AI og ML i historiefaget

Johan Heinsen, professor, ph.d.

Computational Social Science &
Humanities at Aalborg University

Historiefaget og computeren

- Digital historie er på en håndfuld år eksploderet. Er et af vores fags definerende tendenser nu.
- Afspejler sig i alt fra konferenceprogrammer til studieordninger.
- Kunstig intelligens og machine learning spiller en del af denne interesse, men har ikke drevet den.
- Det har derimod grundlæggende ændringer i forudsætningerne for at kende fortiden.

KØBENHAVNS museum
museum of COPENHAGEN

Rigsarkivet
Søg i samlingerne

Arkivalier og datasæt

Afskrifter af arkivalier

Beskrivelse af arkivmaterialer

Søg i noter, beskrivelser, arkivskaber, arkivserie

Arkivskabers navn

Søg på navn på myndighed, virksomhed, organisation

Periode

Fra år - Til år

Arkivmaterialets titel

Søg på titel på register, journal, protokol

Tilgængeligt online

Eksakt match

Nulstil

Søg

Rigsarkivet
Kalvebod Brygge 34
1560 København V

CVR-nr: 60208212
EAN-nr: 5798000794382

Hjælp til søgning
Vejledning i samlinger
Kontakt os
Log ind i Daisy

Cookiepolitik
Privatlivspolitik
Tilgængelighedserklæring
Rigsarkivet.dk

Fortidens levn arrangeret efter proveniens

- Det traditionelle arkiv havde hverken fritekstsøgning eller emneindeksering.
- I stedet handlede det om at følge papirsporets singularitet.
- Tvang os langt, langt, langt ned i tempo.
- Erkendelsen var altid stedbunden, fordi det enkelte dokument var uløseligt knyttet til en serie og en arkivskaber.

Rigsarkivet
4/240
København Stokhus
Signalementsprotokoller
1807 - 1832
1832 - 1849

Stam No.	Navne	Signalement
Ulig 837	Niels Michael Allerup	indkom: Nørrevold den 12te September 1832 fød: Solberg 23 Aar gammel, 5 f. 6½ Tommer høi af Mad, grezerkommet af Lysgivning, mørkt Haar, blaae Øine, hvor høire Ørie har faaet et tvertaaer Ar; paa Hagen et Ar paa Venstre Haand er installeret af Chirurg og et Ærr; paa venstre Arm tagne ubetydelige Lysstemmer.
Uulig 196	Johan Peder Jensen	indkom: Nørrevold den 16 September 1832 fød: Nørre Lyngby under Frederiksborg Slotsfeenberg 28 Aar gammel, 5 f. 6½ Tommer høi kommen af Mad, smalbet Legemsbygning, blaae Øine, mørkbrunt Haar. hvor høire Øie har faaet det stillet et Ærr, ved gnidning Ligner og Nummer: N. K. Jensen og J. S. Jensen. paa venstre Arm: en Brandes mand Bagstykke: J. N. J. S. A. L. S. J. M. E. J. A. M. L. S. 1826 2 Ligner, ved venstre og en gnidning samt Lagstammen: L. A. J. D. M. B. paa Bryst: Christian og Dorthe, 2 Engle, Sol, Maane og 3 Stjerner.

Stam No.	Navne	Signalement
Ulig 838	Jørgen Andersen	indkom: Nørrevold den 22 September 1832 født: Danmarke Bøgen og Vaya paa Tolland, Boende ved 35 Aar gammel, 5 f. 6½ Tommer høi. Middel Mad, grezerkommet Legemsbygning, brune Øine, brunt Haar, fort Skæg: Venstre Øie er blind, skelet.
Uulig 197	Johannes Wilhelm Johansen alias Gudum	indkom: Nørrevold den 11te October 1832 født: Gudum byen under Frederiksborg Amt: Solberg Amt. — 35 Aar gammel, 5 f. 6½ Tommer høi: Middel Mad, grezerkommet Legemsbygning, brune Øine, fort Haar, Skægvoxen og Skæg.
Ulig 839	Alexander Ludvig Meier	indkom: Nørrevold den 19 October 1832 født: Kiøbenhavn 28 Aar gammel, høi 5 f. 8½ Tommer, sterk sammenbygt Legemsbygning, hvor blaae Øine og blondt Haar: rigtig
Ulig 840	Anders Nielsen Kiøbenhavn	indkom: Nørrevold den 23 October 1832 født: Kiøbenhavn 30 Aar gammel, 5 f. 6½ Tommer høi: Middel Mad, grezerkommet Legemsbygning, blaae Øine, mørkt Haar, Lazaret Mærke. paa venstre Arm er installeret et 5 Ørte med Bestanden: N. R. D. paa venstre Haand en Krans af 1825. paa venstre Fod: A. N. S. K. D. L. D. med nogle flere ubetydelige Tegn.

