ICT Solutions for Brilliant Minds

# Aihemallinnus ja muut menetelmät

## Bibliometriikkaseminaari 2022

Aino Ropponen
aino.ropponen@csc.fi

# Tässä esityksessä

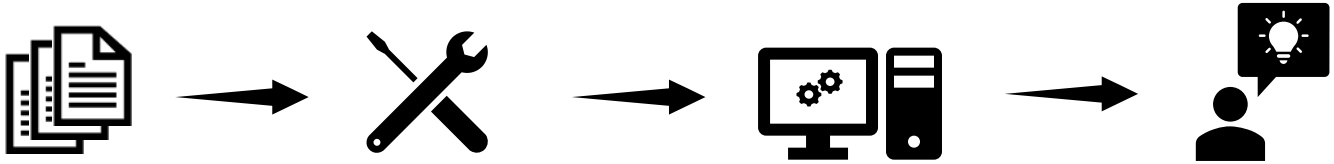
Tekstianalytiikan peruskäsitteitä

- tekstiaineiston esikäsittely
- aihemallinnus
- tekstiaineiston luokittelu

Akatemian lippulaivaohjelman teemojen analyysi

- Alustavia tuloksia ja huomioita

# Tekstianalytiikka



# Tekstiaineiston esikäsittely

**Kielen tunnistaminen** ja vieraskielisten dokumenttien poistaminen

**Tarpeettomien sanojen ja merkkien poistaminen, esim.**

- Hukkasanat (engl. *stopwords*) esim. *mutta, ei, olla, ja*
- Välimerkit
- Numerot
- Internet- ja sähköpostiosoitteet
- Erisnimet (esim. ihmisten nimet)

**Perusmuotoistaminen** eli lemmatisointi (tai stemmatisointi)

- Sanan muuttaminen perusmuotoonsa eli siihen muotoon, jossa se esiintyy sanakirjassa.
- Esimerkiksi sana *on* muuttuu muotoon *olla* ja sana *koulutuksen* muotoon *koulutus*. Perusmuotoisia sanoja kutsutaan nimellä lemma.

# Sanayhdistelmät eli N-grammit

- N-grammi tarkoittaa sanojen (tai kirjainten) muodostamia yhdistelmiä
  - 2-grammi (bigram): kahden sanan (tai kirjaimen) yhdistelmää
  - 3-grammi (trigram): kolmen sanan (tai kirjaimen) yhdistelmä
- Tärkeää etenkin englanninkielisten dokumenttien analyysissa:
  - ”machine”, ”learning” vai ”machine learning”

- **Alkuperäinen virke:** “*Natural language processing is a subfield of linguistics, computer science, and artificial intelligence*.”
- **Esikäsittelyn jälkeen:** natural language processing subfield linguistics computer science artificial intelligence
- **2-grammit** : natural language, language processing, processing subfield, subfield linguistics, linguistics computer, computer science, science artificial, artificial intelligence

# Sanojen merkityksellisyys (TF-IDF)

**TF-IDF:** *term frequency–inverse document frequency*

**Perusidea**: lasketaan painokertoimet dokumentin sanoille perustuen sanojen esiintymismääriin tarkasteltavassa dokumentissa ja kaikissa dokumenteissa

- Sanat, jotka ovat hyvin yleisiä joissakin dokumenteissa, mutta harvinaisia toisissa ovat yleensä merkityksellisiä.
- Sanat, jotka ovat hyvin yleisiä kaikissa dokumenteissa eivät ole yleensä merkityksellisiä.

$$\mathrm{tf}(t,d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}},$$

Sanan *t* esiintymismäärä dokumentissa *d* suhteessa dokumentin *d* kaikkiin termeihin.

