Automaattinen sisällönkuvailu

Henri Ylikotila

Finton laajennetun projektiryhmän kokous
31.1.2018

Automaattisesta sisällönkuvailusta

- Erilaisten aineistojen automaattinen kuvailu on vauhdilla etenevä trendi niin Suomessa kuin maailmallakin.
- Esimerkiksi Suomi.fi-portaalissa kuvaillaan julkishallinnon organisaatioiden palveluita puoliautomaattisen annotoinnin avulla.
- Saksan Kansalliskirjasto puolestaan kuvailee automaattisesti suuren osan aineistosta.

Kansalliskirjaston suunnitelmat

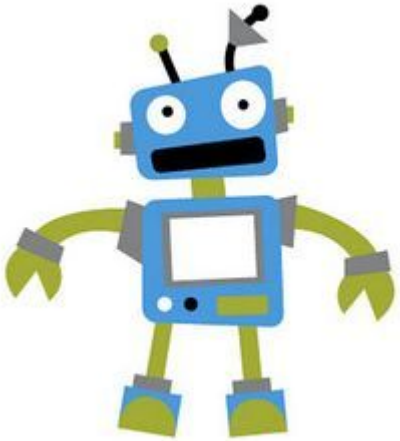
- Automaattisen sisällön kuvailun edistäminen on osa myös Kansalliskirjaston strategian mukaista toimintaa.
  - Työtä tehdään yhteistyössä mm. YLEn, Kansallisarkiston ja Aalto-yliopiston kanssa
    - Esimerkiksi Digitalia-projektissa pyritään tunnistamaan toimijoita ja paikkoja tekstistä
  - Elektroniset aineistot ovat luonnollinen lähtökohta valmiiksi saatavilla olevien kokotekstiensä ansiosta

Annif

- Osma Suominen on kehittänyt prototyypin automaattista sisällönkuvailua tuottavasta Annifiksi nimetystä “robotista”.

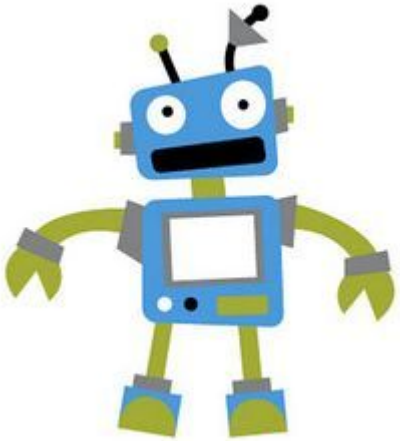
Annif

- Osma Suominen on kehittänyt prototyypin automaattista sisällönkuvailua tuottavasta Annifiksi nimetystä “robotista”.
- Annifia voi kokeilla vapaasti osoitteessa http://annif.org/ ja sen lähdekoodi on vapaasti saatavilla CC0-lisenssillä.



Annif

- Osma Suominen on kehittänyt prototyypin automaattista sisällönkuvailua tuottavasta Annifiksi nimetystä “robotista”.
- Annifia voi kokeilla vapaasti osoitteessa http://annif.org/ ja sen lähdekoodi on vapaasti saatavilla CC0-lisenssillä.
- Hyödyntää tilastollisin menetelmien avulla Finnan teksti-indeksiä.



Annifin jatkokehitys

- Tutkittavia uusia suuntia Annifin jatkokehityksessä ovat
  - Neuroverkot ja koneoppiminen
    - MAUI-työkalu sekä uudemmat lähestymistavat
    - Näiden yhdistäminen nykyisiin tilastollisiin menetelmiin

Annifin jatkokehitys

- Tutkittavia uusia suuntia Annifin jatkokehityksessä ovat
  - Neuroverkot ja koneoppiminen
    - MAUI-työkalu sekä uudemmat lähestymistavat
    - Näiden yhdistäminen nykyisiin tilastollisiin menetelmiin
  - Automaattinen laadunarviointi
    - Tarkoituksena luoda prosessi, jossa työkaluja ajetaan säännöllisesti koeaineistoilla
    - Mahdollistaa Annifin tuottaman kuvailun laadunvalvonnan valittujen laatumittareiden avulla (esim. precision, recall, rolling).

Linkitetyn datan hyödyntäminen

- YSOn LCSH-linkitystä voidaan hyödyntää sisällönkuvailukäsitteiden tuottamisessa kopioluetteloinnin yhteydessä

Automaattisen sisällönkuvailun testi Kansalliskirjaston aineistoilla

- Tavoitteena on saada alustavaa kokonaiskuvaa siitä, kuinka pitkälle ainestomme sisällönkuvailua olisi mahdollista automatisoida.
- Tavoitteena on myös selvittää minkälaisin prosessein automatisointi olisi toteutettavissa osana kirjastojen työarkea.

Automaattisen luokittelun testaus

- Selvitetään aineistojen luettelointiin kuuluvaa UDK-luokitusta automatisoida samaan tapaan kuin asiasanoitusta.
- Saksassa automaattinen sisällönkuvailu aloitettiin juuri luokittelusta
- UDK-luokitus julkaistaan Fintossa

Kiitos

https://tinyurl.com/laapro18-autom