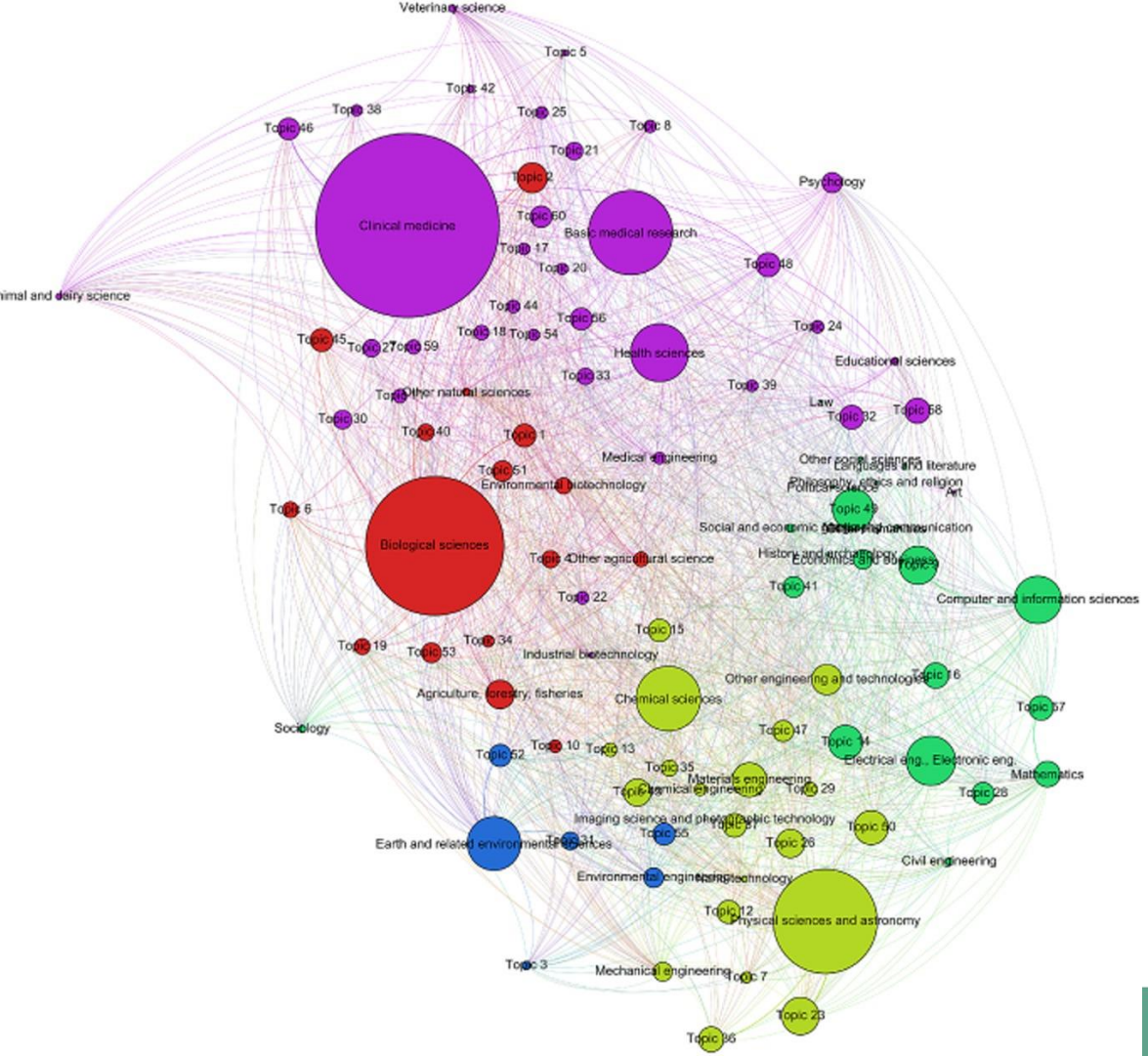


Aihemallinnus Horizon 2020 – tutkimusprojekteille

Pyörni Vartiainen

2.9. 2020

- Aineisto: 16 144 EU:n rahoittamaa tutkimusprojektia ajalta 2014-2020, kustakin tutkimuksesta noin sadan sanan yhteenveto.
- Tavoite: mallintaa tutkimuksissa esiintyviä aiheita ja tutkia aiheiden ominaisuuksia, kuten rakennetta, keskinäisiä suhteita ja yleisyyttä. Samalla arvioidaan menetelmien toimivuutta kyseisen aineiston tapauksessa.
- Metodi: Latent Dirichlet Allocation (LDA). Vaihtoehtoisia toteutuksia esim. Latent Semantic Analysis (LSA) ja Nonnegative Matrix Factorisation (NMF).

LDA-metodin toiminta

- Ohjaamaton algoritmi, jossa kukin dokumentti ajatellaan aiheiden sekoitukseksi ja kukin aihe sanojen sekoitukseksi
- Mallioletuksena oletetaan (virheellisesti), että dokumentit on generoitu satunnaisprosessilla aihe- ja sanajakaumista, jolloin jakaumien parametrejä on mahdollista estimoida “takaisinmallinnuksella” (reverse engineering).
- Aiheiden määrän päättää itse analyytikko vertailemalla mallien ns. probabilistista koherenssia ja tulkitsemalla kvalitatiivisesti aiheiden “järkevyyttä”.

