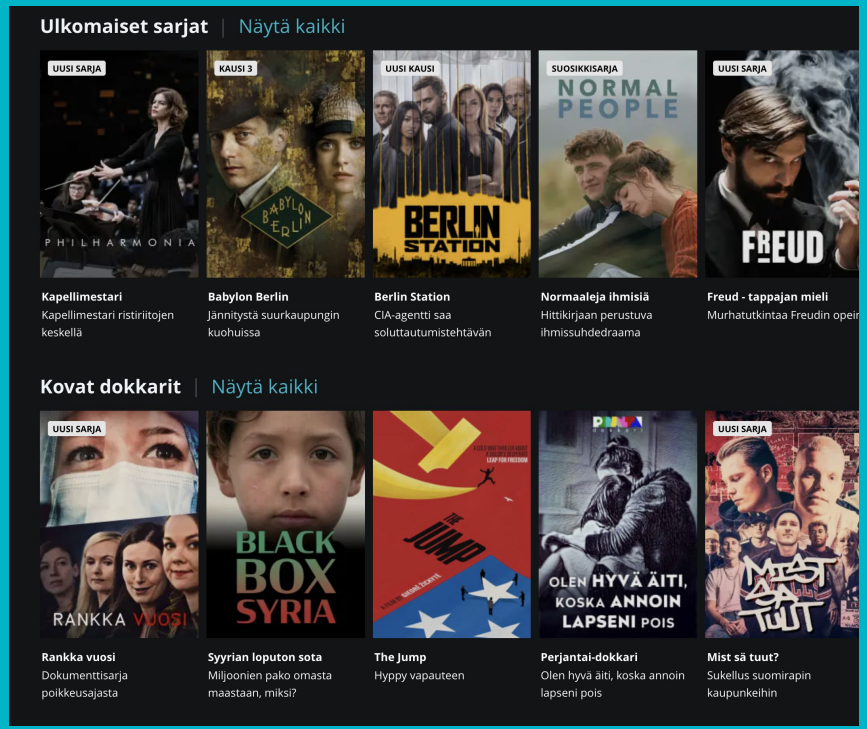
Annif <3 Yle 2.0

Annifin osittainen käyttöönotto artikkeleiden koneavusteisessa asiasanoituksessa

Irene Nikkarinen
Data Scientist, Sisältömetadatatiimi

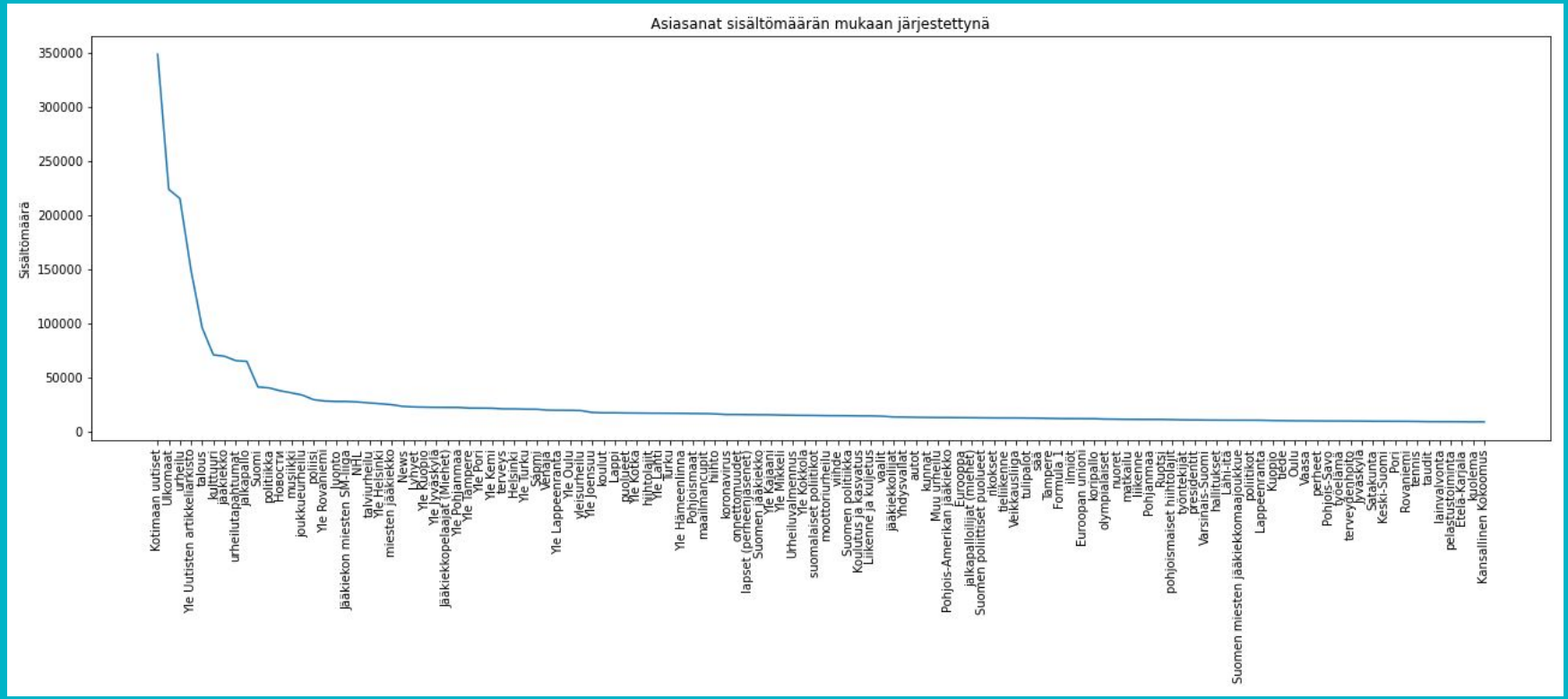
Asiasanoitus Ylellä

Sisältöä asiasanoitettu jo jonkin aikaa koneavusteisesti käyttäen Leikiä



Mikä tekee Ylen käyttötapauksesta haastavan?

Asiasanojen käyttö on epätasaista ja niillä ei vain kuvailla sisältöä



Mikä tekee Ylen käyttötapauksesta haastavan?

Asiasanaston laadussa on ongelmia

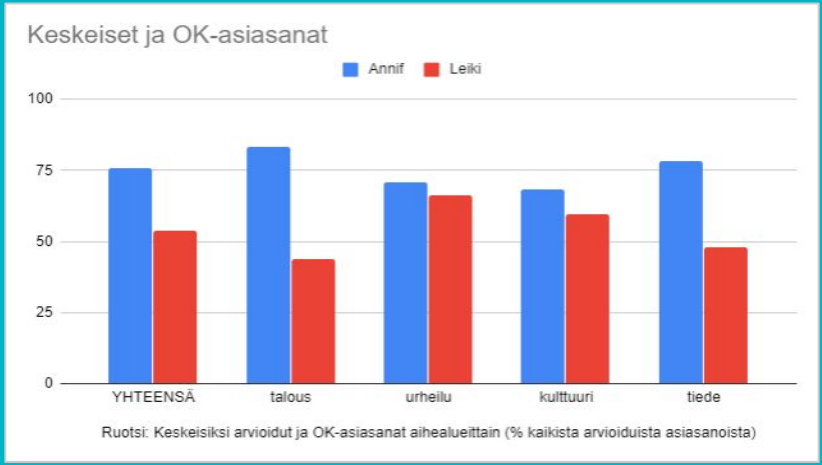
- Paljon osittain tai täysin päällekkäisiä asiasanoja
  - Asiasanastossa syntyy luontaisesti hierarkia, jota emme tällä hetkellä mallinna
    - Asiasanat tuotu eri lähteistä ja ontologioista (Leiki, Wikidata, KOKO)
- Asiasanastoon lisätään uusia asiasanoja noin 200 per viikko
  - Duplikaattien siltaaminen käsityötä

```
['vesillä sattuneet onnettomuudet’, 'vesiliikenneonnettomuudet', 'veneilyonnettomuudet’]

['palo- ja pelastustoiminta', 'palo- ja pelastustyö', 'pelastuspalvelu', 'pelastustoimi', 'pelastustoiminta', 'palotoimi']

[‘pohjalaisia’, ‘pohjalainen’, ‘pohjalaiset’]
```