$$\mathrm{idf}(t,D) = \log \frac{N}{|\{d \in D : t \in d\}|}$$

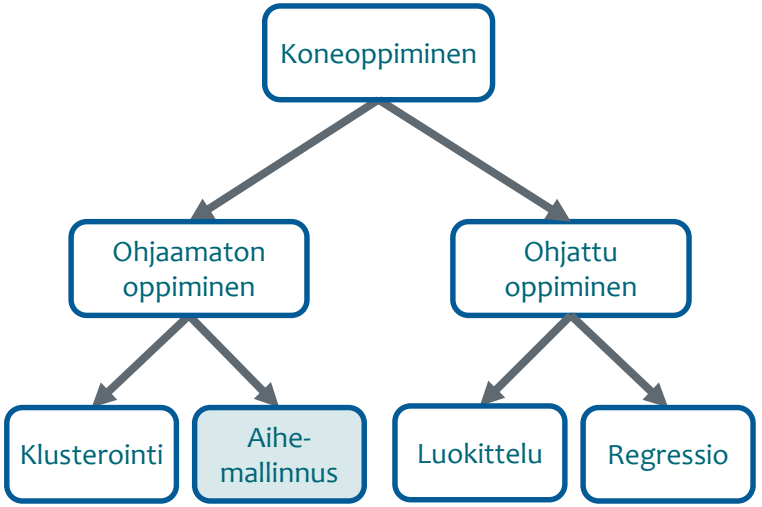
Kaikkien dokumenttien lukumäärä suhteessa dokumenttien määrään, joissa sana *t* esiintyy.

$$\mathrm{tfidf}(t,d,D) = \mathrm{tf}(t,d) \cdot \mathrm{idf}(t,D)$$

TF-IDF painokerroin sanalle t dokumentissa d

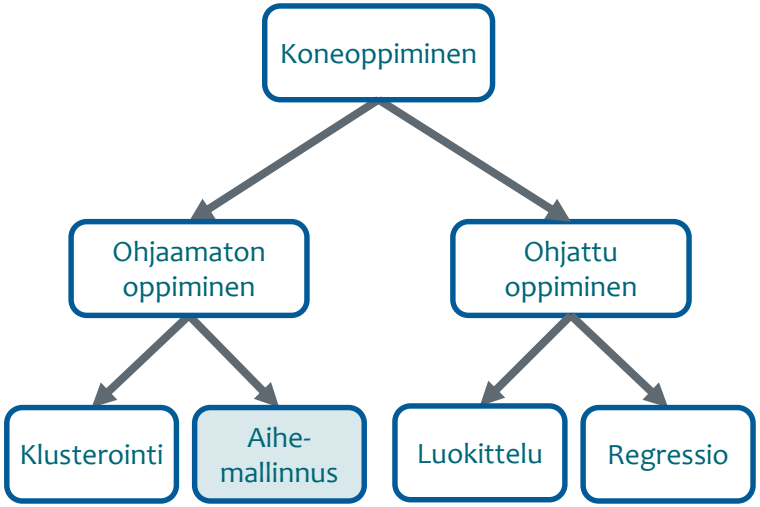
# Aihemallinnus

- Tekstidokumenttien ryhmittely-menetelmä, joka pyrkii tunnistamaan keskenään samankaltaiset dokumentit.
- Ohjaamattoman koneoppimisen menetelmä
  - Ei hyödynnetä opetusaineistoa.
  - Ryhmittelee tekstidokumentit ennalta määrättyyn määrään aiheita, mutta ei ennalta määrättyihin aiheisiin.

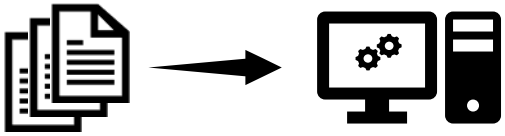


# Aihemallinnus

- Perustuu sanojen esiintymistiheyteen dokumenteissa. Menetelmiä mm.
  - LDA (latent Dirichlet allocation)
  - NMF (non-negative matrix factorization)
- Aihemallinnuksen tuloksena saadaan
  - Sanalistat kunkin aiheen yleisimmistä sanoista.
  - Kunkin dokumentin kuuluminen aiheisiin.
- Huom. aiheiden määrä pitää valita tapauskohtaisesti. Ymmärrettävyyden vuoksi aiheet voidaan nimetä.



# Aihemallinnus



| Aihe | Tyypilliset sanat |
|---|---|
| Aihe 1 | kissa, koira, hevonen |
| Aihe 2 | maito, jogurtti, juusto |
| … | … |
| Aihe N | jalkapallo, tennis, hiihto |

| Aiheen nimi |
|---|
| ”eläimet” |
| ”maitotuotteet” |
| … |
| ”urheilu” |

Aihe 1: 0,3
Aihe 2: 0,3
Aihe 3: 0,4

Aihe 1: 0,8
Aihe N: 0,2

Aihe 4: 0,99

*”Dokumentti 2 liittyy eniten eläimiin, mutta hieman myös urheiluun”.*

Esimerkki

# Aihemallinnus suomalaisille julkaisuille

Suomen Akatemian ja CSC:n yhteistyöhankkeessa vuonna 2021 kartoitettiin Suomessa tutkittavia aiheita ja ilmiöitä.

