KANSALLISKIRJASTO
NATIONALBIBLIOTEKET

Julkaistu kulttuuriperintö tutkimuksen tiedonlouhinnassa ja tekoälykehittämisessä

Johanna Lilja
Palvelujohtaja, Kansalliskirjaston Tutkimuskirjasto

Digital Archiving Futures
Digiarkistojen tulevaisuuskuvat 6.9.2024

Sisältö

- Käsitteitä
- Julkaistu kulttuuriperintö - mitä se on ja miten se karttuu Kansalliskirjastoon?
- Julkaistun kulttuuriperinnön asiakaskäyttö
- Julkaistun kulttuuriperinnön tiedonlouhinta tekijänoikeuslain puitteissa
- Kansalliskirjaston datapalvelut
- Tiedonlouhintaa tutkimushankkeissa
- Julkaistun kulttuuriperinnön käyttö kielimallien kehittämisessä
- Tiedonlouhinnan tulevaisuuden haasteita

Käsitteitä: tiedonlouhinta

- **Määritelmä** joukko menetelmiä, joilla pyritään oleellisen informaation löytämiseen suurista tietomassoista
- **Selite** Tiedonlouhintaa voidaan soveltaa hyvin laaja-alaisesti, sillä lähtökohdaksi tarvitaan ainoastaan raakadataa. Tyypillisesti tiedonlouhinnassa käytetty tietoaineisto on esimerkiksi mittauksia teollisuusprosessista, otteita asiakastietokannasta tai vaikkapa web-palvelimen loki-tiedostoja. Määritelmänä tiedonlouhinta ei rajaa käytettäviä menetelmiä.

Lähde: Tieteen termipankki

Käsitteitä: tekoäly

- Kielikuva, jolla tietokoneohjelman toiminta rinnastetaan inhimilliseen älykkääseen toimintaan
- Tieteenala, joka tutkii fiksusti toimiva koneita (tietojenkäsittelytiedettä, filosofiaa, kognitiotiedettä)
- Laaja ja hajanainen tietokoneohjelmien joukko
  - ohjelmat ihmisen kehittämiä
  - kullakin ohjelmalla on tietty tarkoitus

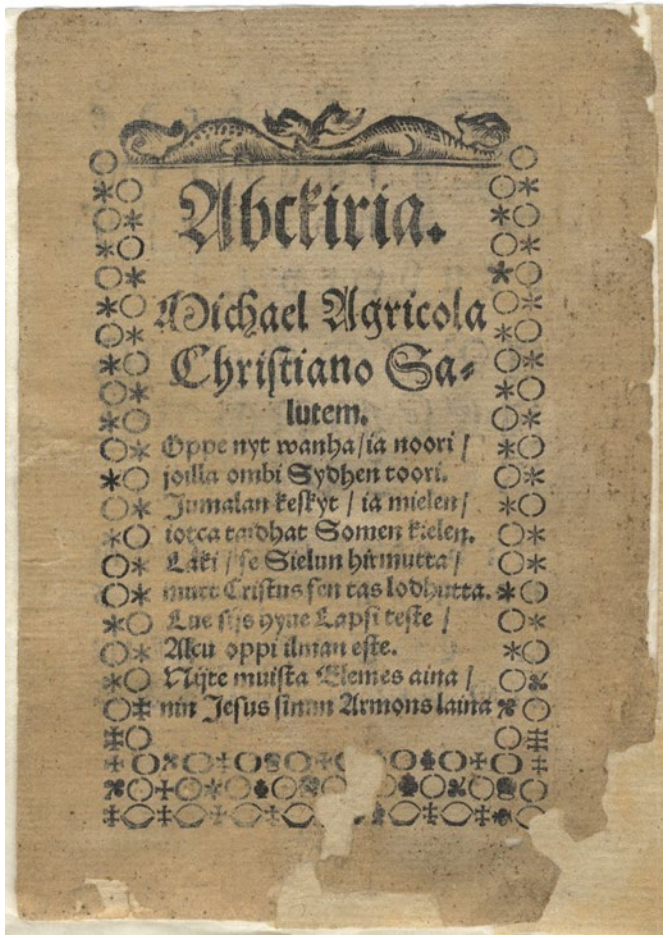
Lähde: Hannu Toivonen, Mitä Tekoäly on? 2.p. Helsinki: Teos, 2023.

