Asiasanoitus-komponentin kuvaus

Tausta, hyöty ja käyttötarkoitus

Komponentti pohjaa Kansalliskirjaston kehittämään Annif-ohjelmistoon, joka on jatkuvan ja aktiivisen kehityksen kohteena. Voit käydä tutustumassa aiheeseen https://annif.org/ . Asiasanoitusta käytetään tiedon etsimiseen aihetunnisteiden perusteella ilman, että koko dokumenttia tarvitsee käydä läpi. Näin esim. valitun asiasanan sisältämät dokumentit löydetään tehokkaasti. Alun perin Annif kehitettiin erilaisten lopputöiden ja tieteellisten artikkelien asiasanoittamiseen, mutta tulokset osoittavat, että se on varsin yleiskäyttöinen. Annif-ohjelmistoa käyttävät esim. Yleisradio ja Saksan kansalliskirjasto.

Mitä komponentti tekee?

Komponentti kokoaa yhteen erilaisia asiasanoitusmalleja hyödyntäessään Annif-ohjelmistoa, joka pohjautuu koneoppisen ja kieliteknologian ratkaisuihin. Lopputuloksena on dokumentin kuvaus asiasanoittain.

Tiedostoformaatit ja mahdolliset esikäsittelyt

Tuetut tiedostoformaatit ovat tällä hetkellä: .pdf, .jpg, .tif, .tiff, .xml ja .txt. Mikäli tiedosto tulkitaan digisyntyiseksi, sille ei suoriteta OCR-käsittelyä. Mikäli kyseessä on kuvatiedosto, OCR-käsittely suoritetaan, sillä Annif osaa tulkita vain tekstisisältöä. Käyttämällä Apache Tika -ohjelmistoa https://tika.apache.org/ tämä onnistuu.

Tulokset

Komponentti palauttaa X-määrän (oletuksena YY) asiasanaa .json-muotoisena datana, jonka voi käyttöliittymästä ottaa ulos (export) myös .csv muotoisena.

Komponentin koulutuksessa käytetty aineisto

Annifin koulutuksessa on käytetty seuraavia sanastoja: YSO-suomi, YSO-english ja ALLFO-svenska.