Transkribering
2024-05-23

Arkivpodden

Transkribering av AI och arkiven, avsnitt 1

Medverkande: Erik Lenas och Andrëa Grängsjö

Produktion och musik: Copyfabriken. Musiken kommer från DayFox via Pixabay och är ”royalty free”.

[Musik]

Erik Lenas: Jag skulle säga att Riksarkivet, när det gäller AI, har en ganska unik position i Sverige, för vi är vad man kallar en data-hub.

Vi har enormt mycket data. Och all artificiell intelligens och alla tillämpningar inom artificiell intelligens drivs av data.

Andrëa Grängsjö: Välkommen till Arkivpodden från Riksarkivet.

Jag heter Andrëa Grängsjö och med mig i studion har jag Erik Lenas, som är Lead Data Scientist på myndigheten. Vi ska prata om AI och arkiven.

Erik, vad gör en Lead Data Scientist?

Erik: Ja, jag får nog växa in lite i den rollen först innan jag kan svara definitivt på det. Det är en ganska ny roll för mig. Men man kan väl säga att en Lead Data Scientist har ett strategiskt ansvar, delverkar i det tekniska arbetet, i projektplanering och även i resursfördelning. Men exakt hur rollen kommer att utformas, det kanske vi vet bättre om ungefär ett halvår.

Andrëa: Vad spännande! Hur kommer det sig att du är intresserad av AI?

Erik: Jag började plugga datavetenskap ganska sent faktiskt, när jag var 36 år. Och jag märkte direkt att jag drogs till det teoretiska, till algoritmer och till matematik. Och det är vad AI bygger på, kan man säga. Så jag drogs till AI tidigt och inriktade mig själv på det.

Andrëa: Vad spännande. Men hur passar det intresset ihop med Riksarkivet?

Erik: Från början är jag litteraturvetare. Så jag har ett starkt intresse för humaniora. Och den här tjänsten på Riksarkivet verkade perfekt för mig. För där fick jag chansen att kombinera mitt intresse för AI och mitt intresse för humaniora. Och humanistisk forskning.

Andrëa: Vilken tur! Vilken match!

Om vi nu går över till ämnet för dagen, AI och arkiven. Du jobbar på Riksarkivets AI-labb. Men varför jobbar Riksarkivet med AI?

Erik: Jag skulle säga att Riksarkivet när det gäller AI har en ganska unik position i Sverige. För vi är vad man kallar en data-hub.

Vi har enormt mycket data. Och all artificiell intelligens och alla tillämpningar inom artificiell intelligens drivs av data. Vi har autentisk data, vi har intressant data och vi har unik data för Sverige och för hela världen.

Andrëa: Wow!

Men hur mycket data har vi egentligen, vet du det?

Erik: Vi har ungefär 270 miljoner skannade bilder i vårt bildlager. Och det här är endast 5 procent av de dokument vi har i våra arkiv.

Och på detta så kommer den enorma mängd Born Digital-arkiv vi har som vi knappt har börjat utforska än med hjälp av AI. Så vi har väldigt mycket data kan man säga.

Andrëa: Ja, det kan man verkligen säga. Men hur använder vi AI då på myndigheten?

Erik: Man kan säga att vi har två huvudspår. Det ena är effektivisering av vårt uppdrag kan man säga, av vår handledningstid. Att hjälpa till med en intern effektivisering, så kan vi säga.

Och det andra spåret är att tillgängliggöra material för forskning. Och då har vi fokuserat hittills mycket på det historiska materialet, alltså att tillgängliggöra historiska arkiv för humanistisk forskning. Det är de två huvudspåren skulle jag säga.

Andrëa: Ja, men jättespännande. Jag tänker på indexeringsprojekt…

Erik: Ja, jag kan förklara kort idén bakom ett indexeringsprojekt. När man skannar ett stort arkiv, då har vi bara bilderna i så att säga pixelformat. Ett program som visar de bilderna har ingen aning om vad som står på de här bilderna. Alltså är de här bilderna helt osorterade när man skannar dem kan man säga. De går inte att söka i. Det går inte att söka på en text i bilderna och få fram den bilden.

