KB-labb

Workshop om språkteknologi för ett flerspråkigt Europa

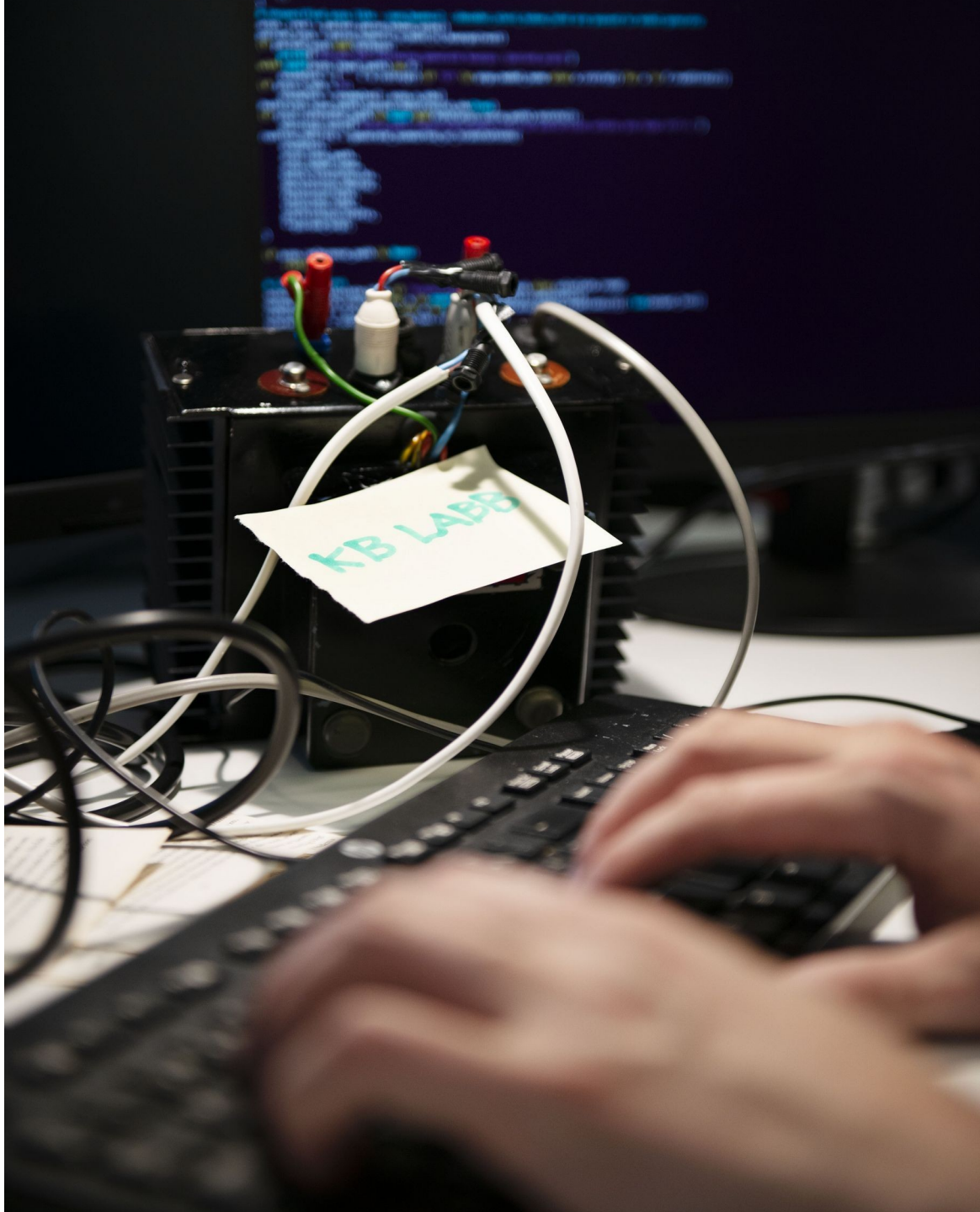
Leonora Vesterbacka 16/5 2024

Kungliga Biblioteket

- Samlar in, bevarar, beskriver och tillgängliggör det svenska kulturarvet
- Pliktlagen omfattar böcker, dagspress, vardagstryck, radio, TV, podcasts, datorspel mm.
- 18 miljoner objekt