Navne	Signalement
Jörgen Andersen	indkom i Slavenrit paa 22. September 1832 fört. Dannemare Byen og Sogn paa Lollands. Sönderherred. 35 Aar gammel /: 62 Tommer höi :/. Middel Maxt, grogort conerat Lu. gemerbygning, brune Øine, brunt Hæer, fort Skiæg. Tyfus begynder at blive skaldet.

Arkivets forandring

- I løbet af det sidste årti har en ny virkelighed meldt sig: arkiverne er enten født digitale (fordi vores samfund er blevet det) eller bliver retrodigitaliseret (hvis de startede som papir).

- Først gik det rigtigt langsomt – og var klart orienteret imod det som havde kommercielt interesse. Eks. google books.

- Nu går det hurtigt og er i stigende grad drevet af behov hos dem, som benytter arkivalierne.

- Med arkiver transformeret til tekst kan vi pludselig bruge computationelle tilgange til den tekst.

Hvordan bliver et arkiv retrodigitaliseret?

- Tidligere brugte man OCR (optical character recognition), der forsøgte at afkode den enkelte bogstavsform. Klare begrænsninger - næsten ubrugeligt på ældre dansk og håndskrift.

- Neurale netværk har revolutioneret tekstgenkendelse.

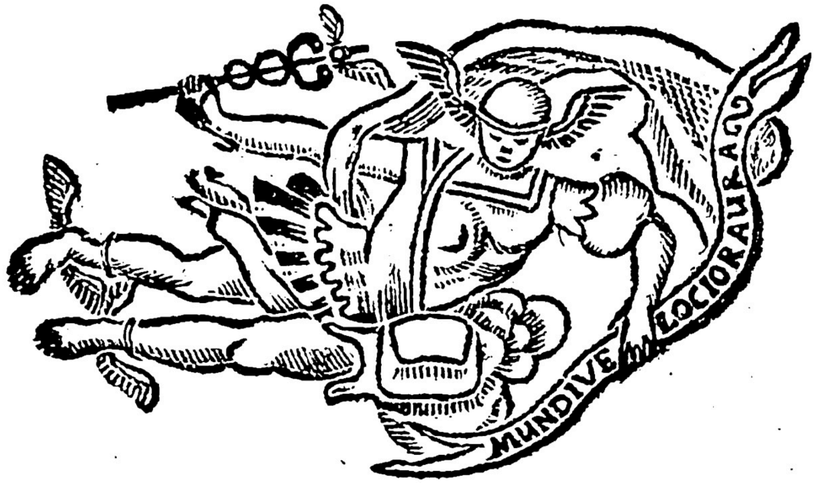
- Anthropics Claude 3.7 (anset som bedste llm til formålet):
  “ind født i Dannemare Sogn en 22° Septbr 1832 født. Dannemark Polgt og Degn søn i Hollands Jord, førend 25 Aar gml men d. 62 tommerfød. Middel Høgt, proportioneret Lemgaardbyggning, brund Øine, brunt Haar, sort Skæg. Sund bygget et blind Halvdel.” <— faktisk lovende.

- Periodespecifik neuralt netværk bygget af mig i Transkribus:
  “indkom i Slaveriet den 22e September 1832, födt i Dannemarke Bye og Sogn paa Lollands Sonde Herred, 25 Aar gammel :/: 62 Tommer höi :/: Middel Væxt, proportioneret Leigemsbygning, brune Øine, bruunt Haar, sort Skjæg. Arsen begynder at blive skaldet.” <— Afhængigt af, at nogen har bygget en model.

