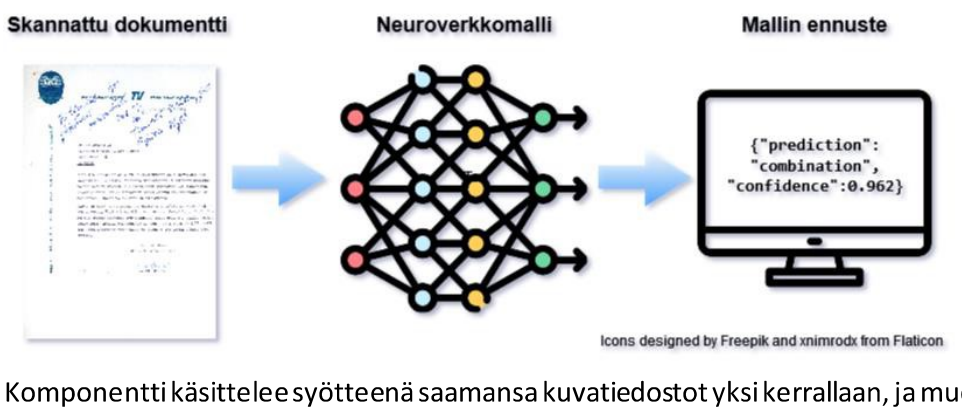
Kirjoitustyypin tunnistus -komponentin kuvaus

Tausta, hyöty ja käyttötarkoitus

Digitoidut asiakirjat voivat sisältää pelkästään kone- tai käsinkirjoitettua tekstiä, tai vaihtelevassa määrin molempia kirjoitustyyppejä sekaisin. Yleisiä esimerkkejä jälkimmäisistä ovat konekirjoitetut dokumentit, jotka sisältävät käsin kirjoitetun allekirjoituksen, sekä erilaiset käsin täytetyt lomakkeet.

Dokumentin luokittelulle sen sisältämien eri kirjoitustyyppien perusteella voi olla tarvetta esimerkiksi, kun halutaan paikantaa aineistosta ne sivut, jotka sisältävät käsin kirjoitettuja merkintöjä. Luokittelua voidaan myös hyödyntää silloin, kun halutaan kohdennetusti ohjata konekirjoitetut dokumentit konekirjoitetun tekstin tunnistusprosessiin (optical character recognition, OCR) ja käsin kirjoitetut dokumentit vastaavasti käsin kirjoitetun tekstin tunnistukseen (handwritten text recognition, HTR).

Mitä komponentti tekee?



Komponentti käsittelee syötteenä saamansa kuvatiedostot yksi kerrallaan, ja muodostaa koulutuksen aikana oppimiensa parametrien pohjalta ennustuksen kuvan sisältämistä tekstityypeistä. Mikäli komponentti pitää todennäköisimpänä vaihtoehtona sitä, että kuva sisältää vain konekirjoitettua tekstiä, sijoitetaan se konekirjoitettujen luokkaan, ja vastaavasti toimitaan kahden muun luokan (käsin kirjoitettu, sekamuoto) tapauksessa. Komponentti palauttaa käyttäjälle ennustetun luokan.

Miten komponenttia käytetään?

Komponentti on saatavilla sekä osana Arkkiivi-käyttöliittymää (http://www.arkkiivi.fi/) että itsenäisenä, ohjelmointirajapinnan (API) tarjoavana sovelluksena DALAI-hankkeen GitHub-sivulla (https://github.com/DALAI-project/WritingtypeAPI). Tarkemmat ohjeet komponentin käyttöön löytyvät oheisilta verkkosivuilta.

Komponentin koulutus

Komponentin koulutukseen on käytetty .jpg-muotoisia kuvatiedostoja, joiden koko on yhdenmukaistettu 224 x 224 pikseliin. Pelkästään käsin kirjoitettuja tekstejä sisältäviä kuvatiedostoja on koulutuksessa käytetty n. 22 000 kappaletta, pelkästään konekirjoitettua tekstejä sisältäviä tiedostoja n. 15 000 kappaletta ja tekstityypin suhteen sekamuotoisia tiedostoja n. 19 000. Ajallisesti mukana on aineistoa 1700-luvun lopulta 2000-luvun alkuun asti.

Vaikka koulutusaineistoon on pyritty kokoamaan monipuolisesti erilaisia esimerkkitapauksia, aineiston määrä on rajallinen ja komponentin luokittelussa tapahtuu myös virheitä. Yhdistelmäluokan tunnistaminen tuottaa komponentille eniten haasteita, ja virheluokitukset ovat mahdollisia erityisesti, jos käsin kirjoitettu teksti on heikosti havaittavaa ja/tai sitä esiintyy dokumentissa vähän.