



Maarit Cederberg, Tuula Hurskainen, Eija Kajava,
Juha Leppämäki, Kristiina Lähdesmäki, Seija Rajaniemi

Tekoäly kirjastotyössä tänään - Mitä tekoäly mahdollistaa?

STKS:n Aineiston saatavuus -työryhmä järjesti 1.2.2024 webinaarin otsikolla Kirjasto kohtaa tekoälyn. Monipuolinen ohjelma keräsi runsaan osallistujajoukon ja herätti paljon keskustelua. Tässä artikkelissa esitellään webinaarin antia. Esityksissä tuotiin esiin muun muassa ChatGPT:n yleisiä piirteitä ja huomioitavia seikkoja. Webinaarissa kuultiin myös eräiden yliopistokirjastojen kokemuksia tekoälysovelluksista.

Tekoäly-webinaarin ensimmäinen puhuja oli tietokirjailija ja yrittäjä Antti Merilehto. Hän puhui siitä, miten tekoälytyökaluja voi käyttää apuna arjessa ja työssä.

Kielimalli ChatGPT on tällä hetkellä tunnetuin tekoälysovellus, ja suorastaan synonyymi tekoälylle. Voi-

daan sanoa, että ChatGPT toi tekoälyn kaikille.

Sanojen arvontakone

Kun pohtii ChatGPT:n luotettavuutta, voi samalla miettiä, voiko internetin sisältöön ylipäänsä luottaa. Lähteiden tarkastus on tärkeää tekoälyä käytettä-

essä. Esimerkiksi valokuvista ei pysty aina erottamaan, onko kuva aito vai tekoälyn avulla tuotettu. Tekniikka kehittyy koko ajan, joten on entistä hankalampaa erottaa aito epäaidosta.

Hyvä neuvo on, että julkisiin tekoälytyökaluihin kannattaa syöttää vain sellaista tietoa, jonka uskaltaa jättää esille printtinä työpaikan lounaspöytään. Toisaalta tekoäly ei kuitenkaan ole vain vaaroja ja ansoja. Tekoäly tehostaa työtä, tästä esimerkkinä Merilehto kertoi tekoälyn näppäryydestä työhakemuksen teossa.

ChatGPT ei ole hakukone eikä se toimi samalla tavalla kuin Google, vaan kyseessä on Merilehdon mukaan ”sanojen arvontakone”. ChatGPT osaa tuottaa tekstiä arvaamalla aina seuraavan sanan sille syötetyn datan perusteella. Joka kerta kun kielimalli tuottaa vähänkin pidemmän tekstin, on lopputulos erilainen. Sitä mukaa kun datan määrä lisääntyy, tulokset parantuvat.

Aiemmin ajateltiin, että tekoäly ja robotit vaikuttavat työelämässä lähinnä mekaaniseen, suorittavaan työhön. Generatiivisen tekoälyn nousun myötä on kuitenkin käynyt selväksi, että tekoäly vaikuttaa vahvasti myös luovaan työhön.

Promptaaminen eli kehotteiden (eng. prompt) antaminen tarkoittaa tekoälyn ohjaamista tai aktivointia kohti haluttua lopputulosta antamalla sille käsky tai ohje. Kun antaa tekoälylle käskyn, täytyy antaa myös tavoite ja tehtävä. Samalla voi tarkentaa myös promptin sävyä.

Merilehto lopetti esityksensä Polanyiin paradoksiin: ”Me tiedämme enemmän kuin osaamme kertoa”, ja kehotukseen ottaa tekoäly käyttöön kokeilemalla.

Kuinka ohjeistaa tekoälysovellusten käytössä?

Tampereen yliopiston kirjaston tietoasiantuntija Taina Peltonen kertoi, kuinka kirjasto ohjeistaa asiakkaitaan ilmaisten tekoälysovellusten käytössä.

Seuraaviin kysymyksiin tarjotaan vastauksia kirjaston koulutuksissa ja koulutusmateriaaleissa: Pitääkö tekoälysovellusten käytöstä työssään mainita? Mitkä ovat tekoälyä käyttävät artikkelihaun työkalut? Voiko ChatGPT:tä hyödyntää opintoihin ja tutkimukseen liittyvässä tiedonhaussa? Millainen tietoturva on tekoälyohjelmistojen käytössä? Entä tekoälysovellusten käytön eettisyys? Eettisinä kysymyksinä tekoälysovellusten käytössä on pohdittu seuraavia teemoja: integriteetti, saavutettavuus, ajankohtauaisuus, mis- ja disinformaatio, tekijänoikeudet sekä yksityisyyden suoja.