Käsitteitä: Suuret kielimallit

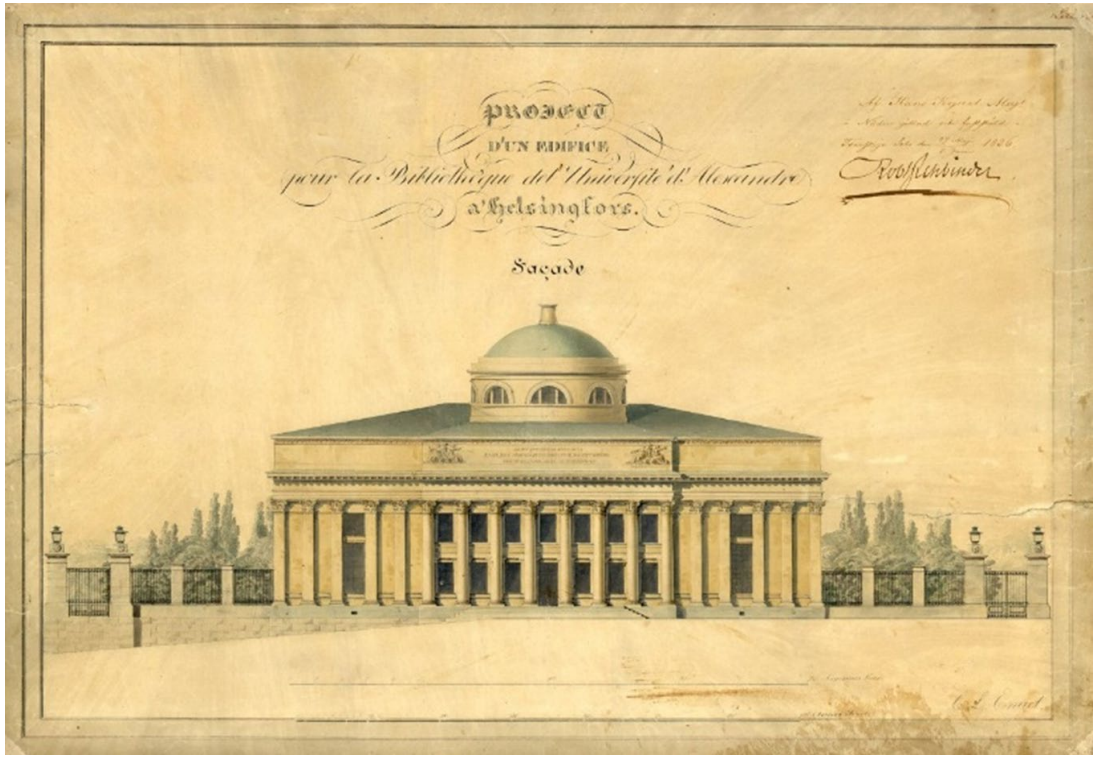
- Suuri kielimalli (LLM) on laskennallinen malli, jonka avulla voidaan ennustaa tekstin jatkuvuutta
- Esim. GPT:t (generative pre-training transformers) vastaavat kysymyksiin, tekevät käännöksiä ja tiivistelmiä tai kirjoittavat koodia)
- Kielimalli perustuu merkkijonojen todennäköisyyksiin, ei tekstin ymmärtämiseen
- Taustalla opetusaineisto (korpus)
- Kielimallin virheet voivat perustua opetusaineistoon (vinoutunut tai puutteellinen aineisto) tai mallia ohjaaviin algoritmeihin

Julkaistu kulttuuriperintö: Ruotsin aika

- Vapaakappaleina tallennetut julkaisut
  - Sensuuri- ja vapaakappalesäädökset Ruotsissa 1661
  - 1707 vapaakappaleoikeus laajennettiin Turun Akatemian kirjastoon -> Ruotsin valtakunnassa painetut julkaisut
- Suurin osa vapaakappalekokoelmasta tuhoutui Turun palossa 1827
- Meneillään digitointiprojekti, jossa Ruotsin ajan fennica-kokoelma pyritään luomaan digitaalisena hyödyntämällä lähialueiden kirjastojen kokoelmia



