Fra fjærpenn til maskinlæring

Tidvis sine erfaringer med å jobbe med Transkribus

Prosjektet "Fra fjærpenn til maskinlæring" har utforsket muligheten for å lære opp en maskin til å lese gamle arkivmanuskripter og muligheten for å formidle dette. Prosjektet har vært finansiert av Arkivverket og Kulturetaten (kulturavdelingen) i Oslo kommune og Kulturrådet. Prosjektet startet i januar 2020 og ble avsluttet i oktober 2021. Resultatene blir publisert på en CC-by-SA-lisens (open access). Personer som har vært sentrale i prosjektet er Ragnhild Hutchison, prosjektleder, Amund Pedersen, Gaute Remman Gunleiksrud, Anne-Sofie S. Skaar, Jon Christian Brekke og Sveinung Næss.

Resultatet er publisert på https://tidvis.no/historiske_databaser/lensregnskapene/ Vi håper med tid å få det lagt inn som del av digitalarkivet.no når dette blir gjort teknisk mulig av Arkivverket.

Målet med "Fra fjærpenn til maskinlæring" har vært å utforske om og hvordan maskinlæring kan brukes til å gjøre gammelt arkivmateriale tilgjengelig for et bredere publikum. Dette er en del av et bredere mål om å demokratisere kunnskap, i vårt tilfelle å finne måter å gjøre kunnskap om fortiden mer tilgjengelig på. I prosjektet har det også vært sentralt å utforske muligheter for hvordan materialet kan formidles.

Prosjektet omhandler "gammelt arkivmateriale", som vi i dette tilfellet forstår som dokumenter fra 1500- og 1600 -tallet som opprinnelig er håndskrevet med fraktur og/ eller gotisk håndskrift, med både romerske og/ eller arabiske tall. I denne rapporten fokuseres det på våre erfaringer med å trene opp AIen til Transkribus.

Gjennomføringen

Vi har brukt lensregnskaper for tiden 1557-1630 som case, med særlig fokus på Østlandet og Oslo. Disse er bevart i Riksarkivet. Deler av disse er tilgjengelig digitalt i digitalarkivet, og noen har vi fotografert selv.

Lensregnskapene er regionale regnskaper, og inneholder dermed de regionale myndighetenes utgifter og inntekter (f.eks. bøter, skatt og toll). De kan avsløre mye om f.eks. spredningen av en pengeøkonomi og handelsnettverk, eller gi innsikt i tidligere oppfatninger om moral og verdi. De tidligste lensregnskapene ble transkribert og utgitt mellom 1887 og 1930-tallet. I dag er disse tilgjengelige som bøker, men også digitalt på bokhylla.no (og der også søkbare). Vi har arbeidet med lensregnskapene for årene som ikke er publisert, og har hatt hovedfokus på tollprotokollene i disse arkivene. Fokuset er motivert av behovet for å begrense materialet som er studert, men også fordi vi som nevnt har laget en database med norske tollregistre for det lange 1700- tallet som vi håper kan utvides tilbake i tid.

Hva har vi gjort?
Vi har trent vår AI på 479 bilder av sider som er transkribert. Det transkriberte materialet er en kombinasjon av

- 121 sider fra de tidligere utgitte Lensregnskapene
- 358 sider med nytt materiale har blitt transkribert

Resultater

Hva har vi funnet?

Resultatet, etter at de ovennevnte sidene matet inn i Transkribus, er

- 6,52% CER (feilprosent) på treningssettet
- 7,49% CER (Faliur rate) på valideringssettet

Med mer transkripsjon matet inn i det kan dette antas å falle litt lenger, men vi kan si oss fornøyde med dette.

Merk: feilraten er det Transkribus beregner er sin egen feilrate. Det kan bare brukes som en indikasjon. Enhver transkripsjon bør gjennomgås manuelt og evalueres med hensyn til pålitelighet.

Hvis brukeren ønsker en perfekt, feilfri transkripsjon, vil ikke resultatet fra Transkibus AI-en være tilstrekkelig. Trolig vil det aldri være tilstrekkelig. Men, hvis brukeren bruker tid på å rette opp feilene, så vil AI-en ha hjulpet henne godt på vei til å få en feilfri transkripsjon.

Hvis brukeren ikke trenger en feilfri transkripsjon, men det i stedet er tilstrekkelig med en transkripsjon hun kan skumme, men ser etter detaljer, kan dette resultatet være nok. Siden Transkribus publiserer resultatene som en PDF med bildet av den originale siden etterfulgt av Transkribus sin transkripsjon, er det også mulig å bruke PDF-ordsøk for å finne spesifikke ord, f.eks. navn, steder, varer. Siden bildet av siden og transkripsjonen av siden er sidestilt i det grafiske brukergrensesnittet, er det også relativt enkelt for brukeren å se på transkripsjonenes pålitelighet og utføre korrekturlesing. Dette forutsetter likevel at brukeren kan lese håndskriften i tilstrekkelig grad.

