Automaattinen asiasanoitus Radio- ja tv-tietokanta Ritvassa

- Automaattista asiasanoitusta on kokeiltu KAVIn ja Ylen yhteisprojektissa 2019
- Ylen omatuotantoisia asiaohjelmia asiasanoitettu manuaalisesti 313 ohjelmaa touko-elokuun aikana, Yle arkisto + KAVI/RTVA
- Tämän jälkeen aineistoa on käytetty Annif-työkalun koulutukseen ja testaukseen
- Tulosten vertailu syys-lokakuussa 2019
- Artikkeli kokeilusta alkuvuodesta 2020

Työvaiheet


Ylellä asiasanoitetaan 100 ohjelmaa ja KAVIssa 213 ohjelmaa
KAVI poimii Ylen asiasanat
Koulutusaineisto jaetaan viiteen osaan testausta varten
Ylen ohjelmien Feenix- ja Metro-ID:t kerätään erilliseen taulukkoon
Ylen ja KAVIn aineisto siistitään
Dokumentoidaan testitulokset erilliseen taulukkoon
KAVIlle annetaan oikeudet Ylen tietokantoihin ja ID-numerotaulukko
Ylen ja KAVIn asiasanoittamista ohjelmista muodostetaan koulutusaineisto
Analysoidaan tulokset

Taustaa

- Pohja-aineistona ohjelmien tekstitystiedostot, jotka on saatu Yleltä
- Asiasta konsultoitu Kopiostoa: lupa käyttää tekstitystiedostoja tv-aineiston järjestämiseen
- Suomenkieliset, kuulovammaiselle tarkoitetut tekstitykset
- Tarkoituksena hyödyntää asiasanoja ohjelma-aineiston sisällönkuvailussa, jotta tutkijat ja opiskelijat löytäisivät haluamaansa tutkimusaineistoa entistä paremmin

Annif

- Tuottaa sisällönkuvailua eli asiasanoja tekstianalyysin perusteella
- Avointa ohjelmakoodia
- Perustuu yhdistelmiin olemassa olevien kielten käsittely- ja koneoppimistyökaluja
- Osma Suomisen luoma, Kansalliskirjasto sitoutunut jatkokehityksen
- Pohja-aineistona Finnasta poimittua materiaalia
- Hyödynnetään sekä leksikaalisia että assosiatiivisia menetelmiä

Annif

- Annifin lisäkoulutus 313 ohjelman avulla, jotka jaettiin viiteen samankokoiseen alinäytteeseen
- Kutakin alinäytettä käytettiin kerran validointitiedostona, loppuja ohjelmia käytettiin harjoitustiedostoina
- Ristiinvalidointiprosessi toistettiin viiteen kertaan ja kaikkien kertojen keskiarvo laskettiin
- Myöhemmin testattiin myös Venetsia-ohjelmatietojärjestelmästä saatujen ohjelmakuvausten käyttöä

Taustaohjelmat

- TF-IDF ja fastText: assosiatiivisia menetelmiä
- Maui: leksikaalinen menetelmä
- PAV ja nn_ensemble: eri menetelmiä yhdisteleviä taustaohjelmia
- Nn_ensemble uusin taustaohjelma, joka hyödyntää neuroverkkoja (nn = neural networks)

Taustaohjelma	Testiaineisto	Tarkkuus (Precision) jos on 5 samaa	Tarkkuuden Precision (kuinka monta oikein tarjotuista) ja saannin Recall (kuinka monta oikein kaikista) keskiarvo
		P@5	F1 score doc average@5 (jatkossa F1 viittauksissa)
tfidf	tekstitykset	0,1303	0,1015
tfidf	tekstitykset + kuvaukset	0,1375	0,1164
fasttext	tekstitykset	0,1818	0,1410
fasttext	tekstitykset + kuvaukset	0,1903	0,1452
maui	tekstitykset	0,2923	0,2243
maui	tekstitykset + kuvaukset	0,3322	0,2555
PAV	tekstitykset	0,3027	0,2335
PAV	tekstitykset + kuvaukset	0,3277	0,2493
nn_ensemble	tekstitykset	0,3747	0,2910
nn_ensemble	tekstitykset + kuvaukset	0,4071	0,3174