One-click arkivet

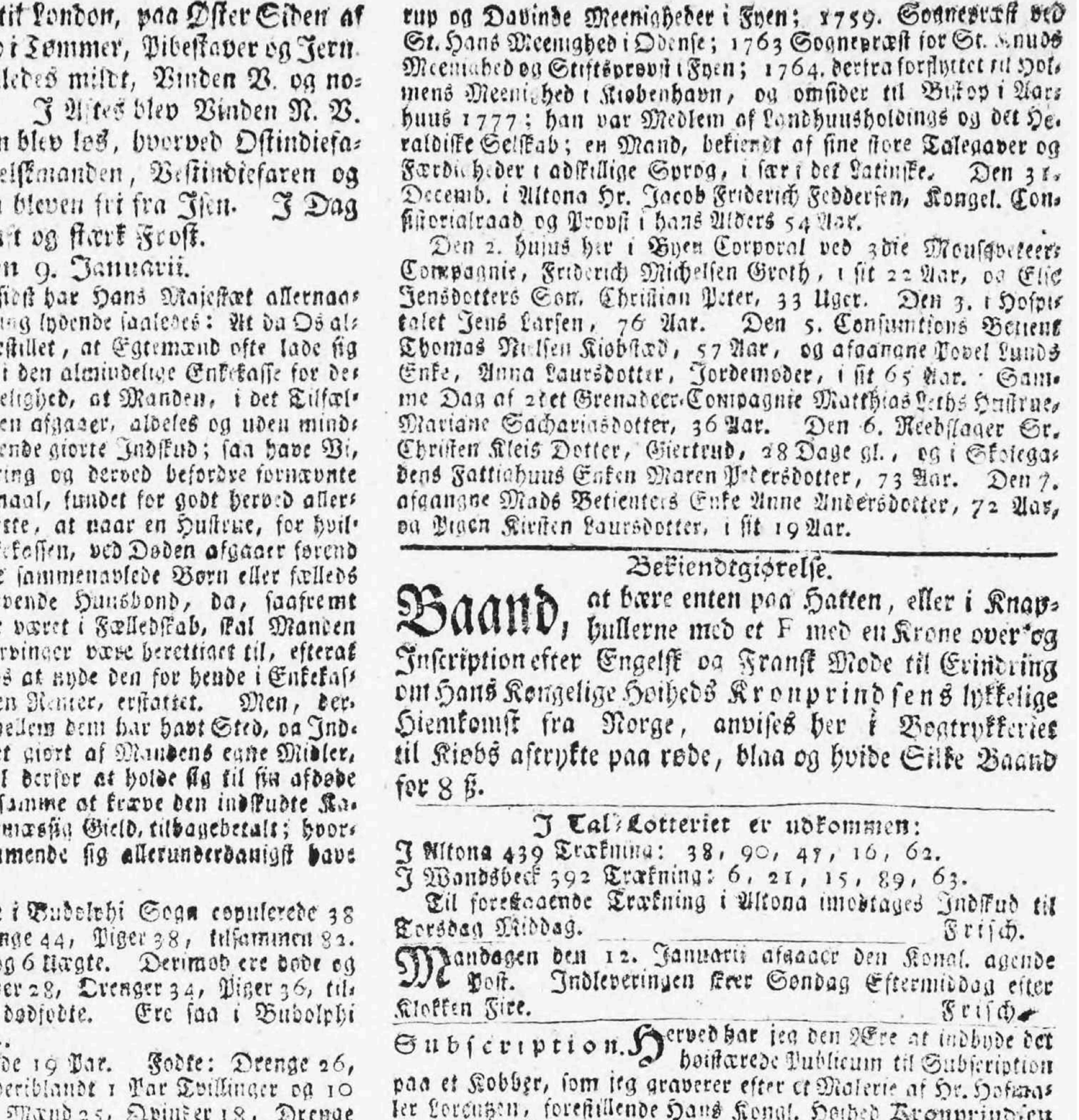
- Digitaliseringen har været et løfte om et arkiv, ubundet af læsesalens åbningstider, provenienspricip, ulæselig skrift…

- Mange forskere har advaret om tabet af kontekstfornemmelse.
- … og drømt om mulighederne for at berige det digitale arkiv med metadata, nye indekseringsmuligheder mm.
- Konturerne tegner sig af et A og B historiefag baseret på evnen til at arbejde computationelt. Hvor digitaliseringen måske ikke er en fordel for dem, som ikke bruger det refleksivt.



