kreodi

Annif asiasanoittajana – kokemuksia Theseuksesta

06.06.2021

TUTKIJAN TYÖPÖYDÄLTÄ. KUVA TIINA TOLONEN.

Usein opinnäytetyön tallentamisen pullonkaulaksi muodostunut asiasanojen lisääminen tallennuslomakkeelle muuttui helpommaksi, kun Theseus-julkaisuarkiston Annif-integraatio otettiin käyttöön syyskuussa 2020. Nyt kun lähes lukuvuoden mittainen aika Annifin kanssa on takana, on hyvä tutkia kuinka se on onnistunut tehtävässään, eli automaattisessa asiasanojen tarjoamisessa opiskelijalle. Tekemässäni tutkimuksessa kävin läpi 200 opinnäytetyötä, jotka oli ensimmäisinä tallennettu Annifin käyttöönoton jälkeen eräässä ammattikorkeakoulussa.

Annif on tekstiaineistojen automaattiseen sisällönkuvailuun tarkoitettu työkalu, jonka toiminta perustuu koneoppimista ja kieliteknologiaa hyödyntäviin algoritmeihin sekä ohjelmistokirjastoihin. Se ehdottaa asiasanoja tai luokkia kontrolloidusta sanastosta, kuten YSO tai YKL. Annif toimii yleisesti ottaen parhaiten asiatekstien kuvailuun, mutta sitä on mahdollista käyttää myös kaunokirjallisuuden asiasanoitukseen. Annif on käytettävissä myös itsenäisenä työkaluna, palvelu on nimeltään FintoAI.

Annif-työkalu on avointa lähdekoodia ja sitä voi kuka tahansa muokata sekä hyödyntää. Algoritmien toiminta perustuu koneoppimiseen ja työkalua voidaan kouluttaa valmiiksi asiasanoitetuilla tai luokitelluilla opetusdatoilla. Se sisältää myös toiminnon tuotettujen asiasanojen laadun arviointia varten, tämä tapahtuu vertaamalla näitä tuotettuja asiasanoja ihmisten antamiin asiasanoihin.

Tähän mennessä Annifia on käytetty ainakin suomen-, ruotsin-, englannin- sekä hollanninkielisille teksteille. Kansalliskirjaston ylläpitämistä julkaisuarkistoista Annif on ollut kytkettynä Vaasan yliopiston Osuva-julkaisuarkistoon maaliskuusta 2020 alkaen sekä Tampereen yliopiston Trepo-julkaisuarkiston ja ammattikorkeakoulujen Theseus-julkaisuarkiston syöttölomakkeisiin syyskuusta 2020 alkaen.

Pidempi kokemus Annifin käytöstä on Jyväskylän yliopiston JYX-julkaisuarkistossa, jossa integraatio tehtiin jo vuonna 2018. Jyväskylän yliopistossa on myös tutkittu Annifin onnistumista asiasanoittajana, ja onnistumisprosentti hieman alle 400 pro gradun otoksessa oli seuraavanlainen:

- noin 85 % opiskelijoista kelpuutti ainakin yhden Annifin ehdotuksen
- yli kolmasosa opiskelijoista valitsi yli puolet Annifin tarjoamista termeistä
- kaikki Annifin tarjoamat termit kelpasivat 14 opiskelijalle.

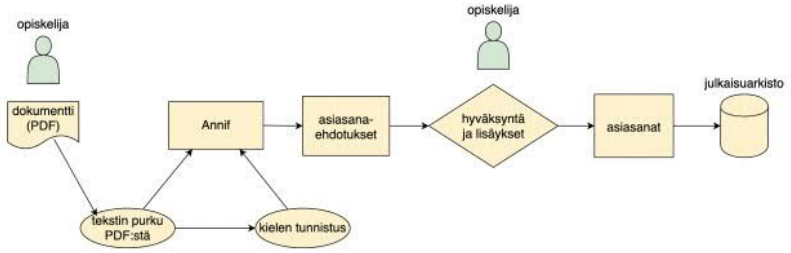
Kirjastoverkkopäivillä syksyllä 2019 järjestettiin työpaja, jossa tarkasteltiin sisällönkuvailun laatua eri näkökulmista ja vertailtiin automaattisesti, puoliautomaattisesti ja manuaalisesti tuotettuja erilaisten materiaalien sisällönkuvailuja. Työpajassa esiteltiin myös muutamien esimerkkien avulla dokumenttien kuvailua ja niiden saamia keskimääräisiä laatuarvosanoja.