Tiedonhakija on vastuussa käyttämiensä lähteiden luotettavuudesta. ChatGPT:n käyttäjän täytyy pohtia seuraavia kysymyksiä: Jos kysymykseen vastataan tuottamalla tekstiä, minkälaisen dataan vastaus perustuu ja ovatko viitatus viitteet oikeita vai hallusinoituja? Onko palvelun tarjoaman vastauksen oikeellisuudella merkitystä?

Kokemuksia ChatGPT:stä ja Keeniouksesta

Tampereen yliopiston kirjastossa on testattu ilmaista ChatGPT 3.5. versiota ja todettu, että se ei sovellu varsinaiseen tiedonhakuun. Vastaus vaihtelee eri kerroilla vaikka kehote olisi sama.

Kielimalleihin syötetyt tiedot opettavat tekoälyä ja tiedot jäävät sovelukseen. Käyttäjän tulisi tarkistaa kaikkien sovellusten käyttöehdot. Henkilökohtaisia ja luottamuksellisia tietoja ei pidä syöttää eikä keskusteluhistoriaa kannata tallentaa.

Tekoälysovellusten käytöstä opinnoissa tulee mainita. Tutkimusjulkaisuissa tekoäly ei voi olla kirjoittaja tai kanssakirjoittaja.

Akateemisia artikkelihakutyökaluja ovat muun muassa Elicit, Semantic Scholar ja Keenious. Elicit hakee noin seitsemän relevantteinta artikkelia sille annettuun tutkimuskysymykseen tai artikkeliin perustuen, mutta ei kerro hakutulosten kokonaismäärää. Semantic Scholarin haku on samantyyppinen sanahaku kuin Google Scholarissa. Hakutulos näyttää artikkelien viittaukset ja pyrkii löytämään viitatuimpien artikkelien lisäksi myös kaikkein merkityksellisimmät artikkelit.

Tietoasiantuntija Leena Huovinen Helsingin yliopiston kirjastosta esitteli tekoälypohjaista Keenious-artikkelihakutyökalua, jonka maksullinen versio on hankittu Helsingin yliopiston käyttöön vuosi sitten. Työkalun data tulee OpenAlexista, joka sisältää valtavan

Henkilökohtaisia ja luottamuksellisia tietoja ei pidä syöttää kielimalleille eikä keskusteluhistoriaa kannata tallentaa.

määrän sekä avoimien että maksullisten artikkeleiden metadatala. Palvelu toimii myös ilman kirjautumista, mutta oman tilin tekemällä saa enemmän hakutuloksia näkyviin ja niitä voi tallentaa.

Keeniousta käyttäessä ei tarvitse miettiä sopivia hakusanoja ja rakentaa hakulausekkeita, vaan tiedonhakija voi aloittaa valitsemallaan tekstillä. Tekoäly analysoi tekstin ja suosittelee saman aiheisia artikkeleita lisää. Tutkijoita on ilahduttanut tiedonhaun tulosten poikkitieteellisyys.

Tiedonhaku murroksessa

Tilastokeskuksen kehityspäälliköiden Riitta Pielan ja Markku Huttusen yhteisestä esityksestä kävi ilmi, että Tilastokeskus avasi ensimmäisenä tilastovirastona Euroopassa internet-palvelunsa vuonna 1995. Markku Huttusen osuus valotti digitaalisten tiedonhakupalvelujen murrosta 1980-luvulta internet-ajan alkuun. Historiaa on luettavissa Huttusen artikkelista *Digitaalisten tietopalveluiden 30-vuotinen taistelu*, joka liittyy Suomen digitalisoinnin historia / Suomen 100 -hankkeeseen. Tiedonhaun kehitys CD-ROMeista internetin hakukoneiden kautta Googleen on tuonut

vähitellen tiedon kaikkien ulottuville. Sosiaalinen media puolestaan on mahdollistanut yksilöllisen tiedonhaun ja samalla polarisoinut viestintäympäristöä. Huttunen korostikin informaatiolukutaidon tärkeyttä mullistuneessa mediaympäristössä.