ENO

Enevældens Nyheder Online



Avissamlingen

Omkring 1 million sider fra 1666 til 1849.

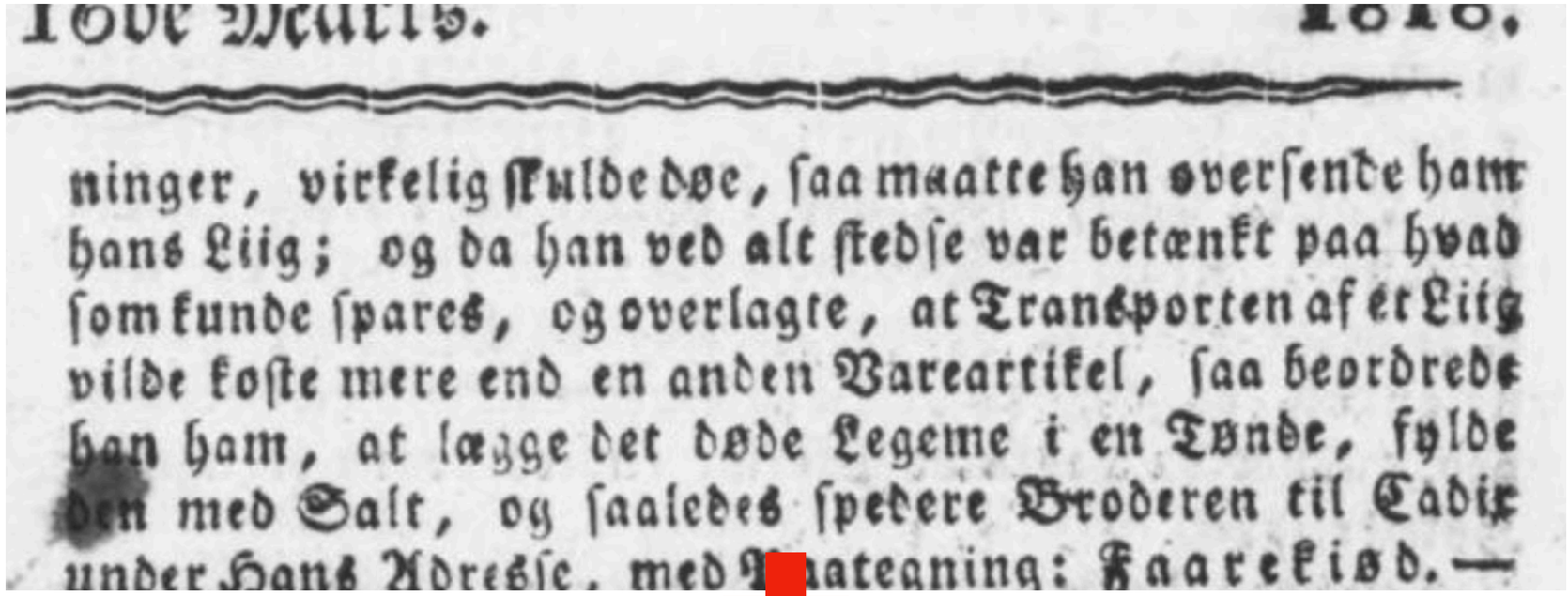
- Hverdagsmedie med en bred læserskare og kontinuerlige oplysninger om alt muligt.
- Måske mest alsidige tekstkorpus fra perioden.
- Meget materiale om tvang og bevægelse.

= uvurderligt stykke trykt kulturarv.

Tilgængeligt som affotografering af mikrofilm via Mediestream.