JYX-julkaisuarkiston graduaineistoa tarkastellessa Annifin tuottamista kuvailuista todettiin, että ne vaikuttavat olleen suppeampia kuin ihmisten itsenäisesti tai Annif-avusteisesti tekemät ja että niissä näkyy myös enemmän toisteisuutta.

Annifin todettiin kuitenkin suoriutuneen graduaineiston kuvailussa keskimäärin hyvin.

Miten Annif toimii?

Annif-prosessin kulku esitetään alla olevassa kaaviossa. Opinnäytetyön tallennusprosessi alkaa tallentamalla ensimmäiseksi opinnäytetyön tiedosto PDF-muodossa. Tästä puretaan raakateksti ja tunnistetaan dokumentin pääasiallinen kieli. Teksti lähetetään Finto AI:n rajapinnan kautta Annifille, joka ehdottaa tekstille asiasanoja. Ehdotetut sanat näytetään tallennuslomakkeella, josta opiskelija voi joko yksitellen hyväksyä tai hylätä ne sekä myös lisätä omia asiasanojaan. Annif tuo opiskelijalle syöttölomakkeella valittavaksi kymmenen asiasanaehdotusta.



Kuva 2. Annifin prosessikaavio.

Toimiiko Annif?

Ennen Annifin käyttöönottoa Theseus-asiantuntijoiden sähköpostiin tuli runsaasti viestejä asiasanoitusta koskien. Erityisen hämmentäväksi tilanteen opiskelijoiden näkökulmasta teki todennäköisesti se, että aikaisemmin Theseuksessa oli opiskelijoilla ollut mahdollisuus lisätä omia termejään eli ns. avainsanoja. Nämä sanat olivat luultavimmin suureksi osaksi asiasanastojen ulkopuolisia termejä, mutta ovat näkyvissä Theseuksen asiasanalistauksessa kirjastohenkilökunnan lisäämien virallisten asiasanojen joukossa.

Kun avainsanojen lisäämisen mahdollisuus poistui ja valittavana olivat ainoastaan YSOn mukaiset termit, syntyi ihmetystä. Erityistä kummastusta herätti se, että Theseuksen asiasanalistasta löytyi sellaisia sanoja, joita opiskelija olisi halunnut käyttää oman työnsä kuvailemiseen mutta ei pystynyt kuitenkaan valitsemaan. Tämän selittäminen opiskelijalle sähköpostin välityksellä on ollut paikoitellen