Aineistona Web of Science -tietokannan julkaisut vuosilta 2008-2019

- Tarkasteltiin englanninkielisiä julkaisuja, joissa vähintään yksi tekijöistä suomalaisesta organisaatiosta.
- Yhteensä 106 736 julkaisua (artikkeli, katsausartikkeli, kirje lehden toimituskunnalle, kirja).

Aihemallinnus tuotti 1026 aihetta

- Keskimäärin 104 julkaisua aihetta kohti (vaihtelu 18-862).
- Isoimmat aiheista laajoja kuten koulutus ja uusiutuva energia, osa pienistä aiheista hyvin spesifejä liittyen esim. tiettyyn eläinlajiin tai sairauteen.

Esimerkki

# Aihemallinnus suomalaisille julkaisuille

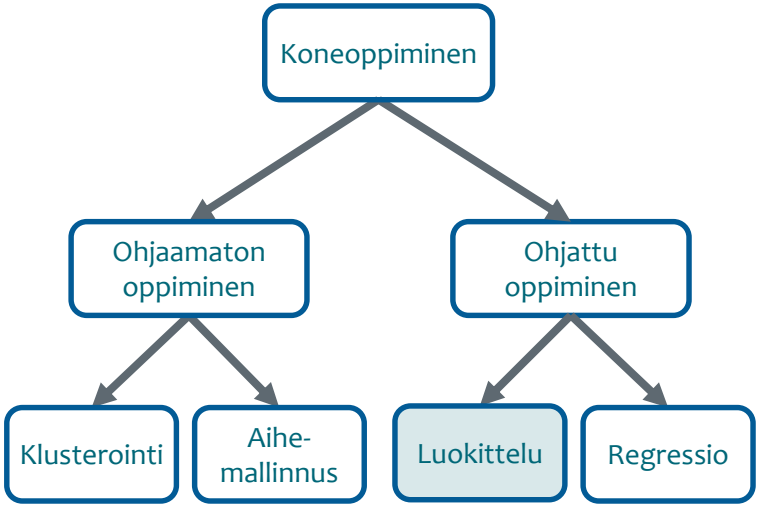
Esimerkki muutamasta löydetystä aiheesta ja aiheiden tyypillisistä sanoista.

| Topic | Human-labelled keywords | Model keywords | tf-idf keywords |
|---|---|---|---|
| 299 | dementia, Alzheimers | impairment dementia alzheimers vascular_risk alzheimer mild_cognitive disease_ad midlife cognitive mci old_age geriatric decline older mild cognition older_adults frailty disease_amyloid apoe ad adults mortality_older brain_atrophy apolipoprotein | disease dementia risk alzheimer cognitive alzheimers |
| 184 | probabilistic models | mcmc chain_monte markov_chain monte carlo monte_carlo metropolis markov bayesian inference likelihood graphical bayes estimation gibbs unbiased parameter stochastic approximation statistical kalman serpent kalman_filter estimating computation | carlo monte markov monte_carlo bayesian chain |
| 14 | service and customer value creation | dominant_logic creation customer value_co value_creation customer_value dominant propositions perceived_value offerings service marketing servitization business service_systems business_models value logic approach_paper methodology service_design product capabilities innovation conceptual | service value creation business customer dominant |
| 242 | microRNA | mirna mirnas micrornas microrna mir rnas transcriptional transcription mrna suppressor noncoding microarray androgen differentially gene_expression regulators cancer_cells chromatin rna expression cancer_cell messenger adipocyte castration transcriptome | mir cancer expression microrna mirna gene |
| 691 | quartz crystal microbalance, nanocellulose | qcm microbalance quartz_crystal quartz polyelectrolyte multilayers cationic afm cellulose cellulose_nfc nanofibrils nfc nanofibrillar plasmon adsorption films ultrathin atomic_force dissipation anionic carboxymethyl adsorbed mfc nanofibrillated chitosan | cellulose qcm quartz crystal films adsorption |
| 15 | machine learning, classification | support_vector classifiers vector_machines classifier vector_machine supervised machine kernel clustering feature discriminant unsupervised classification dimensionality neural_networks regression neural learning vector algorithms neural_network bayes speaker robust prediction | learning data classification analysis machine based |