Utfordringer

Vi har identifisert følgende utfordringer ved å bruke maskinlesing til å transkibere eldre arkivmateriale

a) Krever mye opplæringsmateriell.
   Det har krevd en veldig stor mengde manuelle transkripsjonstimer å nå denne relativt lave feilraten. Vi har brukt både det tidligere publiserte materialet og selv transkribert ytterligere (se over for detaljer). Dette arbeidet har blitt utført av folk som er dyktige til å lese håndskriftene, men også med kunnskap om arkivet og tidsperioden. Denne kompetansen er spesielt viktig når det har vært nødvendig å ta valg som påvirker transkripsjonene (f.eks. avgjørelser knyttet til myntenheter, tall, mynter og vekter/ mål)

   Vi må være ærlige og si at vi aldri ville vært i stand til å trene AI til dette nivået hvis vi ikke hadde vært i stand til å inkludere det tidligere transkriberte materialet

b) Mye arbeid går med å forberede de digitale bildene av originalen slik at linjene og bokstavene kan mates inn i Transkribus. Vi har brukt mellom 5-15 minutter på å

justere hver side. Det bør bemerkes at regnskap, slik mye av vårt materiale har vært. ofte har veldig kompliserte linjesystemer. Dette kan være mindre tidkrevende for andre typer kilder.

c) "Spesielle tegn" er et problem. De gamle dokumentene har forkortelser og spesialtegn som ikke lenger er i bruk. Disse må forstås og finnes løsninger for. F.eks. Ort-tegnet endte vi opp med å skrive som ort

d) Gitt feilprosenten, og også den manuelle korrekturlesingen som viser feilaktige transkripsjoner, er det viktig at hvis materiale som skal brukes i enhver videre analyse eller publisering av noen form må korrekturleses. Også dette tar tid.

Konklusjon

Alt i alt konkluderer vi med at:
ja, vår AI kan nå brukes til å gjøre gammelt arkivmateriale tilgjengelig for et bredere publikum. For å gjøre det krever imidlertid betydelige ressurser i form av folk med høy kompetanse i historie og med mye kunnskap om arkiver og håndskrift. Resultatet er likevel at det kreves betydelig arbeid dersom transkripsjonen skal analyseres eller publiseres på noen måte.

Men, dersom målet bare er å skumme raskt mens du søker etter detaljer, kan resultatene være tilstrekkelige. Det er fortsatt viktig å understreke at all bruk av de spesifikke detaljene som et søk finner, må bli korrekturlest og gjennomgått for å sikres at det er pålitelig.

Ytterligere tanker:

- Det ville hjelpe veldig hvis vi kunne dele våre AI -er! På den måten ville vi ha mer opplæringsmateriell. I dag går dette med programmet Transkribus, men det er litt for omstendelig.
- Å bruke arkivmaterialet vi har jobbet med krever betydelig historisk kunnskap. Skal maskinlæring være en satsning hos Arkivverket er det derfor bekymringsfullt at det i denne sektoren reduseres andelen ansatte med høy historiefaglig kompetanse.
- Vi vil i den gjenværende tiden utforske forskjellige måter å gjøre resultatene våre tilgjengelige for forskjellige målgrupper; mer spesifikt forskere, amatørhistorikere og skoler
- Utfordringene knyttet til demokratiseringen av arkivmateriell når det publiseres digitalt. Hvor feilfritt skal materialet være, og hvem er ansvarlig for å sikre kvaliteten når materialet gjøres tilgjengelig- skal det være den som publiserer eller den som bruker? Velges det første vil det bli en svært kostbar affære å gjøre arkivmateriale tilgjengelig. Velges det siste må brukerne gjøres oppmerksom på ansvaret de har knyttet til å gjøre grundig korrektur på materialet før de analyserer det.

Formidling

Det er store utfordringer knyttet til å formidle lensregnskapene til et bredt publikum. Utfordringene er først og fremst knyttet til at språket er svært annerledes vårt i dag, at de ofte brukte andre ord enn vi gjør i dag og at de også skrev tall annerledes. Dette kommer

med i maskintranskriberingen. Som nevnt over gjør dette at historikere som kjenner perioden vil få mye ut av materialet. Gjennom dem og deres analyser vil materialet bearbeides og analyseres, og konklusjonene deres kan nå ut bredt.
For å nå denne gruppen har vi henvendt oss til historikere og forskere som vi vet har interesse av materialet. Det gjelder særlig historikere og arkeologer ved universitet og høyskoler som arbeider med perioden og historikerforeningen, men også gruppen som arbeider med en ny Oslohistorie i forbindelse med byjubileet i 2024. Erfaring fra tidligere prosjekt, som Historiske toll- og skipsanløpslister tilsier at gjennom forskernes analyser vil materialet så nå ut til et bredt publikum. Man må likevel ha et visst tidsperspektiv fremover og ta i betraktning at analysene av materialet også trenger litt tid.

For andre som ikke har spesialkompetanse vil det være svært vanskelig å bruke materialet. Skal lensregnskapene brukes i undervisning vil man måtte kuratere materialet kraftig. Vi har gjort et forsøk på å gjøre dette ved å lage en undervisningsoppgave for historie VG3 Studiespesialisering der fokus er på metode og kildebruk. Oppgaven går da ut på å se på originalkilden, på den direkte transkriberingen og på den modernisert transkribering av det samme, og gjøre refleksjoner omkring hvordan og hvorfor kilden er utfordrende å arbeide med, og hva slags kunnskap man trenger for å kunne bruke den. Oppgaven gir praktisk erfaring med å arbeide med kilder, og setter fokus på de spennende utfordringene som eldre tids historikere og arkeologer møter.