KANSALLINEN AUDIOVISUAALINEN INSTITUUTTI
NATIONELLA AUDIOVISUELLA INSTITUTET
NATIONAL AUDIOVISUAL INSTITUTE

Kansallinen audiovisuaalinen instituutti

Radio- ja televisioarkisto (RTVA)

Tekstintunnistus radio- ja televisioarkiston videotiedostoille

Nykyinen prosessi:

- TV-ohjelmien tekijätietoja tallennetaan tietokantaan tällä hetkellä mm. käyttämällä ohjelmien lopputekstejä lähteenä.
- Lopputeksteissä olevista tekijätiedoista otetaan kuvakaappaus toistojärjestelmässä (VLC) olevalla toiminnolla ja tallennetuista kuvista poimitaan tiedot manuaalisesti.
- Keskeiset tekijät ja näyttelijät/esiintyjät linkitetään ohjelmaan/sarjaan käyttäen metadatan hallintajärjestelmää.

Tekstintunnistus radio- ja televisioarkiston videotiedostoille

- Manuaalista prosessia halutaan keventää automatiikalla, joka tunnistaa videon kuvasta tekijätietoja ja tallentaa ne tekstimuodossa käsiteltävän ohjelman yhteyteen.
- Hankittiin kehitystyötä Hanselin DPS-hankintajärjestelmällä (hankintapäätös 29.4.2022)
  - Proof of Concept (PoC) -kehitystyö, jonka tavoitteena on automatisoida videotiedostojen manuaalista käsittelyä videokuvasta tekstiksi –toiminnolla.
- Pääpaino on suomenkielisen tekstin tunnistamisessa. Algoritmin pitäisi kuitenkin pystyä tunnistamaan yleisimmät vieraskieliset tekijätiedot (tekijät ja näyttelijät), koska toisinaan tekijätiedot ovat englanninkielisiä (esim. voice over).

Tekstintunnistus radio- ja televisioarkiston videotiedostoille

- Analyysi:
  - Tunnistaa lopputekstien sisältämät sekvenssit ja niihin sisältyvät tekstit.
  - Ratkaisu tunnistaa vähintään ohjelman tekijätietoihin liittyvät suomenkieliset ja yleisimmät vieraskieliset tekstit.
  - Analyysin tuloksena tuotetussa tekstissä on selkeästi esillä, mikä ohjelman tuotantorooli liittyy tiettyyn henkilön tai yhtiön nimeen.
- Automaatio:
  - Löytää analyysiin valittavat tiedostot.
  - Siirtää tiedostot automaation työjonoon.
  - Analyysi, raportointi, lopputulos.

Muuta ajankohtaista automaatiota

- Automaattinen laadunvalvonta hyväksyttiin maaliskuussa 2022 ja otettiin käyttöön päivittäisessä työssä toukokuussa viikolla 18.
  - Arvioidut työaikasäästöt n. 1-1,5 htv
- KAVI mukana LAREINA-hankkeessa in-kind-partnerina (2023-2026)
  - Hankkeessa kehitetään FIN-CLARIN-tutkimusinfrastruktuuria (RI) tukemaan tutkimuksen, elinkeinoelämän ja koko yhteiskunnan tarpeita tarjoamalla innovaatioalusta luonnollisen kielen käsittelyyn sekä avoimesti saatavilla olevia tietokokonaisuuksia ja datapalveluita niin tutkimuksen, koulutuksen, yritysten kuin julkisten organisaatioiden tarpeisiin.
  - Infrastruktuurihankkeen tavoitteena on levittää ja vahvistaa johtavaa asiantuntemusta kielenkäsittelyn alalla.
  - Erillinen PoC KAVIn ja Aalto yliopiston sekä Helsingin yliopiston kanssa syksyllä 2022
- Annif taas kehitystyöhön syksyllä 2022 – keväällä 2023?