

Databeredskap för språkteknologiska tillämpningar

Fallstudier från offentlig sektor

Annika Jonsson (Sveriges Kommuner och Regioner)

Tord Karlsson (Arbetsförmedlingen)

Roger Löfgren (Strängnäs Kommun)

Danila Petrelli (AI Sweden)

Oskar Weinberger (Ekonomistyrningsverket)

Detta arbete är ett resultat av projektet "Databeredskapsverkstad" som finansierats av Vinnova (Sveriges Innovationsmyndighet) under bidraget 2021-03630.

Innehåll

Introduktion.....	4
Huvudsakliga lärdomar.....	5
Fallstudier - Sveriges Kommuner och Regioner.....	8
Om SKR.....	8
Fallstudier.....	8
Bakgrund till projekten.....	8
Fallstudie 1. Kommunstyrelseprotokollen.....	9
Om projektet.....	9
Databeredskapsbedömning.....	9
Erfarenheter av att genomföra arbetet i samarbete med konsulter.....	10
Beskrivning och bilder från webbapplikationen (vårt provtryck / PoC).....	11
Lärdomar.....	13
Fallstudie 2: Styrkedjan.....	13
Om projektet.....	13
Databeredskapsbedömning.....	13
Om ett test med annotering.....	14
Lärdomar.....	15
Fallstudie 3: Biblioteksplaner.....	15
Om projektet.....	15
Databeredskapsbedömning.....	17
Lärdomar.....	17
Samlade lärdomar från de tre fallstudierna.....	19
Fallstudier - Granska Annons (Arbetsförmedlingen).....	21
1 Vad handlade projektet om?.....	21
2 Projektmål.....	21
2.1 Prestationsmål Arbetsförmedlingen.....	22
2.2 Mål i Granska Annons (VO-Direkt).....	22
2.3 Mål i projektets textanalys (VO-IT).....	23
2.4 Mätning i Granska Annons.....	23
3 Hur ser lösningen ut?.....	24
3.1 Problemformulering.....	24
3.2 Dataförberedelser.....	25
3.3 Träna modell.....	26
3.4 Utvärdering.....	27
4 Övergripande arkitektur.....	29
5 Verktyg vi använde oss av.....	29
6 Vilka utmaningar har vi stött på?.....	30
6.1 Organisatoriska utmaningar.....	30
6.2 Utmaningar med vårt data.....	31
6.3 Hantering av data.....	31
6.3 Vidareutveckling av modellen.....	31

7	Resultat av databeredskapsmätningar.....	32
8	Vad kan göras för att öka databeredskapen?.....	32
8.1	Förslag till förbättringar kopplat till de frågeställningar som tas upp i verktyget för Databeredskap.....	32
8.2	Andra saker som kan förbättra vår databeredskap.....	35
9	Viktiga lärdomar.....	35
9.1	Lärdomar från att ha arbetat med databeredskapverktyget.....	36
9.2	Ställer arbete inom offentlig förvaltning speciella krav inom området databeredskap?.....	36
9.3	Generella lärdomar från att arbeta med data.....	37
10	Projektets framtid.....	37
	Appendix.....	38
	Mer om off-line och on-line feature store.....	38
	Varför använde vi ett Feature Store?.....	39
	Mer om confusion matrix.....	40
	Vidare läsning.....	41
	Fallstudie - Strängnäs kommun.....	43
1.	Bakgrund.....	43
2.	AI piloten.....	44
3.	Data Readiness Assessment Method.....	45
4.	Framgångsfaktorer för piloten.....	45
5.	Frågor att lösa för att gå vidare.....	46
6.	Flera frågor från metoden.....	47
7.	Flera viktiga frågor.....	47
8.	De största utmaningarna inför kommande projekt.....	48
	Fallstudie - Ekonomistyrningsverket.....	50
	Projektbeskrivning.....	50
	Sammanfattning.....	50
	Bakgrund.....	50
	Databeredskap.....	52
	Datainhämtning.....	54
	Lagring.....	58
	Processering.....	59
	Utvärdering.....	62
	Data readiness assessment.....	63
	Huvudbilaga.....	65

Introduktion

Skriven av Danila Petrelli

(Detta avsnitt skrevs ursprungligen på engelska och maskinöversatt till svenska. Mänskliga granskare har godkänt översättningen).

Betydelsen av kvalitetsdata inom artificiell intelligens och arbetet med databeredskap är uppenbar men ofta underskattad i teknikutvecklingen. För organisationer inom den offentliga sektorn är det en praktisk fråga att se till att data är förberedda och redo för AI-tillämpningar.

Inom ramen för [Data Readiness Lab-projektet](#) som koordineras av AI Sweden presenterar vi en samling fallstudier som undersöker frågorna om datatillgänglighet, validitet och användbarhet, utifrån erfarenheterna från fyra offentliga enheter: Strängnäs Kommun, Sveriges Kommuner och Regioner (SKR), Arbetsförmedlingen och Ekonomistyrningsverket (ESV). De presenterade användningsfallen fokuserar på de utmaningar som härrör från språkteknologi, på engelska Natural Language Processing (NLP), och de gäller därför hanteringen av text i maskininlärnings syfte.

Dessa fall belyser utmaningarna med att skaffa och använda lämplig data. Berättelserna betonar nödvändigheten av ett välstrukturerat datasystem och de praktiska konsekvenserna av att navigera databeredskap. Genom dessa exempel vill vi erbjuda insikter och uppmuntra till genomtänkta diskussioner om bästa praxis, bidra till bättre beslutsfattande i AI-tillämpningar inom den offentliga sektorn och öka medvetenheten om dessa ofta förbisedda frågor.

För att tillhandahålla ett gemensamt ramverk baseras analysen av databeredskap för dessa fyra fallstudier på den metod som föreslagits i artikeln "[We Need to Talk About Data: The Importance of Data Readiness in Natural Language Processing](#)"¹. Denna metod består av en rad frågor, speciellt utformade för att förstärka diskussioner mellan intressenter i ett projekt och lyfta fram kritiska punkter.

Några av de frågor som projektmedlemmarna fick ta itu med är:

Vilken data ska användas?

Vilken data kan användas?

Hur hämtar vi uppgifterna?

Är mängden data tillräcklig?

Den fullständiga listan med frågor som hänvisas till i fallstudierna finns i [Huvudbilagan](#).

De presenterade fallstudierna fokuserar på populära Natural Language Processing-uppgifter, som att träna en språkmodell från en viss uppsättning data för att känna igen tydliga mönster i texter eller till exempel identifiera semantiska likheter.

I Arbetsförmedlingens tillämpning används historiska platsannonser från webben för att träna en språkmodell för kunna identifiera diskriminerande uttryck i nya platsannonser.

¹ Olsson, Fredrik, and Magnus Sahlgren. "We Need to Talk About Data: The Importance of Data Readiness in Natural Language Processing." *arXiv preprint arXiv:2110.05464* (2021).

I Strängnäs kommuns fallstudie används anonymiserade historiska data från journaler i ärenden om Orosanmälan för att träna en språkmodell att känna igen och förutse tecken på våld eller konflikt i nya orosanmälningar.

Ekonomistyrningsverkets (ESV) projekt syftar till att utveckla en tjänst som med hjälp av språkteknologi kan hjälpa myndigheter i deras arbete att effektivt svara på remisser av statliga utredningar genom att rekommendera relevanta tidigare utredningar, förslag och remissvar.

Sveriges Kommuner och Regioners (SKR) fallstudie undersöker om AI kan användas för att samla information från textdokument som finns på kommunala hemsidor som ett alternativ till att skicka ut enkäter.

Innan vi går vidare till den djupgående analysen av varje fallstudie presenterar vi de viktigaste tipsen, som syftar till att främja förståelsen för de återkommande frågorna hos beslutsfattare och praktiker i AI-initiativ inom den offentliga sektorn, även om liknande överväganden kan gälla för andra intressenter i den privata sektorn och akademien.

Huvudsakliga lärdomar

Skriven av Annika Jonsson, Tord Karlsson, Roger Löfgren, Oskar Weinberger

- **Text är data.** Det behövs en ökad förståelse inom offentlig sektor att data inte är synonymt med sifferdata utan att även text är data, och att textdata inte är samma sak som textdokument. Textdata behöver extraheras från dokument, som kan ha olika grad av maskinläsbarhet beroende på filformat och olika grad av användbarhet beroende på vilken metadata som texten är taggad med.
- **Lägg tid på valet av data.** Eftersom NLP-projekt arbetar med data kan valet av data för att uppnå ett specifikt projektmål vara det viktigaste i hela projektet och behöver övervägas ur flera aspekter. Om än inte uttömmande så ger *Data Readiness Assessment Method* en bra utgångspunkt.
- **Bedöm tidigt i projektet om, och hur du kan komma åt data.** Det räcker inte att veta att data finns. Om det är ett internt system, kan du komma åt det via ett internt API? Om det är data på webben, finns det ett öppet API eller måste du skrapa data? Det sätt på vilket data är tillgängligt avgör både tidsplanen för projektet och den tid som krävs för att komma åt data.
- **Textdokument i PDF-format har låg grad av maskinläsbarhet.** Textdata behöver extraheras från dokumentet och PDF är inte ett format som underlättar detta. Samtidigt så är PDF ett av de absolut vanligaste formaten för publicering av textdokument i offentlig sektor. När PDF ändå används för att lagra textdata bör det vara i tillgänglighetsanpassat format (PDF/UA) med metataggar som tydligt indikerar textens struktur och uppdelning.

- **Vi behöver låsa upp gammal data ur befintliga system och framtidssäkra ny data.** Att utveckla en kapacitet att hantera befintliga datakällor och extrahera och behandla textdata från dem är ett första steg i god databeredskap. I ett längre perspektiv är det viktigt att också se över och utveckla de processer där ny textdata skapas och lagras så att framtida data är av bättre kvalitet, mer heltäckande, mer tillgänglig, mer maskinläsbar och har mer användbar metadata.
- **NLP-tillämpningar kan inte bli bättre än den data de bygger på.** Avancerade metoder kan bara delvis och i vissa fall kompensera för brister i datakvalitet. Med rätt data i tillräcklig kvalitet kan samma problem ofta lösas på betydligt enklare sätt samtidigt som det möjliggör mer avancerade tillämpningar.
- **Spårbarhet och transparens för modellen behövs om den ska användas i offentlig verksamhet.** Det är viktigt att veta vilken data modellen har tränats på. Vilka datakällor som använts och om datan har förändrats på något sätt. Man ska spara källkoden som använts för att bygga modellen samt alla valda konfigurationer. Det gäller vilka hyperparametrar som använts och vilket värde de hade för den gällande modellen. Det måste också vara tydligt vilka mätvärden (metrics) som har valts för att utvärdera modellen. Man ska närsomhelst kunna träna om modellen och den ska ge samma resultat.
- **Ha realistiska förväntningar.** Om din organisation är ny inom området är det viktigt att skapa en gemensam förståelse med intressenterna om realistiska förväntningar på omfattningen och resultatet av ett språkteknologiprojekt.
- **Uppdatering av projektmål är en naturlig del i explorativa projekt.** Det är viktigt att förstå, och att få förståelse hos intressenterna, att projektmål särskilt i explorativa projekt kan behöva uppdateras flera gånger under ett projekts gång.
- **Sträva efter tydliga avtal.** Eftersom licensavtal om system eller andra datakällor är en verksamhet som ofta saknar tydlig reglering om vad kunden får göra med data utöver de ordinarie användarfunktionerna, är det bra att sträva efter avtal som är tydligare med att annan användning av kunddata också är tillåten.
- **Involvera fler perspektiv än det tekniska.** NLP-projekt innehåller många överväganden utöver de tekniska. Se därför till att uppmärksamma fler perspektiv, till exempel juridiska och etiska frågor. Involvera flera kompetenser i projektet såsom juridiskt stöd och kommunikatörer.
- **Samarbete över myndighetsgränser behövs.** Språkteknologiska tillämpningsmöjligheter och intressanta textdatakällor är ofta inte tydligt kopplade till endast en offentlig aktör. Det har därför hänt, och kommer fortsätta att hända, att olika aktörer bedriver snarlika projekt utan att vara medvetna om varandra. Detta innebär ett ineffektivt dubbelarbete och en missad möjlighet till samverkan. Det behövs därför bättre kommunikation inom det offentliga kring vilka språkteknologiska projekt som drivs och planeras samt bättre styrning och ökade medel för gemensamma projekt.

New section follows

Fallstudier - Sveriges Kommuner och Regioner

Skriven av Annika Jonsson

Om SKR

Sveriges Kommuner och Regioner (SKR) är en medlems- och arbetsgivarorganisation där alla Sveriges kommuner och regioner är medlemmar. SKR:s uppgift är att stödja och bidra till att utveckla kommuner och regioners verksamhet.

SKR verkar i den offentliga sektorn men till skillnad från kommuner och regioner som är myndigheter så definieras SKR juridiskt, som en ideell organisation. Detta gör att situationen när det gäller frågor om data och databeredskap kan skilja sig åt mellan SKR och organisationer som definieras som myndigheter.

Fallstudier

Vi har under åren 2020-2022 gjort tre piloter där det övergripande syftet varit att analysera text med hjälp av AI i policydokument publicerade i pdf på kommunernas webbplatser.

Bakgrund till projekten

Antalet enkäter och andra uppgiftssamlingar som görs riktade till kommuner och regioner har ökat trendmässigt sedan 2000-talet.

Det är en tidskrävande arbetsuppgift för kommuner och regioner att besvara enkäter och de får heller ingen ekonomisk ersättning för detta. En av orsakerna till det ökade antalet enkäter (och därmed den ökade uppgiftslämnarbörda för kommuner och regioner) kan vara övergången från pappersenkäter till webbenkäter. Det har gått från att vara relativt kostsamt att göra en pappersenkät (med kostnader för tryck, kuvert, porto, stansning etc.) till nästan gratis att genomföra en enkät via webben sett till avsändaren av enkäten. För den som besvarar enkäten är tidsåtgången och kostnaden däremot densamma oavsett om det är en pappers- eller webbenkät. Detta gör att utvecklingen med det ökade antalet enkäter knappast heller lär vända.

SKR har enligt SFS (1982:668) en roll i att företräda kommuner och regioner i samråd om uppgiftssamlingar och att också minska uppgiftslämnarbördan.

Parallellt med den trendmässiga ökningen av antalet uppgiftssamlingar har samtidigt webben blivit en självklar plats för att dela information. Kommuner och regioner publicerar till exempel många av sina styrdokument och policies på sin webbsida. Denna semistrukturerade information ville vi undersöka om man kunde samla in, strukturera och analysera med hjälp av AI för att om möjligt erbjuda ett alternativ till att göra en uppgiftsinsamling med hjälp av en enkät.

Vi valde att fokusera på protokoll och policydokument såsom program, policy, strategi, handlingsplan och riktlinjer.

Alla kommuner har inte alla typer av policydokument och de kan heta olika saker men många kommuner har samma typ av policydokument. En del av kommunerna har samlat alla policydokument på en sida med en tydlig definition av respektive dokumenttyp. Andra har inte samma tydliga struktur.

Ett exempel på en kommun som har alla policydokument samlade på en sida och en förklaring av vad olika policydokument innehåller är Örebro kommun. På länken nedan finns Örebros samlade policydokument – ett hundratal dokument i pdf.

<https://www.orebro.se/kommun--politik/dokument-statistik--undersokningar/dokument--undersokningar.html>

Fallstudie 1. Kommunstyrelseprotokollen

Om projektet

Vi ansökte till Vinnovas utlysning *Starta er AI-resa* för att brett börja utforska den frågeställning vi hade: om det skulle gå att samla in textdokument från webben och besvara frågor med hjälp av dem istället för att ställa frågor i enkäter. I samband med ansökan till utlysningen fick vi kontakt med en konsultfirma verksam inom tillämpad AI. Vinnova-ansökan fick vi avslag på, men vi valde ändå att gå vidare och utforma ett pilotprojekt tillsammans med konsultfirman.

Projektet påbörjades strax efter Corona-pandemins utbrott.

Vii ville undersöka brett om vi kunde extrahera information från kommunstyrelseprotokollen och för att se på vilket sätt kommunens olika verksamheter påverkades av pandemin, löpande. Tanken var också att kunna föregripa de förväntade enkäterna med uppföljning om hur kommunerna hanterat pandemin och därmed minska uppgiftslämnarbördan för kommunerna. Piloten var vårt första försök att praktiskt undersöka vår frågeställning om man istället för ett enkätutskick kunde få svar på sina frågor genom att söka information i de dokument som kommunerna redan publicerat på sina webbplatser.

Frågeställningen vi ville undersöka var alltså om vi kan hitta andra metoder än enkäten, för att få svar på de frågor som ställs till kommunerna genom att nyttja den information som finns tillgänglig på kommunernas webbplatser?

Databeredskapsbedömning

När projektet påbörjades hade vi inte kännedom om ramverket för *Data Readiness Assessment*. Bedömningen gjordes istället i efterhand i arbetet inom Vinnovaprojektet *Databeredskapsverkstaden*.

Kommunstyrelseprotokollen ligger publicerade på kommunernas webbplatser, totalt 290

stycken och vi hade inte programmatisk tillgång till data (Q1), utan använde oss av webbskrapning för att samla in dem.

De flesta kommunstyrelseprotokoll publiceras i pdf-format och innehåller i huvudsak endast text (Q5). Ett undantag utgjorde de protokoll som skannats in och därmed blir en bild vilket gör det svårare att extrahera text.

Den explorativa och deskriptiva analysen av data som gjordes var att följa upp webbskrapningen (Q6). Hur många av de 290 kommunerna webbskrapade vi in dokument ifrån? Hur många dokument blev det totalt och hur fördelades de över datum och över kommuner var några av frågorna som besvarades. Förekomst av dubletter (av nästan identiska/reviderade) kommunstyrelseprotokoll identifierades också i den explorativa analysen (Q7).

Initialt ville vi avgränsa webbskrapningen till att endast omfatta kommunstyrelseprotokoll från 2020 och framåt men det visade sig vara svårt att lösa bland annat eftersom datum-formaten på protokollen saknar standard. Vi hade alltså mer än tillräckligt med data för vår fallstudie. (Q10).

Eftersom projektet bygger på dokument som samlats in med hjälp av webbskrapning från 290 kommuner så hade det varit svårt för SKR att hantera mer data utanför det här pilotprojektet (Q12). Kommunerna uppdaterar med nya kommunstyrelseprotokoll löpande (varannan vecka är en vanlig kadens) och utöver det förekommer också löpande förändringar av kommunernas webbplatser där kommunstyrelseprotokollen som gör att webbskraporna måste uppdateras.

Erfarenheter av att genomföra arbetet i samarbete med konsulter

Vi anlät en konsultfirma för att utföra arbetet med webbskrapning, textanalys och för att bygga en webbapplikation som proof-of-concept (PoC).

Detta var nödvändigt eftersom vi inte hade tillgång till den kompetens som krävdes för att genomföra uppdraget internt i organisationen. En av utmaningarna med att upphandla konsulter för att genomföra ett sånt här projekt var framförallt att i förväg formulera ett skriftligt uppdrag i ett undersökande projekt där det var svårt att på förhand veta vad som skulle funka och hur lång tid olika delar i projektet skulle ta.

Säkerligen var detta en utmaning för konsulterna också, framförallt kanske att uppskatta tidsåtgången.

En annan utmaning är kommunikationen där teknisk expertis och övrig expertis ska förstå varandras synvinklar för att uppnå bästa möjliga resultat. Det krävs att beställaren kan förklara behoven, och konsulten behöver förstå de tekniska möjligheter och utmaningar som finns specifikt för fallstudien, beskriva dessa i tekniska termer, och i sin tur kommunicera detta till beställaren.

Ytterligare en utmaning är att äga resultaten av projektet när projektet är avslutat och vi inte längre har tillgång till den tekniska expertisen. Det kan delvis hjälpa med skriftlig dokumentation, men alla frågeställningar är inte lätta att hitta svar på i efterhand.

En självklar fördel med att anlita konsulter är att man kan få specialiserad kunskap under en begränsad tid och investering, och som i det här fallet möjliggöra ett projekts genomförande.

Beskrivning och bilder från webbapplikationen (vårt provtryck / PoC)

Skärmbild 1

Träffar i kommunstyrelseprotokollen på sökorden covid, corona och pandemi.

Search results for "covid corona pandemi".

Protokoll: 700, Antal meningar: 5457

Document Title	Date	Size	Pages	Actions
Kalmar län - Emmaboda/2020/593 - 2020-11-23.pdf	2020-11-23	937 kB	14	Preview, Download
Kalmar län - Emmaboda/2020/590 - 2020-11-23.pdf	2020-11-23	928 kB	14	Preview, Download
Värmlands län - Hagfors/2020/48863 - 2020-05-21.pdf	2020-05-21	370 kB	16	Preview, Download
Dalarnas län - Gagnef/2020/71590 - 2020-11-11.pdf	2020-11-11	160 kB	27	Preview, Download
Stockholms län - Tyresö/1990/15146 - 1990-01-01.pdf	1990-01-01	700 kB	16	Preview, Download
Stockholms län - Tyresö/1990/15307 - 1990-01-01.pdf	1990-01-01	744 kB	16	Preview, Download
Västmanlands län - Sala/2020/59690 - 2020-08-31.pdf	2020-08-31	3.43 MB	53	Preview, Download
Kalmar län - Vimmerby/2020/723 - 2020-11-03.pdf	2020-11-03	1 MB	46	Preview, Download

Skärmbild 2

Träffar i kommunstyrelseprotokollen på sökorden covid, corona och pandemi illustrerat i ett diagram.