Annifin ja Ylen historia

Annifin ensitestit keväällä 2020

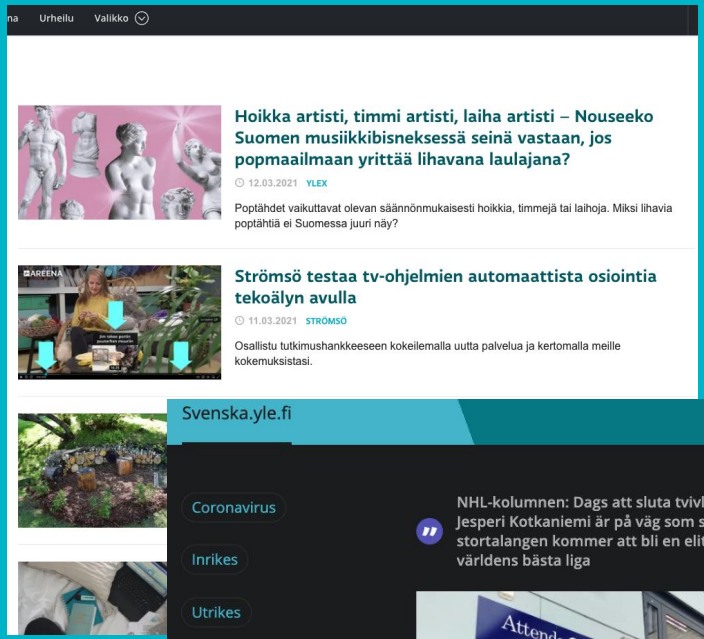
- Annifista Leikin korvaaja?
- Alkuperäisessä kokeilussa yhteensä noin 30 arvioijaa ja noin 100 artikkelia per kieli
  - Annif vaikutti toimivan vähintään yhtä hyvin kuin Leiki suurimmalla osalla aihealueista
- Edelleen käytössä silloin hyväksi havaitut backendit
  - Omikuji Attention
  - Maui
  - Yhdistetään nn ensemblellä
    - Omikujin paino 2, Mauin 1



Seuraavaksi laajempi testi

Miten Annif toimii sisällöntuottajan arjessa?

- Otetaan Annif ensin käyttöön rajatussa joukossa julkaisujärjestelmiä
  - Mahdollisuus vertailla Leikiin
  - Konkreettisia lukuja joilla perustella Annifin käyttöönotto
- Ensimmäiseksi testijoukouksi valikoituivat artikkelijulkaisujärjestelmät SYND ja FYND
  - FYND tuottaa suomenkielistä aihesivustoa https://yle.fi/aihe
    - Aihesivujen tyyli ja aiheet eroavat jonkin verran uutis- ja ajankohtaistoimituksen sisällöstä
  - SYND tuottaa kaiken Ylen ruotsinkielisen sisällön: https://svenska.yle.fi/



Tutkimusasetelma

Miten saadaan parempaa tietoa Annifin toiminnasta?

1. Testataan Annifia ensin SYND/FYNDissä
2. Tallennetaan asiasanaehdotukset sekä Leikiltä että Annifilta
3. Tarjoillaan toimittajalle Annifin asiasanat (n. 15 ehdotusta)
4. Tarkistetaan kerran päivässä molempien mallien asiasanoille, mitkä tarjotuista asiasanoista on otettu julkaistuissa artikkelissa käyttöön
    a. Haetaan tällöin ehdotukset molemmilta malleilta myös sellaisille artikkeleille joille ei haettu aikaisemmin
    b. Tallennetaan analytiikka

- Asetelma on Annifille edullinen, sillä sisällöntuottaja näkee vain Annifin ehdotukset
- Artikkelin ehdotusten ja lopullisten asiasanojen yhdistämistä ei voida tehdä täysin luotettavasti
    - Tehdään otsikon perusteella
- Annif koulutetaan viikottain AWS:ssä käyttäen Ylen artikkelidataa

yle


yle
Tuloksia

Kerätty analytiikka

Dataa Annifin käyttäytymisestä kerättiin noin kolme kuukautta

	Ihminen **varmasti** nähnyt Annifin ehdotukset	Ihminen **mahdollisesti** nähnyt Annifin ehdotukset	Yhteensä
Suomi	114	337	451
Ruotsi	1173	2406	3579
Yhteensä	1287	2743	4030

yle

Kerätty analyt

Näiden kahden sarakkeen värinen ero syntyy tilanteista, joissa artikkelin otsikkoa on muutettu asiasanaehdotusten hakemisen jälkeen

Dataa Annifin käyttäyty me kuukautta

	Ihminen **varmasti** nähnyt Annifin ehdotukset	Ihminen **mahdollisesti** nähnyt Annifin ehdotukset	Yhteensä
Suomi	114	337	451
Ruotsi	1173	2406	3579
Yhteensä	1287	2743	4030

yle

Paljonko Annifin ehdotuksia valitaan?