Julkaistu kulttuuriperintö: Venäjän aika



- 1809 jälkeen Ruotsista ei enää saatu vapaakappaleita
- 1828 määräys, jolla kaikki Venäjän valtakunnassa painetut, sensuurin läpäisseet julkaisut luovutettiin Helsingin yliopiston kirjastoon

➔Fennica-kokoelman jatkuvuus

➔Slaavilainen kokoelma

➔Vähemmistökielten kokoelmat

Julkaistu kulttuuriperintö: itsenäisyyden aika

- Painovapauslaki 1919
  - Vapaakappaleet Helsingin yliopiston kirjastoon ja neljään muuhun kirjastoon
- Vapaakappalelaki 1980 – kulttuurin tuotteiden säilymiseksi, tilastoimiseksi ja luetteloimiseksi
  - Ääni- ja kuvatallenteet lain piiriin
- Kulttuuriaineistolaki 1433/2007
  - Verkkoaineistot ja e-julkaisut tallennuksen piiriin
  - Tallentaja Kansalliskirjasto

Julkaistu kulttuuriperintöä karttuu myös

- Lahjoituksin
  - Esim. Turun palon jälkeen Fennica-kokoelmaan lahjoittivat kirjoja suomalaiset,ulkomaisia arvokokoelmia taas Venäjän tiedeakatemia ja venäläiset aateliset.
- Ostoin
  - Esim. A.E. Nordenskiöldin kokoelma, joka ostettiin kirjastolle 1900-luvun alussa, on nykyisin Unescon Memory of the World -rekisterissä

Kuva [Cosmographia. La Sphera], 01.01.1480, s. 35. https://digi.kansalliskirjasto.fi/teos/binding/2767605?page=35. Kansalliskirjaston digitaaliset aineistot (Nordenskiöldin kokoelma, Ptolemaios-atlakset)

Kulttuuriaineistojen asiakaskäyttö

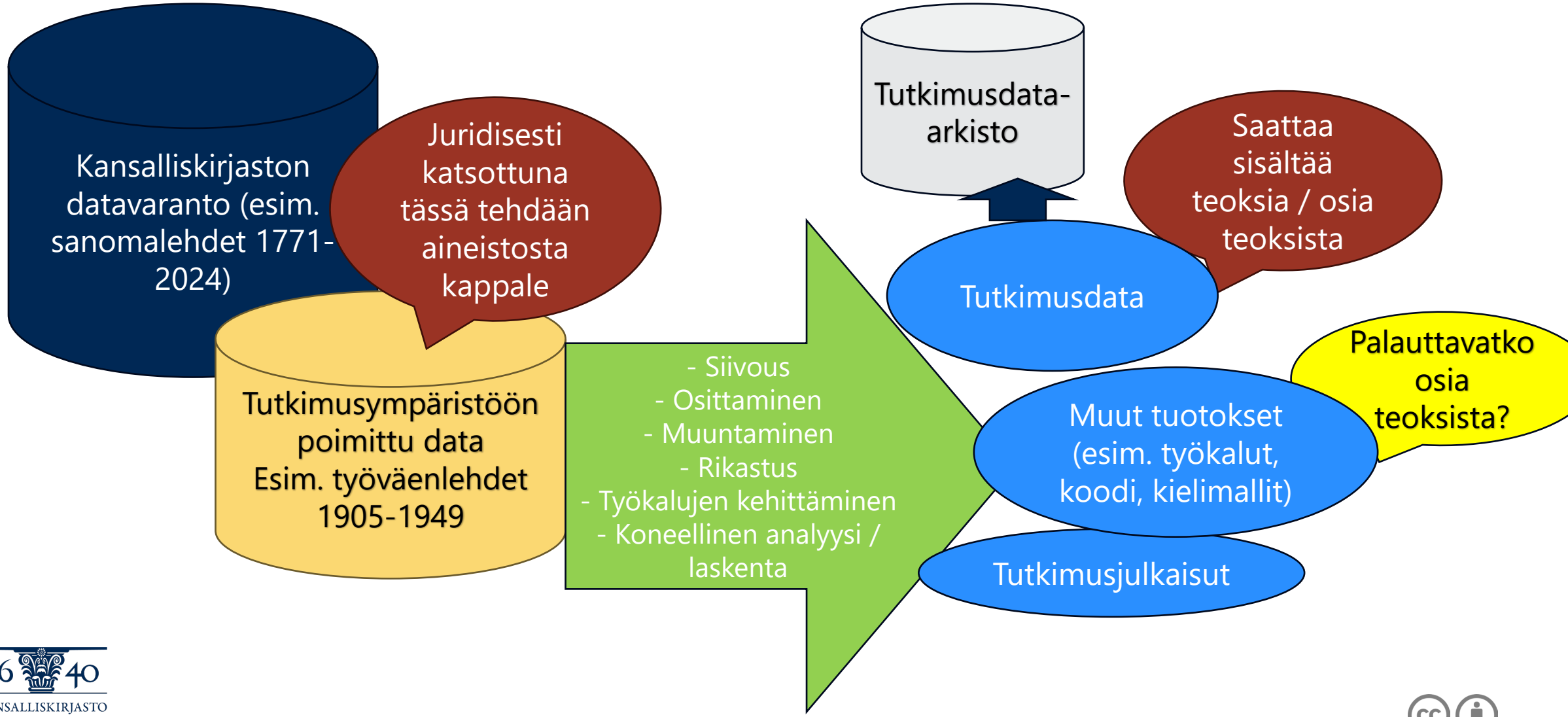
- Käyttö määritellään tekijänoikeuslaissa 404/1961
- Painettuja teoksia käytetään Kansalliskirjaston erikoislukusalissa ja vapaakappalekirjastojen lukusaleissa
- E-kirjoja, e-lehtiä ja e-musiikkia, digitoituja tekijänoikeudenalaisia aineistoja sekä verkkoarkistoa käytetään vapaakappaletyöasemilla ja äänitteitä Kansalliskirjaston kuunteluhuoneessa,
- Digitoituja tekijänoikeudesta vapaita tai lisensioituja aineistoja voi käyttää verkossa digi.kansalliskirjasto.fi-palvelussa, äänitteitä Raita-tietokannassa