KB-labb

- Nationell forskningsinfrastruktur
- Svensk språkmodellsfabrik möjliggjord av
  - laglig tillgång till de största samlingarna svensk text-och ljuddata
  - egen beräkningskapacitet samt tillgång till några av världens bästa superdatorer via EuroHPC JU



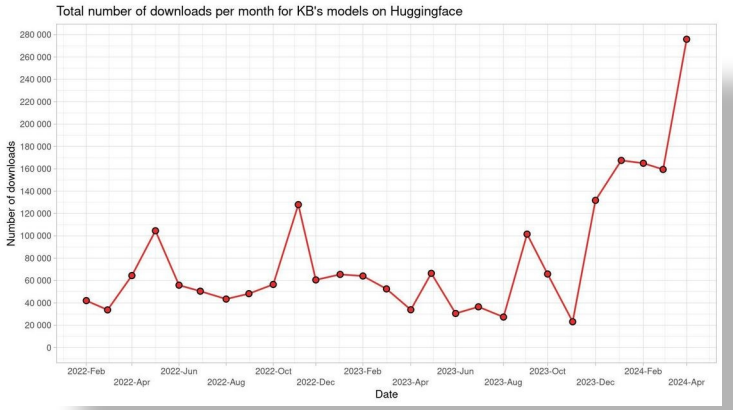
Hur tränar KB-labb modeller på svenska?

- Vi tränar modeller utvecklade av tech-jättarna (Meta, Google, OpenAI...)
  - men ger dessa **svensk data** istället
  - främst s.k. prediktiva

National Library of Sweden

Hur tränar KB-labb modeller på svenska?

- Vi tränar modeller utvecklade av tech-jättarna (Meta, Google, OpenAI...)
  - men ger dessa **svensk data** istället
  - främst s.k. prediktiva
- Modellerna publiceras på huggingface och **laddas ned tusenals gånger** i månaden



Skatteverket

Bolagsverket

Institutet för språk och folkminnen

Domstolsverket
Sveriges domstolar

Försäkringskassan

Finansinspektionen

Arbetsförmedlingen
Swedish Public Employment Service

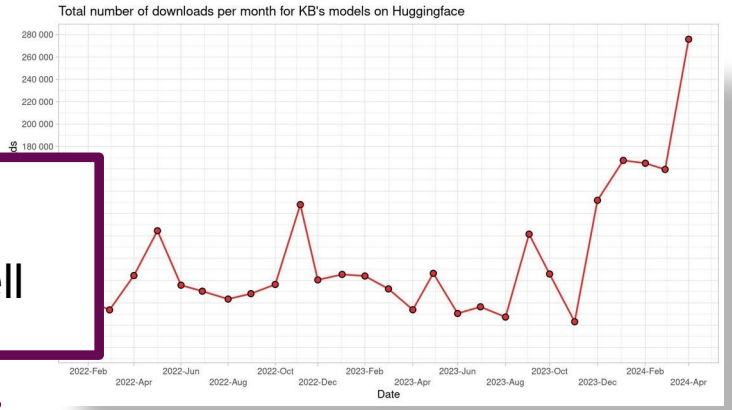
ESV
Ekonomistyrningsverket

tillväxtverket

Kungl. Biblioteket
National Library of Sweden

Hur tränar KB-labb modeller på svenska?

- Vi tränar modeller utvecklade av tech-jättarna (Meta, Google, OpenAI...)
  - men ger dessa **svensk data** istället
  - främst s.k. prediktiva
- Modellerna publiceras på huggingface och **laddas ned tusenals gånger** i månaden