Artikkeleille, joille ihminen on varmasti nähnyt Annifin ehdotukset

malli	kieli	montako prosenttia asiasanoista mallin ehdotuksia keskiarvo	asiasanamäärä artikkelissa keskiarvo	valittujen ehdotusten keskiarvo	lisättyjen asiasanojen (ei ehdotettu mutta julkaistu) keskiarvo	artikkeleita
annif	sv	87%	10.13	8.70	1.33	1173
annif	fi	72%	10.46	7.27	3.66	114
leiki	sv	45%	10.13	4.46	5.53	1173
leiki	fi	39%	10.46	3.93	6.36	114

yle

Mille asiasanoille Annif toimii?

Annifin ja Leikin vertailua asiasanakohtaisesti

- Asiasanoista ei juurikaan luotettavaa tietoa → analyysi jouduttiin tekemään pitkälti käsin
    - Mikäli asiasanoista tunnettaisiin hierarkia, voitaisiin tutkia aihekohtaisesti
- Laskettiin jokaiselle asiasanalle Leikin ja Annifin onnistumisprosentin erotus, ja tutkittiin asiasanoja jossa erot olivat suurimmat
- Mallit ehdottaneet kokeilussa vain n. 5% kaikista asiasanoista
    - Ei voida tehdä lopullisia päätelmiä

yle

- Leiki saattaa toimia paremmin erisnimissä
  - Urheilu ja urheilijat
- Asiasanat joissa Annif on toiminut Leikiä paremmin ovat pitkälti yleiskäsitteitä

Pohjois-Lapin seutukunta
Kysyntä ja tarjonta
purjehdus
virkamiehet
lahjonta
nopeusrajoitukset
kevyt liikenne
rikosilmoitukset
ammatilliset oppilaitokset
lahjoitukset
teleoperaattorit
julkisuuden henkilöt
asuntopalot
ekoenergia
siivous
mellakat
taloudelliset ennusteet