Skärmbild 3

Enskilda meningar i kommunstyrelseprotokollen med träffar på sökorden covid, corona och pandemi illustrerat i ett diagram.

Länsstyrelsen bedömer att **följderna av pandemin** fortsätter över lång tid

Samtidigt har Kalmar län färre **allvarligt sjuka covid -19 patienter inlagda** på länets sjukhus jämfört mot de senaste veckorna

Sedan tidigare finns oro för att **coronakrisen** bidrar till **oönskade sociala konsekvenser** såsom ökad brottslighet, missbruk och psykisk ohälsa

Pågående arbeten och insatser **Provtagning för covid -19** som omfattar personal inom hälso- och sjukvård, kommunal vård och omsorg samt personlig assistans startade måndagen 25 maj

Arbetet fortsätter med att analysera och planera för **hur coronakrisen kan påverka andra parallella krishändelser** i länet

I och med att **Covid 19** kom **skärptes riktlinjer och rutiner** upp kring vårdhygien

Lärdomar

En stor utmaning i projektet som tog lång tid i anspråk var att samla in kommunstyrelseprotokollen då de ligger publicerade på kommunernas enskilda webbplatser, totalt 290 stycken.

Vi samlade in protokollen med hjälp av webbskrapning och även om många webbsidor liknar varandra i teknisk mening behövdes ändå ett antal skraddarsyddas webbskrapor byggas.

Om vi hade velat driftsätta webbapplikationen hade lösningar varit svår att förvalta när nya protokoll publiceras löpande samtidigt som kommunernas webbsidor uppdateras med jämna mellanrum.

Initialt ville vi avgränsa insamlingen av kommunstyrelseprotokoll från februari till april 2020. Men det var svårt att begränsa vilka dokument som skrapades hem med hjälp av datum på dokumenten då dessa inte är standardiserade. En annan svårighet var dubletter med reviderade dokument.

De flesta protokoll publiceras i PDF vilket till viss del gjorde det utmanande att extrahera texten.

Personuppgifterna i form av mötesdeltagarnas namn och partitillhörighet utgjorde också ett hinder.

Vårt slutresultat i vår PoC/webbapplikation var en applikation där kommunstyrelseprotokollen vara sökbara ned på meningsnivå (se skärmbilderna). För att webbapplikationen skulle vara användbar hade vi behövt något sätt att aggregera upp resultaten av de enskilda meningarna, exempelvis tematiskt/ämnesvis. Med enkäten som förebild hade sedan velat kunna påstå något generellt utifrån detta. Till exempel: Minst hälften av kommunerna hade under en viss tidsperiod besöksförbud på äldreboendena.

Fallstudie 2: Styrkedjan

Om projektet

Projektet ingick i Vinnovas *Språkdata* labb och SKR var behovsägare.

Att undersöka den så kallade styrkedjan, det vill säga vilka mål kommunen sätter upp för olika verksamheter i en verksamhetsplan och hur de sedan analyserar måluppfyllelsen i årsredovisningen.

Databeredskapsbedömning

Databeredskapsbedömningen utifrån ramverket *Data Readiness Assessment* gjordes i samband med Vinnovaprojektet *Databeredskapsverkstaden* efter att projektet slutförts.

Årsredovisningarna och verksamhetsplanerna ligger publicerade på kommunernas webbplatser, totalt 290 stycken och vi hade inte programmatisk tillgång till data (Q1).

Årsredovisningarna och verksamhetsplanerna är allmän handling och vi bedömde därför att vi hade laglig tillgång till dem (Q2).

Personuppgifter förekom i viss utsträckning i årsredovisningarna i form av namn, bilder och foton (Q3).

De flesta årsredovisningar och verksamhetsplaner publiceras i pdf-format och innehåller såväl text, som diagram, tabeller, bilder, illustrationer och spalter vilket vi bedömer skulle ha försvårat maskinläsbarheten om vi gått vidare med projektet (Q5).

Om ett test med annotering

Vi började projektet med att manuellt samla in ett 20-tal årsredovisningar och verksamhetsplaner.

Utifrån insamlade dokument och den frågeställning vi hade skapade vi en annoteringsmall i Excel där följande frågor skulle besvaras:

- 1) Vilka är de uppsatta målen?

Utifrån de olika målen skulle varje mål följas upp med frågorna:

- 2) Sker en uppföljning av målet?
- 3) Har målet uppfyllts?
- 4) Görs en analys varför målet uppfyllts eller ej?

När vi började annotera materialet som vi samlat in från de 20 kommunerna – vi tre personer som jobbade i projektet – insåg vi ganska snart att alltför många frågor lämnades obesvarade för att vi skulle kunna gå vidare med annoteringen.

Målformuleringarna var inte alltid tydliga och framförallt var inte uppföljningen av de uppsatta målen lätt att urskilja i materialet. Exempelvis kunde ett mål vara uttryckt men inte i kvantitativa termer och därför svår att svara på om målet var uppfyllt. T.ex: "alla elever i grundskolan ska få en god utbildning", och inte "alla elever i grundskolan ska ha minst godkänt betyg i matematik, svenska och engelska" Det som oftast saknades var en analys av varför målet uppfyllts eller ej. Det faller sig naturligt då svaret på den frågan också bygger på att man uttryckt ett tydligt mål som är uppföljningsbart, att man faktiskt följt upp målet och att man till sist analyserar utfallet.

Efter den inledande annoteringen kunde vi konstatera att datamaterialet vi hade (årsredovisningar och verksamhetsplaner) inte svarade upp mot de frågeställningar vi hade och vi bestämde oss för att avsluta projektet.

Lärdomar

Det är svårt för behovsägare att identifiera lämpliga användningsfall då ett lämpligt sådant ska uppfylla flera kriterier som innefattar såväl domänkunskap, datatillgång, datakvalitet samt kunskap om vad som är möjligt eller inte möjligt att göra med AI.

Även dokument som man kanske inte förväntar sig ska innehålla personuppgifter kan göra det som till exempel årsredovisningarna.

Dokument som många tror ska finnas lätt tillgängliga på kommunens webbsidor kan vara svåra att hitta eller finns inte publicerade.

Fallstudie 3: Biblioteksplaner

Om projektet

Enligt bibliotekslagen ska alla regioner och kommuner anta biblioteksplaner.

I december 2022 hade 98 procent av kommunerna en aktuell biblioteksplan enligt Kungliga biblioteket (KB). En biblioteksplan har en giltighetstid på cirka 3-4 år. Kommunerna mejlar in sina aktuella biblioteksplaner till KB som publicerar dem samlat på sin webbplats.

[Aktuella biblioteksplaner – Kungliga biblioteket – Sveriges nationalbibliotek – kb.se](https://www.kb.se/samverkan-och-utveckling/biblioteksutveckling/biblioteksplaner.html)

Att samla in biblioteksplanerna med hjälp av webbskrapning var då enklare än i de två föregående projekten. Dels för att dokumenten är publicerade på en webbplats istället för som tidigare 290, och dels för att antalet dokument är begränsat till max 290 dokument, då varje kommun bara kan ha en aktuell biblioteksplan.

<https://www.kb.se/samverkan-och-utveckling/biblioteksutveckling/biblioteksplaner.html>

Personuppgifter förekommer i princip heller inte i biblioteksplanerna och dokumenten består huvudsakligen av text. Tabeller, diagram, bilder och annat som försvårar extraheringen av text förekommer begränsat.

Kungliga biblioteket (KB) har ett särskilt uppdrag att följa upp biblioteksplanernas användning och innehåll. Uppföljningen görs genom att läsa och analysera ett urval av biblioteksplanerna.

Den senaste uppföljningsrapporten *Trender i Biblioteksplanerna 2020* omfattade ett urval av 78 biblioteksplaner antagna under 2020.

Liksom i de två föregående projektet (Kommunstyrelseprotokollen och Styrkedjan) ville vi undersöka om det skulle gå att samla in textdokument från webben och analysera dem med hjälp av metoder inom Natural Language Processing (NLP) för att hitta svar på frågor man normalt skulle ställa i en enkät.

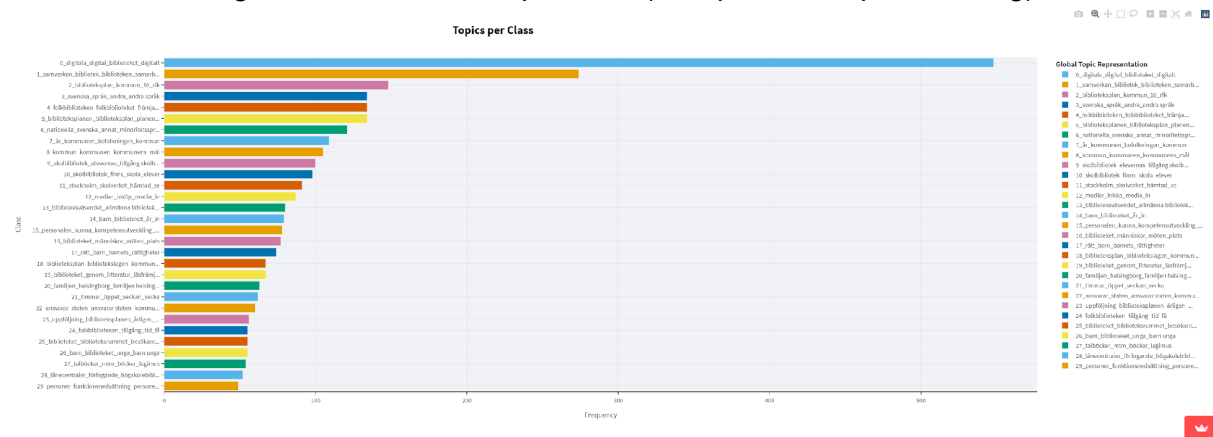
Beskrivning av webbapplikationen

En prototyp av en webbapplikation skapades i Streamlit för att kunna interagera med biblioteksplanerna. Streamlit är ett Python-ramverk för att enkelt skapa webbappar fokuserade på data och AI.

Nedan presenteras tre skärmdokument som beskriver huvuddelarna av funktionaliteten i webbapplikationen.

Skärmdokument 1

Tematisk indelning av texten i biblioteksplanerna (unsupervised topic modeling).

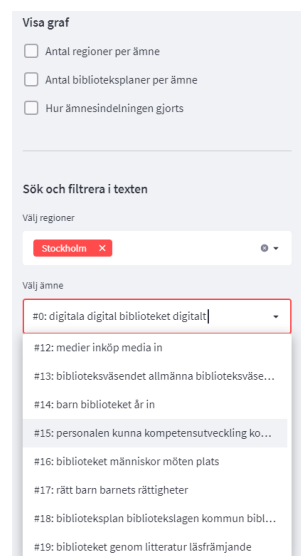


På en övergripande nivå delades biblioteksplanerna in i 29 teman såsom exempelvis tillgång till skolbibliotek, digitalisering, läsfrämjande, nationella minoritetsspråk, medieinköp och öppettider.

Skärmdokument 2

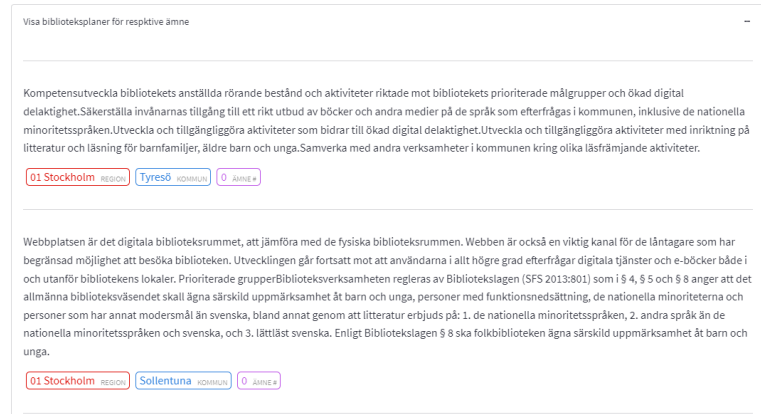
Till vänster: En meny för att välja län och meny för att välja ämne/tema.

Till höger: Textavsnitt ur biblioteksplanerna inom ämnet digitalisering filtrerat på kommuner i Stockholms län.



Texter från varje ämnen

[Link to bottom](#)



Skärmlapp 3

Ett textavsnitt på temat *Tillgång till skolbibliotek*.

Huvudmannen är enligt skollagen ansvarig för att skapa goda förutsättningar för att kvalitativa skolbibliotek utformas samt främja likvärdigheten.

Kommunstyrelsen ansvarar för att skolbiblioteken utformas på ett lämpligt sätt som uppfyller lagens krav. Rektor har enligt läroplanen ansvaret för att skolbiblioteket används i undervisningen för att stärka elevernas språkliga förmåga och digitala kompetens. Idag finns skolbibliotek på Tranemo

Gymnasieskola och Tranängskolan. Övriga skolor har boksamlingar och såväl elever som pedagoger servas av bokbussen som besökskolorna enligt fastställd tidtabell.

14 Västra Götaland REGION

Tranemo KOMMUN

9 ÄMNE #

1.0 P

Databeredskapsbedömning

Databeredskapsbedömningen utifrån ramverket *Data Readiness Assessment* gjordes i samband med Vinnovaprojektet *Databeredskapsverkstaden* efter att projektet slutförts.

Biblioteksplanerna ligger publicerade på Kungliga bibliotekets webbplats och vi hade inte programmatisk tillgång till data (Q1), utan använde oss av webbskrapning för att samla in dem. Biblioteksplanerna är allmän handling och vi bedömde därför att vi hade laglig tillgång till dem (Q2 och Q3).

De flesta biblioteksplanerna publiceras i pdf-format och innehåller i huvudsak endast text (Q5).

En övergripande explorativ och deskriptiv analys gjordes (Q6 och Q7). Resultatet av webbskrapningen följdes upp med hur många av de 290 kommunerna som hade en aktuell biblioteksplan och hur många dokument det blev totalt. För insamlade dokument gjordes också en övergripande validering. Då i stort sett alla kommuners biblioteksplaner fanns publicerade och kunde samlas in var data tillräckligt i kvantitativ mening (Q10).

SKR hade haft möjlighet att hantera mer data utanför det här pilotprojektet (Q12). Varje kommun ska ha en biblioteksplan och de uppdateras ungefär vart tredje eller fjärde. De mejlas in av kommunerna till KB och KB publicerar biblioteksplanerna på sin webbplats.

Lärdomar

I jämförelse med pilotprojektet för Kommunstyrelseprotokollen hade projektet med biblioteksplanerna två fördelar. Den ena fördelen var att alla dokument låg samlade på en

webbplats vilket gjorde webbskrapningen mindre krävande (om än inte problemfri) och den andra fördelen var avsaknaden av personuppgifter.

Jämfört med styrkedjeprojektet så hade vi en bredare och ett mer övergripande syfte med vad vi ville ha ut av applikationen vilket var både en styrka och svaghet.

I praktiken ville vi i biblioteksplaneprojektet brett undersöka om man kunde göra den typ av uppföljning som görs manuellt skulle kunna underlättas med hjälp av metoder inom NLP (även om just denna uppföljning inte görs med hjälp av enkäten som metod).

Frågeställningen i styrkedjeprojektet var specifik på så sätt att man ville koppla ett uppsatt mål i ett dokument till uppföljningen av det målet i ett annat dokument.

Det bredare syftet vi hade i biblioteksplanerna gjorde att vi ändå kunde genomföra projektet och bygga en prototyp trots att våra primära idé att använda fråga-svar (question-answering) som metod inte fungerade. Istället delade vi in textavsnitten i teman med hjälp av en oövervakad ämnesmodellering (unsupervised topic modeling) och visade på vilka teman som dominerar biblioteksplanerna.

Det som vi saknade med den webbapplikation vi tog fram var möjligheten att aggregera resultaten för att kunna göra generella påståenden om ett specifikt tema. Efter att ha byggt denna applikation insåg vi ämnesindelningen var kraftfull men bara första steget mot att få en övergripande bild av biblioteksplanerna. I praktiken saknas flera steg för att det ska vara användbart.

En annan sak som hade varit intressant att undersöka är skillnaden mellan vad som kommunerna förväntas skriva om i biblioteksplanerna jämfört med vad de faktiskt skriver om.

Vad kommunerna förväntas skriva om i biblioteksplanerna finns beskrivet i KB:s vägledning: <https://www.kb.se/samverkan-och-utveckling/biblioteksutveckling/biblioteksplaner/vagledning.html>

En annan källa till vad som förväntas innehållsligt av en biblioteksplan är KB:s uppföljningsrapport med utgångspunkt i bibliotekslagen och händelser i omvärlden. Den senaste uppföljningsrapporten tar upp bl.a. prioriterade grupper, bibliotekets demokratiska uppdrag, skolbiblioteken och digital delaktighet.

Vad de faktiskt skriver om får man en ganska tydlig bild av med hjälp av vår ämnesmodellering.

Samlade lärdomar från de tre fallstudierna

Kommuner publicerar många textdokument som skulle kunna lämpa sig för språkteknologiska analysmetoder men att tillämpa sådana metoder försvåras av framförallt tre orsaker.

Den ena orsaken är att dokumenten publiceras i pdf som inte är ett maskinläsbart format. Den andra orsaken är att dokumenten publiceras på 290 webbplatser. Den tredje orsaken handlar om att samma dokument kan ha olika namn (t.ex, biblioteksplan vs biblioteksprogram) och olika sätt att skriva datum (t.ex. 2023-08-29 vs 29 Aug -23) datum på

dokumenterna och också ha förhållandevis olika innehåll även om det är samma typ av dokument (t.ex. biblioteksplan eller årsredovisning).

För att underlätta användningen av språkteknologiska analysmetoder ska:

- dokumentet vara i ett maskinläsbart format
- dokumenterna ska kunna nås samlat
- dokumenterna ska helst inte innehålla personuppgifter
- dokumenterna ska helst inte innehålla tabeller, diagram, bilder etc. då det försvårar extrahering av text
- Jämför man flera liknande dokument ska innehållet vara strukturerat på någorlunda samma sätt (titel, datum, innehåll etc.)

Om alla, eller delar av dessa, förutsättningar är uppfyllda så underlättas tillämpningen av AI väsentligt, och tidsåtgången för att förbereda datamaterialet kan kortas avsevärt.

New section follows

Fallstudier - Granska Annonser (Arbetsförmedlingen)

Skreven av Tord Karlsson

1 Vad handlade projektet om?

Projektet startade i början av 2021 och hade två huvudsyften. Det ena var att minimera risken för att publicera annonser som kan vara diskriminerande och det andra var att se till att arbetsgivarna inte bryter annonsvilkoren som Arbetsförmedlingen har satt upp för att få publicera jobbannonser. (Arbetsförmedlingen publicerar jobbannonser på platsbanken där Platsbanken är Arbetsförmedlingens sida som listar de lediga jobb som anmälts till arbetsförmedlingen. <http://platsbanken.arbetsformedlingen.se>)

I denna case study går vi igenom den del av projektet som hade att göra med diskriminerande annonser.

Arbetsförmedlingen har ansvaret för att innehåll och text i platsannonser överensstämmer med annonsvilkoren och att platsannonsen inte är av sådan art att den ska undantas från publicering. Arbetsförmedlingen behöver därför följa upp kvalitetssäkra annonser på Platsbanken genom att granska och avpublicera annonser som bryter mot villkoren som framgår i diskrimineringslagen², DL, (2009:567) eller som på annat sätt bryter eller missbrukar platsbankens annonseringsvillkor.