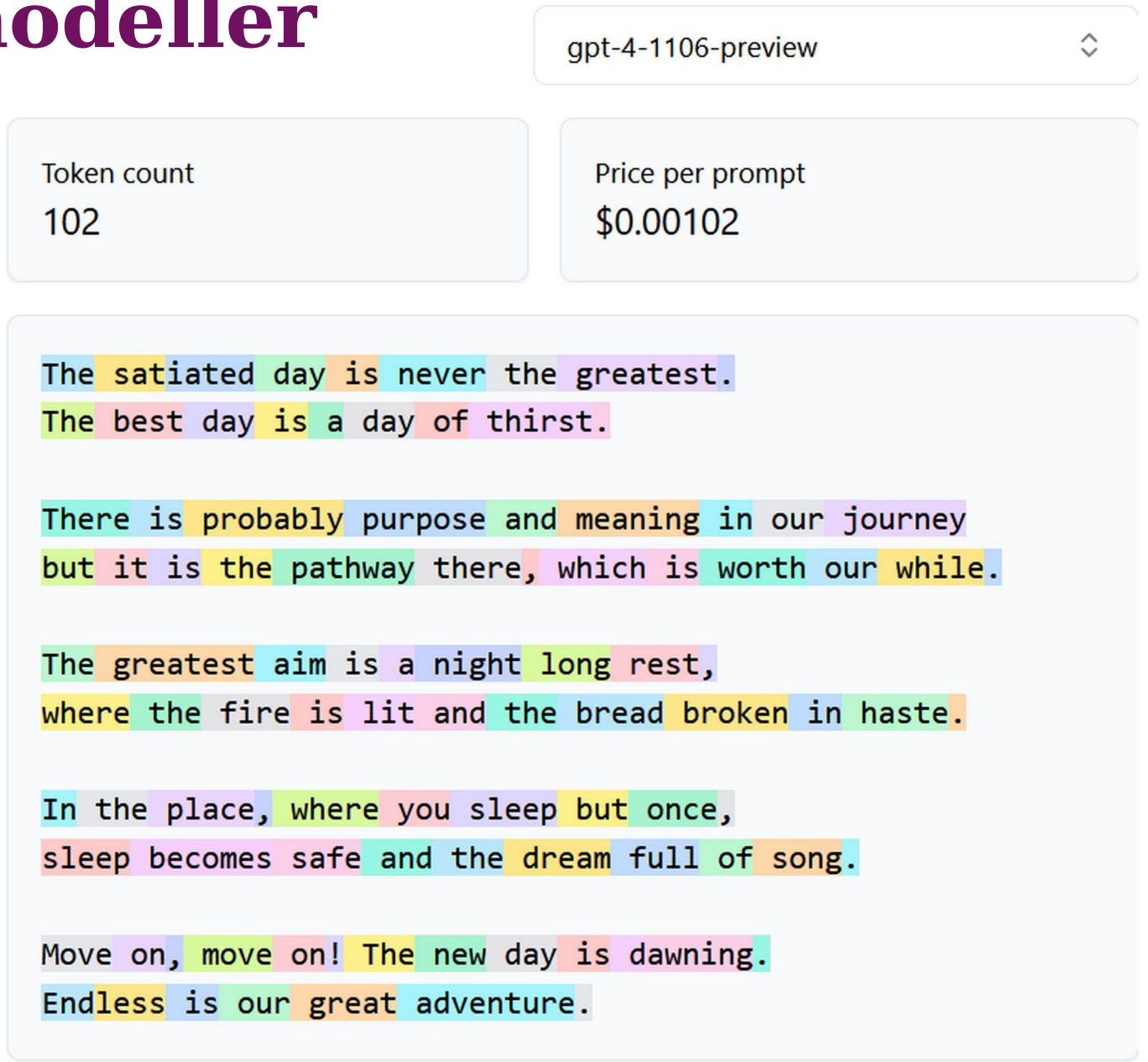
KB-labbs NER-modell

[Kungliga biblioteket]^{ORG} kallas oftast för [KB]^{ORG} och är [Sveriges]^{LOC} nationalbibliotek.

Skatteverket
Bolagsverket
Institutet för språk och folkminnen
Domstolsverket
Sveriges domstolar
Försäkringskassan
Finansinspektionen
Arbetsförmedlingen
Swedish Public Employment Service
ESV
Ekonomistyrningsverket
tillväxtverket
Kungl. Biblioteket
National Library of Sweden

Varför behövs språkmodeller på svenska?

- Tokeniserare i stora språkmodeller är optimerade för engelska, eller massiv flerspråkighet.



Varför behövs språkmodeller på svenska?

- Tokeniserare i stora språkmodeller är optimerade för engelska, eller massiv flerspråkighet.

Varför behövs språkmodeller på svenska?

- Tokeniserare i stora språkmodeller är optimerade för engelska, eller massiv flerspråkighet.
- GPT4 är 35 % till 50 % dyrare att använda på svenska.