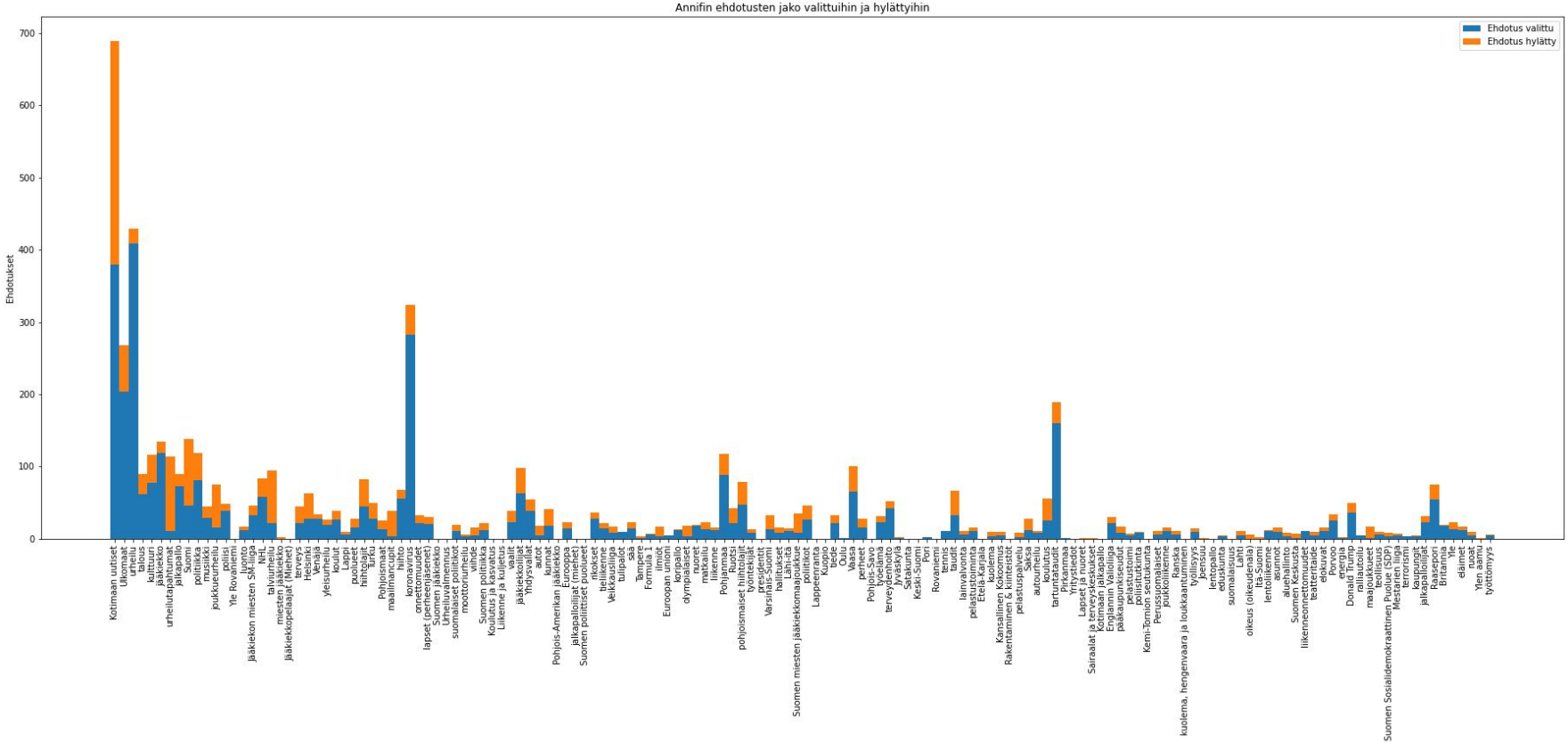
Asiasanat, joille **Annifin** ennustukset ovat valittu useimmin verrattuna Leikiin (asiasanoista, joita molemmat mallit ovat ennustaneet)

Anaheim Ducks
Nigeria
markkinat
viinit
Saaristomeri
Kilpailu- ja kuluttajavirasto
Nepal
syntymäpäivät
Mika Lintilä
Kiekko-Vantaa
Linda Sällström
pseudotiede
apartheid
Martin Luther King, Jr.
Sex Pistols
Nikolaj Ehlers
Alexandria Ocasio-Cortez

Asiasanat, joille **Leikin** ennustukset ovat valittu useimmin verrattuna Leikiin (asiasanoista, joita molemmat mallit ovat ennustaneet)

yle

Annifin ehdotusten jako valittuihin ja hylättyihin



Mitä asiasanoja lisätään eniten?

Suosittuja asiasanoja, koska luku on absoluuttinen

Lisäksi paljon ajankohtaisasiasanoja
- Virustaudit
- Marinin hallitus
- Capitolin valtaus

Annif
COVID-19
koronavirus
Kotimaan uutiset
Pohjanmaa
Koronarokote
uutiset
Ulkomaat
kulttuuri
sää
sääennusteet
Euroopan unioni
Niko Kytösaho
tietokilpailut
Keski-Pohjanmaa

Leiki
Kotimaan uutiset
Ulkomaat
urheilu
koronavirus
COVID-19
tartuntataudit
kulttuuri
talous
rokotus
Yhdysvallat
virustaudit
Koronarokote
Vaasa
jääkiekkoilijat