Arbetsförmedlingens rättsavdelning vid tillfället för projektets start skrev: "Det är arbetsgivaren som har att följa och tillämpa diskrimineringslagen, DL. Även om arbetsgivaren har delegerat arbetsgivarfunktionen eller beslutanderätten till annan part så är det alltid arbetsgivaren som har ansvaret för felaktigheter av diskriminerande karaktär. Myndigheten har dock ett ansvar för vilken information som läggs upp på sina webbplatser. Myndigheten har således en skyldighet att granska och kontrollera och kvalitetssäkra annonserna och innehållet. Arbetsgivaren har att följa de villkor som är uppsatta för Platsbanken. Om oegentligheter upptäcks ska myndigheten i första hand kontakta arbetsgivaren för korrigerande åtgärder. Hur denna granskning ska gå till, via app eller på annat lämpligt sätt är upp till verksamheten att bestämma". (Af, 2020)

Det kommer in ca 3 200 nya annonser per dag i Platsbanken. Platsbanken är en internetjänst för att marknadsföra lediga jobb för hela Sverige (<https://arbetsformedlingen.se/platsbanken/>)

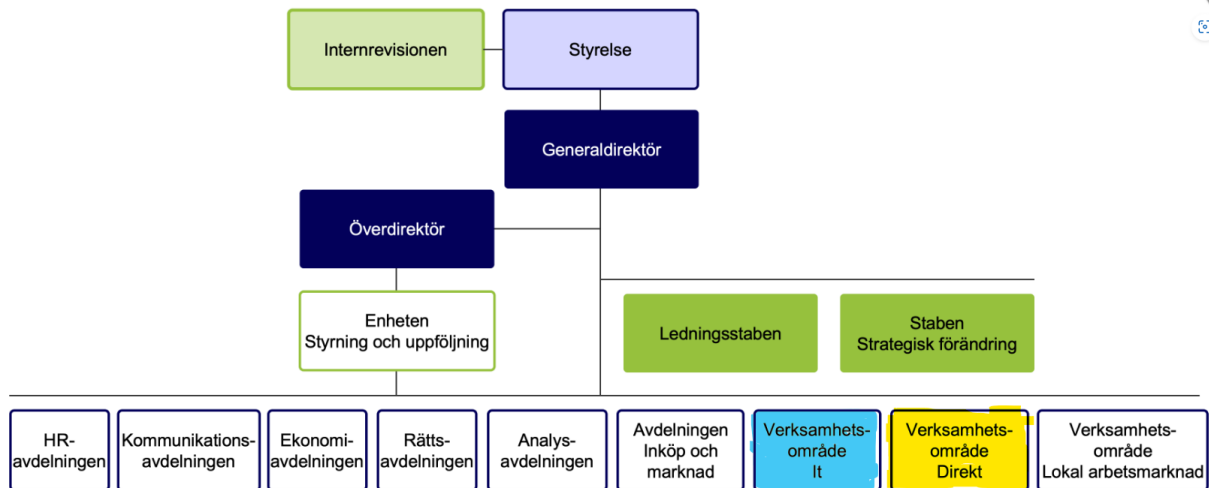
2 Projekt mål

Arbetsförmedlingen är en stor organisation och man har därför mål på olika nivåer i organisationen så även vid detta projekt. Det var två verksamhetsområden som var inblandade i att ta fram lösningen: Verksamhetsområde Direkt (VO-Direkt) och

²

https://www.riksdagen.se/sv/dokument-och-lagar/dokument/svensk-forfattningssamling/diskriminering-slag-2008567_sfs-2008-567/

Verksamhetsområde IT (VO-IT). Nedan är målen för de olika verksamhetsområdena kort beskrivna.



Figur 1. Övergripande organisation på Arbetsförmedlingen (2023-06-07 <https://start.arbetsformedlingen.se/organisation/arbetsformedlingens-organisation>)

2.1 Prestationsmål Arbetsförmedlingen

På översta nivån för Arbetsförmedlingen som myndighet, var ett av målet "Öka antalet deltagare i matchningstjänster" (Arbetsförmedlingens verksamhetsplan 2022. Version 1.0 Diarienummer: Af-2021/0011 6302). Om vi minskar antalet diskriminerande annonser, kommer vi att få ökat förtroende för Platsbanken och därmed minska risken för att tappa besökare. Fler besökare leder till ökat antal deltagare i matchningstjänster.

2.2 Mål i Granska Annonser (VO-Direkt)

VO-Direkt är ansvariga för Platsbanken samt det som publiceras i platsbanken. VO-Direkt satte upp följande mål:

1. Förbättrad annonskvalitet: färre diskriminerande Annonser/Platser. Färre annonser i Platsbanken som är diskriminerande eller bryter mot annonsvillkoren på andra sätt.
 - a. Indikator: Antal borttagna annonser
2. Förbättrad annonskvalitet: förkortad tid som potentiellt felaktiga annonser visas.
 - a. Indikator: Andel annonser som granskas i tid
3. Ökad kundnöjdhet
 - a. Indikator: Flera kunder som anger att de är nöjda med Annonsera och platsbanken.

2.3 Mål i projektets textanalys (VO-IT)

Projektet Granska Annonser bedrevs av två olika parter (VO-IT och VO-Direkt) som ansvarade för olika delkomponenter. VO-IT ansvarade för den del som gjorde den maskinella textanalysen. I bland går denna del av projektet under namnet Textanalys. Det var i delprojektet Textanalys vi byggde AI-modeller som skulle hitta potentiellt diskriminerande meningar. Kort beskrivet skulle Textanalys erbjuda ett REST API dit annonser skickas för analys. Till svar från API:t skulle man sedan få en lista med meningar som bedömts vara potentiellt diskriminerande samt den diskrimineringsgrund som har högst sannolikhet för att vara diskriminerande i meningen.

I den version projektet tog fram tittade man endast på de individuella meningarna som annonsen är uppbyggd av. Men om en mening är diskriminerande eller ej beror på kontexten meningen befinner sig i. Det är till exempel acceptabelt att specificera att man söker en man som är i viss ålder om arbetet innebär att vara personlig assistent. I lösningen tittade inte textanalysen på hela annonsen utan skulle endast hitta om det finns potentiellt diskriminerande text i de meningar som utgör annonsen. Det fanns även en avgränsning att endast leta efter diskriminerande text i annonser skrivna på svenska.

Tillvägagångssätt och indikatorer var att:

1. Identifiera annonser som är skrivna på svenska
 - a. Indikator: Andel annonser där svenska språket har identifierats korrekt.
2. Bryta ner annonsen i meningar på korrekt sätt. (Då olika personer skapar meningarna kommer strukturen att vara väldigt olika. Det är då svårt att dela upp en annons i meningar med enkla metoder.)
 - a. Indikator: Andel annonser där texten har brutits ner i meningar på korrekt sätt.
3. Identifiera meningar som innehåller potentiellt diskriminerande text.
 - a. Indikator: Antal diskrimineringsgrunder som identifieras av modellen.

2.4 Mätning i Granska Annonser

För att veta hur stort problemet var, samt mäta hur väl vi lyckas med att hitta diskriminerande annonser behövs ett mått på hur många annonser som är diskriminerande.

En initial studie utfördes några år innan projektet där domän experter involverades. Resultatet från studien presenteras i tabellen nedan..

Klass	Antal meningar	Andel i procent
Ej diskriminerande	3672	85,6%
Ålder	217	5,1%
Kön	155	3,6%

Språk	152	3,5%
Etnicitet	57	1,3%
Religion	16	0,4%
Könsöverskridande	12	0,3%
Funktionsnedsättning	10	0,2%
Sexuell läggning	1	0,0%

Tabell 1. Antal meningar i de olika diskrimineringsgrunderna.

Tyvärr lyckades inte Projektgruppens medlemmar återfinna det data som låg till grund för denna bedömning. Inte heller lyckades man hitta namn eller någon referens till de personer som gjort bedömningen.

För att kunna skaffa sig en egen bild av om storleken på siffrorna är rimliga, gjordes som ett första steg ett urval av 385 annonser från alla annonser som Arbetsförmedlingen har tagit emot under år 2022. Därefter delades annonser i fyra mängder och efter det gick fyra personer manuellt genom 97 annonser var och markerade ut om annonstexten inte var på svenska samt om annonsen har något åldersuttryck relaterat till person i sig. Resultatet från denna genomgång var att 5,6% av annonserna innehöll potentiellt åldersdiskriminerande uttryck, det vill säga helt oberoende av om hela annonsen kunde anses som diskriminerande eller ej.

3 Hur ser lösningen ut?

För att ta fram AI lösningar på Arbetsförmedlingen har vi tagit fram en checklista³ där ett antal steg listas för att ta fram en AI lösning. För att förenkla har vi delat upp arbetet i fyra större steg. De fyra stegen var: problemformulering, ta fram data, träna modell och utvärdera modell.

3.1 Problemformulering

Vi delade upp problemet i två delar, 1) hitta problematiska uttryck och 2) avgöra diskriminering. Med denna uppdelning blev det tydligare vem som ansvarade för vad. Alternativet var att göra allt i ett svep, det vill säga gå direkt mot diskriminering. Att avgöra om en annons är diskriminerande skulle ställa krav på att annotören har juridisk kunskap. Då vi istället valde att leta efter väldefinierade uttryck (krav på ålder, krav på kön) blev uppgiften enklare.

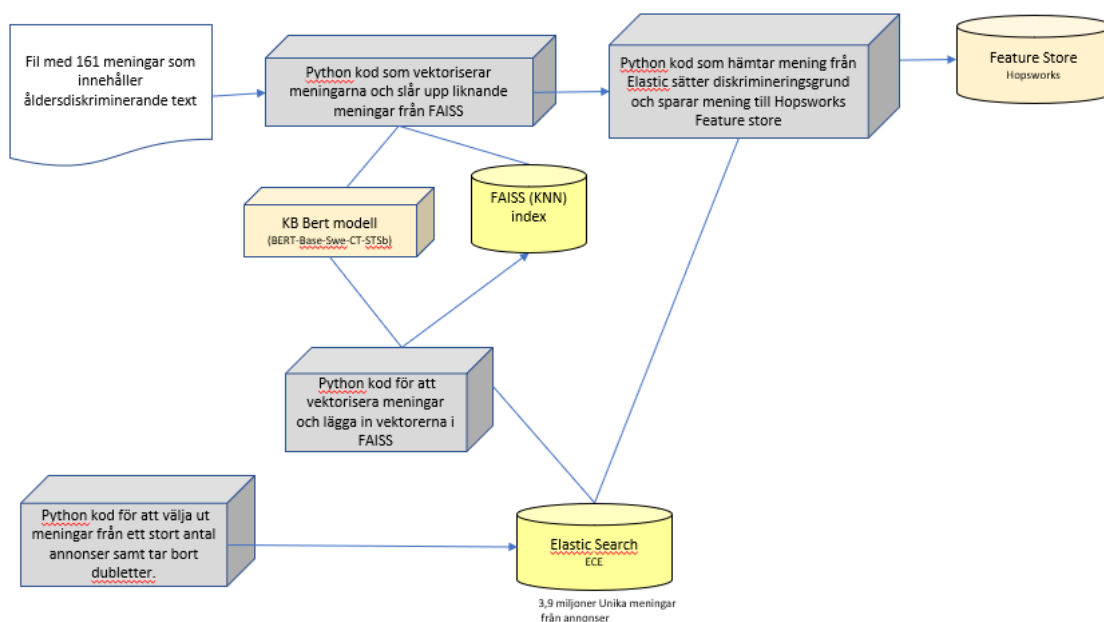
Beroende på diskrimineringsgrund ligger nyckeln till att avgöra om det rör sig om diskriminering i hela annonsens kontext.

³ Checklisten vi tog fram bygger till stor del på idéer från boken. Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow : Concepts, Tools, and Techniques to Build Intelligent Systems, skriven av Géron, Aurélien och utgiven av O'Reilly Media, Incorporated.

I uppdelningen av ansvar och de avgränsningar som listats ovan, var VO-IT ansvariga för att hitta en lösning som tar en mening och klassificerar meningen i någon av de 9 klasserna. Man bedömde endast om en mening var potentiellt diskriminerande helt oberoende av sammanhanget meningen befinner sig i. VO-Direkt var ansvariga för att sedan bedöma om annonsen som helhet var diskriminerande.

3.2 Dataförberedelser

Granska Annons Dataflöde hitta många representativa meningar fas 1.



Figur 2. Flödet för att hitta fler potentiellt diskriminerande meningar.

Projektet behövde ta fram en bra datamängd med representativa meningar för de olika klasserna samt annotera dessa meningar med rätt klass (typ av uttryck som kan kopplas till diskrimineringsgrund).

Vi kunde inte bara träna modellen på diskriminerande fraser, inte heller på enbart oproblematiske fraser. Vi behöver dessutom tillräckligt många av varje klass så att modellen faktiskt bryr sig om att försöka hitta skillnaden.

När vi startade projektet hade vi få meningar som exemplifierade de olika diskrimineringsgrunderna. Under projektet tog vi därför fram ett stort antal annonser, delade upp dessa annonser i meningar och vektoriserade dessa meningar med hjälp av en "Contrastive Tension (CT)" modell för Semantic Textual Similarity (BERT-Base-Swe-CT-STSb). Contrastive-Tension är ett sätt att ta fram en Transformer-modell som är bra på att ta fram inbäddningar som är lämpliga för att hitta liknande texter/meningar. Utifrån en Contrastive-Tension modell får man en representation där liknande meningar ligger nära varandra.

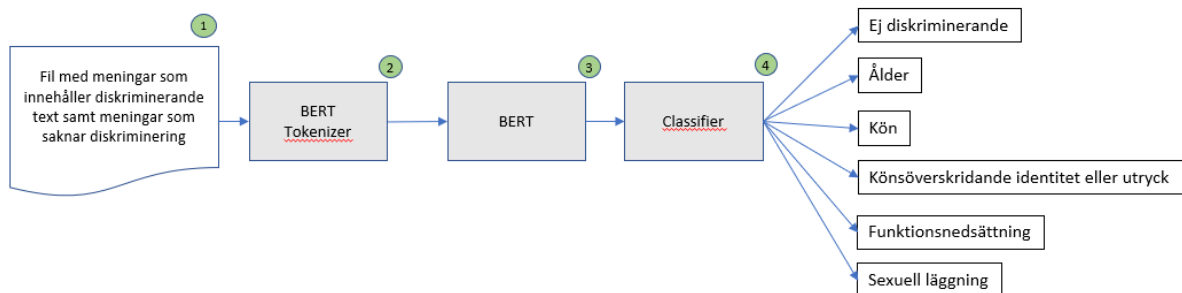
De vektoriserade meningarna lästes in till ett FAISS index (FAISS är framtaget av Facebook och är ett KNN index som erbjuder förmågan att väldigt snabbt söka upp vektorer som ligger nära varandra i en vektorrymd.)

Vi började således med att ta fram ett fåtal potentiellt diskriminerande meningar och sökte upp fler likartade meningar med hjälp av "similarity search" i FAISS index och gick därefter igenom de meningar som hittades manuellt och annoterade efter bästa förmåga, det vill säga, vi satt en etikett som identifierar till vilken klass en mening tillhör.

CT-användning kan liknas vid "Upsampling Minority Class"⁴. Vi duplicerade dock inte, vi förde in fler exempel av samma klass.

För att annotera de nya meningarna som vi plockade fram med hjälp av "similarity search" använde vi en produkt som heter Label Studio (<https://heartex.com/product>).

3.3 Träna modell



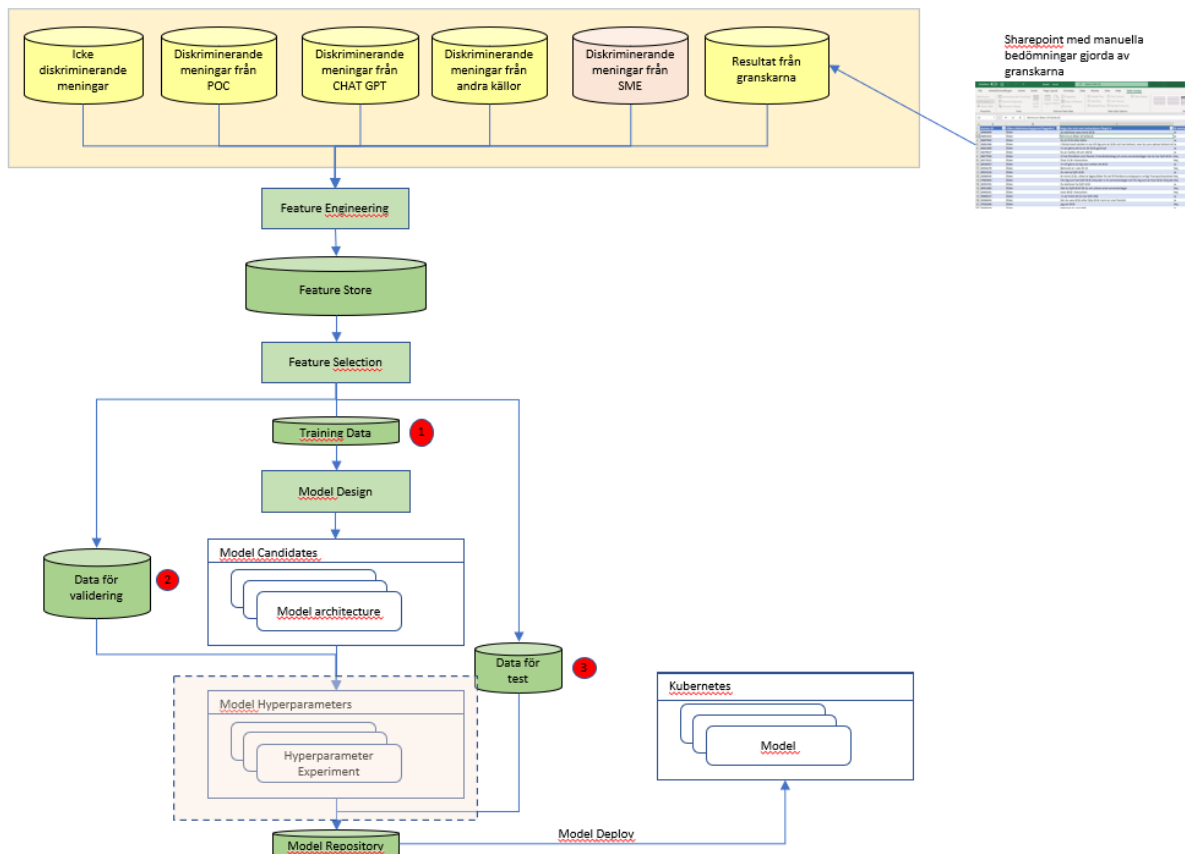
Figur 3. Huvudkomponenter som är involverade i träning av modellen.

1. **Dataförberedelser:** Det datat vi plockade fram, och som beskrivs i avsnitt 3.2, delade vi upp i tre olika mängder, markerade med de röda cirkarna 1, 2 och 3 i figur 4 nedan. En mängd som används för träning av modellen, en mängd som används för validering och en mängd som används för testning av modellen. I vårt fall valde vi att dela upp den ursprungliga datamängden till en träningsmängd som innehöll 60%, en valideringsmängd som innehöll 20% och en testmängd som innehöll 20%. Det är viktigt att man i varje mängd har en lika stor andel av de olika klasserna, därför gjorde vi ett slumpvis stratifierat urval från grunddata mängden till de tre olika mängderna.
2. **Tokenisering:** En dator har svårt att förstå text och kan endast analysera siffror. I detta projekt valde vi en Transformer baserad modell som heter BERT (swe-BERT). För att få rätt siffror in till den modellen användes först BertTokenizer som ger varje ord en siffra.
3. **Vektorrepresentation:** Efter att orden i meningen hade fått siffror tilldelade till sig, så skickades dessa siffror (även kallad vektor) in till BERT modellen som skapar en ny vektor för varje mening i en multidimensionell rymd (768 dimensioner). Denna vektor

⁴<https://towardsdatascience.com/5-techniques-to-work-with-imbalanced-data-in-machine-learning-80836d45d30c>.

kallas för en embedding vektor⁵. I princip kan man säga att varje ord i meningen ges en siffra där siffran ett ord kan få tilldelad till sig varierar med betydelsen av meningen. Det innebär att meningar som har likartad betydelse kommer att ligga nära varandra i vektorrymden.

4. **Modellträning**: Som nästa steg tog vi den inbäddade vektorn (embedded vector) och tränade ett neuralt nätverk där vi använde en Softmax aktiveringsfunktion vid näst sista nivån och fick som resultat en vektor som hade samma storlek som antalet klasser vi önskade identifiera.



Figur 4. Bilden ovan visar hur vi tar fram modellen med hjälp av vår plattform Hopsworks.

3.4 Utvärdering

Att testa och utvärdera en AI/ML-modell är en iterativ och ofta tidskrävande process och så även för detta projekt, det kräver en noggrann analys, experiment och finjustering för att uppnå önskad prestanda.

I detta projekt sparade vi följande metrics för varje träning:

- roc_auc
- matthews_corrcoef
- precision
- recall

⁵ <https://vaclavkosar.com/ml/transformer-embeddings-and-tokenization>
<https://developers.google.com/machine-learning/crash-course/embeddings/video-lecture>

- accuracy
- F1-score

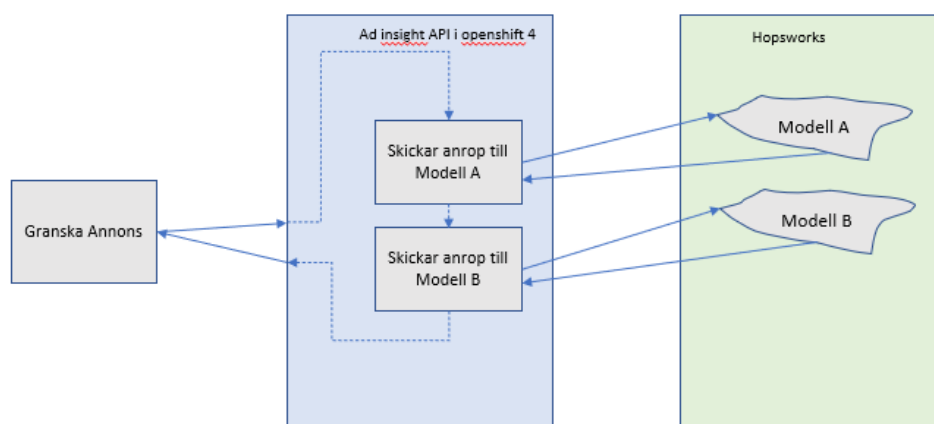
Vid jämförelse av modellerna hade matthews_corrcoef högst prioritet. Vi tittade även noga på confusion matrix vilket fungerade bra när det bara var två klasser.

