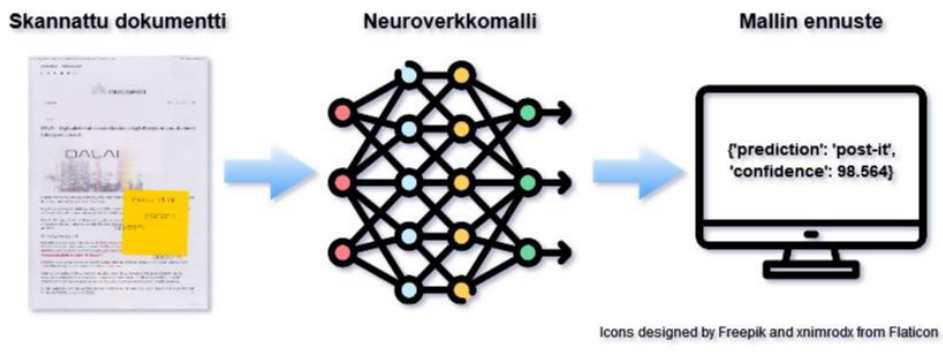
Post-itien tunnistus –komponentti

Komponentin tausta, hyöty ja käyttötarkoitus

Asiakirjojen digitointiprosessissa skannatuksi saattaa tulla dokumentteja, joiden päälle on kiinnitetty yksi tai useampi post-it-lappu. Mikäli post-it-lapun sijainti vaikuttaa asiakirjan sisällön luettavuuteen, on hyvä, jos tällaiset tapaukset voidaan tunnistaa ja tarvittaessa skannata uudelleen. Post-it-lappujen koneellinen tunnistus auttaa näin parantamaan digitaalisen aineiston laatua ja vähentää tarvetta manuaaliseen laaduntarkistukseen.

Mitä komponentti tekee?



Komponentti käsittelee syötteenä saamansa kuvatiedostot yksi kerrallaan ja muodostaa koulutuksen aikana oppimiensa parametrien pohjalta ennustuksen kuvan sisällöstä. Mikäli komponentti tunnistaa kuvasta post-it-lapun yli 50 % todennäköisyydellä, kuva luokitellaan virheelliseksi. 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/PostitAPI). Tarkemmat ohjeet komponentin käyttöön eri ympäristöissä löytyvät oheisilta verkkosivuilta.

Komponentin koulutus

Komponentin koulutukseen on käytetty .jpg-muotoisia kuvatiedostoja, joiden koko on yhdenmukaistettu 224 x 224 pikseliin. Koulutuksessa käytettyjä dokumentteja on yhteensä n. 55 000 kappaletta, joista post-it-lappuja sisältäviä dokumentteja on n. 4 000. Post-it-lappuja sisältäviä kuvia on sekä tehty itse että poimittu Kansallisarkiston massadigitointiprosessin yhteydessä manuaalisesti virheelliseksi luokitelluista kuvista. Dokumentit sisältävät vaihtelevan määrän eri kokoisia ja eri värisiä, eri puolille dokumenttia sijoitettuja post-it-lappuja, joissa voi olla myös tekstiä.

Vaikka koulutusaineistoon on pyritty kokoamaan monipuolisesti erilaisia esimerkkitapauksia, aineiston määrä on rajallinen ja komponentin luokittelussa tapahtuu myös virheitä. Esimerkiksi post-it-lappuja muistuttavat sisältöelementit, kuten värilliset neliskulmaiset tekstikentät, saattavat johtaa virheelliseen luokitukseen.