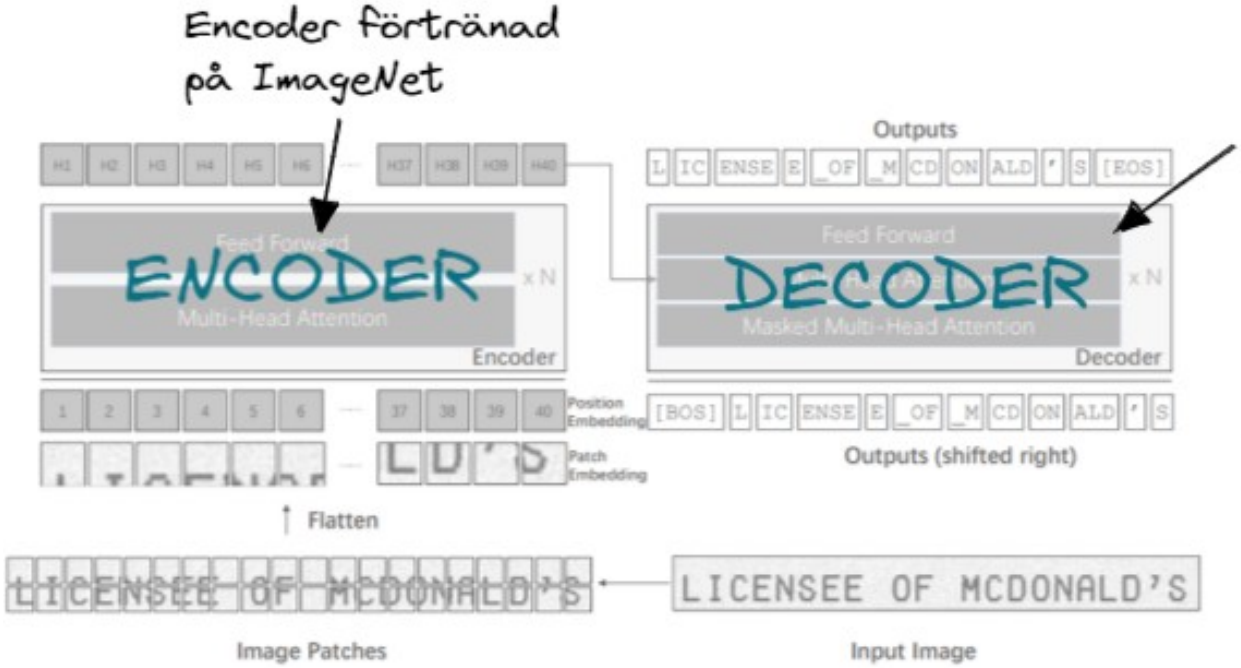
- S - antal utbyten
- D - antal borttagningar
- I - antal insättningar
- N - antal tecken i GT

WER - samma fast för ord

$$CER = \frac{S + D + I}{N}$$

$$WER = \frac{S_w + D_w + I_w}{N_w}$$

“TrOCR: Transformer-based Optical Character Recognition with Pre-trained Models” (2022, arxiv)



Decodern är en liknande BERT-Modell

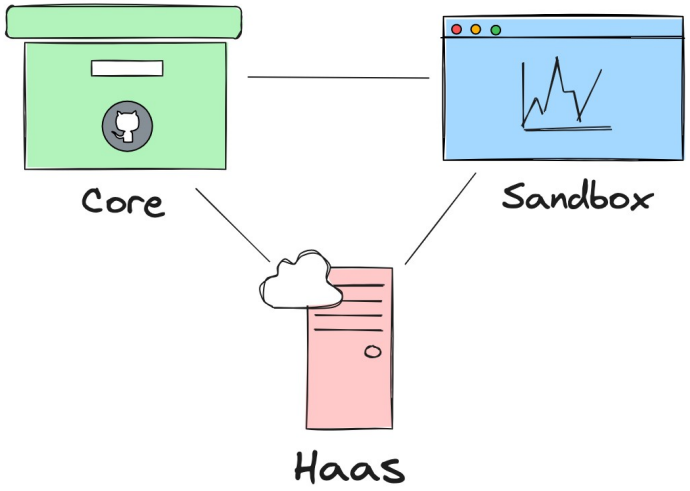
Kan enkelt fine-tunas för bli mer domänspecfik