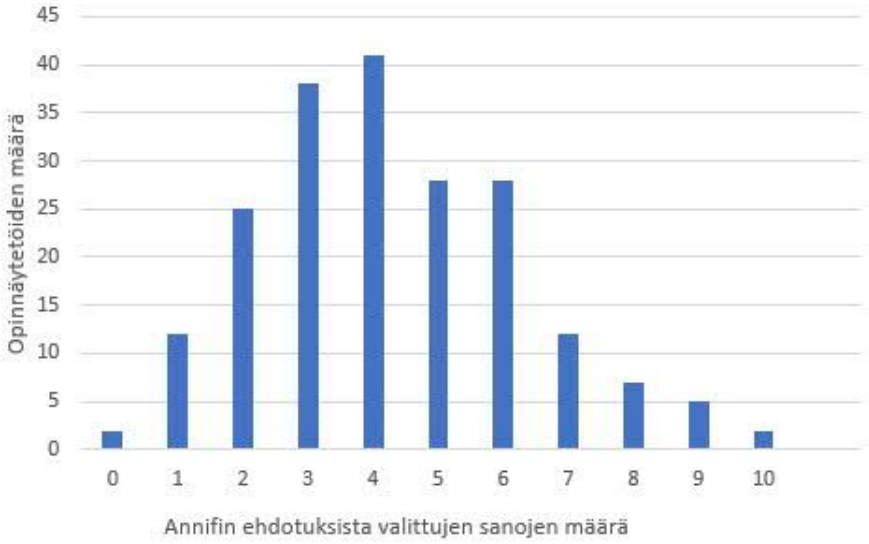
hieman haastavaa, kun täytyy yrittää selittää asia ilman että käyttäisi valtavaa määrää kirjastojargonia eikä myöskään ryhtyisi syvällisesti avaamaan Theseuksen kehityshistoriaa. Opiskelijan käsitys yleensä on ollut se, että kun sanaa oli Theseuksessa kuvailuun jo käytetty, niin hänenkin täytyy sitä pystyä sitä omalle työlleen käyttämään.

Annif-integraation kytkemisen jälkeenasiasanoihin liittyneet yhteydenotot ovat vähentyneet huomattavasti. Tästä on voitu päätellä, että automaattinen asiasanojen ehdottaminen on ollut onnistunut ratkaisu ja kenties helpottanut ja nopeuttanut opinnäytetyön tallentamisen todennäköisesti haastavinta vaihetta. Nythän parhaassa tapauksessa on pystynyt suoraan ruksimaan syöttölomakkeelta juuri ne termit, joilla työtään on halunnut kuvailla.

Miten Annif onnistui Theseus-otoksessa?

Tein pienen tutkimuksen koskien sitä, kuinka paljon opiskelijat kelpuuttavat Annifin tarjoamia termejä. Keräsin myös ylös termejä, joissa Annif vaikutti olevan aiheesta täysin pihalla. Otokseeni valikoitui 200 opinnäytetyötä yhdestä ammattikorkeakoulusta. Selvitin myös, onko eri koulutusalojen välillä määrällisiä eroja siinä, kuinka opiskelijat kelpuuttavat Annifin tarjoamia termejä. Toisin sanoen halusin selvittää, missä kohtaa Annifilla saattaisi olla opettamisen tarvetta.

200 opinnäytetyön joukossa oli vain kaksi opinnäytetyötä, joissa opiskelija ei ollut hyväksynyt yhtään Annifin tarjoamaa termiä. Molemmat näistä olivat tekniikan alan opinnäytetöitä. Näin ollen ainakin yhden Annifin ehdotuksista hyväksyi 99 % opiskelijoista. Toisessa ääripäässä tilanne oli täsmälleen sama, eli kaksi opiskelijaa hyväksyi kaikki kymmenen Annifin tarjoamaa termiä, toinen näistä kulttuurialan ja toinen liiketalouden opinnäytetyö. Toiseen näistä opiskelija oli lisännyt vielä itse valitsemiaan termejä peräti seitsemän, eli tässä opinnäytetyössä asiasanoja oli kaikkiaan 17.



Kuva 3. Annifin ehdottamista termeistä opiskelijan kelpuuttamien sanojen määrä per opinnäytetyö.

41 % eli hieman yli viidesosa opiskelijoista valitsi vähintään viisi Annifin tarjoamaa termiä. Pelkästään Annifin tarjoamia termejä oli kaikkiaan 114 opinnäytetyössä eli noin 57 % tutkituista opinnäytetöistä. Loppuosaan otoksen opinnäytetöistä oli lisätty muita YSOn termejä 1–9 kpl.