Vi sparade även confusion matrix för varje träning. I en confusion matrix får man bra överblick över hur modellen presterar. Men i detta fall då vi har obalanserade klasser samt många klasser, kan det vara utmanande att använda en confusion matrix. Under Appendix är confusion matrix beskriven i mer detalj. Vid ett fåtal klasser kan det dock vara ett bra hjälpmedel för att diskutera med andra parter inom verksamheten för att förstå vad de värderar högst och vad deras mål och prioriteringar är. För att ge ett kort exempel så kan vår modell generera något vi kallar för False Negative vilket i detta fall innebär att en mening innehåller problematiskt uttryck som inte upptäcks. Modellen kan även ibland ge resultat som är False Positive vilket i detta fall innebär att en mening har pekats ut som problematisk men i själva verket så innehåller meningen inte något problematiskt uttryck, men den har skickats till granskning i alla fall.

Det mätvärde som var lite ovanligt var matthews_corrcoef. Om vi har att göra med klasser som inte är balanserade så kan precision, recall och accuracy ge missvisande värden. I dessa fall kan matthews_corrcoef (MCC) vara ett bättre mätvärde att använda för att jämföra olika modellers prestanda.

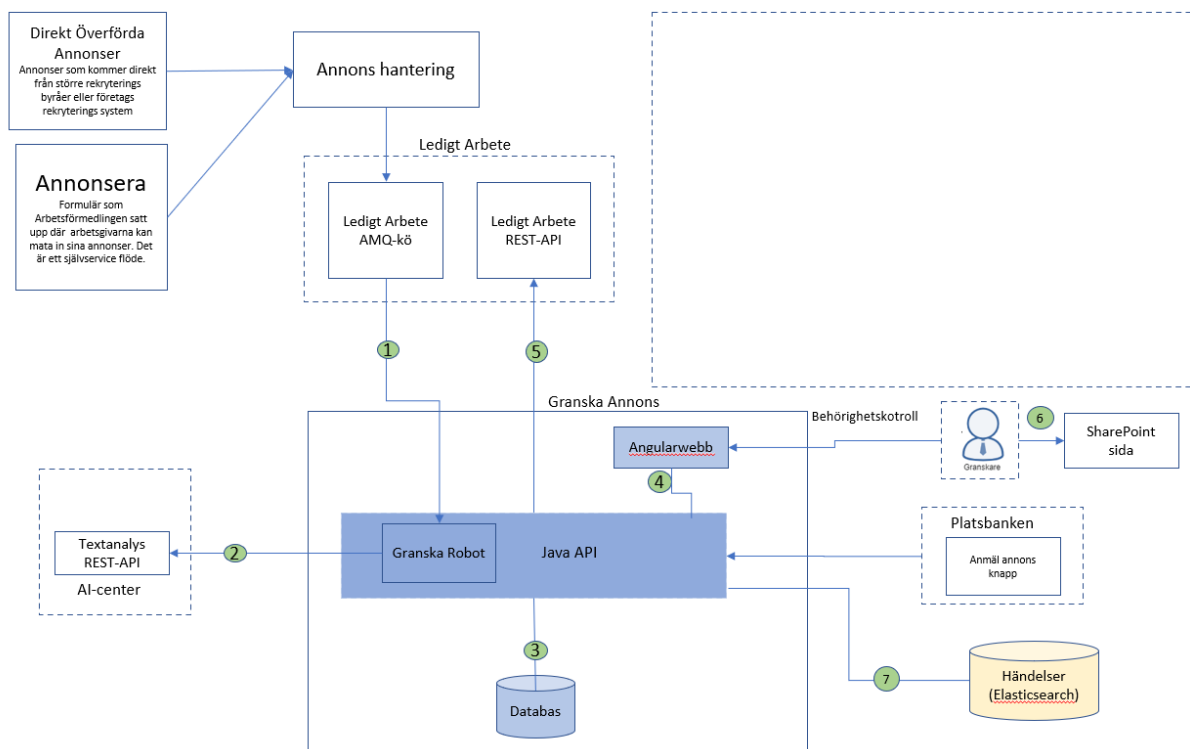
<https://towardsdatascience.com/the-best-classification-metric-youve-never-heard-of-the-matthews-correlation-coefficient-3bf50a2f3e9a>

Vi satte även upp ett flöde där vi tog varje anrop och skickade det till både den produktionssatta modellen samt till den "nya" kandidatmodellen. Vi kunde då i realtid jämföra den nya modellen med den gamla och förvissa oss om att den inte kommer att ge sämre resultat än den modell som redan är i drift. Vi körde även ett stort antal (21 000) historiska annonser mot den nya modellen och jämförde resultatet med den "gamla" modellen innan vi satte den nya modellen i produktion.



Vi kunde även mäta kontinuerligt prestandan på modellflödet i form av antal anrop per sekund som hanteras samt hur lång tid det tog för varje anrop att få ett svar (latency).

4 Övergripande arkitektur



1. Granska Robot lyssnar på alla nya och förändrade annonser i Ledigt Arbete och gör en automatisk granskning av annonsen.
2. För varje automatisk granskning använder Granska Robot en komponent hos AI-center för textanalys. Komponenten returnerar meningar i annonsen som eventuellt strider mot diskrimineringslagen.
3. För varje automatisk annonsgranskning sparas resultatet i en databas, även om inga felaktigheter hittades.
4. Manuell granskning sker, av alla annonser där textanalyskomponenten har hittat ett eller flera potentiellt diskriminerande uttryck, genom en webbaserad granskningsapplikation.
5. Vid beslut att avpublicera annonsen sker anrop till Ledigt Arbete
6. Vid granskning skriver granskaren in resultatet av granskningen på en SharePoint sida i Excel format
7. Händelser som sker i bedömningen av en annons sparas till Elasticsearch databas.

5 Verktyg vi använde oss av

På Arbetsförmedlingen lagras data i flera olika typer av databaser. Vi har no-SQL databaser (Elasticsearch och Opensearch), rena relationsdatabaser (MS SQL Server och Oracle), data warehouse (Informatica produkt ovanpå Oracle), datalake (Cloudera) samt streaming lagring (KAFKA)

Annonserna lagras dels i Oracle Databas, i Elasticsearch och i en Data Lake baserad på en paketering av ett stort antal produkter gjord av Cloudera (<https://www.cloudera.com/>). I den plattform vi använde för att bygga modeller lagrade vi data i något som kallas för en Feature Store. Feature store är uppdelad i två datalagringsteknologier, dels en Hadoop baserad lagring för offline data och dels en Relationsdatabas som är en klustrad version av MySQL (RonDB) för on line data.

I projektet flyttade vi sedan över data till en plattform som idag är baserad på en "Enterprise" version av Hopsworks.

Hopsworks plattformen (<https://www.hopsworks.ai/>) består av en Feature Store, Jupyter Notebook Server, Model Registry, Model Serving, stöd för experiment och jobflöden (Airflow, <https://airflow.apache.org/>) och datavalidering (Great Expectations, <https://greatexpectations.io/>) I Hopsworks så är data, modeller mm. indelat i projekt. En person kan vara med i flera projekt, men det är endast via medlemskap i ett projekt som man når data och kod som ligger under projektet. Detta innebär att vi har bättre kontroll på vem som har tillgång till vilket data, kod och modeller.

Koden är till stor del skriven i Python och för att utveckla koden använde vi dels Jupyter Notebooks och på senare tid i allt större utsträckning Visual Studio Code, vilken vi anslöt via fjärruppkoppling in mot Jupyter Notebook servern i Hopsworks. Det innebar att vi kunde bygga saker utan att behöva ladda ner data till den lokala datorn.

Några reflektioner från projektet var att:

1. Det är viktigt att vi har spårbarhet på tre nivåer. Kod, modell och data:
 - a. Vi lagrade och versionshanterar vår kod i Bitbucket, vilket är ett git baserat verktyg.
 - b. Data lagrade vi i Hopsworks Feature Store för att få spårbarhet, mer om det nedan.
 - c. Modeller lagrade vi i Hopsworks model registry tillsammans med metadata och kopplingar till Feature Store.

6 Vilka utmaningar har vi stött på?

Vi har stött på utmaningar inom flera områden. Man kan dela in dessa i organisatoriska utmaningar, utmaningar med data, hantering av data och vidareutveckling av modellen.

6.1 Organisatoriska utmaningar

Att endast titta på de enskilda meningar en annons är uppbyggd av innebär att vi inte kan bedöma om en annons är diskriminerande utifrån hela annonsens sammanhang. Enligt lagen så kan det vara OK att vissa annonser är diskriminerande baserat på vilket jobb annonsen avser. Om en diskriminering är kopplad till ett lagrum och kopplingen beskrivs i annonsen så kan det vara OK att den innehåller diskriminerande text. Exempel på jobb där viss typ av diskriminering kan vara tillåten är personlig assistent. Utmaningen har här legat i att få personer med hög kompetens runt lagrum och hantering av diskriminerande annonser

att avsätta tid för att annotera annonser samt att gå igenom de annoteringar vi i projektet gjort.

6.2 Utmaningar med vårt data

Vi står inför en obalanserad datamängd där endast en andel om mindre än 6 % av meningarna kan betraktas som åldersdiskriminerande. Det framgår tydligt av tabellen ovan att antalet meningar inom varje klass avsevärt varierar, vilket medför utmaningar vid jämförelse av modellens prestanda.

Vi önskar erhålla ytterligare exempel på diskriminerande meningar. Inom vissa områden av diskriminering är de exempelmeningar vi för närvarande har likartade i sin struktur och uppbyggnad.

Det är av yttersta vikt att en annons är uppdelad på korrekt sätt i meningar. Vi har utfört flera försök med olika metoder för att uppnå en korrekt segmentering av annonsen i meningar. Dessa metoder inkluderar kombinationen av reguljära uttryck med NLTK samt användning av en mer omfattande BERT-modell.

6.3 Hantering av data

Uttag av träningsdata, valideringsdata och testdata har initial skett till separata filer. I vissa fall har spårbarheten till vilket data en modell har tränats på saknats. Visst data har vi inte lyckats återfinna, exempel på det är det data som legat till grund för tabellen ovan (Tabell 1. Antal meningar i de olika diskrimineringsgrunderna)

6.3 Vidareutveckling av modellen

Ingen automatisk återkoppling baserad på den bedömning granskarna gör av de annonser som modellen pekat ut som potentiellt diskriminerande finns på plats. Då modellen har gjort en prediktion så skickas prediktionen till en handläggare. I dagsläget så antecknar handläggaren en del av sina bedömningar i ett Excelark men de lägger inte in alla bedömningar i excelarket. Det innebär att vi inte fullt ut kan använda handläggarnas arbete till att förbättra modellen över tid.

Taggen som används för att annotera meningarna tillåter inte identifiering av flera typer av diskriminering i samma mening. För att klargöra vad som menas, låt oss anta att vi har tre olika klasser: ok, ålder och kön. För närvarande har dessa klasser fått taggarna 0, 1 och 2. Men vad händer om en mening både innehåller könsdiskriminering och åldersdiskriminering? Det behöver kunna annoteras på lämpligt sätt.

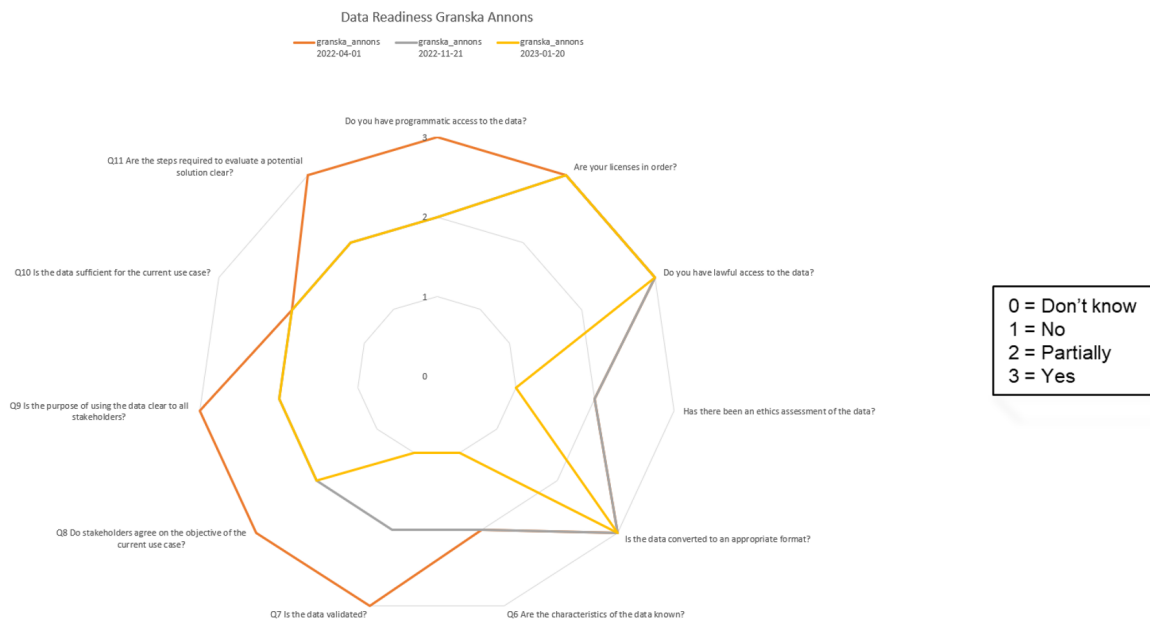
	ålder	kön	Tal
ok	0	0	0
ålder	1	0	1
kön	0	1	2
ålder och kön	1	1	3

7 Resultat av databeredskapsmätningar

Tre mätningar av projektets databeredskap har utförts. De tre mätningarna har utförts vid följande tidpunkter.

1. Mätning 1: 2022-04-01
2. Mätning 2: 2022-11-21
3. Mätning 3: 2023-01-20

Att diagrammet har gått åt fel håll beror inte på att vi fick in sämre data under projektet löptid utan det berodde på att vi fick mer förståelse för datat och dess brister under projektet löptid. Under 8.1 är de olika frågeställningarna listade i mer detalj.



8 Vad kan göras för att öka databeredskapen?

8.1 Förslag till förbättringar kopplat till de frågeställningar som tas upp i verktyget för Databeredskap.

Följande tabell är disponerad efter de frågeställningar som finns i *Data Readiness Assessment method*:

Frågeställning	Bedömning	Primärt problem	Förslag till åtgärd

1	Har du programmatisk tillgång till data?	Delvis	<p>1.) Resultatet av alla granskares bedömning av en annons har vi inte programmatisk åtkomst till. Vi</p> <p>2.) Resultatet lagras i ett excelark</p> <p>3.) Vi lagrar inte bedömningen av varje annons / mening</p> <p>4.) Vi får inte in annonsid via nuvarande api</p> <p>5.) Vi har endast tillgång till annonser som publiceras på platsbanken</p>	<p>Applikationen som granskar annonser lagrar information om granskarnas bedömning av dels hela annonsen och dels en bedömning av den eller de meningar som modellen pekat ut som potentiellt diskriminerande. Vi får åtkomst till den databasen.</p> <p>API ändras så att även annonsid skickas med i anropet till prediktions APIet</p>
2	Är dina licenser i ordning?	ja		
3	Har du laglig tillgång till uppgifterna? (ställs några speciella krav på att hantera uppgifterna? GDPR)	ja	All information som kan identifiera en person bör tas bort ur annonserna och meningarna vi sparar. Exempel på sådana uppgifter kan vara e-postadress till facklig representant, telefonnummer till kontaktpersoner. Namn på kontaktpersoner.	Om någon form av personlig information lagras så måste denna information plockas bort ur de meningar vi sparar.
4	Har det gjorts en etisk bedömning av uppgifterna?	Nej	Ingen etisk bedömning har gjorts av annons texterna	Annonserna är som de är. Vi skulle med fördel kunna gå in i ett tidigare skede av flödet och hjälpa till med att identifiera potentiella problem i annonstexten samt guida arbetsgivarna i utformningen av annonstexten.
5	Har data ett lämpligt format?	ja		
6	Är egenskaper kända?	Nej	<p>1.) Vi vet inte vilka texter som är diskriminerande</p> <p>2.) Frekvensen av förekomst av olika diskriminerande texter i annonser är inte känd</p> <p>3.) Vi har bristfällande information om olika diskriminerande uttryck för de olika diskrimineringsgrunderna</p>	<p>1.) Genom sampling av annonser från ett stort urval och manuell genomgång av annonserna så har vi tagit fram siffror på ett par diskrimineringsgrunder.</p> <p>2.) Vi behöver få fler varierande exempel på diskriminerande uttryck från domän experter</p>
7	Är uppgifterna validerade?	Nej	<p>1.) Data innehåller dubletter</p> <p>2.) Uppdelningen av en text i meningar har brister</p> <p>3.) Meningar på andra språk än svenska var med i den initiala träningen av modellen.</p>	<p>1) Ta bort dubletter</p> <p>2) Ta endast med svenska annonser genom att detektera vilket språk en annons är skriven på</p> <p>3) Prova olika sätt och modeller för att utreda vilket är det optimala sättet att dela upp en text i meningar.</p>

8	Är intressenterna överens om syftet med det aktuella användningsfallet?	Delvis	I den aktuella lösningen har man endast tittat på att hitta meningar som har ålders uttryck. Man önskar bredda detta till andra diskriminerande områden. Man kan inte uttala sig om en annons är diskriminerande eller ej genom att endast titta på enskilda meningar. En mening kan vara diskriminerande eller ej beroende på kontexten. Intressenten vill att man tar med så många olika diskrimineringsgrunder som möjligt först och därefter tittar på precisionen inom varje diskrimineringsgrund	
9	Är syftet med att använda uppgifterna tydligt för alla intressenter?	Delvis	Vi får idag inte information om annonsen som helhet är diskriminerande. Vi får ingen prioritet för att få en noggrann databaserad återkoppling på om annonstexten som helhet är bedömd som diskriminerande samt hur granskarna bedömt de enskilda meningarna.	Vi får tillgång till informationen som gäller för hela annonsen
10	Är uppgifterna tillräckliga för det aktuella användningsfallet?	Delvis	Vi har för närvarande för få exempel på texter som diskriminerar inom de olika diskrimineringsgrunderna. Vi har tillräckligt med exempel på annonstexter, flera miljoner. Vi har endast tillgång till de annonser som Arbetsförmedlingen samlar in, vilket är ungefär 30% av de jobbannonser som produceras i Sverige. Det kan vara så att vi skulle få bättre och fler exempel att träna våra modeller på om vi får åtkomst till alla jobbannonser som produceras i Sverige.	Vi måste utöka tränings datamängden. Ett nytt flöde är under utveckling där vi sparar den bedömning områdes experter gör på varje mening, det vill säga information om meningens kan anses vara diskriminerande som en fristående mening.
11	Är stegen som krävs för att utvärdera en potentiell lösning tydliga?	Delvis	Vi vet inte hur vi ska mäta om vi har uppnått det som kunden önskar. Det innebär att vi vet inte när vi är klara.	Genom att arbeta med datat och domän experter hoppas vi kunna få fler exempel på texter som potentiellt kan vara diskriminerande. Därefter får man manuellt gå igenom tillräckligt stort urval av annonser för att för varje diskriminerande grund får fram en rimlig siffra på frekvensen av diskriminerande annonser för varje diskrimineringsgrund.

12	Är din organisation beredd att hantera mer data som denna utanför projektets ram?	Delvis	Historiska annonser lagras i en Data Sjö och det är lite komplicerat att söka och hämta relevant information. Index på ingående ord i annonsen saknas.	Om vi ser behov av att söka och analyser på ingående ord i annonserna vore det lämpligt att lagra alla annonser i Elasticsearch där vi idag lagrar alla våra loggar och annan textdata. Ett alternativ till att lagra informationen i Datasjön är att lagra den i en vanlig SQL databas, i det fall det inte finns någon ambition att indexera upp ingående text.
13	Är uppgifterna säkrade	ja		
14	Är det säkert att dela data med andra?	ja		
15	Får du dela uppgifterna med andra?	ja		Annonserna är publika, men rensning av kontaktuppgifter behöver göras innan annonserna kan spridas i ett vidare sammanhang.

8.2 Andra saker som kan förbättra vår databeredskap

Som beskrivet under 5, har vi ibland haft problem med att återfinna det data som används för att träna en specifik modell samt beskrivning av det datat. En teknik att förbättra denna hantering är att använda ett Feature Store. Det data som används för att träna en AI modell kallas för features.

Med hjälp av ett feature store kan vi versionshantera vårt data, beskriva datat med Metadata och på ett enklare vis hålla kontroll på vilket data som använts för att träna en viss version av vår modell.

9 Viktiga lärdomar

Det är bra att etablera goda principer för sårbarhet tidigt! Det finns en risk att saker glöms bort om det inte dokumenteras på en gång. Personer byts ut och saker glöms över tid.

Vi har även noterat att när vi arbetat utforskande så tar det inte lång tid innan man glömmer varifrån datat kommer, vad som gjorts med datat i form av tvättning, borttagning av dubletter och transformeringar. Även koden som profileringen av datasetet gjordes med är snart glömd.