Ett indexeringsprojekt går ut på att hitta någon identifierare, till exempel en fastighetsbeteckning eller ett aktnummer i de här bilderna. Och sen tolka det med hjälp av en handskriftsigenkänningsmodell kanske om de är handskrivna. Och sen knyter vi bilden till den här fastighetsbeteckningen till exempel.

Och nu har vi plötsligt möjlighet att söka efter den bilden med hjälp av en fastighetsbeteckning. Och det är här den här väldiga förkortningen i handledningstid kommer in. För nu kan man snabbt få upp den bild man vill ha och den bild som efterfrågas.

Andrëa: Härligt! Verkligen. Kan du berätta lite om något indexeringsprojekt som pågår?

Erik: Just nu pågår ett stort projekt på Riksarkivet som heter DF-projektet. Och det handlar om att indexera fastighetshandlingar.

Den största delen av det här projektet är skanningen. Det är en dyr process. Vi kommer att skanna 70 miljoner bilder.

Och AI-delen av det här projektet är att indexera upp alla de här bilderna med ett handskrivet aktnummer som finns på bilderna. Så att AI-delen av projektet hittar det här aktnumret och tolkar det med hjälp av en handskriftsigenkänningsmodell.

Andrëa: Det måste ju göra att det går mycket snabbare att hitta.

Erik: Det gör det. Det går mycket snabbare att lösa fastighetsärenden helt enkelt.

Andrëa: Ja, verkligen. Jag tänker på det här med slitage och så på handlingar. När vi skannar in och tillgängliggör det digitalt, då är vi också lite räddare om originalhandlingarna, eller hur?

Erik: Ja, precis. Jag kan ju inte så mycket om bevarandetekniken, men det är många som är väldigt duktiga på det på Riksarkivet. Och det är en jätteviktig del, såklart, att bevara originalen. Men ibland går det inte, kanske. Ibland degraderar dokumenten ändå. Då är det ju viktigt med digitiseringen, det vill säga skanningen.

Det vi gör på AI-labbet är mer vad som händer efter skanningen. Hur vi utvecklar de skannade bilderna till någonting som går att söka i.

Andrëa: Ja, men toppen. Jag tänker på det här med automatisk arkivassistent…

Erik: Ja, det är ju ett ämne som är väldigt stort just idag, med ChatGPT som kom för ett par år sedan. Att skapa automatiska chatbotar, kan man säga, som är tränade på, eller som kan använda sig av företagets eller myndighetens egen data. Det är idén.

Och man kan ju då tänka sig en chatbot som bygger på någon ChatGPT-liknande modell, men som använder sig av Riksarkivets arkivguider eller arkivbeskrivningar för att kunna svara på en fråga på naturligt språk, och kunna berätta direkt svaret på den frågan, det vill säga här och här kan du titta om du vill hitta svaret på din ursprungliga fråga gentemot arkiven. Det skulle ju vara fantastiskt om det kunde funka på ett bra sätt.

Vi har gjort ett Proof of Concept-projekt, men det skulle behöva utvecklas mycket mer för att bli en fullt fungerande tjänst.

Andrëa: Ja, men vad spännande, då är det på gång!

Erik: Absolut!

Andrëa: Haha, det hoppas vi på! Ja, jag tänker på, du har ju talat om nyttan med att använda AI i myndigheten, effektiviseringar och att tillgängliggöra och så. Jag tänkte, det är ju vad vi ser för nytta.

Vad ser vi för nytta för samhället? Vad ser du för nytta för samhället?

Erik: Vad gäller tillgängliggörandet av historiskt material så tror jag att det kan ge oss en annan typ av forskningsfrågor gentemot det här materialet. Det kan ge oss en bättre kontakt med vår skrivna historia genom att erbjuda sökningar direkt mot materialet istället för att man närläser. Det kommer aldrig ersätta närläsning och närstudering av

ett specifikt litet material. Men det kommer erbjuda nya möjligheter till datadriven forskning.