Tekoäly käsittelee luonnollista kieltä, ja se yksinkertaistaa loppukäyttäjien hakuprosessia. Tekoälylle on mahdollista esittää monimutkaisia kyselyitä, jotka tekoäly ymmärtää ja tulkitsee ja ajan myötä myös pystyy räätälöimään tuloksia yksittäisten käyttäjien ja hakuhistorian mukaan. Personoidut suositukset voivat rikastuttaa hakutulosta ja edistää tiedon levittämistä, mistä on joko hyötyä tai haittaa.

Tilastokeskuksessa on tehty kokeiluja OpenAI:n kehittämällä kielimalleilla. Tekoälyä haluttaisiin myös käyttää datakatalogiin, joka lukisi suoraan aineistoihin liittyviä metatietoja aineistojen tallennuspaikasta ja tarjoaisi API-rajapinnan, luokittelisi aineistoja automaattisesti sekä parantaisi metatietojen laatua ja tiivistäisi aineistot lyhyiksi ja ymmärrettäviksi kuvauksiksi. Tilastokeskuksen haluttaisiin oma chatbot isojen chatbottien tuottaman tiedon rinnalle. Isot AI Chatbotit eivät useinkaan perustu avoimiin vaan suljettuihin malleihin, joiden taustasta ei ole tietoa. Kannattaa esimerkiksi monitoroida, miten tunnetuimmat chatbotit tulkitsevat ja esittävät oman organisaation tietoja.

Tekoäly tulee mullistamaan tutkimustiedon jakeluun liittyvät palvelut.

Suositusjärjestelmät, aineistokokoelmien hallinta ja luokittelun ja analysoinnin automatisointi sekä tallennus- ja hakuprosessien optimointi tulee muuttamaan työtä tehokkaammaksi. Älykkäät asiakaspalvelubotit pystyvät tarjoamaan sekä tutkimusapua että personoitua käyttäjätukea olennaisena osana tutkimusprosessia.

Nimientiteettinen tunnistus arkistoaineistosta

Kansallisarkiston koneoppimisen pääsuunnittelija Mikko Lipsanen esitteli nimientiteettien tunnistusta. Nimientiteettien tunnistus eli NER (named-entity recognition) tarkoittaa tekstidatasta ennalta määriteltyihin kategorioihin kuuluvien elementtien tunnistamista. Yleisiä entiteettejä ovat esimerkiksi paikannimet, organisaatiot ja henkilöiden nimet. NER-tunnistus helpottaa digitoidun aineiston automaattista metatiedottamista ja dokumenttien ja aineistojen linkittämistä. Se myös nopeuttaa tiedonhakua.

Nimientiteettien tunnistuksessa on edetty sääntöpohjaisesta tunnistamisesta koneoppimista ja kielimalleja hyödyntävään tunnistukseen. Suomenkielisistä NER-malleista FINER on sääntöpohjainen malli, joka on kehitetty Helsingin yliopistossa FIN-CLARIN-konsortiolle. TurkuNLP-tutkimusryhmän kehittämä NER-tunnistin puolestaan hyödyntää BERT-kielimallia.

Suomenkieliset NER-aineistoja

ovat muiden muassa FINER-korpus, joka sisältää teknologia-alan artikkeleita Digitoday-lehdestä sekä Turku NER korpus, joka sisältää mm. lakitekstejä ja uutisartikkeleita. Aineistot ovat pitkälti digisyntyisiä, eivätkä sisällä arkistoaineistoa, joten arkistoaineiston NER-tunnistuksen kannalta korpuksot ovat rajoittuneita.


Arkistoaineistot, jotka ovat pääosin digitoituja asiakirjoja, tuottavat omat haasteensa. Tarvitaan NER-malli erityisesti digitoitua asiakirja-aineistoa varten. Kehitysprojektit DALAI sekä FIN-CLARIAH olivat mukana kehittämässä erityistä NER-mallia arkistoa-aineistoille. Tavoitteena on malli, joka toimisi paitsi digitoitulla arkistoa-aineistoilla, mutta myös erityyppisillä teksteillä. Arkistoa-aineistolle räätälöidyn NER-mallin kehitystyön tulos on saatavilla ja testattavissa [Kansallisarkiston Hugging Face](#) -sivulla.