Både OCR- & HTR-modeller förtränad vikter finns tillgängliga publikt

HTRFLOW

Den mångfacetterade kodbasen

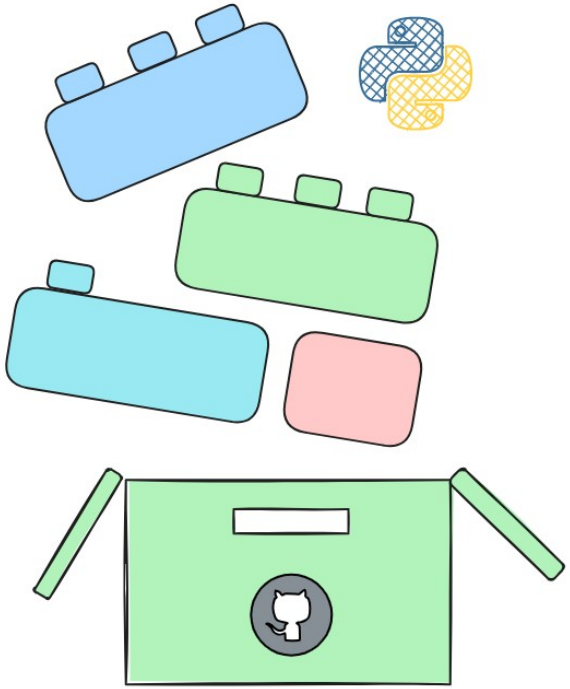
HTRFLOW



Vi har utvecklat HTRFLOW för att göra OCR/HTR processen enklare att förstå, anpassa och implementera i produktion

HTRFLOW-Core

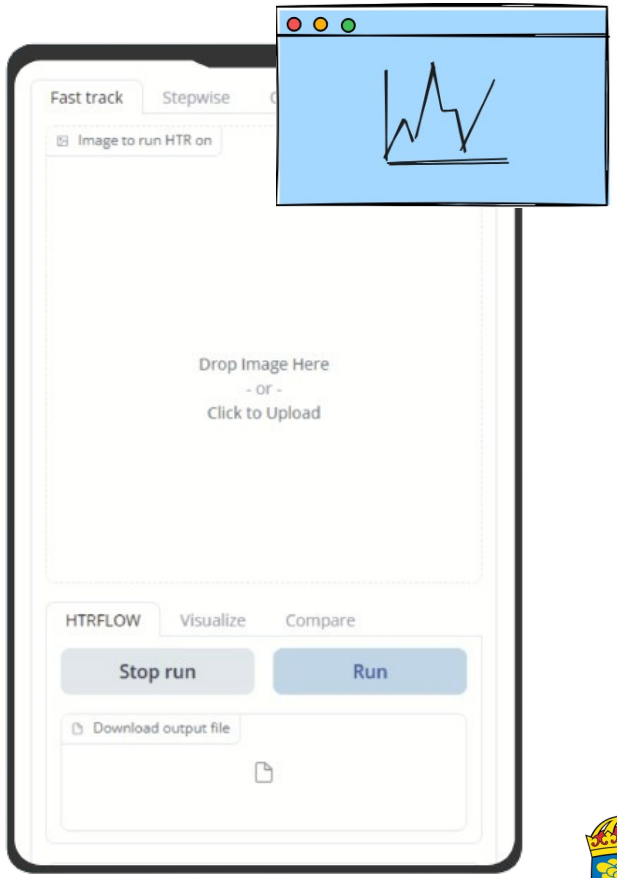
- Standardiserar OCR/HTR processen så att hjulet inte behöver uppfinnas på nytt med varje nytt projekt
- Underlättar samarbete internt och externt.
- "Legoprincipen" - sätt ihop din egen pipeline med modeller och andra steg i en YAML config fil



HTRFLOW-Sandbox

- För att visualisera och dema OCR/HTR processen
- Testköra våra egna modeller och snabbt jämföra dom med andras
- Open-source

Testa själv:



Activities Visual Studio Code

Riksarkivet

localhost:7000

Keycloak Acc... Wiki leonardo

Version 0.1.0

HTRFLOW

Explore AI models for Handwritten Text Recogntion developed by the Swedish National Archives