MEN: OCR-genkendelsen har kun en præcision på c. 50% – “OCR-bingo”

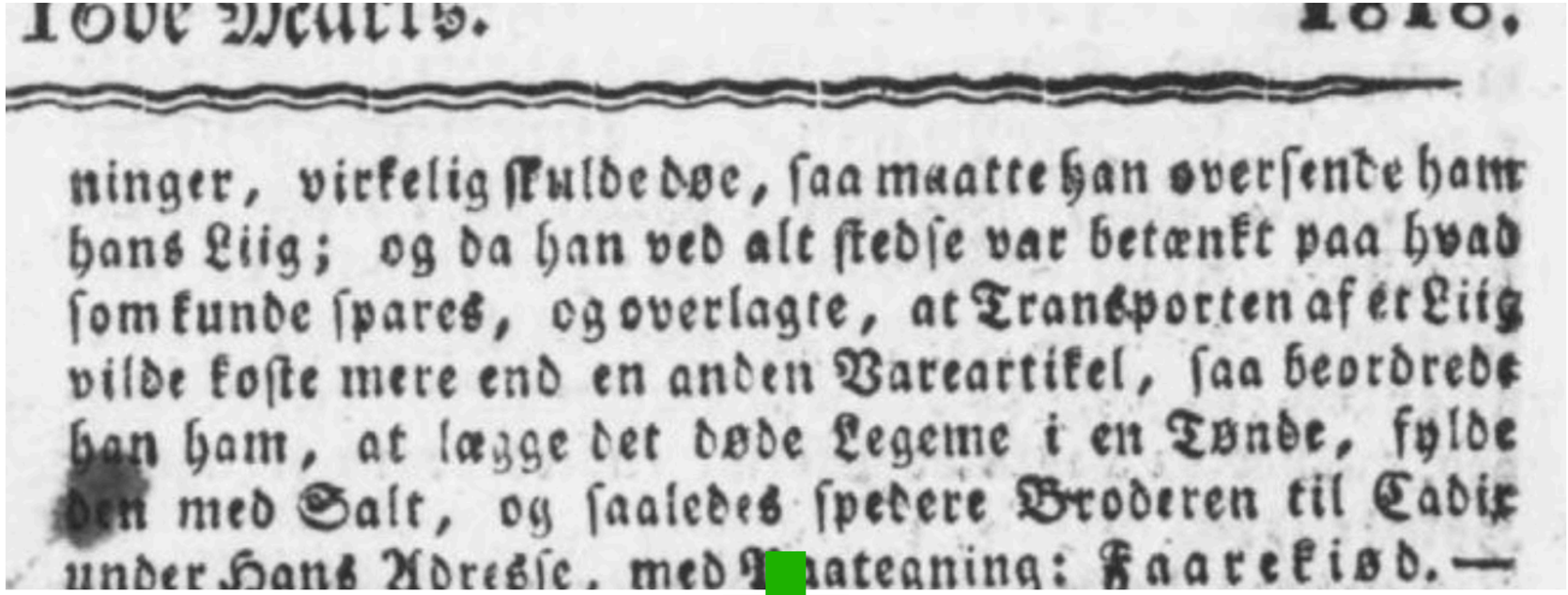
Norge har en tilsvarende samling med tilsvarende udfordringer.



ninger, virkelig skulde døe, saa maatte han oversende ham
hans Liig; og da han ved alt stedse var betænkt paa hvad
som kunde spares, og overlagte, at Transporten af et Liig
vilde koste mere end en anden Vareartikel, saa beordrede
han ham, at lægge det døde Legeme i en Tønde, fylde
den med Salt, og saaledes spedere Broderen til Cadix
under Hans Adresse, med Paateaning: Faarekiød.—

ninger, virkeligfluldeboe,saamaattehanoversendehanr
han« Liig ; og da han ved alt stedse var betankt paa hvad
somkunde spare«, ogoverlagte, atTransportenafstLiig
vilde koste mere end en anten Bareartikel, saa beordrede
Kam ham, at lagge tet dode Legeme i en Tonde, fylde
M med Talt, og saalcde« sptdere Broderen til Cadhr
underHan« Adresse. medPaategning: þÿ F a a r e k

Den klassiske OCR
- Best case scenario



ninger, virkelig skulde døe, saa maatte han oversende ham hans Liig; og da han ved alt stedse var betænkt paa hvad som kunde spares, og overlagte, at Transporten af et Liig vilde koste mere end en anden Vareartikel, saa beordrede han ham, at lægge det døde Legeme i en Tønde, fylde den med Salt, og saaledes spedere Broderen til Cadix under Hans Adresse, med Paategning: Faarekiød. —