Och vad gäller effektiviseringsprojekten så handlar det om att få ut så mycket som möjligt för de skattepengar vi har hand om. Att kunna fullfölja vårt uppdrag på ett så bra sätt som möjligt.

Andrëa: Ja, men absolut.

Riksarkivets uppdrag är att säkerställa samhällets arkiv och information och göra den möjlig att använda över tid. Vi tar emot och bevarar arkiv. Vi gör arkivinformationen tillgänglig och vi besvarar förfrågningar om arkiven och deras innehåll. Och det här gör vi på många olika sätt.

Men vad menar vi med att tillgängliggöra våra arkiv med hjälp av artificiell intelligens?

Erik: Ja, man kan ju säga att vi tillgängliggör ett historiskt arkiv bara genom att ha det i våra arkiv. Och så kan man gå till läsesalen och så kan man beställa upp det arkivet och läsa själv och hitta det man själv letar efter.

Men det här tillgängliggörande betyder att man måste nästan veta exakt vad det är man letar efter innan man går in i läsesalen och gör sin beställning. Och man måste veta exakt var man ska hitta det.

Det vi tänker är att man kan lägga mycket mer i ordet tillgängliggöra. Man kan dels digitisera det så att det finns skannat. Då behöver man till exempel inte vara i Sverige och gå in i en läsesal utan man kan accessa de här dokumenten globalt via vår bildvisartjänst.

Andrëa: Det är ju bra.

Erik: Det är jättebra.

Sen kan vi göra en handskriftsigenkänning på de här arkiven. Det vill säga att vi omvandlar formatet från att vara en skannad bild till att vara digital text. Vilket då gör att man kan söka mot den här texten.

Och ponera nu att vi gör handskriftsigenkänning på väldigt stora mängder material. Då kommer man kunna söka mot det här materialet. Och vidare om vi tränar historiska språkmodeller. Jag ska förklara det närmare sen.

Men om vi utvecklar vidare AI-baserad funktionalitet så kommer man kunna chatta med det här materialet. Man kommer kunna ställa

direkta frågor på naturlig svenska och få sitt svar och få referenserna visade i det här materialet. Så att det erbjuder nya möjligheter.

Och det är ett annat sätt att se på ordet tillgängliggöra. Man kan lägga mycket i det eller man kan lägga lite i det.

Med det sagt, det är inte lite att ha det i våra arkiv. Det är en väldigt viktig uppgift. En väldigt stor uppgift. Men vi kan göra ännu mer.

Andrëa: Ja, och just det här att kunna nå informationen oavsett tid eller plats är ju också väldigt värdefullt.

Erik: Precis.

Andrëa: Men du sa tidigare hur mycket som finns digitalt.

Erik: Ja.

Andrëa: Och det var fem procent, ungefär.

Erik: Fem procent av våra, precis…

Andrëa: Ja.

Erik: … av våra dokument.

Andrëa: Ja, och då tänker jag, de övriga 95…

Erik: Ja.

Andrëa: Vad är ambitionen med det här? Och vilka begränsningar finns det?

Erik: Vi säger ju på AI-labbet att vår ambition är att HTR:a allt.

Andrëa: Ja.

Erik: Men vi kommer nog antagligen inte göra det. Därför så måste det finnas en prioritetsordning. Och den ska inte vi på AI-labbet sitta och bestämma. Det har inte vi kompetens till.

Utan den prioritetsordningen ska komma uppifrån. Den ska komma från forskarsamhället. Och den ska arbetas fram gemensamt.

Och sen, olika material erbjuder olika svårigheter. Till exempel formulär eller historiska tabeller erbjuder helt andra svårigheter och möjligheter också än till exempel löptext. Så vi kommer att ta det i steg. Det kommer att vara en flerstegsraket, helt klart.

Men ambitionen är att HTR:a så mycket som möjligt. Och tekniken finns för att göra det.

Andrëa: Ja, men det är ju väldigt mycket material. Så man får ju också förstå att det behöver prioriteras och det tar tid. Och självklart så kostar det pengar.

Erik: Det gör det, absolut.

Andrëa: Det behöver vi också tänka på.