Alojen välisestä jakaumasta

Tutkimukseni otos koostui siis 200 opinnäytetyöstä, jotka oli tallennettu Theseukseen heti Annifin käyttöönoton jälkeen. Ko. ammattikorkeakoulussa on aloituspaikkoja eniten tekniikan- ja luonnonvara-alan yksikössä, joten valmistuneet opinnäytetyöt painottuvat selvästi sinne.

Koulutusala	Määrä	Vain Annif	Annif+oma	vain oma
Kulttuuriala	16	2	14	0
Liiketalous	47	30	17	0
Sosiaali- ja terveysala	54	21	33	0
Tekniikka- ja luonnonvara-ala	83	59	22	2

Taulukko 1. Valintojen jakautuminen koulutusaloittain.

Määriä tutkimalla osoittautui, että tekniikan ja luonnonvara-alan opiskelijat olivat tyytyväisimpiä Annifin ehdotuksiin, joten luultavasti ne vastasivat hyvin niitä

termejä, jotka opiskelija oli lisännyt työnsä tiivistelmäsivulle tietämättä Annifin olemassaolosta. Kaksi opiskelijaa oli kuitenkin hylännyt kaikki Annifin ehdotukset. Vain Annifin ehdotuksiin luotettiin suuresti myös liiketalouden opiskelijoiden keskuudessa. Omia termejä haluttiin lisätä enemmän sosiaali- ja terveysalan opiskelijoiden ja erityisesti kulttuurialan opiskelijoiden keskuudessa.

Annifin erikoisuuksia

Annif käyttää kokotekstiä ehdotusten pohjana. Tämän takia tuntui erikoiselta, että Annif tarjosi valittavaksi termiksi opiskelijan sukunimeä, jos se sattui olemaan esimerkiksi eläinlaji tai paikkakunta. Opinnäytetyö itsessään ei millään tavalla käsitellyt noita aiheita eikä työn tiedostosta ko. sanaa ei muista yhteyksistä löytynyt. Annif yllätti myös muodostamalla termin opiskelijan etunimestä ja sukunimen ensimmäisestä kirjaimesta (Santeri A. => santeriat). Opinnäytetyön aihe ei millään tavalla viitannut santeriaan eli afrokuubalaiseen voodoosukuiseen uskontoon, joten arvelin logiikan löytyvän opiskelijan nimestä. Annif myös mielellään tarjosi valittavaksi eri Microsoftin ohjelmien nimiä, kuten Word for Windows, Excel, Power Point tai käyttöjärjestelmää Windows10.

Betonirakentamiseen liittyvässä opinnäytetyössä oli Annif tarjonnut valittavaksi termiä Virumaa, joka on virolainen muinaismaakunta. Tämän ehdotuksen taustalla lienee betoniin kohdistuva pakkovoima eli viruma. Muutamassa tapauksessa en pystynyt keksimään miten Annif oli päätynyt tarjoamiinsa termeihin, kuten hirsitalojen vientiä käsittelevään opinnäytetyöhön ehdotettu Uusi-Ruotsi, joka oli vuonna 1638 perustettu siirtokunta Pohjois-Amerikan itärannikolla. Samaten Annifin ehdottama termin edet, joka on aasialaisen heimon nimi, yhteys murskauslouhintaan ja pilvipalveluun jäi hämärän peittoon. Työhyvinvointia käsittelevässä opinnäytetyössä oli Annifin tarjoamien hyvin ymmärrettävien termien joukossa ehdotuksena termi kiilteet, jonka liittymistä aiheeseen en myöskään pystynyt havaitsemaan.

Nykyisin opiskelijat tekevät opinnäytetöinään usein erilaisia opetus- tai ohjevideoita. Näissä Annif tarjosi opiskelijalle valittavaksi termiä kuvanauhurit, joka todennäköisesti perustuu sanan video esiintymiseen opinnäytetyön tekstissä. Eräässä EEG-rekisteröintiä käsittelevässä opinnäytetyössä Annifin tarjoamissa termeissä oli sekä urkujen rekisteröinti että kuvanauhurit, kumpikin aivan yhtä kaukana aiheesta.