Julkaistun kulttuuriperinnön tiedonlouhinta ja tekijänoikeus

- Digital Single Market -direktiivin (2019) artikla 3 edellyttää, että tutkijoilla on oikeus tehdä tiedonlouhintaa aina, kun aineistoon on laillinen pääsy
- Suomen tekijänoikeuslaki päivitetty tämän mukaiseksi 3.3.2023 13§
  - *Se, jolla on laillinen pääsy teokseen, saa valmistaa siitä kappaleita käytettäväksi tekstin- ja tiedonlouhintaa varten ja säilyttää kappaleita yksinomaan kyseistä tarkoitusta varten, jollei tekijä ole nimenomaisesti ja asianmukaisella tavalla pidättänyt tätä oikeutta.*
  - *Tutkimusorganisaatiot ja kulttuuriperintölaitokset, joilla on laillinen pääsy teokseen, saavat valmistaa siitä kappaleita tieteellisessä tutkimuksessa tapahtuvaa tekstin- ja tiedonlouhintaa varten ja säilyttää niitä tieteellistä tutkimusta varten, mukaan lukien tutkimustulosten todentamiseen myöhemmin edellyttäen, että teoksen kappaleet ovat vain siihen oikeutettujen saatavilla. Siitä, mitä edellä tässä momentissa säädetään, ei voida poiketa sopimuksella, eikä sitä voida estää teknisin keinoin.*
- Kuitenkin saman lain 16b§ kieltää kappaleen valmistuksen kulttuuriaineistolain nojalla luovutettujen aineistojen osalta. Toistaiseksi on epäselvää, miten näitä säännöksiä tulisi rinnakkain soveltaa.