Det är viktigt att Domän-expertiser är med under projektets hela gång och avsätter tid rimlig tid för arbetet. I detta use case så ska vi bedöma om en mening är potentiellt diskriminerande eller ej. För vissa diskrimineringsgrunder går det att göra en hyfsad bedömning utan att man innehar domänexpertis men för exempelvis språk och etnicitet och kopplingen däremellan är bedömningen väldigt utmanande. Vi skulle gärna ha större närvaro av domän experter som gör bedömning av de olika exemplen på diskriminerande meningar och annonser.

9.1 Lärdomar från att ha arbetat med databeredskapsverktyget

Ur ett databeredskaps perspektiv så har radardiagrammet krympt (datat ser sämre ut) ju mer vi har arbetat med applikationen. Den spontana känslan är att databeredskapsnivån borde öka för varje mätning. Vi har noterat ett par faktorer som påverkat bedömningen.

1. Personer som gör uppskattningen har olika läggning, en del ser mer problem än andra.
2. För oss har det även varit så att ju mer vi arbetat med datat ju fler brister har vi upptäckt.

Det är viktigt att man arbetar aktivt med data och att man ofta stämmer av med alla berörda parter inom projektet kontinuerligt. Risken för missförstånd är stor.

Databeredskap är ett abstrakt begrepp och det kan vara svårt att närma sig området. Det är då bra att få ett verktyg som strukturerar arbetet.

Detta är ett viktigt steg på vägen för att öka medvetenheten om databeredskap. I detta projekt behöver vi prioritera de områden som ger störst värde.

9.2 Ställer arbete inom offentlig förvaltning speciella krav inom området databeredskap?

Beroende på i vilket sammanhang modellen används så kan högre krav ställas på modellen, dess dokumentation samt modellens förklarbarhet. Om modellen används för att fatta ett beslut är kraven högt ställda redan idag. . Inom snar framtid, dvs. december 2023 (detta är skrivet i april 2023) kommer EU AI Act att klubbas och troligt är då att de flesta applikationerna inom offentlig förvaltning kommer att klassas som högrisk AI. Om en applikation klassas som högrisk AI kommer det då att ställas krav på att man kan följa data från källa till användandet av modellen samt krav på att data berikas utifrån kontexten och av domänexperter. Man måste veta vilken kvalitet datat har, vilket kräver att man kan mäta kvalitet i alla steg som datat tar. Man måste kunna spåra om datat har ändrats på vägen, vem som är ansvarig för datat och om datat är rätt datat för den modell som ska tränas. Många av dessa områden täcks av databeredskaps verktyget.

För närvarande ställer även lagstiftningen krav på hur data som innehåller information som kan bidra till att identifiera en person kan lagras. Det är t.ex. Inte tillåtet att lagra datat på servrar som ägs av företag som är baserade i USA eller Kina. Det finns även krav på var rent geografiskt servrarna står placerade.

I ett samhälle med en offentlig förvaltning som betalas av medborgarna är det väldigt viktigt att medborgarna kan lita på det arbete som förvaltningen utför. Det ställer högre krav på att modellen och dess prediktioner kan förklaras.

9.3 Generella lärdomar från att arbeta med data.

Använd inte för komplicerad teknik. Det måste vara tydligt vad den nya tekniken kan bidra med till projektet. Det är allmänt känt att stora projekt dras igång som kostar mycket pengar, kräver stora resurser och tar lång tid utan att man tydligt har definierat vilket värde projektet eller produkten ska leverera. Exempel på ett sådant projekt är införandet av en Data Lake först baserad på HortonWorks och sedan Cloudera. Innan man inför en Data Lake är det viktigt att först förstå vilken typ av data man har, är det primärt ostrukturerad eller strukturerad data? Hur stora volymer av data är det som ska hanteras och sparas?

Man startar många "lokala" initiativ utan att se till hur de bidrar till helheten och vilket värde som ska levereras till när i tiden. Därpå är uppföljningen bristfällig.

Myndigheten använder i stor utsträckning konsulter inom analysområdet och konsulterna är inhyrda under en begränsad tid. Man underskattar hur viktigt det är att ha erfarenhet från att ha arbetat med data under en längre tid samt ha god förståelse för verksamheten för att kunna dra korrekta slutsatser av analyserna.

Ovan tillsammans med att kunna identifiera mönster och beroenden samt nästan viktigast att ställa frågor om datat gör området till en utmaning på myndigheten.

Varje team har idag relativt stor frihet att välja de verktyg de önskar för olika uppgifter, i styrdokumentet står det att ska vara lönsamt men då uppföljningen är bristfällig resulterar det i att en uppsjö av olika ramverk programspråk och databaser används vilket försvårar när man ska få till att dataströmmar smidigt genom våra system och gör automatisk kontroll och styrning mycket svårt att få till

10 Projektets framtid

Inom ramen för projektet kommer vi arbeta vidare med följande:

Angreppssätt för att ta vara på kontextuella faktorer som kan vara avgörande för hurvida en platsannons är diskriminerande.

Involvering av fler sakkunniga med relevant kompetens inom diskriminering och bias.

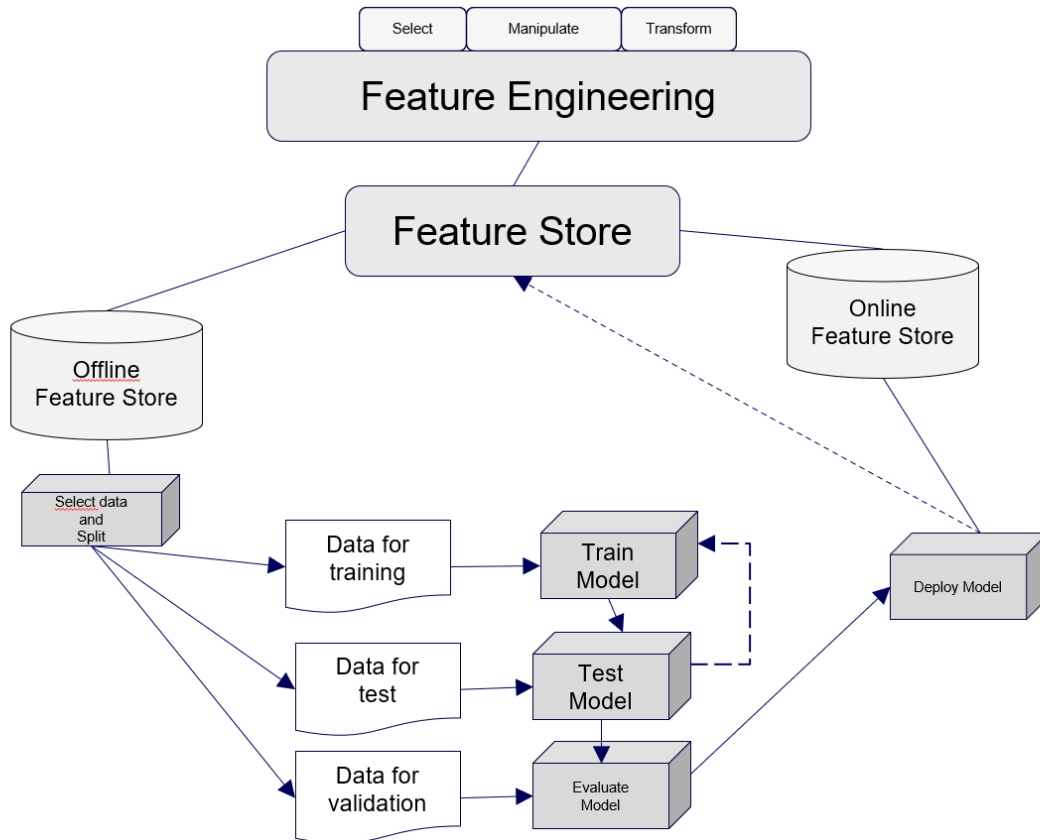
Appendix

Mer om off-line och on-line feature store

Ett Feature Store kan implementeras antingen som en off-line eller en on-line lösning, beroende på behoven och kraven hos organisationen. Låt oss utforska båda typerna närmare:

1. Off-line Feature Store: En off-line Feature Store är vanligtvis en batchorienterad lösning där dataattribut förbereds och lagras i förväg innan de används i modellträning eller prediktion. Här är några fördelar med en off-line Feature Store:
 - a. Dataförberedelse och transformation: Genom att förbereda och transformera dataattribut i förväg kan man undvika upprepad förberedelse för varje modellträning. Detta sparar tid och resurser.
 - b. Reproducerbarhet: Genom att ha en off-line Feature Store kan man säkerställa reproducerbarheten av modellträning och prediktion. Eftersom attributen är förberedda och lagrade i förväg kan man säkerställa att samma attribut används varje gång.
 - c. Skalbarhet: Genom att ha en off-line Feature Store kan man förbereda och lagra attribut för stora mängder data i förväg. Detta möjliggör skalbarhet och hantering av stora datamängder utan att behöva generera attribut i realtid.
2. On-line Feature Store: En on-line Feature Store är en realtidsorienterad lösning där dataattribut beräknas och tillhandahålls vid behov för modellträning och prediktion. Här är några fördelar med en on-line Feature Store:
 - a. Realtidsattribut: Genom att beräkna attribut i realtid kan man använda de senaste och mest aktuella data värdena för modellträning och prediktion. Detta är särskilt användbart när attribut kräver uppdaterade data för att vara effektiva.
 - b. Anpassningsbarhet: Med en on-line Feature Store kan man dynamiskt anpassa attributen baserat på förändrade affärsregler eller nya datakällor. Det ger flexibilitet och möjlighet att snabbt anpassa sig till förändrade krav.
 - c. Realtidsövervakning: Genom att ha attributen tillgängliga i realtid kan man övervaka och spåra modellprestanda kontinuerligt. Detta underlättar upptäckt av driftproblem eller behovet av att uppdatera modeller baserat på förändrade förhållanden.

Både off-line och on-line Feature Stores har sina egna fördelar och kan vara värdefulla beroende på användningsområde och kraven hos organisationen. Det kan också vara möjligt att ha en hybridlösning som kombinerar båda tillvägagångssätten för att dra nytta av deras styrkor.



Varför använde vi ett Feature Store?

På Arbetsförmedlingen använder vi idag ett Feature Store från Hopsworks.

Ett Feature Store är en centraliserad plattform för att hantera, organisera och distribuera dataattribut (features) som används för maskininlärning och AI-modeller. Det finns flera fördelar med att använda ett Feature Store:

1. Återanvändning och konsistens: Ett Feature Store möjliggör återanvändning av dataattribut över olika projekt och team. Istället för att skapa och underhålla attribut i varje projekt kan teamen använda befintliga attribut från Feature Store. Detta förbättrar konsistensen i attributen och minskar risken för duplicering eller inkonsekvenser.
2. Snabbare modellutveckling: Feature Store underlättar snabbare utveckling av AI-modeller. Genom att tillhandahålla ett centralt bibliotek av dataattribut kan utvecklare och data scientists enkelt hitta och använda de relevanta attributen för sina modeller. Det minskar tiden som spenderas på att tillrättalägga datat och möjliggör snabbare iterationscykler.
3. Datakvalitet och övervakning: Genom att ha en centraliserad plats för dataattribut kan ett Feature Store hjälpa till att säkerställa data.
4. Genom att Feature Store är en integrerad del av vår plattform för utveckling av AI får vi även kontroll på vilket data som har använts för att träna vilken modell.

Mer om confusion matrix

En confusion matrix, även känd som en felmatris, är en tabell som används för att utvärdera prestandan hos en klassificeringsmodell. Den visar antalet korrekta och inkorrekta förutsägelser uppdelade i olika kategorier. Här är en förenklad illustration av en confusion matrix:

	Förutspått (Predicted)	
	Class A	Class B
Class A	TP	FN
Class B	FP	TN

Där:

TP (True Positive) representerar antalet korrekta förutsägelser för klass A. FN (False Negative) representerar antalet inkorrekta förutsägelser där modellen förutsåg klass B istället för klass A.

FP (False Positive) representerar antalet inkorrekta förutsägelser där modellen förutsåg klass A istället för klass B.

TN (True Negative) representerar antalet korrekta förutsägelser för klass B.

Fördelar med en confusion matrix

Klar översikt: En confusion matrix ger en tydlig översikt över hur modellen presterar för varje enskild klass. Det ger en mer detaljerad bild än en enkel prestandamätare som noggrannhet (accuracy).

Exempel:

Anta att vi har en klassificeringsmodell för att skilja mellan bilder med hundar och katter. Vi utvärderar modellen på en testuppsättning som innehåller 100 bilder, varav 60 är hundar och 40 är katter. Modellen ger följande resultat:

TP (True Positive): Modellen förutsäger korrekt att 50 bilder är hundar.

FN (False Negative): Modellen missar 10 bilder och förutsäger felaktigt att de är katter istället för hundar.

FP (False Positive): Modellen gör 5 felaktiga förutsägelser och tror att vissa katter är hundar.

TN (True Negative): Modellen korrekt identifierar 35 bilder som katter.

Utifrån denna information kan vi analysera modellens prestanda mer detaljerat och se var den fungerar bra och var den kan förbättras.

Svårigheter med att använda en confusion matrix för utvärdering av en modell

Trots sina fördelar finns det vissa svårigheter med att använda en confusion matrix för utvärdering av en modell:

1. Obalanserade datamängder: Om datamängden har en obalans mellan klasserna kan confusion matrix vara missvisande. Det kan leda till överdriven fokus på de dominerande klasserna och negligera de underrepresenterade klasserna.
2. Multiklassklassificering: Confusion matrix blir mer komplext och svåräst när det gäller modeller med fler än två klasser. Det kan vara svårt att dra slutsatser om modellens prestanda för varje enskild klass.
3. Fokuserar inte på prediktionsosäkerhet: En confusion matrix ger information om rätt och felaktiga klassificeringar, men den ger inte insikt i modellens osäkerhet i sina förutsägelser. För att bedöma modellens prediktionsosäkerhet kan man behöva använda andra metoder, som kalibrering av förutsägelser eller probabilistiska klassificeringsmodeller.

Vidare läsning

[CRISP-ML\(Q\) \(ml-ops.org\)](https://ml-ops.org/)

[MLOps at a Reasonable Scale \[The Ultimate Guide\] - neptune.ai](https://neptune.ai/blog/ml-ops-at-a-reasonable-scale-the-ultimate-guide)

<https://github.com/FreddeFrallan/Contrastive-Tension>

https://openreview.net/pdf?id=Ov_sMNau-PF
[semantic-sentence-stockholm-ai-presentation.pdf. \(103\) Fredrik Carlsson: Sentence Embeddings with Contrastive Tension - YouTube](#)

New section follows

Fallstudie - Strängnäs kommun

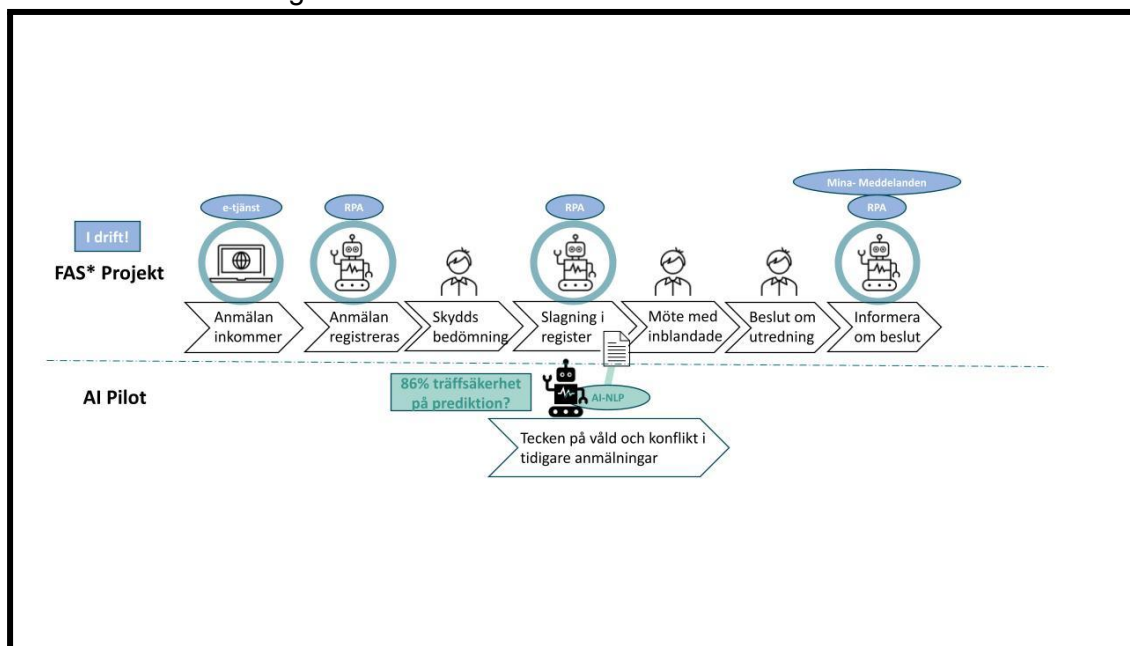
Skriven av Roger Löfgren

1. Bakgrund

I takt med en ökande befolkning och på sikt färre anställda finns det ett behov hos Sveriges kommuner att med stöd av bland annat digitalisering förbättra sina processer utifrån invånarperspektiv, kvalitét och tidsåtgång.

Strängnäs kommun har idag ett antal ärendeprocesser där kommunen ger sina invånare möjlighet att inkomma med ansökningar eller anmälningar från e-tjänster. Den information som sänds in till kommunen läses sedan in i kommunens system för handläggning. I vissa ärendeprocesser tar kommunen hjälp av RPA (Robotic Process Automation) för att underlätta hämtning och sammanställning av de uppgifter som behövs för beslutsunderlag. Införande av RPA i ärendeprocesserna förkortas av kommunen med "FAS" för förändrat arbetssätt.

En av de processer som använder RPA är ärendeprocessen för orosanmälan. Orosanmälan regleras kap. 11 och 14 i socialtjänstlagen (2001:453) och innebär att alla som får kännedom om att ett barn far illa bör anmäla det till kommunens socialnämnd som ska utreda frågan och kan besluta om åtgärder.



Schematisk bild över ärendeprocessen för orosanmälan

Den övre delen av bilden (FAS* projekt) visar ärendeprocessen för orosanmälan med RPA, Mina Meddelanden och ändringar i verksamhetssystem som driftsattes 2021. RPA används för att:

- 1) föra över uppgifter i anmälan som inkommer i XML-fil från e-tjänsten till databas för verksamhetssystemet,

2) efter en initial skyddsbedömning av handläggare, vilket måste göras enligt lag, gör RPA en slagning i ett register om barnet förekommer i tidigare ärendejournaler och ger då handläggaren resultatet av slagningen.

3) föra över information om beslut till kommunens förmedlingstjänst för digital myndighetspost, till exempelvis Kivra.

RPA är till stor hjälp, särskilt vid stora ärendemängder, för att hämta och lämna information som handläggaren annars hade behövt göra manuellt och där API:er (Application Programming Interface) inte finns tillgängliga. För att med teknik kunna stödja handläggaren med att leta efter mönster eller indikationer i texter (strukturerad text eller löptext) räcker det inte med endast RPA eller API.

Att lära en maskin eller ett program att tolka texter var en ny teknik för kommunen och 2021 fattades ett beslut om att genomföra en pilot som systerprojekt till FAS projektet för orosanmälan för att lära sig om det. Pilotens mål var inte att driftsätta nya lösningar. Syftet för kommunen var att lära sig mer om ny teknik för att kunna dra slutsatser inför kommande projekt. Piloten gick under namnet "AI piloten".

2. AI piloten

Vid en orosanmälan är tidsfaktorn och faktabaserade underlag kritiska. Utöver det stöd den RPA som implementerats i FAS projektet ger, skulle det vara till ytterligare hjälp med maskinella prediktioner på om inkomna orosanmälningar innehåller tecken på våld och konflikt. De maskinella prediktionerna behöver ha hög träffsäkerhet så att handläggaren kan få meddelande om brådskande anmälningar även när denne är upptagen med annat.

För att skapa bra förutsättningar till prediktioner på våld och konflikt behövde projektet hitta de datamängder som var bäst lämpade för att träna en språkmodell (AI XLM modell som vid tiden för piloten tillhandahölls av Peltarion). Den tränade språkmodellen skulle kunna hitta textavsnitt som indikerade våld och konflikt. Valet av data att träna språkmodellen på föll på tidigare journalanteckningar i verksamhetssystemet.

Piloten valde journaldata från historiska journaler för att få träna och pröva en språkmodell eftersom det var där de större datamängderna fanns och att datat var realistiskt. Hade uppdraget varit att driftsätta modellen, så hade ett naturligt nästa steg varit att testa modellens förmåga att identifiera tecken på våld och konflikt på data i de XML-filer som nya anmälningar inkommer i.

Projektmålet sattes till att träna och pröva språkmodellen för tecken på våld och konflikt, inte att driftsätta ett automatiserat flöde. För att uppnå projektmålet behövde inte någon ytterligare RPA föras in utan hämtning och lämning av journalanteckningarna utan det kunde ske manuellt genom att markera och kopiera text. Journalerna var läsbara direkt i verksamhetssystemet och det räckte för projektets del att kopiera text därifrån, manuellt anonymisera (radera namn, adresser etc.) och kopiera in den i ett excelark där manuell annotering (märkning) av stycken som innehöll tecken på våld eller konflikt.