Fast track | Stepwise | Overview | How to use

1. Region segmentation | 2. Line segmentation | 3. Text recognition | 4. Explore results

Image to region segment

Drop Image Here
- or -
Click to Upload

Settings

Clear | Run

Example images

Göteborgs poliskammare, 1877

Göteborgs poliskammare, 1886

Göteborgs poliskammare, 1865

Bergskollegium, Relationer och skrivelser anggående utländska bergverk, 1698

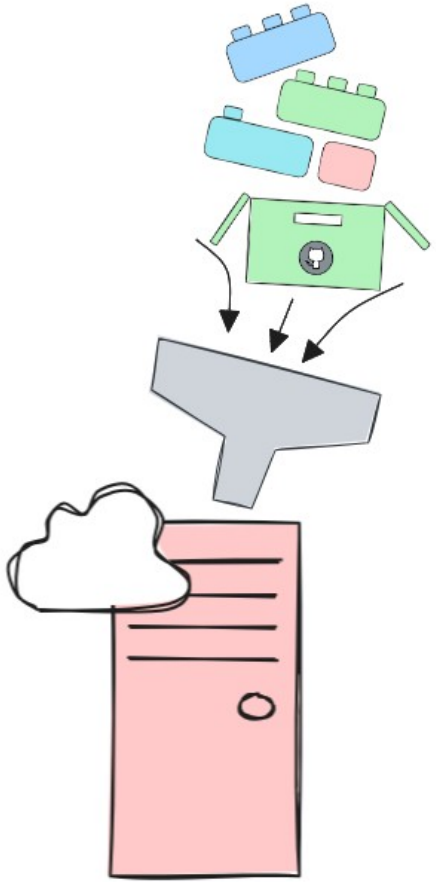
Bergskollegium, Relationer och skrivelser angående utländska bergverk, 1784

Pages 1 2 3

Segmented regions

HTRFLOW-as-a-Service (HaaS)

- En orealiserad idé om att serva våra egna modeller via ett API
- Skulle låta våra egna användare bestämma vilka arkiv som ska digitaliseras
- Kösystem, autentisering, gränser etc.
- Kräver resurser, är det en del av vårt uppdrag?

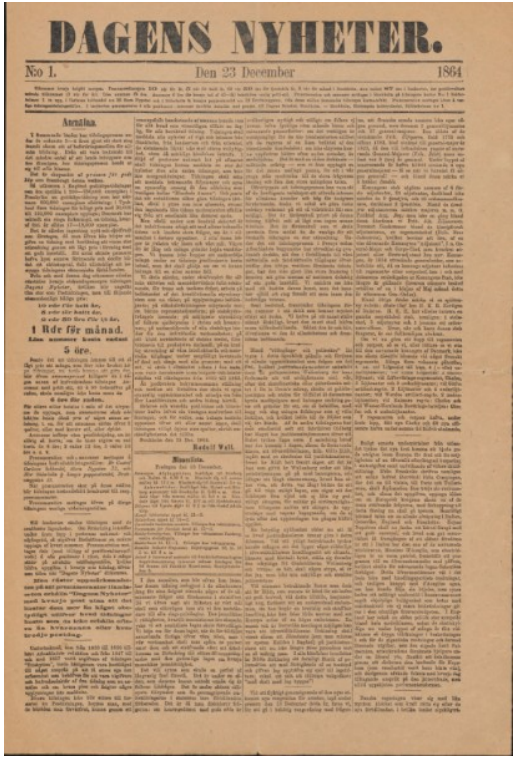


OCR:a om hela kubhist2?

5 miljarder ord 1800-talstext

Kubhist2

- Svenska tidningar 1645-1926
- Skannat av Riksarkivets på uppdrag av KB
- OCR av Kungliga Biblioteket
- 5.5 miljarder tokens
- Referensset
  - Svenska tidningar 1818-1870
  - Svenska tidningar 1871-1906


DAGENS NYHETER.

No 1. Den 23 December 1864

Annalka.