Kansalaisten ja organisaatioiden datalukutaito

Webinaarin viimeisenä puhujana Helsingin yliopiston dosentti ja ohjelmajohtaja Karoliina Snell herätteli tietoisuuttamme datalukutaidon tarpeellisuudesta otsikolla *Datalukutaito vastuullisen päätöksenteon edellytyksenä*. Hän on mukana monitieteisessä DataLit-hankkeessa, jossa kehitetään yhdessä sidosryhmien kanssa ratkaisuja ja työkaluja tietoon perustuvan datalukutaidon tueksi. Tutkimusryhmän keskeisiä tavoitteita ovat ymmärrettävien ja luotettavien käytäntöjen luominen sekä datalukutaidon edistäminen yhteiskunnassa.

Kiteytettynä datalukutaito on ymmärrystä siitä, miten dataa tuotetaan, käsitellään, analysoidaan ja esitetään. Toisin sanoen se on dataan liittyvien keskeisten teknisten, oikeudellisten ja sosiaalisten kysymysten tiedostamista ja ymmärtämistä. Datan ulottuvuuksien käsittäminen rajoituksineen ja edellytyksineen vähentää epärealistisia odotuksia. On selvää, ettei tekoäly pysty vastaamaan kysymyksiin silloin, kun sen pohjalla on epätäydellistä ja jalostamatonta raakadataa. Snell viittaa aiheesta David J. Handin teokseen *Dark Data: Why what you don't know matters*, joka käsittelee ”pimeän datan” olemusta.

Organisaatiot tarvitsevat vastuulliseen päätöksentekoon syvempää datalukutaitoa kuin yksittäiset kansalaiset. Kollektiivinen datalukutaito monimutkaisine järjestelmineen vaatii monialaista ja organisaatioiden rajat ylittävää vastuunjakoa. Tietosuoja ja datan käytön etiikka liittyvät oleellisesti vastuulliseen toimintaan. Paras vallitseva tila on sellainen, jossa algoritmit kohtelevat kansalaisia yhdenvertaisesti.

Tekoäly käsittelee tietoa niin kuin sille opetetaan, mutta muokkaa samalla käyttäjiensä ajattelua. Tilanne elää koko ajan. Paljon tietoa ja erilaisia ajattelumalleja jää koneoppimisen ulkopuolelle, mutta emme voi tietää mitä. Avointa ja kriittistä yhteiskunnallista keskustelua on syytä seurata ja ylläpitää. Paljon tietoa ja ajatuksia herättävän esityksen jälkeen huomio kohdistuu kirjastoihin. Onko datalukutaidosta tulossa uusi kansalaistaito muiden monilukutaitojen joukkoon? 

Lähteet ja lisätietoa

<https://docplayer.fi/47316000-Digitaalisten-tieto-palveluiden-30-vuotinen-taistelu.html>

<https://keenious.com>

<https://elicit.com/>

<https://www.semanticscholar.org/>

[Tiedonhaku ja tekoäly - Tiedonhaun opas - Oppaat | Guides at Tampere University Library \(tuni.fi\)](#)

<https://openalex.org/>

<https://libraryguides.helsinki.fi/keenious/fi>

[FINER](#)

[FINER-korpus](#)

[TurkuNLP](#)

[Turku NER-korpus](#)

[DALAI](#)

[FIN-CLARIAH](#)

[Kansallisarkiston NER-malli](#)

Kirjoittajat

MAARIT CEDERBERG

Tampereen yliopiston kirjasto
maarit.cederberg@tuni.fi

TUULA HURSKAINEN

Suomen ympäristökeskus,
Avoin tiede
tuula.hurskainen@syke.fi

EIJA KAJAVA

Eduskunnan kirjasto
eija.kajava@eduskunta.fi

JUHA LEPPÄMÄKI

Helsingin yliopiston kirjasto
juha.leppamaki@helsinki.fi

KRISTIINA LÄHDESMÄKI

Helsingin yliopiston kirjasto
kristiina.lahdesmaki@helsinki.fi

SEIJA RAJANIEMI

VTT, Oulu
seija.rajaniemi@vtt.fi