Med utgångspunkt i den annoterade texten, som lades in i en excelfil och fördes över till språkmodellen, tränades sedan språkmodellen upp. Den tränade språkmodellen prövades sedan mot andra journalanteckningar och en prediktion på över 86 % kunde uppnås.

3. Data Readiness Assessment Method

När AI piloten påbörjades var inte *Data Readiness Assessment Method* känd för projektet. Man kan dock i efterhand konstatera att projektet före, under och efter pilotens genomförande, ändå försökte besvara flertalet av frågorna i modellen. Projektdeltagarna tog del av modellen och använde verktyget efter projektet i en workshop.

Många av frågorna i metoden har en generalitet som sträcker sig utöver AI projekt och kan kännas igen för personer som arbetat med att ta ut data från ett system eller datakällor för en ny användning, till exempel vid tillgängliggörande av öppna data, överföring av data mellan system, överföring av data till Data Warehouse eller arkivering i e-arkiv.

Den följande texten anknyter till de 15 frågorna i metoden.

4. Framgångsfaktorer för piloten

I de närmaste några av framgångsfaktorerna för pilotens genomförande, med hänvisning till närliggande frågeställningar i Data Readiness Assessment Method

- Projektet hade en tydlig avgränsning och en engagerad styrgrupp med rimliga förväntningar. Styrgruppen stödde projektets inriktning att piloten i första hand skulle vara utforskande och lära organisationen inom ett nytt och komplext område, inte i dagsläget leda till en driftsättning (jämför Q8. Do stakeholders agree on the objective of the current use case).
- Projektet valde den data som behövdes för att uppnå projektmålet, det vill säga att träna och pröva språkmodellen för tecken på våld och konflikt för att lära oss om tillämpning av en ny teknik. De journalanteckningar som användes är tagna direkt från handläggningen, var därmed realistiska och kunde tillsammans ge ett sammantaget och historiskt realistiskt textunderlag för annotering. Valet av data och vad det skulle användas för i piloten förankrades i berörd verksamhet, projekt och styrgrupp (jämför Q9 Is the purpose of using the data clear to all stakeholders?). Observera att alla persondata anonymiserades.
- Valideringen av källdata gjordes inte genom särskild metod utan genom att projektet valde rätt data för att kunna nå projektmålet(förre punkten), vilket var tillräckligt för piloten. Risker för snedvrida eller misstolka textdata vid annoteringen av texterna hanterades bland annat med att gjorda annoteringar även diskuterades i grupp (jämför Q7 Is the data validated?).
- Projektet valde den mängd data som behövdes för att ge det en, med styrgruppens godkännande, tillräckligt hög träffsäkerhet i prediktioner. Cirka 600 journaler annoterades för språkmodellen och test av språkmodellen gav en prediktion på över 86 % för tecken på våld och konflikt. Det bör dock noteras att om en driftsättning skulle bli aktuell så behöver träffsäkerheten förbättras ytterligare, även om exakt mål om hur hög träffsäkerhet inte är satt i dag. Fler journaler behöver då annoteras och gärna utifrån ytterligare kategorier (jämför Q10 Is the data sufficient for the current use case?).
- Projektet hade ett bra rättsligt stöd. Pilotens behandling av personuppgifter behövde stöd ha stöd i lagstiftningen, bland annat GDPR, och projektet behövde även känna till vilka begränsningar som fanns (jämför till exempel Q3 Do you have lawful access to the data? Q15 Are you allowed to share the data with others?).

5. Frågor att lösa för att gå vidare

Även om pilotprojektet inte syftade inte till att driftsätta en lösning, kom projektet ändå att diskutera vad som behöver tillkomma eller lösas för att kunna vidare med en driftsättning. En del av detta löstes teoretiskt, men inte prövades alltså inte i praktiken då det föll utanför projektmålet att genomföra. Utgångspunkten var processen för orosanmälan, men det skulle teoretiskt kunna vara andra ärendeprocesser med liknande drag.

Prioriteringar inom verksamheten

Data Readiness Assessment Method ägnar sig inte åt resursfrågor, men alla initiativ inom en verksamhet behöver prioriteras mot andra. Även om AI piloten gav värdefulla insikter, och kan ses som en komplettering till en redan driftsatt process (FAS-projektet), kommer en fortsättning med driftsättning av AI pilotens delarna kräva ytterligare insatser. Annotering (och anonymisering) behöver göras utifrån ett större textunderlag vilket kräver personell tid. Ytterligare RPA eller API behöver införas för att få programmatisk tillgång till journaldata istället för att kopiera text manuellt. För att komma vidare skulle ett konkret nästa steg kunna vara en nytto- och kostnadskalkyl.

Q1. Do you have Programmatic access to the data?

Piloten behövde inte lösa denna fråga för att nå projektmålet eftersom text markerades och flyttades manuellt. Om lösningen ska driftsättas som automatiserad process behöver RPA eller API användas för att hämta upp data direkt från verksamhetssystemets databaser. Databasens uppbyggnad för orosanmälan är känd från ett tidigare projekt kopplat till e-arkiv. Databaserna är välstrukturerade, väldokumenterade och innehåller igen kryptering så teoretiskt finns goda förutsättningar för programmatisk access.

Q2. Are your licenses in order? In the case of a third-party data provider.

Denna fråga var underordnad i piloten eftersom de journalanteckningar som kopierades och användes var läsliga i verksamhetssystemets användargränssnitt, det vill säga var tillgängliga genom den funktionalitet som avtalats med leverantören av systemet. Om lösningen skulle driftsättas som automatiserad process behöver hämtning ske från systemets databaser. Avtalet med leverantören av det verksamhetssystem för orosanmälan reglerar idag inte vare sig om det är tillåtet eller otillåtet vilket ger en osäkerhet vad som gäller.

Frågan om vad avtalet tillåter är större än för orosanmälan och är något organisationer överlag behöver arbeta med. Inför upphandling behöver beställaren ställa krav på avtalstext med rätt till data i de system eller tjänster som upphandlas. Målet kan vara en generell och återanvändbar avtalsskrivelse som tillåter användning av och direkt tillgång till data kopplat till beställarens verksamhet även för ändamål, andra än de gängse med handläggning, statistik och rapporter, som ligger i linje med beställarens behov och lagstiftning, till exempel krav på tillgängliggörande av öppna data för vidareutnyttjande.

Q3. Do you have lawful access to the data, samt närliggande frågor (Q2, Q14 och Q 15).

De frågor som diskuteras under den här rubriken är delvis lösta inom piloten men behöver besvaras på nytt inför en möjlig driftsättning eftersom rättsläge och avtalsskrivelser kan förändras.

De rättsliga frågorna vi kom i kontakt med under piloten var inte endast kopplade till om piloten hade laglig tillgång till datat (Q3) utan till exempel även till om behandlingen av personuppgifterna hade stöd i GDPR vilken bedömningen var att den hade. Andra rättsliga frågor var att säkerställa gallring av de kopierade data som använts i piloten när den avslutats.

En reflektion är att det kan vara till fördel att besvara flera av frågorna i Data Readiness Assessment Method parallellt då de delvis går in i varandra. Till exempel såg vi frågan om laglig tillgång till datat (Q3) med ett för piloten nödvändigt tillägg om tillåten behandling samt gallring av personuppgifter delvis gå in i frågorna "Q2 Are your licenses in order?" "Q14 Is it safe to share the data with others" och "Q15 Are you allowed to share the data with others?".

- Parallellt med avtalsmässiga tillgången (Q2) finns frågan laglig tillgången (Q3), med en följdfråga om vad det är som begränsar tillgången och om det är avtal, vad som går att göra åt det?
- Parallellt med säkerheten vid delning av datat (Q14), till exempel för träning av externt hållen språkmodell, finns både frågor (Q2) om avtalet tillåter sådan delning och frågor om lagen tillåter andra ta del av datat (Q15) utifrån rättsläget (Schrem II).

Den andra punkten blir särskilt aktuell för organisationer som överväger att använda externa språkmodeller, särskilt om en möjlig koppling till ägande i tredjeland kan misstänkas vilket är problematiskt utifrån dagens rättsläge.

Q4. Has there been an ethics assessment of the data?

De etiska aspekterna uppmärksammades tidigt i projektet, inte minst för att piloten skulle hantera känsliga persondata. I dessa frågor saknade vi tyvärr ett bra stöd för hur man bäst kan göra etiska bedömningar. EU:s etiska riktlinjer för tillförlitlig AI gav inte tillräckligt stöd. Det bör komma en bättre reglering och stöd om de etiska aspekterna. Frågan är generell, angelägen och långt ifrån isolerad till en driftsättning av pilotdelarna för orosanmälan. [Ett förslag till en EU lagstiftning](#) har tagits fram.

Q5. Is the data converted to an appropriate format?

Utifrån modellens förklaring av frågan var denna fråga svår att besvara för piloten. Inför hämtning av källdata direkt från verksamhetssystemets databaser såg projektet andra frågeställningar (se "Frågor utöver modellens frågor") som piloten teoretiskt löst avseende det system som används vid orosanmälan.

6. Flera frågor från metoden

Frågorna Q6 till Q12

- Q6. Are the characteristics of the data known? Frågan kunde inte förstås.
- Q7 till Q10 Är frågor som har tagits upp under avsnittet Framgångsfaktorer för piloten.
- Q11. Are the steps required to evaluate a potential solution clear? För piloten var svaret ja på denna fråga, men inget vi kunde vara säkra på förrän efter att projektet genomförts.
- Q12. Is your organization prepared to handle more data like this beyond the scope of the project? Denna fråga skiljer sig från de andra i modellen på så sätt att den inte syftar till det enskilda projektet utan till hur organisationen jobbar och bör jobba med databeredskap. Här behövs en fastslagen strategi och taktiska steg framåt.

7. Flera viktiga frågor

Eftersom kommunen inför driftsättning av lösningen behöver komma åt de databaser där journalanteckningarna lagras blir några ytterligare frågor relevanta för att åstadkomma programmatisk access till data med RPA eller API. Dessa frågeställningar är även relevanta för andra projekt där data hämtas från databastabell och inte från datafil eller webbplats

- Är databasen välstrukturerad, är det tydligt var den data man vill ha för att nå projektmålet ligger; samma tabeller, uppdelat på fler tabeller, konsekvent infört eller inte?
- Är data krypterat? Är svaret ja och krypteringsnyckeln inte är tillgänglig går det inte att komma åt data.
- Är teckenkodning känd? Nödvändigt för att få fram information från lagrat data.
- Ligger datat i egen driftmiljö eller som molnlagring? Molnlagring innebär i regel sämre tillgång.
- Förändringsledning för påverkad verksamhet, god acceptans finns för RPA inslag i de processer det är infört på kommunen men det är inte prövat för möjliga AI-inslag.

8. De största utmaningarna inför kommande projekt

Med kommande projekt kan avses både driftsättning av processen för orosanmälan (piloten) och nya andra processer där kommunen vill ta hjälp av språkmodeller. Utmaningarna kan även ses som generella.

- Prioriteringar mot andra initiativ. Träning av språkmodeller och införande av automatiserade processer för texttolkning med prediktioner kräver resurser och det finns en risk att mer akuta åtgärder går före. Eftersom införande av denna typ av relativt ny teknik handlar om organisationernas förflyttningar behöver en strategi tas fram och beslutas av ledningen.
- Den förflyttning en sådan strategi skulle innebära, behöver taktas med arbeten inom kompetensförsörjning, organisation, teknik, kultur och samverkan i syfte att över tid allt bättre möta upp verksamheternas behov.
- Avtal med leverantörer av verksamhetssystem. Eftersom användning av data för andra ändamål än funktionalitet för handläggning, statistik och rapporter sällan är avtalade behöver organisationer arbeta med generella avtalstexter som inkluderar sådana andra ändamål.
- Språkmodeller med koppling till ägande i tredjeland när behandling av personuppgifter ingår. Detta problem drabbar idag särskilt mindre till medelstora organisationer som behöver ta hjälp av externa språkmodeller för att få kraftfulla program för databehandling och dataanalys. Om alternativ växer fram inom Sverige eller åtminstone inom EU blir detta problem mindre.
- Anonymisering. I piloten skedde manuell anonymisering, men i driftsatta lösningar behöver det göras automatiserat med en tillförlitlig programvara som även kan tränas för att upptäcka indirekta personuppgifter, det vill säga som kan peka ut en viss person genom ett sammanhang utöver uppenbara direkta personuppgifter som personnummer och personnamn.
- Etiska överväganden. Det finns en utspridd farhåga inför utvecklingen av AI teknik, bland annat över hanteringen av persondata i sådan teknik. Kunskap och mognad om vilka risker och hur de kan undvikas behöver förbättras. En balanserad lagstiftning och riktlinjer behöver även tas fram.

New section follows

Fallstudie - Ekonomistyrningsverket

Hitta remissvar - Remisshantering med stöd av AI

Skriven av Oskar Weinberger

Projektbeskrivning

Sammanfattning

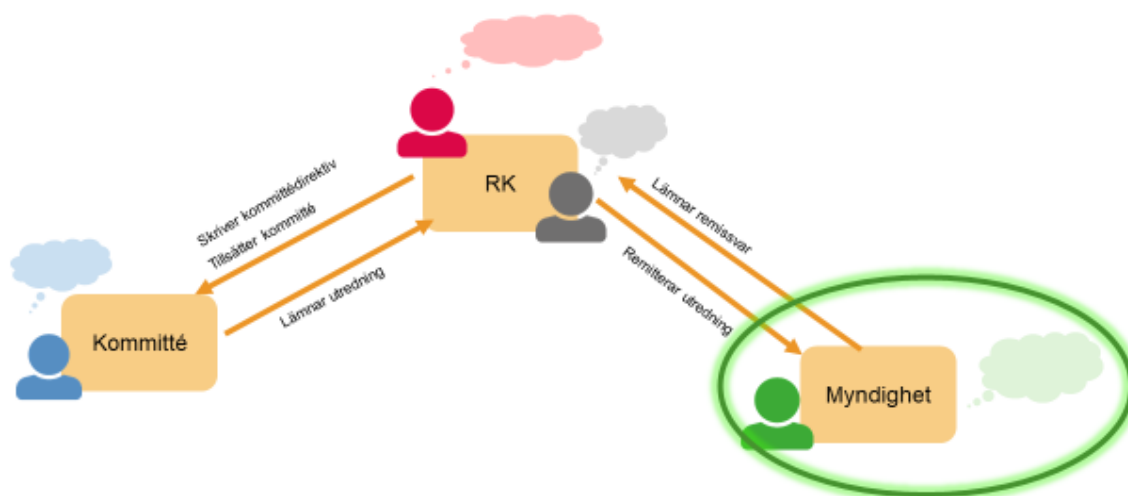
Hitta remissvar är ett pilotprojekt som Ekonomistyrningsverket (ESV) har drivit tillsammans med Tillväxtverket, Statens servicecenter och Statskontoret, myndigheter som alla har gemensamt att vi svarar på många remisser. Projektet har syftat till att utveckla ett verktyg som med hjälp av språkteknologi kan underlätta för myndighetshandläggare i deras arbete med att svara på remisser. En första version av tjänsten finns tillgänglig på [Hitta remissvar \(esv.se\)](https://hitta.remissvar.esv.se).

Applikationen låter användaren söka och välja bland förslag i statliga utredningar. En språkmodell ger sedan rekommendationer på vilka förslag i andra utredningar som kan vara relevanta för handläggaren baserat på semantisk likhet. Kopplat till varje utredning finns också de remissvar som har inkommit till Regeringskansliet. På så sätt kan en handläggare som fått in en ny utredning att skriva remissvar på snabbt hitta vad för liknande förslag som lagts i tidigare utredningar samt vad den egna myndigheten och andra remissinstanser då haft för åsikter.

Bakgrund



Remissförfarandet i staten



Figur 1

Remissförfarandet är en central del i den svenska politiska processen. En stor andel av de förslag till ny politik som läggs fram av regeringen föregås av en statlig utredning. Dessa utredningar skickas typiskt på remiss till berörda myndigheter, intresseorganisationer och näringslivsaktörer. Även organisationer och medborgare som inte blivit ombedda att skriva remissvar kan svara på remissen. Dessa får då möjlighet att ge synpunkter på utredningen kan därigenom påverka eventuella politiska beslut som tas med utredningen som underlag, och remissförfarandet är därför en viktig mekanism för demokratisk återkoppling.

Figur 1 illustrerar hur remissprocessen går till och de olika roller i flera olika organisationer som är involverade. Kortfattat och förenklat går det till på följande vis, där vi fokuserar på myndigheter som huvudsaklig svarsinstans: Tjänstepersoner på Regeringskansliet skriver kommittédirektiv och tillsätter en kommitté. Kommittén i sin tur tar fram en utredning som den lämnar till Regeringskansliet, som sedan bestämmer ifall den ska remitteras och vem/vilka som ska ombedjas att svara på den. Myndigheten tar emot den remitterade utredningen och bestämmer vem som ska ansvara för att ta fram ett svar och en utredare/tjänsteperson får i uppdrag att ta hand om arbetsuppgiften. När myndighetens svar är formulerat och avstämt med myndighetschefen så lämnas svaret till Regeringskansliet, som tar till sig och sammanställer synpunkterna för vidare hantering.

Det finns för var och en av de här stegen och rollerna ett antal behov och utvecklingsmöjligheter som skulle kunna realiseras med hjälp av AI och språkteknologi, men i detta projekt valde vi att fokusera på myndigheter som svarar på remisser. Detta eftersom det är den del vi själva har störst rådighet över och möjlighet att själva utveckla något utan att vara beroende av vilja och resurser från andra.



Projektets prioriterade behov



Figur 2

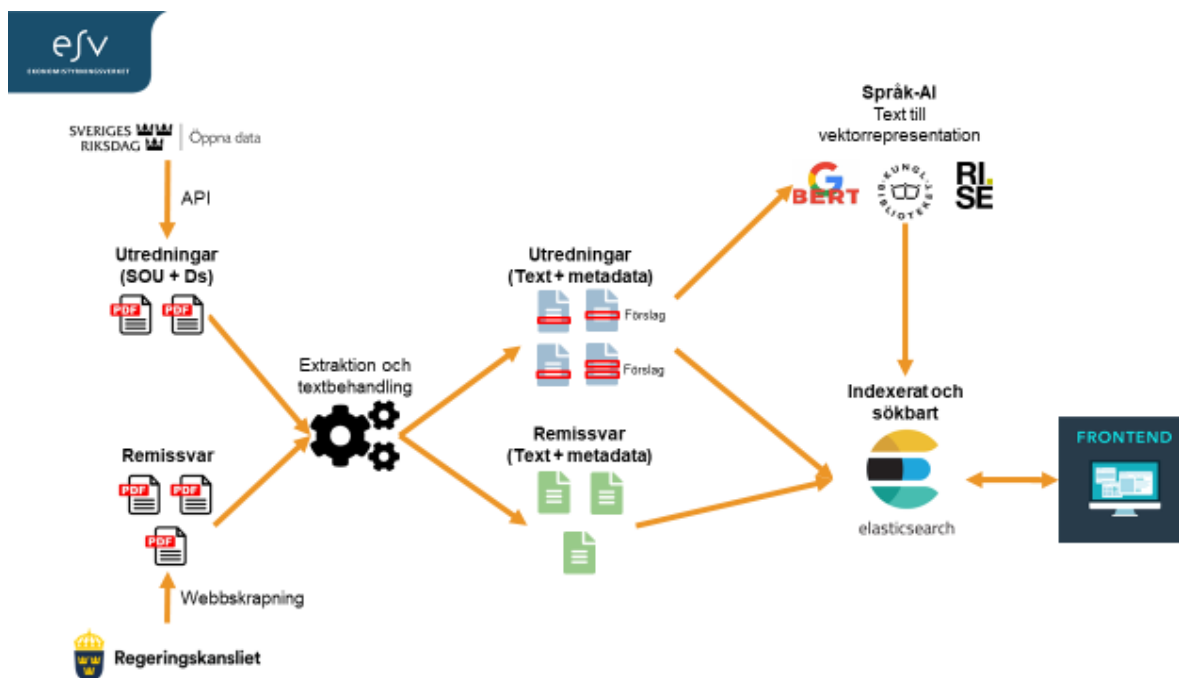
I projektets tidiga skeden hade vi ett antal möten där vi gemensamt kartlade behov och möjliga AI-tillämpningar kopplade till remissförfarandets olika roller. Det gjordes i samverkan mellan å ena sidan personer med insyn och erfarenhet av statlig styrning och utredningsväsendet och å andra sidan personer med kompetens inom programmering och språkteknologi. Vi tror att det är en viktig förutsättning att kombinera kunskap om verksamhetens behov med kunskap om tekniska och datamässiga möjligheter och begränsningar för att identifiera och utveckla potentiella AI-tillämpningar. Det är viktigt att förankra utvecklingen i faktiska behov så att den inte drivs helt av att bara vilja använda ett förutbestämt nytt AI-verktyg. Dock är det i vår erfarenhet inte heller lämpligt att endast utgå från existerande behov, utan det behövs kompletteras med och inspireras av kunskap om de nya tekniska möjligheterna för att hitta innovativa nya lösningar.