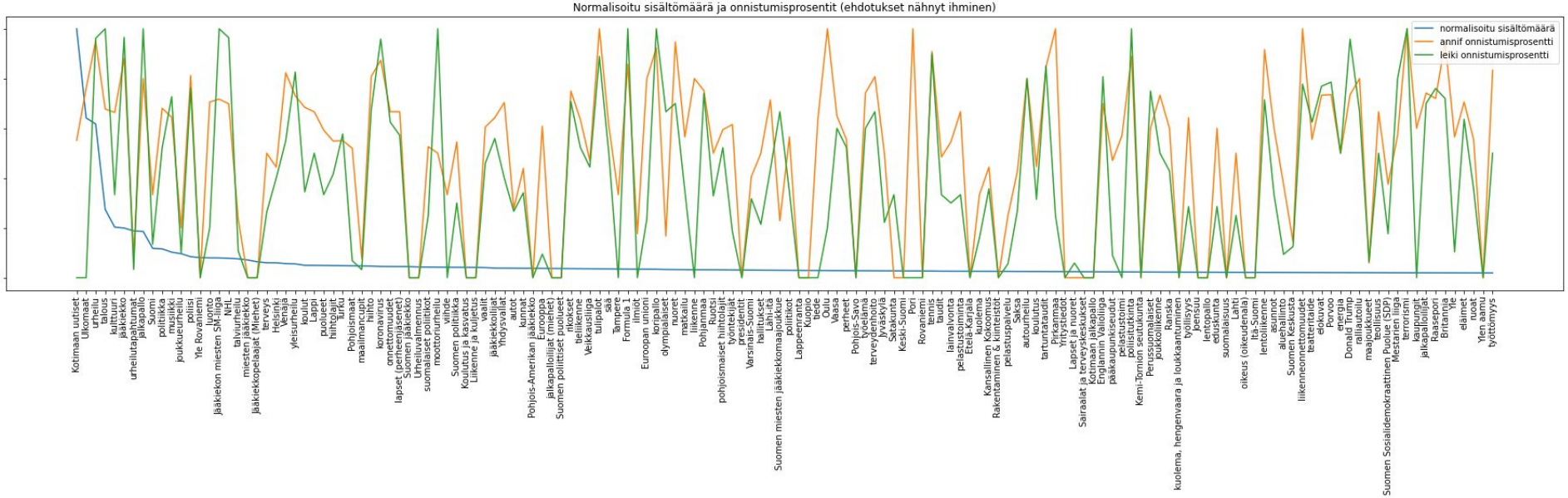
yle

Mitä heikkouksia Annifilla on?

- Annif assosioi asiasanoja, ja saattaa ehdottaa väärää, mutta saman aihealueen asiasanaa
  - *Kerttu Niskanen* tarjotaan asiasanaksi artikkeliin Riitta-Liisa Roposesta
  - *Joulu* ehdotetaan asiasanaksi jalkapalloa käsitteleviin artikkeleihin
    → Uudelleenkoulutus niin, että kannustetaan ennustamaan asiasanoja, jotka esiintyvät yhdessä?
- Oleellisia asiasanoja ehdotetaan matalalla relevanssilla

Annifin ja Leikin onnistumisprosentit

...yhdessä asiasanan normalisoidun sisältömäärän kanssa



Ennustetaanko yleisiä asiasanoja paremmin?

Korrelaatioita sisältömäärän kanssa

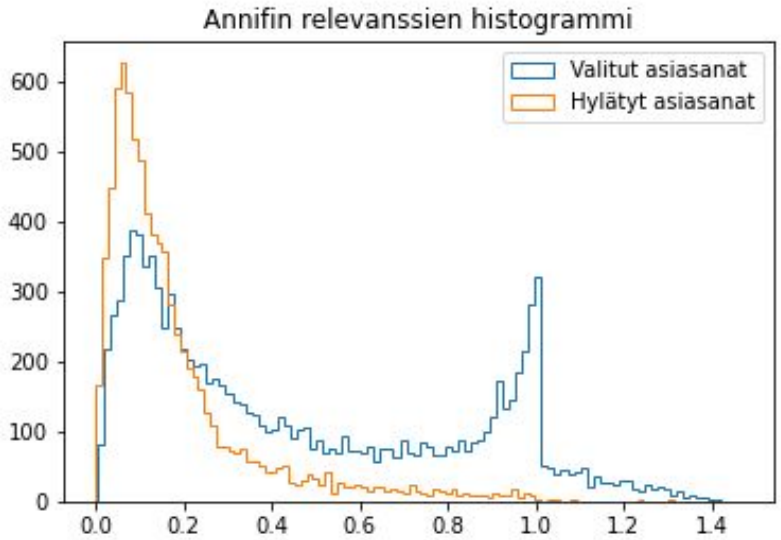
Muuttujat	Malli	Spearman-korrelaatio	p-arvo
Onnistumisprosentti ja sisältömäärä	Annif	0.34	< 0.01
Onnistumisprosentti ja sisältömäärä	Leiki	0.36	< 0.01
Ehdotuskerrat ja sisältömäärä	Annif	0.46	< 0.01
Ehdotuskerrat ja sisältömäärä	Leiki	0.57	< 0.01

Käyttäen artikkeleita, jotka on varmasti nähnyt ihminen
Mallin korrelaatiota laskettaessa käytetty vain sellaisia asiasanoja, joita malli on ennustanut kokeilun aikana

yle

Annifin relevanssit

Ovatko valittujen asiasanojen relevanssit suuremmat kuin hylättyjen?