Vores Transkribus-model

ninger, virkelig skulde døe, saa maatte han oversende ham hans Liig; og da han ved alt stedse var betænkt paa hvad som kunde spares, og overlagte, at Transporten af et Liig vilde koste meere end en anden Vareartikel, saa beordrede han ham, at lægge det døde Legeme i en Tønde, fylde ten med Salt, og saaledes spedere Broderen til Cadix under Hans Adresse, med Paategning: Faarekiød. -

Hvor vi er nået til

- Vi har (gen-)digitaliseret omkring 500 millioner ord.
- Dækker (næsten) alle de store aviser i Danmark og Norge fra midten af 1700-tallet til c. 1840. Og enkelte bevarede ældre aviser.
- Gennemsnitlig præcision er svær at udregne, men vi er et sted omkring 95% på ordniveau.
- Med hjælp af machine learning har vi segmenteret aviserne i forsøget på at rekonstituere den enkelte tekst. Vi er ikke helt i mål endnu.
- Pludselig har vi muligheder for at gøre ting, der simpelthen ikke var mulige før – f.eks. at træne sprogmodeller.
- Vores mål er at bygge gode værktøjer til forskere, undervisere og alment interesserede. Med fokus på opdagelse.

Værktøj 1: Et søgbart arkiv

- Forestil dig, at du er i gang med at undersøge et emne. Hvad hvis du kunne søge (nærmest) ubegrænset?

- Du må gerne vælge en anden avis, hvis du hellere vil et andet sted hen.

ENO Enevældens Nyheder Online Om projektet

Extraordinaire Maanedlige Relationer
Odense Adresse-Contoirs Efterretninger
Viborger Samler
Aarhuus Stifts-Tidende
Lolland Falsters Stifts Adressecontoirs Efterretninger
Den Vest-Sjællandske Avis (Slagelse)
Aalborgs Stifts Adresseavis og Avertissementstidende
Den Nord-Cimbriske Tilskuer (Thisted)
Aalborg Stiftstidende

Aarhuus Stifts Adresse-Contoirs Tidender

Link til fritekstsøgning: https://hislabaau.shinyapps.io/Aarhus/

Udgivelsessted: Aarhus

Periode dækket: 1794-1839

Navne: Aarhuus Stifts Adresse-Contoirs Tidender (1794-1811); Aarhuus Stifts Kgl. allene priv. Adresse-C
Tidender, som forsendes med Brevposten (1812-1820); Den til Forsendelse med Brevposterne Kgl. aller
Aarhuus Stifts-Tidende (1821-1839)

Digitaliseret: September 2024 - Marts 2025

Digitaliseret af: Johan Heinsen

Reuse

https://creativecommons.org/licenses/by-nc/4.0/

Værktøj 2: Et associationsleksikon over tid?

- Hvad hvis man kunne søge på et ord og se hvad, der er nært beslægtet med ordet? Over tid…

- Identificeret ved at træne 3 simple, men periodespecifikke word2vec-modeller, der fanger semantisk betydning af ord (også dem, der er genkendt forkert)

Udforsk periodespecifikke ord-embeddings.

Søg på et ord og se dets nærmeste slægtninge i periodespecikke word2vec-embeddings. OBS: Hvis et ord ikke optræder mindst 20 gange i hvert af de 3 snit produceres en fejl.