Figur 2 visar de behov för myndighetsrollen som vi valde att utgå från i utvecklingen av Hitta remissvar, med det inringade behovet som det huvudsakliga behovet. Vår analys var att detta är frågor som handläggare ofta ställer sig när de skriver remissvar men att det i nuläget inte finns datadrivna lösningar för att få bra svar. De dokument som behövs är svåråtkomliga och spridda över flera system med dålig sökbarhet, och det finns därför ett högt personberoende för att svara på den här typen av frågor. Med detta som utgångspunkt definierade vi därför följande mål med projektet:

- Att kunna göra avancerade sökningar i publicerade Ds- och SOU-utredningar.
- Att göra statliga utredningar mer överskådliga.
- Att skapa smarta matchningar mellan nya förslag och tidigare utredningsförslag.
- Tillgängliggöra alla remissvar för de matchade förslagen.

Databeredskap

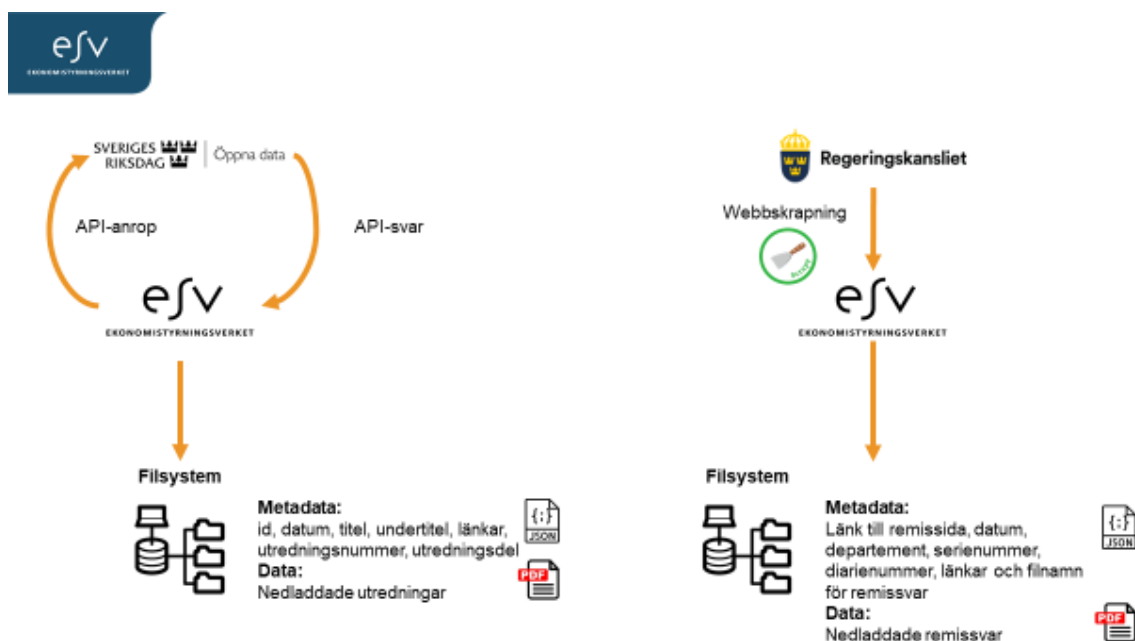
Vi kommer nu presentera applikationens nuvarande data-arkitektur och analysera den utifrån ett databeredskapsperspektiv med hjälp av ramverket för textdataberedskap i artikeln [We Need to Talk About Data: The Importance of Data Readiness in Natural Language Processing](#). Deras metod innefattar ett antal utvärderingsfrågor, och vi kommer att referera till ett urval av dessa löpande genom texten.



Figur 3

Figur 3 är ett förenklat diagram över applikationens olika komponenter och visar dataflödet från externa datakällor via processering till webbgränssnitt. Det visar att vi använt oss av två datamängder, hämtade från två externa datakällor: 1. Statliga utredningar (i nuläget endast serierna SOU och Ds) hämtade via öppet API från Riksdagens öppna data. 2. Remissvar hämtade med webbskrapning från regeringens webbsida. Dessa dokument har i huvudsak varit i PDF-format. Dokumenten har vi sedan processerat för att kunna indexera remissvar samt utredningarnas förslag i sökmotorn Elasticsearch som vi exponerat via ett API mot ett webbgränssnitt där användare kan göra sökningar m.m.

Datainhämtning



Figur 4

Q1 Do you have programmatic access to the data?

Eftersom den data vi behöver för Hitta remissvar inte förvaltas av oss utan ligger i externa datakällor blir frågan om programmatisk åtkomst i första steget en fråga om hur vi kan hämta data in till våra system. I figur 4 visas hur vår lösning för det ser ut. Vi börjar till vänster med inhämtningen av utredningar. För att hämta statliga utredningar från Riksdagens öppna data skickar vi först ett API-anrop till deras [öppet API](#), det vill säga ett gränssnitt där de själva på ett strukturerat och stabilt sätt ger öppen programmatisk åtkomst till riksdagens databaser. Anropet är ett [GET](#)-anrop på formen

GET <https://data.riksdagen.se/dokumentlista/?sok=&doktyp=sou&rm=&from=2022-01-01&tom=&ts=&bet=&tempbet=&nr=&org=&iid=&avd=&webbtv=&talare=&exakt=&planering=&facets=&sort=datum&sortorder=asc&rapport=&utformat=json&a=s#soktraff>

och specificerar vilken data vi vill få tillbaka i svaret. För tydlighets skull är de viktigaste bitarna i anropet rödmarkerade, nämligen att vi i detta fall ber om dokument av typen SOU, från och med 2022-01-01. Svaret som Riksdagens öppna data skickar tillbaka är ett JSON-objekt med en lista av metadata och url-länkar till utredningar. Ett utklipp av svaret skulle i detta fall se ut så här:

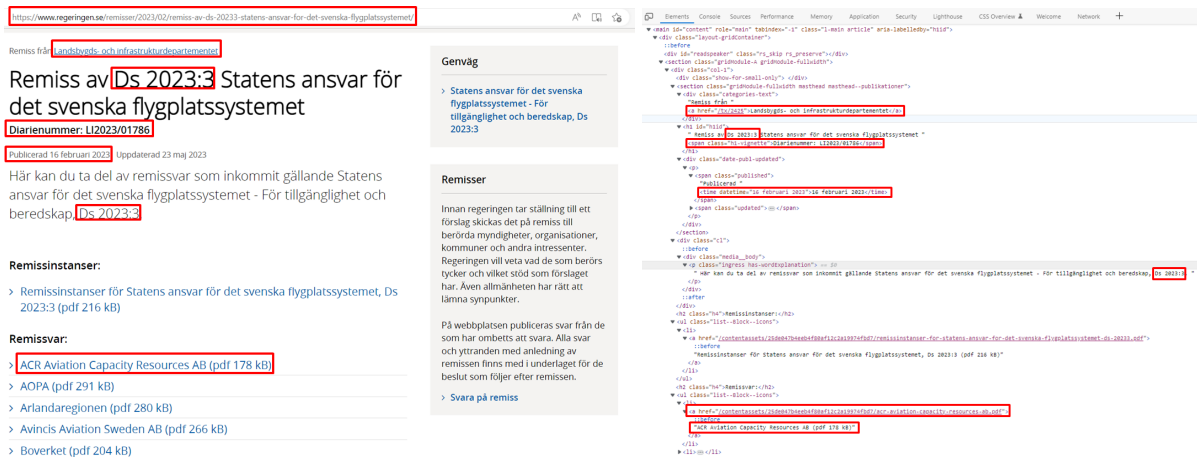
```

{
  "dokumentlista": {
    ...
    "dokument": [
      {
        "datum": "2022-01-01",
        "undertitel": "",
        "dok_id": "H9B36d2",
        "dokument_url_text": "//data.riksdagen.se/dokument/H9B36d2.text",
        "dokument_url_html": "//data.riksdagen.se/dokument/H9B36d2.html",
        "titel": "Hälsa- och sjukvårdens beredskap - struktur för ökad förmåga, del 2",
        "rm": "2021",
        "doktyp": "sou",
        "tempbeteckning": "d2",
        "nummer": "6",
        "filbilaga": {
          "fil": [
            {
              "typ": "pdf",
              "namn": "sou_2021__6_d2.pdf",
              "storlek": "3842534",
              "url": "http://data.riksdagen.se/fil/469765B8-70BF-47C4-9BE0-CAC6D087ADC0"
            }
          ]
        },
      },
      ...
    ]
  }
}

```

Från detta svar får vi metadata (dokument-id, datum, titel, undertitel, länkar, utredningsnummer, utredningsdel) samt url-länkarna som vi använder för att ladda ner PDF-filer och HTML-filer för utredningarna för intern lagring. Metadata sparar och lagrar vi i JSON-format.

Till höger i figur 4 visas schematiskt vår inhämtning av remissvar från Regeringskansliet som vi gör via [webbskrapning](#) av [regeringen.se/remisser](#) med Python-verktyget Scrapy. Det innebär att vi med Scrapy navigerar oss till remissidor på regeringens webbsida, laddar ned HTML-koden för dessa och letar efter länkar till remissvaren och relevant metadata. Figur 5 nedan visar till vänster ett exempel på en remissida med sådan information markerad och till höger var informationen finns att hitta i tillhörande HTML-kod.



Figur 5

I jämförelse med ett öppet API ger webbskrapning en sämre åtkomst till data, av flera anledningar. Till exempel kräver det en betydligt mer omfattande programmeringsinsats att göra datainhämtningen. Eftersom webbsidor är byggda för grafisk presentation och inte datahämtning så kan användbar information vara svåråtkomlig eller helt saknas. Ofta har de som skapat webbsidan inte haft tillgänglighet eller maskinläsbarhet som en hög prioritet och det kan därför krävas mycket behandling av HTML-koden för att extrahera rätt data. Det är heller inte ovanligt med inkonsekvent struktur, både på olika ställen på webbsidan och över tid. I projektet Hitta remissvar har vi haft problem med båda. Dels har olika remissidor kunnat ha något olika struktur och dels så har länkar plötsligt ändrats. Det finns vid webbskrapning alltså alltid en risk att lösningen för datainhämtning slutar fungera eftersom datakällan ändrat sin webbsida på något sätt.

Q9 Is the purpose of using the data clear to all stakeholders?

Vi har tolkat denna fråga som att det handlar om de krav tillämpningen ställer på data och huruvida alla berörda är medvetna och överens om dessa. Vi tar upp den här eftersom det visar på en viktig skillnad mellan inhämtning via API och via webbskrapning som förklarar mycket av nackdelarna med webbskrapning jämfört med ett API. Riksdagens öppna data har med sitt API byggt en tjänst för att tillhandahålla statliga dokument som data som kan hämtas programmatiskt och i stor skala. Även om de inte är medvetna om de specifika datakrav vår tillämpning har så är ett API ett löfte om att upprätthålla en viss struktur, tillgänglighet och stabilitet över tid. Regeringskansliet däremot har inte alls på samma sätt aktivt åtagit sig rollen som datakälla, och webbskrapning är ett försök att arbeta runt det.

Q10 Is the data sufficient for the current use case?

Ett annat viktigt perspektiv att analysera datakällor utifrån är hur heltäckande de är. I vårt fall är frågan hur många av alla statliga utredningar och remissvar vi kan komma åt via Riksdagens öppna data respektive Regeringskansliets webbsida. Det är en alltså en fråga om datas tillgänglighet (fråga Q1), men också om datas användbarhet/tillräcklighet (fråga Q10) eftersom det beror på den specifika applikationen om all data behövs eller om applikationen fortfarande är användbar med bara en delmängd.

Riksdagens öppna data har en relativt komplett samling av statliga utredningar, där utredningar saknas bara i undantagsfall. Däremot har regeringen.se/remisser väldigt få remissvar sparade innan 2015. Vad vi är medvetna om finns heller inga andra publicerade

eller maskinellt åtkomliga datakällor där dessa äldre remissvar går att få tag på. Vi gjorde bedömningen att det finns ett värde i en tjänst som Hitta remissvar trots att historiken över remissvar i all väsentlighet är begränsad till efter 2015 och att det kan förekomma att listorna med publicerade remissvar inte är kompletta.

Q6 Are the characteristics of the data known?

Q7 Is the data validated?

En explorativ analys av datas egenskaper och validering går hand i hand. Validering av data handlar om att kontrollera att data är korrekt, och för att kunna göra det behövs en uppfattning om vilka egenskaper data borde ha. Eftersom utforskning av data och validering behöver göras i flera steg av databehandlingen kommer vi att återkomma till dessa frågor i senare delar av texten också. För sifferdata går det t.ex. att beräkna statistiska mått, rita fördelningar, hitta outliers och saknade värden osv. Textdata däremot är i högre utsträckning kvalitativt och ostrukturerat, så det är inte självklart vad, om något, som kan analyseras och mätas för att kontrollera riktigheten. Vi har därför tyckt det varit svårt att veta vad explorativ analys och validering av textdata kan innebära i praktiken och det har varit en utmaning att jobba med systematiskt. Exempel på frågor som vi ändå identifierat och undersökt är:

- Innehåller filerna det de borde? Det vill säga statliga utredningar och remissvar av rätt typ från rätt avsändare med rätt metadata. En väldigt grundläggande koll som är väldigt svår att automatisera. Istället har vi gjort ett urval av dokument som vi kontrollerat manuellt. Många dokument har vi också kollat av andra anledningar under utvecklingsprocessen. Vi har på så sätt hittat några dokument där metadata inte stämmer överens med innehållet, till exempel en utredning som ser ut att handla om Nato men egentligen är en utredning om äldrevård eller remissvar som innehåller svaret från en annan remissinstans. Vi har då fått korrigera dessa för hand.
- Hur många utredningar och remissvar hämtar vi? Hur många sidor långa är de? Hur många är det per år, dokumenttyp eller remiss? Det här är exempel på enkla kvantitativa frågor som går att beräkna för hela text-korpusen och som kan ge en indikation på om det finns problem med data. Vi kunde till exempel se att regeringen.se har ett relativt stabilt antal remissvar sparade från 2015 och framåt men att det innan det finns väldigt få.
- Är data heltäckande? En fråga som är relaterad till den innan men ganska svår att svara på. För remissvar var det lätt att se att data i princip saknas innan 2015, men hur komplett remisvarsdata är efter det finns inget facit som är maskinellt tillgängligt för oss att jämföra mot eftersom regeringen.se så långt vi vet är de enda som publicerar relativt kompletta sammanställningar med remissvar. Vi har av en slump sett att ett fåtal remissvar försvunnit från regeringens webbsida när vi gjort nya webbskrapningar för att uppdatera vår data och vet därför att inte heller data efter 2015 kan vara helt komplett. För utredningar finns ingen skillnad i datamängd mellan år eller liknande tydlig indikator på att data inte är heltäckande, och vi tror att Riksdagens öppna data har en mer eller mindre fullständig samling av statliga utredningar. En mer noggrann kontroll av detta hade varit möjlig genom att jämföra antalet utredningar med informationen som redovisas i regeringens årsbok, vilket vi dock inte gjort.
- Vilka länkar går inte att öppna? Är dokument korrupta eller bristfälliga på andra sätt? Detta är också frågor som går att försöka besvara kvantitativt. Länkar som inte går att öppna är lätta att registrera och räkna. Exempel på andra dokumentproblem som vi hittat och kvantifierat är dokument med ingen eller väldigt lite text (typiskt

bild-PDFer), råtext med inga eller väldigt få blanksteg (en del äldre utredningar) och dokument med väldigt hög andel specialtecken (på grund av problem med avläsning av teckenkodningen, för en del äldre utredningar).

- Vilka värden antar metadata? Saknas vissa värden? Metadata är mer strukturerad än själva textdatan och därför ofta lättare att undersöka och validera eftersom det går att använda enkla metoder för strukturerad data. Vi har till exempel undersökt distinkta värden för departement och svarsinstanser och kollat om titeldata för utredningarna saknas. Vi har då sett att i princip alla äldre utredningar saknar titel i metadata (ett problem vi valt att inte åtgärda) och att nya utredningar till en början inte har titeldata (vilket vi korrigerar manuellt vid datauppdatering).

Q12 Is your organization prepared to handle more data like this beyond the scope of the project?

Den här frågan handlar om datainhämtning över tid, vilket vi har tänkt på uppdelat i två delar. Den första delen handlar om vilken kapacitet som finns att uppdatera med ny data. Målet är en automatiserad, schemalagd och övervakad process för uppdatering av data. I den nuvarande applikationen är uppdatering av data något som initieras manuellt och också kräver en hel del handpåläggning, men det är under utveckling att ha större delen av uppdateringen automatiserad och schemalagd. Det ställer högre krav på den interna datahanteringen eftersom vi inte bara skriver till data en gång utan flera och det därför finns risk att vi skriver över eller slår samman med gammal data på fel sätt. Antagligen kommer vi också vilja ha någon form av koll på ändringshistoriken av data i dokumentations- och utvärderingssyften. Exempel på utmaningar i automatiseringen av uppdateringen som vi stött på är exempelvis att remissvar ibland ändras till nya versioner eller till och med försvinner från regeringens webbsida. Vi kan därför inte bara skrapa hem alla remissvar och ersätta den gamla datan, utan vi måste jämföra gammal och ny data och slå samman på rätt sätt. För utredningarna har vi haft problemet att utredningar som nyligen lagts till i Riksdagens öppna data ofta saknar titel vilket är svårt att programmatiskt korrigera.

Den andra delen handlar om att förbättra processer för att skapa data från dokument. Vilken metadata behövs? Vilka format borde vi egentligen använda? Vilka system ska data lagras i? Vem lägger in textdata i databasen när, och på vilket sätt? Det är inte bara en teknisk fråga eftersom dokumenten skrivs och används som en del av den vanliga verksamheten, som alltså måste ändra sina arbetssätt och förhållande till dokumenten för att göra dem till bättre data. I fallet där den egna organisationen genererar och förvaltar textdata är detta en viktig del av att framtidssäkra databeredskapen. För Hitta remissvar däremot, där vår data kommer från externa parter, handlar denna punkt snarare om att föra en dialog med dem om hur deras data kan förbättras och framförallt hur det kan undvikas att ny data har samma databeredskapsproblem som gammal data.

Lagring

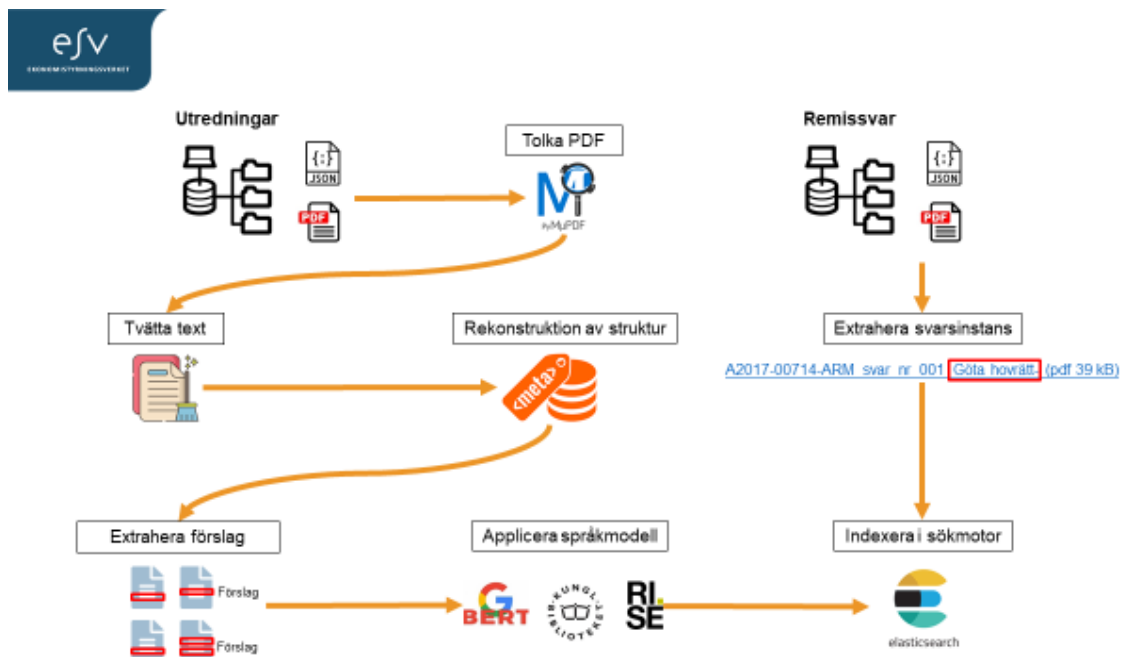
Q1 Do you have programmatic access to the data?

Vi har tänkt på frågan om programmatisk åtkomst som kopplad dels till inhämtning från eventuella externa källor men också till interna system och datalagring. Liksom många andra myndigheter och kommuner lagrar Ekonomistyrningsverket sina textdokument i ett ärendehanteringssystem. Det är byggd för manuell hantering, diarieföring och arkivering

men uppfyller inte de krav på programmatisk åtkomlighet som behövs för att använda text som data i en applikation som Hitta remissvar. Detta eftersom det varken finns ett smidigt gränssnitt för att hämta ut data i bulk eller tillräckligt fullständig eller strukturerad metadata för att kunna sortera och hitta bland dokumenten. Vi hade till exempel stora svårigheter att få ut alla våra egna remissvar ur vårt ärendehanteringssystem och det var faktiskt lättare att hämta dem genom vår webbskrapning av regeringens webbsida. Vi vet också om andra myndigheter som byggt AI-prototyper som valt att skrapa sina egna webbsidor på interna dokument istället för att gå via ärendehanteringssystem.