Nu pratar vi om alla bra saker med AI. Och det ska vi fortsätta göra. Men jag tänkte, vad kan en AI inte göra?

Erik: Idag finns det väldigt mycket en AI inte kan göra. Om du tänker dig en forskare inom humaniora. En forskare har haft ett helt liv fram tills det ögonblick som är nu. En enhetlig, kontinuerlig upplevelse av världen.

Det är klart det finns diskontinuiteter och kaos i den upplevelsen också. Men det här har skapat en unik världsbild som kommer utifrån en individ. Och det är utifrån den här positionen som man gör sin forskning.

En AI fungerar inte så. Den har ingen sån kontinuerlig, unik erfarenhet av världen. Utan den har ett fågelperspektiv på statistiska samband inom det jätteorganism som är språket.

Så en AI kan inte göra forskningen. Den kan inte formulera de intressanta forskningsfrågorna. Och den kan inte presentera en fråga utifrån den unika världsbild som varje forskare har.

Däremot kan en AI vara en fantastisk hjälp i forskningen.

Andrëa: Ja, det är ju spännande. För varje forskare har ju med sig ett helt liv av upplevelser och erfarenheter. Och AI:n är ju ganska,

vad ska man säga, nyfödd. Och lär sig det som vi talar om för den, att den ska lära sig.

Ja, men spännande, spännande. Men då tänkte jag, vad har vi gjort hittills på myndigheten?

Erik: Vi har genomfört ett ganska storskaligt indexeringsprojekt liknande det här DF-projektet, också med fastighetshandlingar. Där vi indexerade upp sju miljoner akter med fastighetsbeteckningar. Det körs nu, och det är en väldig hjälp i ärendehandläggningen.

Vi har även utvecklat Swedish Lion. Dels genom plattformen Transkribus, men också utanför Transkribus. Vi tillgängliggör våra modeller på Hugging Face.

Och Swedish Lion är en handskriftsigenkänningsmodell som kan tolka historisk löptext från 1600-talet till 1900-talet.

Andrëa: Oj, det är ett rejält spann.

Erik: Ja, det är ett rejält spann. Och just nu håller vi på AI-labbet att utveckla en kodbas, som man kallar det. Det vill säga en applikation, kan man säga. Som man kommer kunna använda i en mängd olika handskriftsigenkänningsprojekt.

Jag kommer att berätta mer om hur det funkar, kanske i nästa poddavsnitt.

Men kortfattat så tränar vi modeller, främst handskriftsigenkänningsmodeller. Som vi använder dels i effektiviseringsprojekt, men också i tillgängliggörandeprojekt.

Och så skriver vi Open Source-kod. Open Source betyder att den kod vi skriver är fritt tillgänglig för alla att använda och att medverka i.

Andrëa: Jättespännande. Det är härligt att höra att det redan kommer till nytta. Det är inte så att det här är någonting som är planerat att om fem eller tio år så ska det trilla ut i ett resultat, utan det här ger resultat längs vägen.

Erik: Ja, det gör det, absolut.

Andrëa: Det är jättespännande.

Erik: Och det är viktigt att det gör det, skulle jag säga. Det är viktigt för vårt självberättigande, att det kommer organisationen till nytta direkt. Och även att det kommer våra användare till nytta, det vill säga allmänheten.

Andrëa: Ja, otroligt. Jag blir glad.

Erik: Ja, härligt.

Andrëa: Då vill jag tacka dig Erik för den här pratstunden.

Och till er lyssnare så vill jag säga missa inte nästa avsnitt, för då får du bland annat veta mer om vårt arbete med handskriftsigenkänning som Erik nämnde, datadriven forskning och varför historiska språkmodeller är viktiga.

Tack för att du har lyssnat på Arkivpodden, podden som vill sprida kunskap, tankar, kultur och öka intresset för arkiv och samhällsviktig information.

Vill du veta mer om Riksarkivet, gå in på riksarkivet.se.

Vi som har varit med i det här avsnittet är Andrëa Grängsjö och Erik Lenas från Riksarkivet.

[Musik]