1 Rdr för månad.

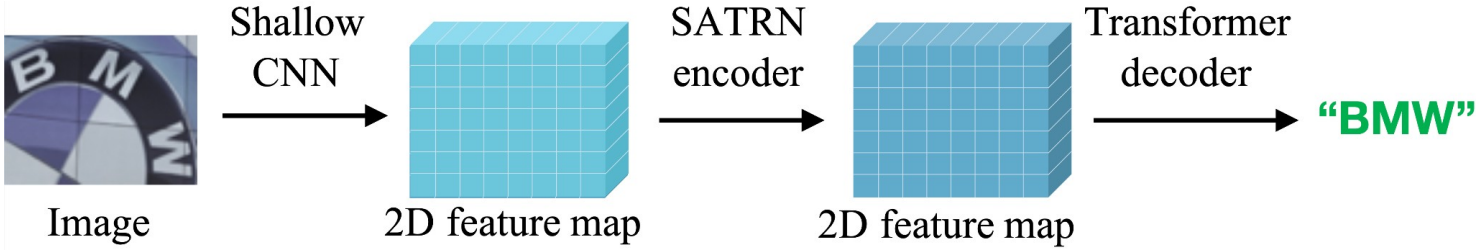
5 öre.

Rudolf Wall.

Minutia.

Vad har vi gjort?

- Bara ett provskott
- Omvandlat referensmängden svenska tidningar 1818-1870 till ett OCR-träningsset (fraktur och antikva)
- SATRN-modell, CER på 0.02



Syntetisk träningsdata

- Utgå från högkvalitativa texter från det diakroniska korpuset
- Passa normaldistribution över radlängd, skiljetecken etc. från ovan nämnda dataset
- Formatera om text enligt dessa distributioner
- Generera ocr-textlinjebilder med fraktur/antikva
- Behöver utvärderas!