För Hitta remissvar har vi istället valt en enkel lösning där vi lagrar rådata i ett filsystem på en gemensam server. Filsystemet används också för lagring av mellansteg i processeringen medan färdigpreparerad data läggs in och görs sökbar för användarna i sökmotorn Elasticsearch. Vi har alltså valt bort att använda en av diverse dokumentdatabaser. Anledningen till det är dels att vi inte har tidigare erfarenhet av att arbeta med dem och därför bedömde det som en för stor ansträngning att implementera i detta skede av projektet. Vi har också gjort bedömningen att vi ännu inte har behov av mer avancerade lagringslösningar eftersom vår data ändras relativt sällan och av få personer. Detta kan vara något vi behöver omvärdera, till exempel om vi blir fler utvecklare i projektet och därför har fler som läser och ändrar i data samtidigt eller som en följd av en ändrad kravbild i takt med att vi automatiserar och schemalägger inhämtningen av data.

Processering



Figur 6

Figur 6 visar flödet av dokumenthantering och datatransformationer från lagring av dokument till färdig och sökbar data. De ingående stegen kan sammanfattas på följande vis:

- Tolka PDF: Vi använder Python-paketet PyMuPDF för att läsa ut råtext från dokumenten, hämta eller gissa textstil och textstorlek och gissa styckesindelningar/textblock.

- Tvätta text: Hantering av blanksteg och specialtecken samt borttagning av stycken med få eller inga bokstäver.
- Rekonstruktion av struktur: Kod för att utifrån råtext och metadata från tolkningssteget inferera rubriker, fotnoter, sidhuvuden, sidfötter, innehållsförteckningar och referenser. Detta är ett större processeringssteg som behövs eftersom PDF-filer typiskt saknar en stor mängd av denna strukturinformation. Se nedan för mer diskussion.
- Extrahera förslag: Kod för att identifiera och plocka ut förslagen ur utredningarna. Vi rensar först bort fotnoter, sidhuvuden och sidfötter. Sen hittar vi textblock som innehåller textsträngen "förslag:" men som inte är rubriker, innehållsförteckningar eller referenser. Det finns också vissa undantag som inte tas med, till exempel "skälen för våra förslag:". Efter att första textblocket i ett förslag hittats slår vi ihop det med nästa textblock tills en sammanlagd maxlängd är nådd eller vi stöter på ett nytt förslag.
- Applicera språkmodell: Transformerar förslagstexterna till vektorer med en modell för semantisk likhet, Contrastive-Tension/BERT-Base-Swe-CT-STSB. Detta gör att förslagens skrivelser kan jämföras semantiskt genom att matematiskt beräkna hur nära deras vektorrepresentationer ligger varandra.
- Extrahera svarsinstans: Plockar ut namnet på svarsinstansen från remissvarets filnamn som vi skrapat från regeringens webbsida.
- Indexera i sökmotor: Här läggs text, metadata och vektorer in tillsammans i Elasticsearch och görs sökbara.

Q5 Is the data converted to an appropriate format?

Mycket av processeringen i figur 6 behövs för att hantera att PDF-formatet är ett grafiskt presentationsformat med dålig metadata för strukturinformation. Med strukturinformation menar vi indikatorer för exempelvis rubriker, underrubriker, brödtext, styckesindelning, läsordning, tabeller, bilder osv. Bristen på sån metadata gör att råtexten från en PDF-fil kan bli en soppa av ord och tecken där rubriker, brödtext, fotnoter, tabeller och till och med läsordning blandas utan att det går att utröna vad som egentligen är menat att stå. Det gör det också svårare att extrahera intressanta delar av texten. Om till exempel utredningarna från början hade behandlats som data och sparats med användbara metataggar i ett mer maskinläsbart format än PDF så hade förslag kunnat ha egna metataggar. Vi hade då kunnat förenkla vår pipeline väsentligt och plocka ut förslag från dokumenten i ett direkt och enkelt steg. Det hade minskat den totala arbetsinsatsen men också eliminerat flera källor till fel i konverteringen från PDF till användbar data.

Remissvaren från Regeringskansliet finns bara tillgängliga i PDF. Riksdagens öppna data har flera format att välja på, bland annat PDF, HTML och JSON. Trots att både HTML och JSON i teorin är bättre format för textdata så har vi valt att använda PDF-filer även för utredningarna. HTML- och JSON-filerna ser nämligen inte ut att innehålla mer strukturinformation än vi kan få ut från PDF-filerna. Vad vi kan bedöma är anledningen att det är PDF-filerna som är ursprungsformatet som sedan konverterats till HTML och JSON. Vi har därför valt att extrahera textdata direkt från PDF-filerna eftersom det låter oss skräddarsy i högre utsträckning vilken information vi vill extrahera. Att använda PDF-filer har också fördelen att PDF-tolkare som PyMuPDF kan använda viss grafisk information för att inferera struktur, till exempel för att gissa textblock/stycken, vilket vi inte vet om konverteringen på Riksdagens öppna data utnyttjat.

En kommentar här är att PDF aldrig är ursprungsformat för textdata. Typiskt skrivs texten i Word eller liknande redigeringsprogram och konverteras först när texten är klar till PDF, vilket sedan är det format som utredare och myndighetshandläggare delar som officiell version till andra. Dessa arbetsdokument i Word finns väldigt sällan sparade eller tillgängliga. Det går att skapa Word-dokument med användbara metataggar genom att exempelvis använda mallar (det finns mallar för hur statliga utredningar ska formateras), så Word är i grunden ett mer maskinläsbart format än PDF. Den informationen går dock förlorad när Word konverteras till PDF istället för ett mer maskinläsbart format.

Om det enda syftet vore att lagra text som data borde inte PDF användas alls, men om det ska användas så är tillgänglighetsanpassad PDF (PDF/UA) att föredra eftersom det använder sig av metataggat innehåll. Sedan 2019 finns på EU-nivå ett tillgänglighetsdirektiv vilket gjort att statliga texter sedan dess oftare använder tillgänglighetsanpassad PDF. Utmaningarna för våra syften är dock fortfarande flera. Bland annat så efterföljs inte alltid direktivet, eller så görs tillgänglighetsanpassningen i ett senare skede efter att en icke-tillgänglig version först spridits externt. När det gäller remissvaren är de väldigt sällan tillgänglighetsanpassade. Ett annat problem är att PDF/UA fortfarande är relativt ovanligt internationellt, vilket gör att det finns få verktyg för att avläsa de tillgänglighetsanpassade metataggarna och därmed fullt ut dra nytta av det bättre formatet. Det gäller särskilt om man som vi vill göra det programmatiskt för tusentals dokument snarare än manuellt för ett fåtal dokument. Slutligen kan inte heller tillgänglighetsanpassad PDF innehålla mer information än den metadata som läggs in. För att få verkligt maskinläsbar textdata krävs därför, förutom bättre lagringsformat, också att användbar metadata taggas när texterna produceras.

Q6 Are the characteristics of the data known?

Q7 Is the data validated?

Nedan går vi igenom den explorativa analys och validering vi gjort för varje steg i figur 6.

- Tolka PDF: För ett urval av dokumenten har vi manuellt kollat följande.
 - Får vi med all text?
 - Hur hanteras bilder, tabeller, osv?
 - Hur blir indelningen i textblock?
 - Hur ser informationen om textstil och storlek ut?
- Tvätta text: Återigen en manuell jämförelse mellan text innan och efter tvätt. Tas för mycket eller för litet bort?
- Rekonstruktion av struktur: För att kunna inferera strukturinformation undersökte vi manuellt ett urval av flera olika utredningar för att se vilka egenskaper de hade och hitta en metod som fungerar för alla. Det var särskilt viktigt att göra ett urval med utredningar spridda över flera år eftersom kvalitet och struktur för utrednings-PDFerna varierat mycket över tid. Vi jämförde manuellt de beräknade metataggarna med ursprungsdokumentet och modifierade sen vid behov metoden för att gissa metataggar och jämförde igen tills vi hade tillräckligt få fel.
- Extrahera förslag: Genom att kolla på ett flertal utredningar över tid kunde vi se att de flesta förslag sedan tidigt 2000-tal dels står i grafiskt markerade förslagsrutor (vilket vi i nuvarande version av applikationen inte utnyttjat) och dels att de oftast indikeras av "förslag:". Vi valde därför extraktionsmetoden som beskrivits tidigare. För att validera och iterativt förbättra metoden har vi undersökt följande frågor.
 - Får vi med alla förslag?
 - Får vi med sånt som inte är förslag?

- Får vi med för mycket eller för litet av förslagen?
- Finns det kvalitetsproblem med extraherade texten?

Utvärdering

Q11 Are the steps required to evaluate a potential solution clear?

Det finns inget facit för vad som är "rätt" liknande förslag, och vi har inte hittat något naturligt att jämföra med som proxy. Exempel på vad som skulle kunna legat till underlag för en sådan proxy är om remissvar på ett strukturerat och förutsägbart brukade refererade till tidigare remissvar och utredningar som haft bäring på svarsinstansens nya yttrande. Istället har vi därför haft några olika spår för utvärdering som vi också hoppas kunna vidareutveckla.

Dels kan vi utvärdera applikationen som helhet genom att fråga användare vad de tycker. Under projektets gång har vi haft flera testrundor både internt och externt där projektdeltagare och myndighetsanställda kunna ge feedback både på hur bra matchningarna är men också annan funktionalitet, webbgränssnittets utseende och användarvänlighet. Vi har också planer på att utforma en enkät för att få ytterligare återkoppling på tjänstens användbarhet från diverse intressenter.

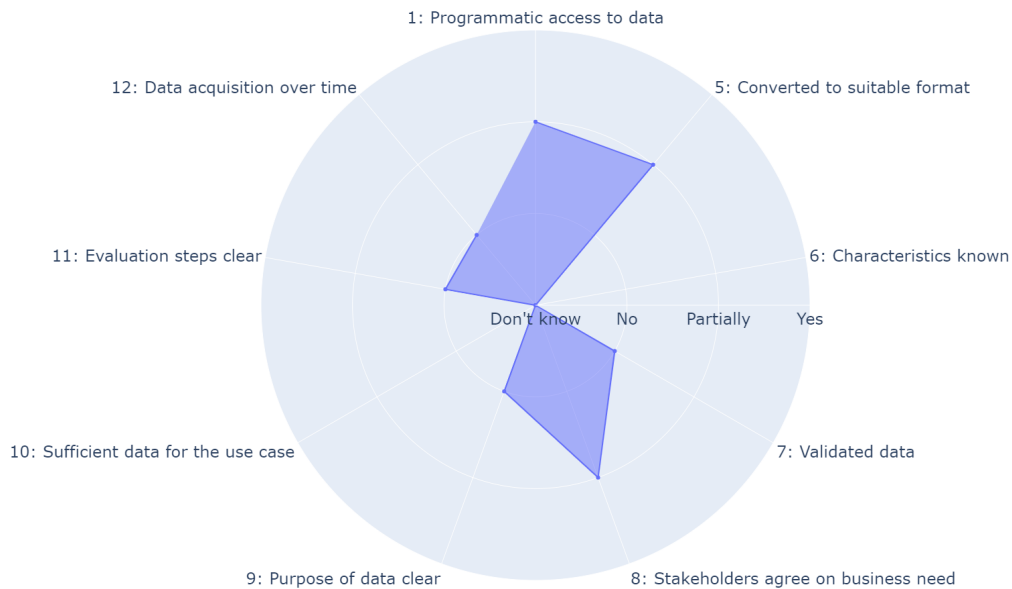
För att mer specifikt utvärdera rankingen av liknande förslag har vi sedan tidigare hållit i en workshop för statliga utredare från projektets partnermyndigheter där de fått i uppdrag att göra egna listor med liknande förslag. Dessa listor har vi sparat i Excel-format och sen använt som jämförelsematerial som vi kan testa vår lösning mot genom att mäta hur stort överlappet är mellan utredarnas listor och de AI-genererade liknande förslagen, inom engelska litteraturen kallat recall. Om överlappet är stort är det en indikator på att våra automatiska matchningar inte är helt irrelevanta.

Att på detta sätt använda recall som mått har vissa brister, där det största är att det inte alls fångar möjligheten att vårt system hittar relevanta liknande förslag som utredarna inte hittar. I någon mening är dessa oväntade men relevanta förslag precis vad vi vill att applikationen ska kunna hitta, men det är svårt att hitta sätt att utvärdera. Ett annat problem är att det är tidskrävande för utredarna att göra manuella matchningar. På ett sätt visar det i sig värdet av applikationen eftersom en lista med matchningar som för en människa tar över en timme att skapa tar mindre än en sekund att göra maskinellt. Det gör dock också att utvärderingsmaterialet från vår workshop fick en ganska låg volym och därför inte kan ge ett tillförlitligt kvantitativt mått på systemets kvalitet. Vårt mål är att i projektets nästa fas genomföra flera workshops för att få mer underlag.

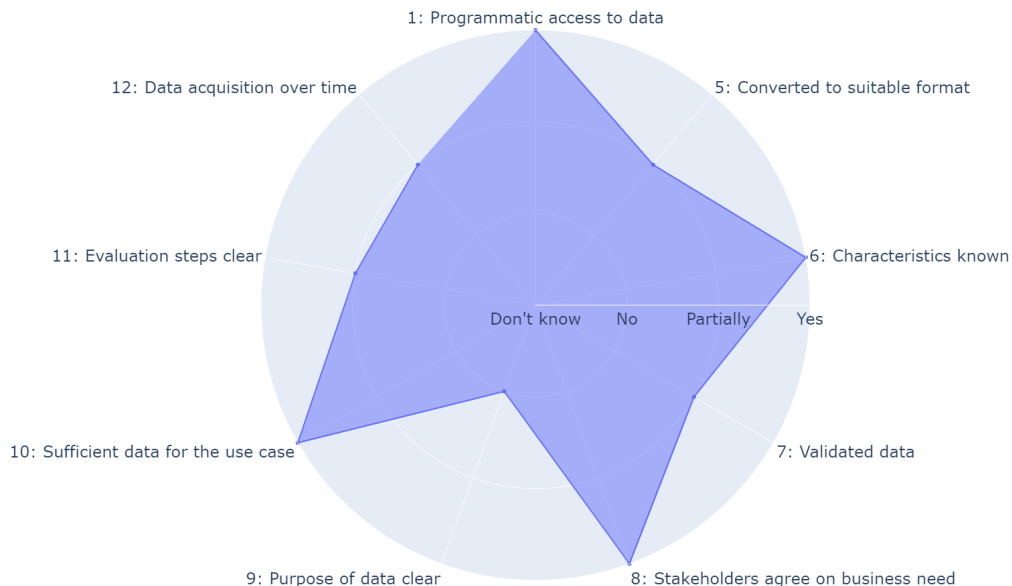
En annan möjlighet vi diskuterat är att bygga in funktionalitet för användarfeedback i webbgränssnittet. Användarna (antingen alla eller endast utvalda testare) skulle då kunna indikera om de listade förslagen är bra eller inte. I den enklaste varianten hade detta kunnat göras med en tumme upp eller en tumme ner. På så sätt skulle vi kunna bygga upp en bank med exempel på bra och dåliga matchningar som kan användas både för att kvantitativt utvärdera nuvarande system men också som träningsdata för att träna språkmodeller på att bli ännu bättre på just vår datamängd och problemställning, det vill säga semantisk likhet mellan förslag i statliga utredningar.

Data readiness assessment

Här följer en sammanställning av våra svar på de frågor ur [We Need to Talk About Data: The Importance of Data Readiness in Natural Language Processing](#) med störst relevans för applikationen Hitta remissvar. Vi visar en uppskattning av hur beredskapen såg ut innan vi hade börjat bygga applikationen, till vänster i figur 7, och en uppskattning för hur den såg ut efter att den första versionen av applikationen hade lanserats, till höger i figur 7.



Figur 7



Figur 8

New section follows

Huvudbilaga

Denna bilaga syftar till att ge en snabbreferens till metoden för bedömning av databeredskapsnivå som används av medlemmarna i detta projekt. Vi redovisar här hela listan med frågor som de finns i artikeln "[We Need to Talk About Data: The Importance of Data Readiness in Natural Language Processing](#)"⁶.

Q1 Do you have programmatic access to the data? The data should be made accessible to the people who are going to work with it, in a way that makes their work as easy as possible. This usually means programmatic access via an API, database, or spreadsheet.

Q2 Are your licenses in order? In the case you plan on using data from a third-party provider, either commercial or via open access, ensure that the licences for the data permit the kind of usage that is needed for the current project. Furthermore, make sure you follow the Terms of Service set out by the provider.

Q3 Do you have lawful access to the data? Make sure you involve the appropriate legal competence early on in your project. Matters regarding, e.g., personal identifiable information, and GDPR have to be handled correctly. Failing to do so may result in a project failure, even though all technical aspects of the project are perfectly sound.

Q4 Has there been an ethics assessment of the data? In some use cases, such as when dealing with individuals' medical information, the objectives of the project require an ethics assessment. The rules for such a probe into the data are governed by strict rules, and you should consult appropriate legal advisors to make sure your project adheres to them.

Q5 Is the data converted to an appropriate format? Apart from being accessible programmatically, and assessed with respect to licenses, laws, and ethics, the data should also be converted to a format appropriate for the potential technical solutions to the problem at hand. One particular challenge we have encountered numerous times, is that the data is on the format of PDF files. PDF is an excellent output format for rendering contents on screen or in print, but it is a terrible input format for data-driven automated processes.

Q6 Are the characteristics of the data known? Are the typical traits and features of the data known? Perform an exploratory data analysis, and run it by all stakeholders in the project. Make sure to exemplify typical and extreme values in the data, and encourage the project participants to manually look into the data.

Q7 Is the data validated? Ensure that the traits and features of the data make sense, and, e.g., records are deduplicated, noise is catered for, and that null values are taken care of.

⁶ Olsson, Fredrik, and Magnus Sahlgren. "We Need to Talk About Data: The Importance of Data Readiness in Natural Language Processing." *arXiv preprint arXiv:2110.05464* (2021).

Q8 Do stakeholders agree on the objective of the current use case? What problem are you trying to solve? The problem formulation should be intimately tied to a tangible business value or research hypothesis. When specifying the problem, make sure to focus on the actual need instead of a potentially interesting technology. The characteristics of the problem dictates the requirements on the data. Thus, the specification is crucial for understanding the requirements on the data in terms of, e.g., training data, and the need for manual labelling of evaluation or validation data. Only when you know the characteristics of the data, it will be possible to come up with a candidate technological approach to solve the problem.

Q9 Is the purpose of using the data clear to all stakeholders? Ensure that all people involved in the project understands the role and importance of the data to be used. This is to solidify the efforts made by the people responsible for relevant data sources to produce data that is appropriate for the project's objective and the potential technical solution to address the objective.

Q10 Is the data sufficient for the current use case? Given the insight into what data is available, consider the questions: What data is needed to solve the problem? Is that a subset of the data that is already available? If not: is there a way of getting all the data needed? If there is a discrepancy between the data available, and the data required to solve the problem, that discrepancy has to be mitigated. If it is not possible to align the data available with what is needed, then this is a cue to go back to the drawing board and either iterate on the problem specification, or collect suitable data.

Q11 Are the steps required to evaluate a potential solution clear? How do you know if you have succeeded? The type of data required to evaluate a solution is often tightly connected to the way the solution is implemented: if the solution is based on supervised machine learning, i.e., requiring labelled examples, then the evaluation of the solution will also require labelled data. If the solution depends on labelled training data, the process of annotation usually also results in the appropriate evaluation data. Any annotation effort should take into account the quality of the annotations, e.g., the inter-annotator agreement; temporal aspects of the data characteristics, e.g., information on when we need to obtain newly annotated data to mitigate model drift; and, the representativity of the data.

Q12 Is your organization prepared to handle more data like this beyond the scope of the project? Even if the data processing in your organization is not perfect with respect to the requirements of machine learning, each project you pursue has the opportunity to articulate improvements to your organization's data storage processes. Ask yourself the questions: How does my organization store incoming data? Is that process a good fit for automatic processing of the data in the context of an NLP project, that is, is the data stored on a format that brings it beyond Band C (accessibility) of the Data Readiness Levels? If not; what changes would need to be made to make the storage better?

Q13 Is the data secured? Ensure that the data used in the project is secured in such a way that it is only accessible to the right people, and thus not accessible by unauthorized users. Depending on the sensitivity of the project, and thus the data, there might be a need to classify the data according to the security standards of your organization (e.g., ISO 27001), and implement the appropriate mechanisms to protect the data and project outcome.

Q14 Is it safe to share the data with others? In case the project aims to share its data with others, the risks of leaking sensitive data about, e.g., your organization's business plans or abilities have to be addressed prior to sharing it.

Q15 Are you allowed to share the data with others? In case the project wishes to share its data, make sure you are allowed to do so according to the licenses, laws, and ethics previously addressed in the project.