arbeide

Søg

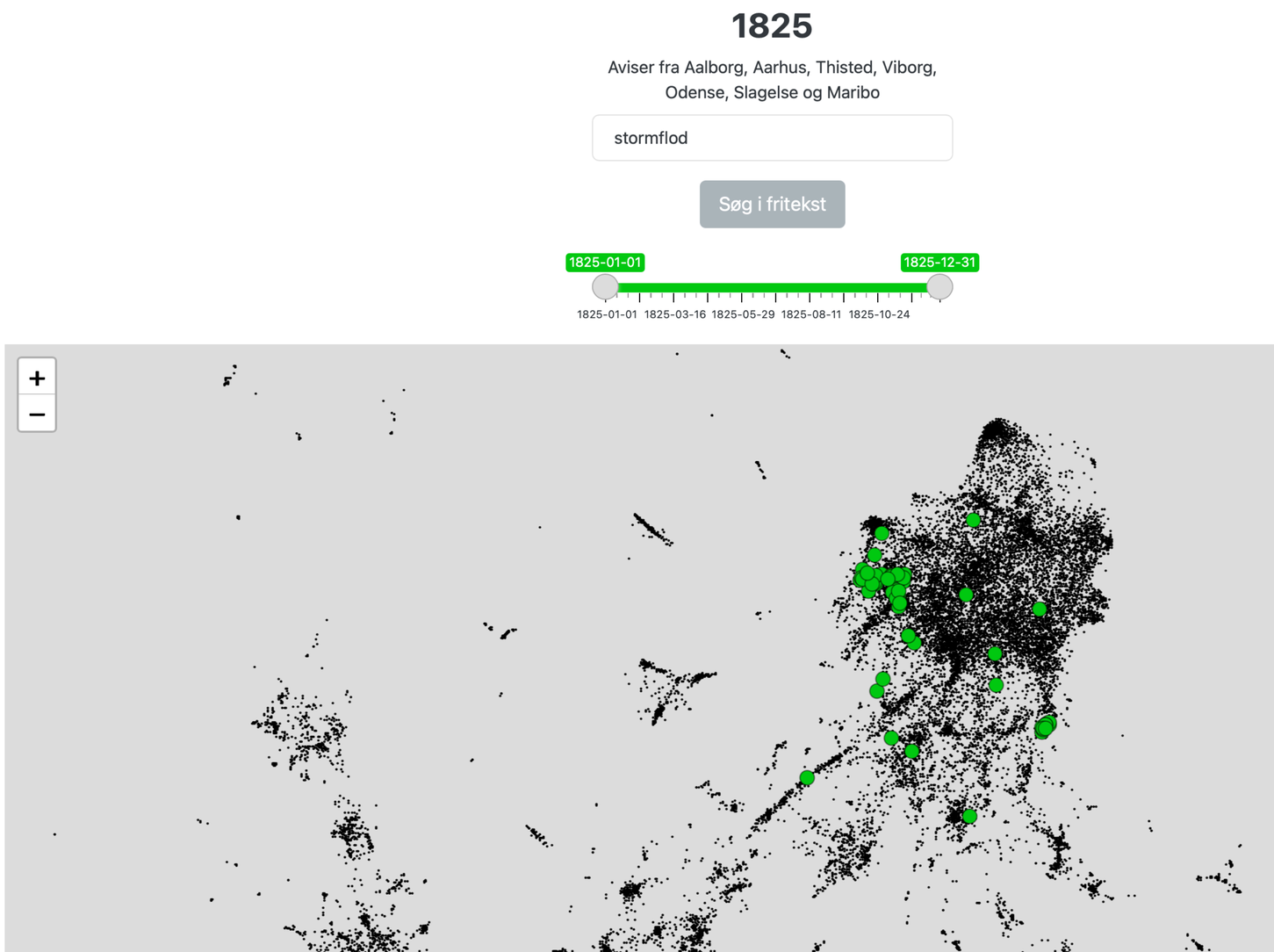
1762-1806	1807-1824	1825-1845
arbeyde	arbeid	arbeid
hensættes	arbejde	solidt
tugthuus	beide	durabelt
forbedringshuus	arbeidet	probert
hensettes	tugthuus	forsvarligt

Værktøj 3: En associationstopografi?

- Hvad hvis man kunne visualisere teksterne relativt til hinanden baseret på semantisk slægtskab? Og så udforske dem som en slags betydningskort?

- Umap-graf baseret på en simpel, men materiale-specifik sprogmodel (doc2vec).

  OBS: Hvis det crasher, er det ikke jeres skyld. Ikke klar til klasselokalet endnu



Værktøj 4: En forestillet geografi?

- Hvad hvis vi kunne læse teksterne baseret på steder, de omtaler?

  Baseret på et sample af 1000 artikler fra Aalborg 1818–1827. Koordinater leveret af Google's Maps API.

Værktøj 1: tinyurl.com/486w932x

Værktøj 2: tinyurl.com/3at8sad8

Værktøj 3: tinyurl.com/4yeh4etk

Værktøj 4: tinyurl.com/527cpd67

Overvej undervejs, hvordan du kunne bruge værktøjerne - i forberedelse og undervisningslokalet.

Og hvis man gerne vil give feedback:
forms.office.com/e/JFFp88ZrKi 🙏