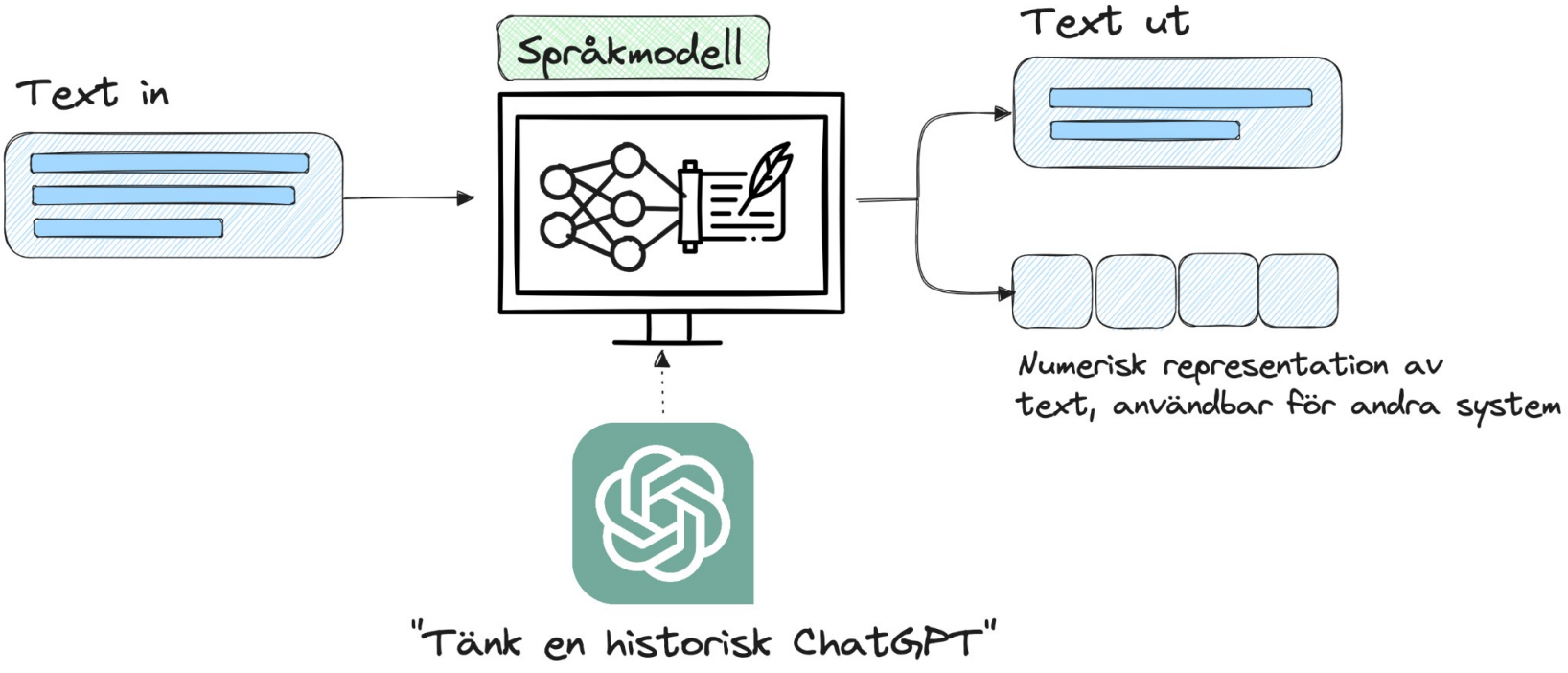
Utmaningar

- Layoutanalys
  - Så automatiserad annoteringsprocess som möjligt
- Läsordning
  - För sammanhängande texter
- Skala upp
  - Optimering av HTRFLOW-Core
  - GPU-inköp
  - Data pipeline

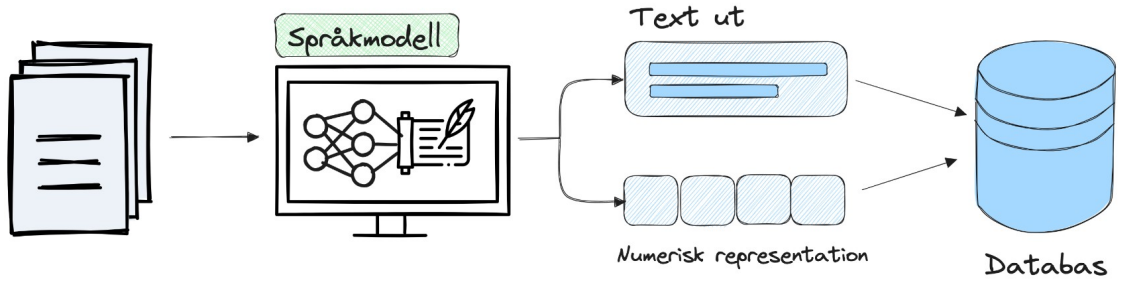
Historiska språkmodeller

Vad vill vi göra med all text?

En historisk Språkmodell



AI-baserad sökfunktionalitet

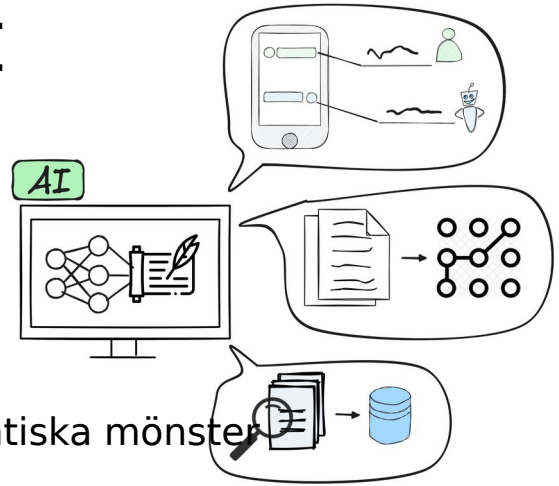


Semantisk sökning!

- Indexerad text + AI = Google mot arkivet
- Semantiskt != Ctrl+f
- Chatapplikation

Användargränssnitt för sök

NLP/NLU för historisk text



- **Hantering av föråldrat språk:** Tolka kontext och avkoda språklig förändring över tid.
- **Förbättrad sökning:** Relatera nutida söksträngar till historisk text semantiskt.
- **Datadriven forskning:** Identifiera statistiska och semantiska mönster i stora textmängder, nya typer av frågeställningar.
- **Social nätverksanalys:** Kartlägg relationer mellan historiska figurer.
- **Influensanalys:** Förstå hur idéer och begrepp sprids och utvecklas över tid.
- **Modernisering av text:** Översätt historiska texter till modernt språk för ökad tillgänglighet.
- **Interaktiva läroplattformar:** Integrera digitala dokument i chattapplikationer, t.ex. chatta med trolldomskommissionen!

Tack för att ni lyssnade!

Riksarkivet