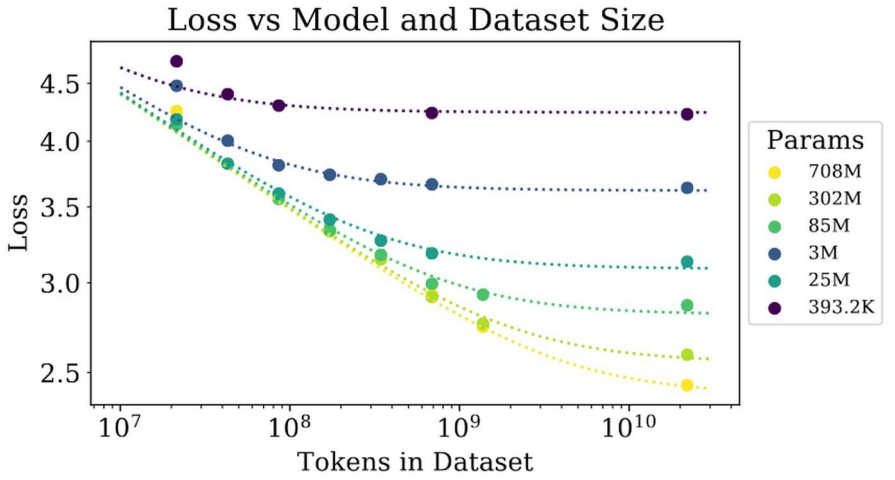
Varför behövs språkmodeller på svenska?

- Tokeniserare i stora språkmodeller är optimerade för engelska, eller massiv flerspråkighet.
- GPT4 är 35 % till 50 % dyrare att använda på svenska.
- Samhällbärande myndigheter efterfrågar
  - transparens och spårbarhet
  - modeller som kan hantera känslig data

I pipelinen: KB-labbs stora språkmodeller på svenska (+ norska & danska)

- Tillsammans med Norska och Danska nationalbiblioteken planerar vi
  - **uppdaterade uppsättning av alla modelltyper:** Encoder (BERT), Encoder-Decoder (T5, UL2), Decoder (Llama, Mistral, GPT-liknande modeller)
  - **användbara modeller** i en modellstorlek som folk kan köra lokalt på sina arbetsstationer
- **BERT:**
  - ny version med längre kontextlängd (512 → 2048)
  - uppdaterade embeddingmodeller för semantiskt sök + RAG
- **T5/UL2:**
  - användbara för maskinöversättning, grammatical error correction, komponent i multimodala modeller
- **Llama/Mistral:**
  - först och främst upp till 7 miljarder parametrar
  - utforska instruktionsfinträning

Stora modeller drar nytta av mer data

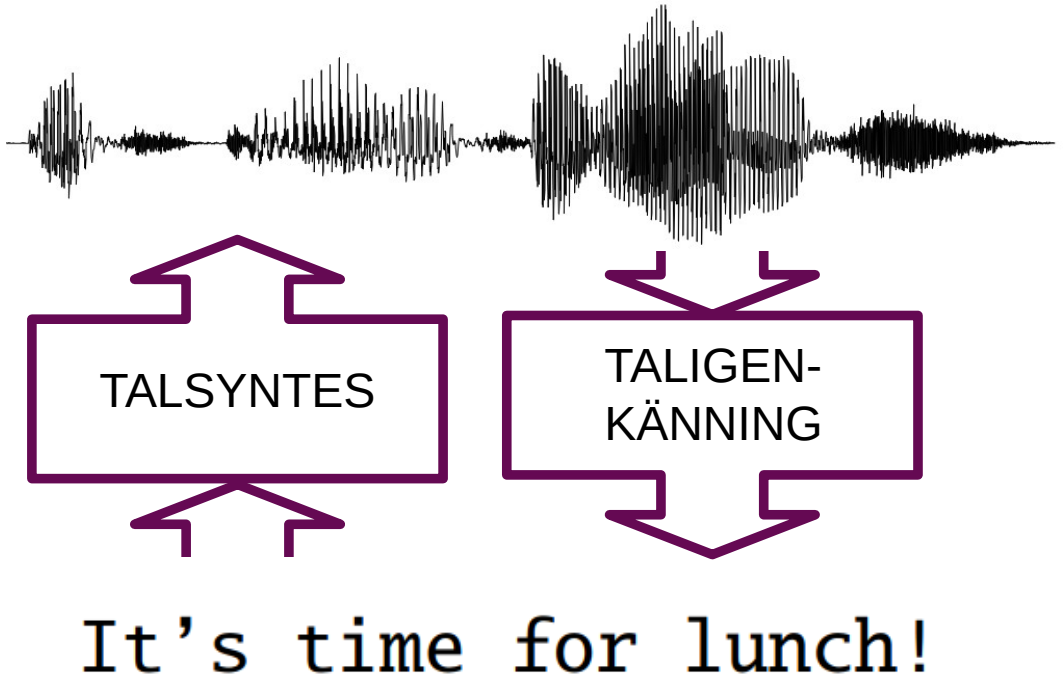


Scaling Laws for Neural Language Models (Kaplan et. al, 2020)