-> Tällä hetkellä kulttuuriaineistoista louhintatarkoituksiin annetaan vain tekijänoikeudesta vapaita aineistoja

Datan kulku tiedonlouhintaa käyttävässä tutkimuksessa



Kansalliskirjaston datapalvelut tekijänoikeudesta vapaalle aineistolle

Lataustyökalu digi.kansalliskirjasto.fi -palvelussa

Datan lataus itsepalveluna

Digi.kansalliskirjasto.fi -palvelun lataushetken mukainen sisältö

Valmiit ladattavat datapaketit

https://data.nationallibrary.fi/

https://digi.kansalliskirjasto.fi/opendata

Staattisia (=datapaketin sisältö tuotantoajan mukainen sisältö)

FIN-CLARIAH-infrastruktuuri

CSC:n alustalle viety digitoidut julkaisut metatietoineen

Mahdollisuus päivittää aineistoa ja tallettaa versioita omaa tutkimusta varten

CSC:n ympäristö mahdollistaa suurteholaskentaa vaativat menetelmät

Kansalliskirjaston dataa myös Kielipankissa

- Kielipankkiin toimitetaan vuosittain datana uudet digitoidut sanoma- ja aikakauslehdet
- Kopiosto-Kielipankki-Kansalliskirjasto – sopimus mahdollistaa uuden (tekijänoikeudenalaisen) aineiston tutkimuskäytön
- Kielipankista aineisto voidaan ladata tutkimusympäristöön

klk-sv-1880-1948-s-vrt	Kansalliskirjaston sanoma- ja aikakauslehtikokoelman ruotsinkielinen osakorpus, 1880–1948, sekoitettu, VRT	PUB	Lataus		klk			B
KLK-sv	Kansalliskirjaston sanoma- ja aikakauslehtikokoelman ruotsinkielinen osakorpus, Kielipankki-versio	PUB	Korp		klk			A
klk-fi-v2-1874-vrt	Kansalliskirjaston sanoma- ja aikakauslehtikokoelman suomenkielinen osakorpus versio 2 (1771-1874), VRT	PUB	Lataus		klk		2024	A
klk-fi-v2-korp	Kansalliskirjaston sanoma- ja aikakauslehtikokoelman suomenkielinen osakorpus versio 2, Korp	PUB	Korp		klk		2023	A
klk-fi-v2-vrt	Kansalliskirjaston sanoma- ja aikakauslehtikokoelman suomenkielinen osakorpus versio 2, VRT	RES	Lataus		klk		2024	A
KLK-fi	Kansalliskirjaston sanoma- ja aikakauslehtikokoelman suomenkielinen osakorpus, Kielipankki-versio	PUB	Korp		klk			A

KANSALLISKIRJASTO
NATIONALBIBLIOTEKET

Tiedonlouhintaa hyödyntäneitä tutkimushankkeita Kansalliskirjastossa: digitoidut aineistot

- Digitalia-hankkeet 2015 – 19
  - Kansalliskirjaston omaa tekstintunnistuksen sekä nimien, kuvien ja artikkelien poiminnan kehitystyötä
- COMHIS-hanke 2016-2019
  - Kansalliskirjasto toimitti datapaketit ja ground truth -aineiston
  - Tutkimusmenetelmäkehitystä; Metadatan harmonisoinnin menetelmät; BLAST-työkalu tekstien kierrätyksen jäljittämiseen (uutisten siirtyminen paikasta toiseen)
- NewsEye-hanke (Horizon)
  - historiallisten sanomalehtien parempi saavutettavuus tekoälyyn perustuvien työkalujen, kuten tekstintunnistuksen, rakenneanalyysin ja monikielisen prosessoinnin (NER ym.) avulla
  - Kansalliskirjasto hyötyi erityisesti tekstintunnistuksen parannuksista
  - Europa Nostra Award 2024 tutkimushankkeiden sarjassa

Meneillään olevia hankkeita

- DHL.FI – Digitaaliset menetelmät kotimaisen kirjallisuushistorian uudistajana (Suomen Akatemia)
  - Kansallisbibliografian metadataa harmonisoimalla ja analysoimalla kartoitetaan ja tutkitaan Suomen kirjallisuushistoriaa uudesta näkökulmasta
  - Laajennetaan merkittävästi vallitsevaa käsitystä Suomen kirjallisuushistoriasta
  - Digitoidut teokset tukevat tutkimusta
- Kuvitellut kotimaat (Koneen säätiö)
  - Kansalliskirjasto tuottaa kuratoidun datapaketin ja kokoelmakuvauksen amerikansuomalaisten lehtien (1876-1923) digitoidusta kokoelmasta, jota on täydennetty Kongressin kirjaston luovuttamilla lehdillä
  - Tutkijat tunnistavat koneellisesti henkilönnimiä, verkostoja, tekstilajeja ja tekstitoisintoja sekä selvittävät, miten mielikuvat vanhasta ja uudesta kotimaasta välittyivät lukijoille sanomalehtien avulla