Lähde: Katja Mankinen ja Yrjö Leino: Identifying research topics and collaboration networks in Finland: topic modelling of scientific publications in 2008–2019 https://www.aka.fi/suomen-akatemian-toiminta/tietoaineistot/tieteen-tila/

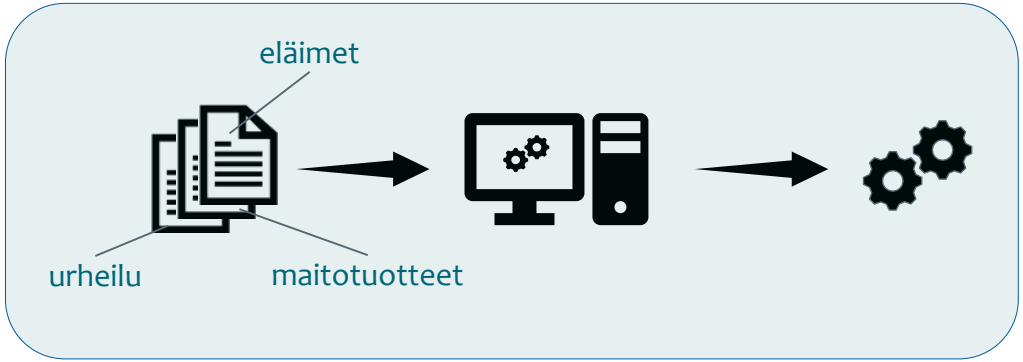
# Tekstidokumenttien luokittelu

- Tekstidokumentteja voidaan luokitella ennalta määrättyihin luokkiin ohjatun koneoppimisen menetelmillä kuten
  - satunnaismetsä (random forest)
  - tukivektorikoneet (support vector machines, SVM)
  - neuroverkot (neural networks).
- Edellyttää opetusdataa eli suuren määrän dokumentteja, joiden luokka on tiedossa.