National Library of Sweden

I pipelinen: KB-labbs tal-till-textmodeller på svenska

- Matchar en **vågform till en textsträng**
- Modeller utvecklade av Meta och OpenAI
  - **Wav2vec2.0 och Whisper**
  - enormt effektiva för tal-till-text
    - har dock tränats på en **mindre mängd svenska** av oklart ursprung
    - därmed svårt med **svenska dialekter**



Model	Company	Total dataset [h]	Only Swedish [h]
Wav2vec2.0 (XLS-R)	Meta	436 000	16 325
Whisper	OpenAI	680 000	2119

KUNGL. BIBLIOTEKET
National Library of Sweden

Träningsdata för tal-till-textmodeller på svenska

- Radio och TV från KBs samlingar
  - allt som sänds levereras till KB tack vare **pliktlagen**
- **Lokalradio (P4)**:
  - grundträningsmaterial för wav2vec2.0
- **Svensk TV med svenska undertexter**:
  - finträningsmaterial för Whisper

National Library of Sweden

Träningsdata för tal-till-textmodeller på svenska

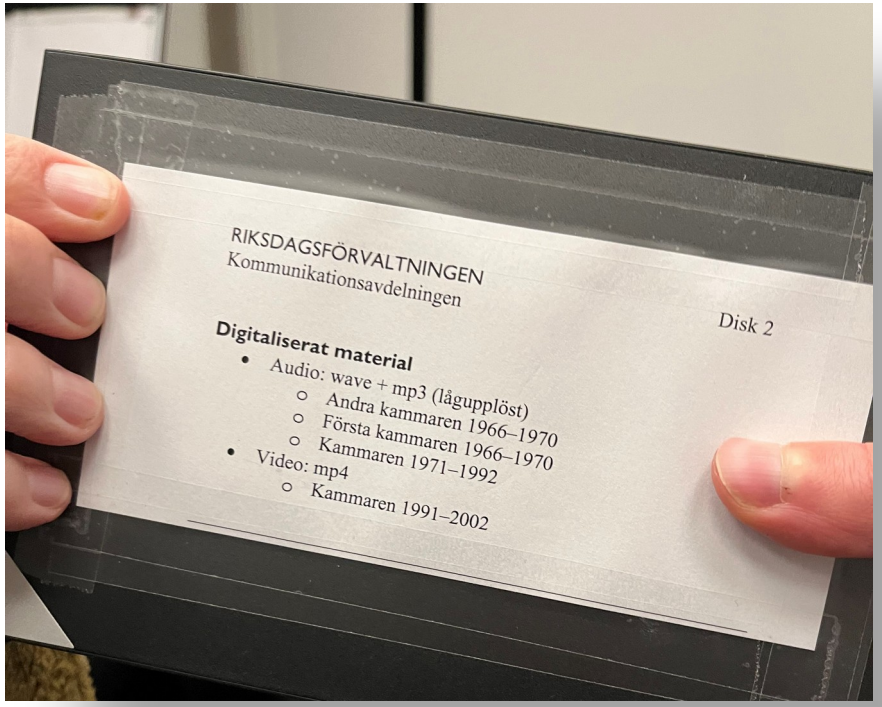
- **SVT** levererar inte undertexter till KB
  - däremot är de intresserade av bättre tal-till-textmodeller för att **automatisera deras undertextning**
  - Vi samarbetar därmed och KB får ta del av annoterat material med information om dialekter, brytningar, talares ålder mm.

svt

KUNGL. BIBLIOTEKET
National Library of Sweden

Träningsdata för tal-till-textmodeller på svenska

- Riksdagsförvaltningen tillgängliggör inspelningar och protokoll för alla anföranden i Riksdagen
  - även de är intresserade av att **automatisera undertextning** och **underlätta protokollförande** mha tal-till-textmodeller
  - vi har tillgång till anföranden från 1966
    - 15 000 - 20 000 timmar transkriberat tal