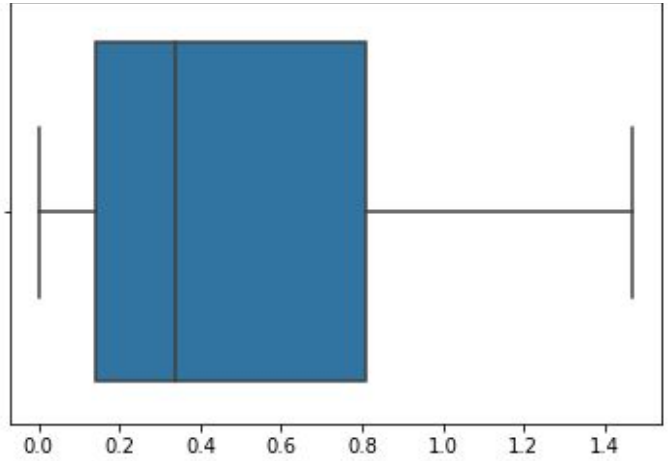
Käytetty artikkeleita, jotka varmasti nähnyt ihminen

yle

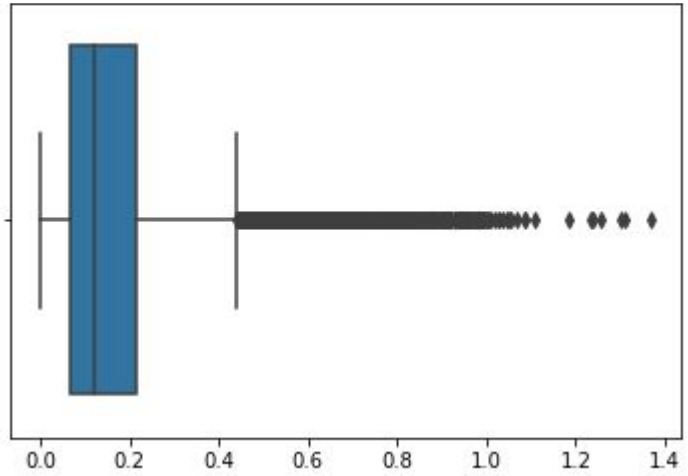
Annifin relevanssit

Ovatko valittujen asiasanojen relevanssit suuremmat kuin hylättyjen?

Boxplot **valittujen** asiasanojen relevansseista molemmille kielille



Boxplot **hylättyjen** asiasanojen relevansseista molemmille kielille



Käytetty artikkeleita, jotka varmasti nähnyt ihminen

yle


yle
Lopputulemat
Mikä on Annifin tulevaisuus Ylellä?

Annifin tulokset ovat olleet lupaavia

Annif tarjoaa Ylelle paljon mahdollisuuksia

- SYND:in ja FYND:in sisällöntuottajilta ei ole tullut valituksia Annifin ehdotuksista
  - No news is good news
- Annifia ollaan myös testattu asiasanoittamaan ohjelmia niiden litteraatioiden perusteella
  - Tulokset positiivisia
  - Tulokset saattavat parantua kouluttamalla litteraatioilla
- Mahdollisuus kouluttaa Annif erikoistumaan erilaisiin sisältöihin
- Tarjottujen ehdotusten rajaaminen
  - Onko tärkeämpää, ettei hyviä asiasanoja leikkaudu pois, vai ettei niitä ole ylimääräisiä?
- Annifin ennustusten laatu sidoksissa Ylen asianaston laatuun
  - Annifille tarjottuja luokkia rajoitettava
- Leikissä mahdollisuus painottaa otsikkoa -- sama toiminnallisuus Annifiin?

yle

Kiitos! Tack!

irene.nikkarinen@yle.fi

yle