Julkaistu kulttuuriperintö kielimallien kehittämisen tukena

- Kielimallien kehittämiseen tarvitaan hyvin laajoja aineistoja hyvää kieltä
  - Pienet kielialueet ongelmallisia, koska kielidataa on vähän
- Digitoidussa aineistossa tekstintunnistuksen virheet voivat heikentää tuloksia

➔Syntysähköinen tekstidata parasta

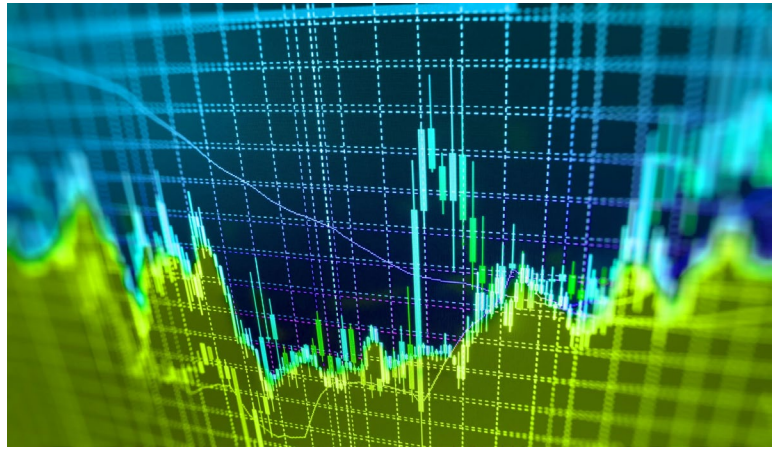
- Turku NLP –tutkimusryhmä hyödyntänyt Kansalliskirjastolta vektoreiksi kryptattuna datana luovutettuja aineistoja pilottihankkeessa
- Pilotti ei ole johtanut pysyvään datapalveluun, koska selvitettäviä kysymyksiä on vielä paljon

Tekoälykehityksen ja tiedonlouhinnan haasteita

- Tekoälykehitys ja kielimallit lisänneet kustantajien huolta datansa väärinkäytöstä https://www.nytimes.com/2023/12/27/business/media/new-york-times-open-ai-microsoft-lawsuit.html
- Vaarantuuko tutkijoiden oikeus tiedonlouhintaan?
- Datan läpinäkyvyys ongelmana globaalien jättien tuottamissa generatiivisen tekoälyn palveluissa

Tiedonlouhinnan edistäminen jatkossa Kansalliskirjastossa

- Tekijänoikeudenalaisen aineiston louhintaratkaisujen suunnittelu tekijänoikeuslain puitteissa
- Tunnistettavaa dataa: kulttuuriaineistodatan kuvailuhanke
  - Tutkija pääsee Kansalliskirjaston hakupalvelun kautta lataamaan digitoituja kokoelmia datana
  - Kokoelmakuvaukset digi.kansalliskirjasto.fi -palvelussa kontekstoivat kokoelmaa
- Versionhallinnan edelleenkehittäminen (CSC)
- Tavoitteena turvata tutkimuksen mahdollisuudet kehittää tiedonlouhintaa tekoälyn ja muiden tulosten saavuttamiseksi läpinäkyvään ja tunnistettavaan dataan perustuen ja tekijöiden oikeuksia kunnioittaen



Loppulausumat

Niin kauan kuin on olemassa ihmisiä, on olemassa sotia.

(Curre Chatin vastaus kysymykseen, mihin sanoihin päättyy Väinö Linnan Tuntematon sotilas – 22.8.2024)

Aika velikultia – nuo GPT:t!

KANSALLISKIRJASTO
NATIONALBIBLIOTEKET

www.kansalliskirjasto.fi