Sanastotestaus: kaksi esimerkkiä

- Syötävät sävelet: Kissaherra Ravel
  Alinäyte: 5, ID: PROG_2017_00717037
  Yle Teema & Fem , 15.8.2017 klo 20.50-20.59
  https://www.rtva.kavi.fi/program/details/program/25683811

- Musiikin voima : Mauno Järvelä
  Alinäyte: 4, ID: PROG_2017_00728430
  Yle Teema & Fem , 10.12.2017 klo 16.45-16.51
  https://www.rtva.kavi.fi/program/details/program/26791648

Syötävät sävelet: Kissaherra Ravel

Luetteloijan antamat	PAV (tekstitykset)	PAV (tekstitykset ja kuvaus)	nn_ensemble (tekstitykset)	nn_ensemble (tekstitykset ja kuvaus)
kulttuurihistoria kulinarismi säveltäjät ruokalajit syöminen juomat ruoanvalmistus	säveltäjät ruoanvalmistus sarjakuvataiteilijat ensimmäinen maailmansota kodinkoneet	säveltäjät ruoanvalmistus ensimmäinen maailmansota sarjakuvataiteilijat lihansyönti	säveltäjät ruoanvalmistus kulttuurihistoria juomat syöminen	säveltäjät ruoanvalmistus syöminen kulttuurihistoria juomat

nn_ensemble: kaikki ehdotetut sanat olivat myös ihmisen antamina
PAV: sarjakuvataiteilijat ja kodinkoneet eivät liity mitenkään aiheeseen

Musiikin voima : Mauno Järvelä

Luetteloijan antamat	PAV (tekstitykset)	PAV (tekstitykset ja kuvaus)	nn_ensemble (tekstitykset)	nn_ensemble (tekstitykset ja kuvaus)
lapset (ikäryhmät) kansansoittajat musiikinopettajat pelimannimusiikki viulumusiikki viulistit	säveltäjät apuvälineet nykymusiikki äänimaisema Musiikin aika	taidemusiikki viestintä taiteet säveltäjät teologit	säveltäjät äänimaisema ruoanvalmistus hiljaisuus ruoat (ruokalajit)	viestintä taide säveltäjät taiteet ruoanvalmistus

Luetteloija on asiasanoittaessaan käyttänyt ennakkotietoaan henkilöstä; ohjelmassa ei mainita viuluista tai pelimannimusiikista mitään.
Termi ”apuvälineet” eikä Musiikin aika –tapahtuma ole sisällöille relevanttia.
Kumpikaan taustaohjelma ei anna yhtäkään samaa kuin luetteloija, ja termit ”ruoanvalmistus” sekä ”ruoat (ruokalajit)” eivät ole ohjelmassa relevantteja. Muut taustaohjelman antamat termit ovat osin relevantteja.
Taustaohjelman antamaan kehnohkoon tulokseen selityksenä on se, että ilmaisu on hyvin visuaalista.

Annif tosielämässä

- Jyväskylän yliopiston kirjaston JYX-tietokannassa Annif on otettu käyttöön 2018.
- Kun opiskelija julkaisee gradunsa verkossa JYXissä, hän näkee lomakkeella Annifin ehdotukset asiasanoiksi.
- Opiskelija voi hyväksyä tai hylätä asiasanoja, on myös mahdollista lisätä omia.
- Informaatikko tarkastaa asiasanoituksen ennen julkaisua.
- Hyväksytyistä ja hylätyistä asiasanoista pidetään tilastoa.

Johtopäätelmiä

- Jos automaattinen asiasanoitus toteutetaan esimerkiksi Ritvaan, toiminnon käyttöä tulisi tilastoida kuten Jyväskylän tapauksessa.
- Hyväksyttyjen ja hylättyjen asiasanojen osuuksista on kerättävä tilastotietoa.
- Pitää myös seurata, kuinka paljon annetut hakusanat ovat käyttäjien tiedonhakukäytössä.

Johtopäätelmiä, jatkoa

- Annifia voi käyttää missä tahansa tekstimuotoisessa aineistossa, jota voidaan asiasanoittaa ja luokitella.
- Yksi mahdollinen käyttötapa voisi olla esimerkiksi puheentunnistuksen yhdistäminen Annifiin.
- Annifin käyttöönottoa Ritvassa tullaan edistämään tulosten pohjalta.
- Fiktiivisten ohjelmien asiasanoitusta on aikomus testata keväällä 2020.