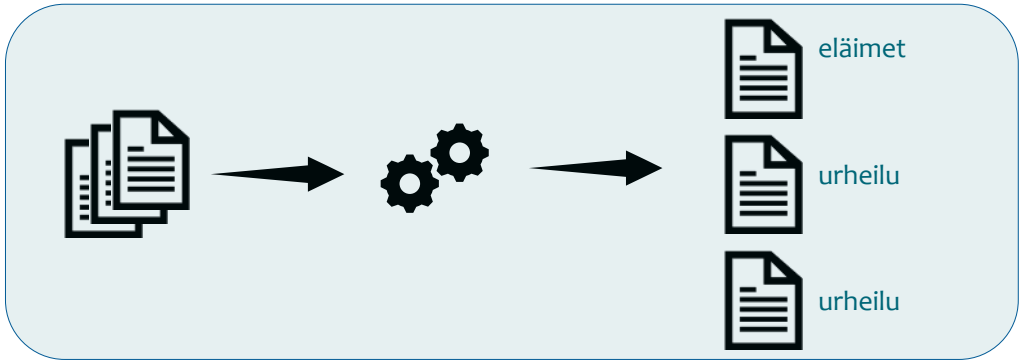


# Tekstidokumenttien luokittelu

Luokittelumallin opettaminen opetusdatalla



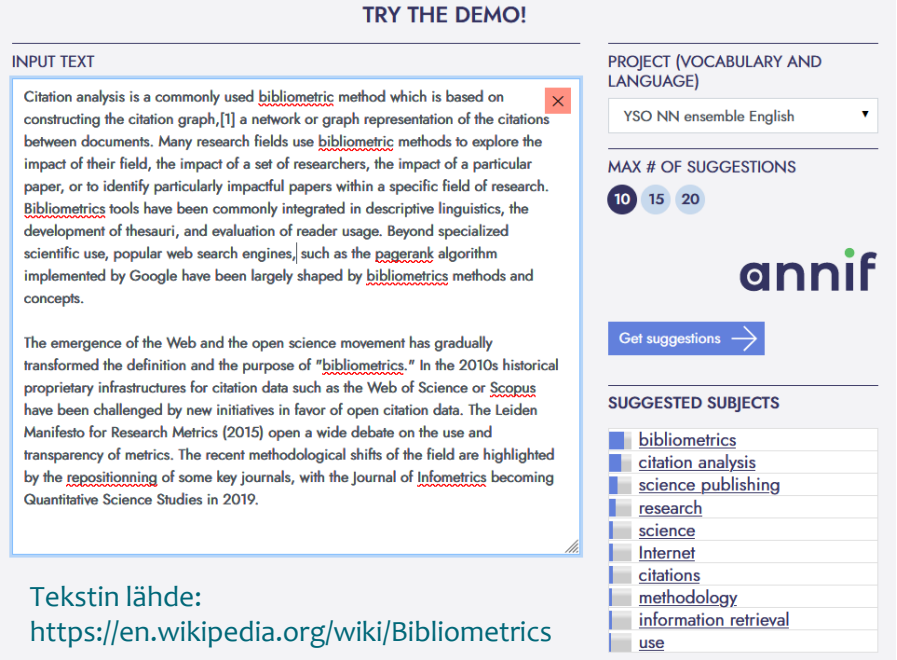

Mallin käyttäminen uusien dokumenttien luokitteluun



Esimerkki

# Aihesanoitus (ANNIF)

- Annif on valmiiksi opetettu luokittelija suurelle määrälle luokkia. Sitä voidaan käyttää automaattiseen sisällönkuvailuun.
- Käytössä mm. tiedejatutkimus.fi-sivustolla, jossa sitä käytetään rahoituspäätösten aihesanoittamisen apuna. Käytettävä sanasto on YSO (yleinen suomalainen ontologia).




Tekstin lähde:
https://en.wikipedia.org/wiki/Bibliometrics

# Analyysi suomalaisista julkaisuista Akatemian lippulaivojen teemoissa 2022

# Lippulaivaohjelman tutkimusaiheet

Suomen Akatemian ja CSC:n yhteistyöhankkeessa tarkastellaan lippulaivaohjelman tutkimusaiheita. Mukana tarkastelussa lippulaivat: INVEST, 6G, FinnCERES, PREIN, iCAN ja FCAI.

Tavoitteena tunnistaa lippulaivaohjelman tutkimusaiheisiin liittyviä julkaisuja ja osaamiskeskittymiä ajanjaksoilla 2010-2014 ja 2015-2018.

Aineistona

- Web of Science -julkaisuaineisto tarkasteluvuosilta
- avainsanat lippulaivakohtaisesti
- esimerkkijulkaisuja lippulaivakohtaisesti.

# Aineisto ja esikäsittely

Alustavassa analyysissä aineistona Web of Science -julkaisut vuosilta 2015-2018:

- vähintään yksi tekijöistä suomalaisesta organisaatiosta,
- julkaisutyyppi: lehtiartikkeli, konferenssijulkaisu, katsausartikkeli
- englanninkielinen

→Yhteensä 40536 julkaisua.

Analyysiin otettu mukaan julkaisujen otsikko, tiivistelmä ja avainsanat.

Tekstin esikäsittely: poistettu välimerkit ja hukkasanat, lemmatisoitu.

Tarkasteltu pelkkiä sanoja sekä 2-grammeja.

# Alustava analyysi ja jatkosuunnitelmat

- Lippulaivojen avainsanojen perusteella tunnistettu lippulaivoihin liittyviä julkaisuja:
  - avainsanojen esiintyminen dokumenteissa
  - avainsanapohjainen aihemallinnus.
- Ensimmäisessä analyysissä 15144 julkaisua tunnistettu kuuluvaksi johonkin lippulaivojen tutkimusaiheista → kynnysarvoa muutettava tiukemmaksi.
  - Esim. SixG-lippulaivaan liittyisi 4485 julkaisua ja FCAI-lippulaivaan 1646 julkaisua.
- Jatkoaskelia:
  - Ohjatun koneoppimisen menetelmien (esim. satunnaismetsä, random forest) hyödyntäminen julkaisujen luokittelussa.
  - Vuosien 2010-2014 julkaisujen analyysi ja vertailu vuosien 2015-2018 tuloksiin.



facebook.com/CSCfi


twitter.com/CSCfi


linkedin.com/company/csc---it-center-for-science


github.com/CSCfi