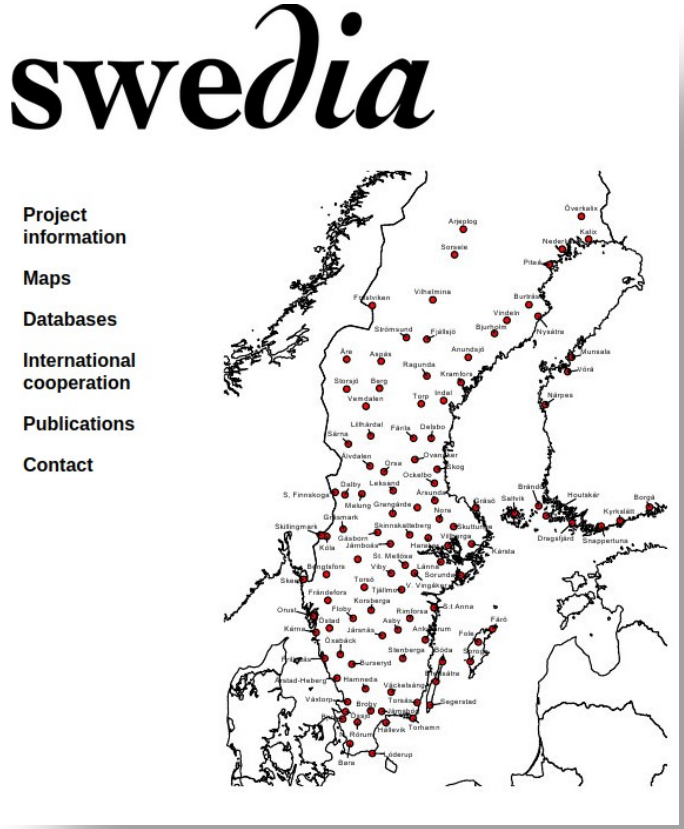


SVERIGES RIKSDAG

KUNGL. BIBLIOTEKET
National Library of Sweden

Träningsdata för tal-till-textmodeller på svenska

- Institutet för språk och folkminnen har ljudarkiv med dialekter och minoritetsspråk
  - Otranskriberat material:
    - grundträningsmaterial för wav2vec2.0
  - Transkriberat material:
    - Finträningsmaterial för wav2vec2.0 och whisper



Institutet för språk och folkminnen

KUNGL. BIBLIOTEKET
National Library of Sweden

Träningsdata för tal-till-textmodeller på svenska

- Institutet för Språk och Folkminne har ljudarkiv med dialekter och minoritetsspråk
  - Otranskriberat material:
    - grundträningsma
  - Transkriberat mate
    - Finträningsmate whisper

Start på finträning av Whisper efter sommaren, stay tuned!

Institutet för språk och folkminnen

KUNGL. BIBLIOTEKET
National Library of Sweden

Tack!

leonora.vesterbackaolsson@kb.se

https://huggingface.co/KBLab

Lag om ändring i lagen (1960:729) om upphovsrätt till litterära och konstnärliga verk

- Lagändring enligt Europaparlamentets och rådets direktiv (EU) **2019/790** av den 17 april 2019 **om upphovsrätt och närstående rättigheter på den digitala inre marknaden** och om ändring av direktiven 96/9/EG och 2001/29/EG, i den ursprungliga lydelsen.
- 15 a §4 Den som har **lovlig tillgång till ett verk** får framställa exemplar av verket **för text- och datautvinningsändamål**. Exemplaren får inte behållas längre än vad som är nödvändigt för ändamålet och får inte användas för andra ändamål. Första stycket **gäller inte om upphovsmannen på lämpligt sätt har förbehållit sig den rätt som avses där**.
- 15 b § **Forskningsorganisationer**, sådana **bibliotek** och museer som är tillgängliga för allmänheten, arkiv samt institutioner för film- eller ljudarvet får framställa exemplar av verk **som de har lovlig tillgång till**, dock inte datorprogram, **för att utföra text- och datautvinning för forskningsändamål**. Exemplaren får inte behållas längre än vad som är nödvändigt för ändamålet och får inte användas för andra ändamål. Exemplaren ska lagras på ett sätt som hindrar obehörig användning. Första stycket hindrar inte att upphovsmannen vidtar proportionerliga åtgärder för att säkerställa integritet och säkerhet i nätverk och databaser som innehåller verk. Avtalsvillkor som inskränker rätten att använda verk enligt denna paragraf är ogiltiga.