Asiasanaksi tarjottiin usein myös termejä ammattikorkeakoulut tai opinnäytteet, jotka lienevät turhan itsestään selviä kuvaamaan opinnäytetöitä, varsinkaan kun opinnäytetyöt eivät kumpaakaan aihetta käsitelleet. Tämä on selvästi toisteisuutta, jota havaittiin myös JYXin graduaineiston tutkimuksessa. Selvästi hankaluutta aiheuttavat myös termit, joilla on eri merkitys eri asiayhteyksissä, kuten istukka tai niska. Autojen moottoreissa käytettäviä lisäaineita käsittelevässä opinnäytetyössä Annif tarjosi termiä golf, kun työssä tutkimusvälineistönä käytettyjen VAG-konsernin valmistamien henkilöautojen joukossa oli yksi, jonka mallinimi on Golf.

Enemmän hyötyä kuin haittaa

Annif on näiden kuluneiden kahdeksan kuukauden aikana osoittautunut oivalliseksi apulaiseksi. Vaikka paikoitellen ehdotukset saattavat mennä ohi aiheen, opiskelijat osaavat olla myös valitsematta niitä eivätkä automaattisesti valitse jokaista ehdotusta.

Tutkin myös hieman sitä, kuinka paljon Annifin ehdotukset vastaavat niitä termejä, jotka opiskelijat ovat itsenäisesti valinneet ja kirjoittaneet ne tiivistelmäsivulle työn tiedostoon. Yllättävän paljon löytyi samoja termejä. Eniten ehkä heittoa oli sosiaali- ja terveysalan osalta, jossa opiskelijat olivat listanneet erilaisten testien ja laitteiden nimiä. Annif joutuu joskus myös sijaiskärsijän osaan, kun YSOn lievä jälkeenjääneisyys käy ilmi. Aivan samaa vauhtia sanastoon ei ilmesty uusia termejä kuin uusia ilmaisuja vakiintuu käyttöön.

Joissakin tapauksissa opiskelija voi myös hämmentyä ehdotusten määrästä. Annif tarjoaa kymmenen ehdotusta jokaisesta tallennetusta tiedostosta, joten syöttölomakkeella voi olla melkoinen määrä valittavissa. Vasta törmäsin opinnäytetyöhön, jossa opiskelijalla oli 40 termiä valittavissa, osa tosin päällekkäisiä liitetiedostojen sisällöstä johtuen.

Lähteet

Lehtinen, M., Niininen, S., Inkinen, J., Lappalainen, M. 2021. Automaattisen kuvailun palvelun integroiminen Kansalliskirjaston järjestelmäkokonaisuuksiin – tietovirrat ja prosessit. Kansalliskirjaston raportteja ja selvityksiä 1/2021. http://urn.fi/URN:ISBN:978-951-51-6986-

0"

Häyrinen, A. 2019. Annif oikeissa töissä – miten Annifia käytetään JYU:n Avoimen tiedon keskuksessa. Ari Häyrisen esitys Kirjastoverkkopäivillä 23.10.2019 Helsingissä. http://urn.fi/URN:NBN:fi-fe2019120445632

Lehtinen, M., Inkinen, J. & Suominen, O. 2019. Aaveita koneessa: Automaattisen sisällönkuvailun arviointia Kirjastoverkkopäivillä 2019. Tietolinja, 2019(2). http://urn.fi/URN:NBN:fi-fe2019120445612

Kirjoittajat

Tiina Tolonen

Informaatikko
Tiina Tolonen on informaatikko Tiedekirjasto Pegasuksessa, joka palvelee Oulun yliopistoa ja Oulun ammattikorkeakoulua.
Kirjoittajan muut artikkelit ›

Artikkelin tiedot

Kirjoittaja: Tiina Tolonen

Numero: 3/2021 Avoin tki & oppiminen

URN: http://urn.fi/URN:NBN:fi-fe2021060634290

Lisenssit