Olkoot M dokumenttien määrä, N sanojen kokonaismäärä sekä $0 < \alpha, \beta < 1$. Dokumentit oletetaan generoituneen seuraavan satunnaisprosessin kautta:

1. Jokaista dokumenttia $i \in \{1,\ldots,M\}$ kohti generoidaan ykköseen summautuva vektori $\Theta_i = \Theta_{i,1}, \ldots, \Theta_{i,K} \sim \text{Dir}(\alpha)$, joka kuvaa dokumentin *i* aihejakauman: $\Theta_{i,k}$ = P(aihe *k* esiintyy dokumentissa *i*).
2. Jokaista aihetta $k \in \{1,\ldots,K\}$ kohti generoidaan vastaavasti vektori $\phi_k = \phi_{k,1}, \ldots, \phi_{k,N} \sim \text{Dir}(\beta)$, joka kuvaa aiheen *k* sanajakauman: $\phi_{k,j}$ = P(sana *j* esiintyy aiheessa *k*).
3. Kunkin dokumentin *i* jokaiseen sanapaikkaan $j \in \{1,\ldots,N_i\}$ generoidaan:
   3.1. aihe $z_{i,j} \sim \text{Multinom}(\Theta_i)$
   3.2. valitusta aiheesta sana $w_{i,j} \sim \text{Multinom}(\phi_{z_{i,j}})$.

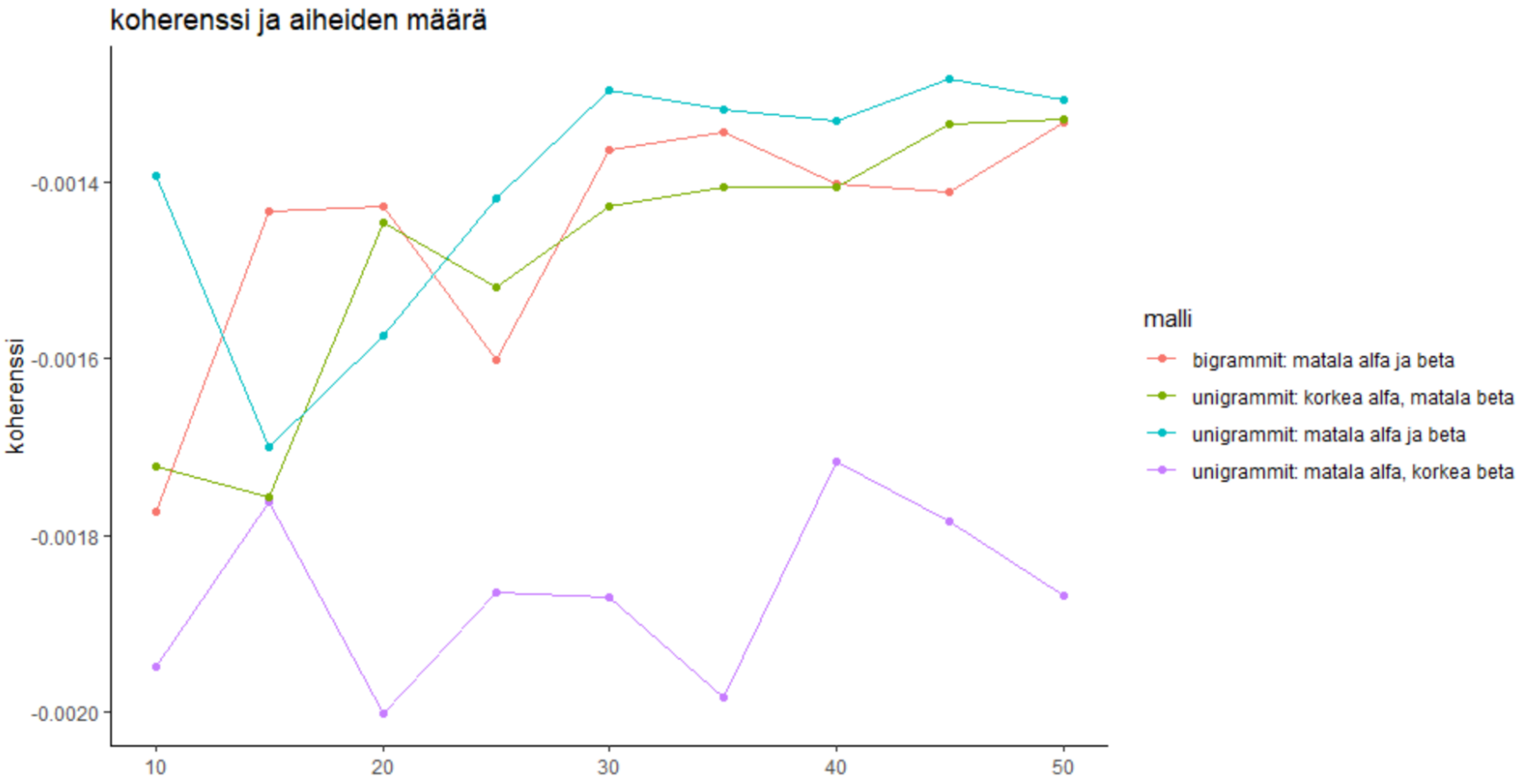
Tämän mallin parametreille Θ ja ϕ muodostetaan SU-estimaatit ns. VEM-iteraatiolla (Variational Expectation Maximation).

Koherenssi ja mallin arviointi

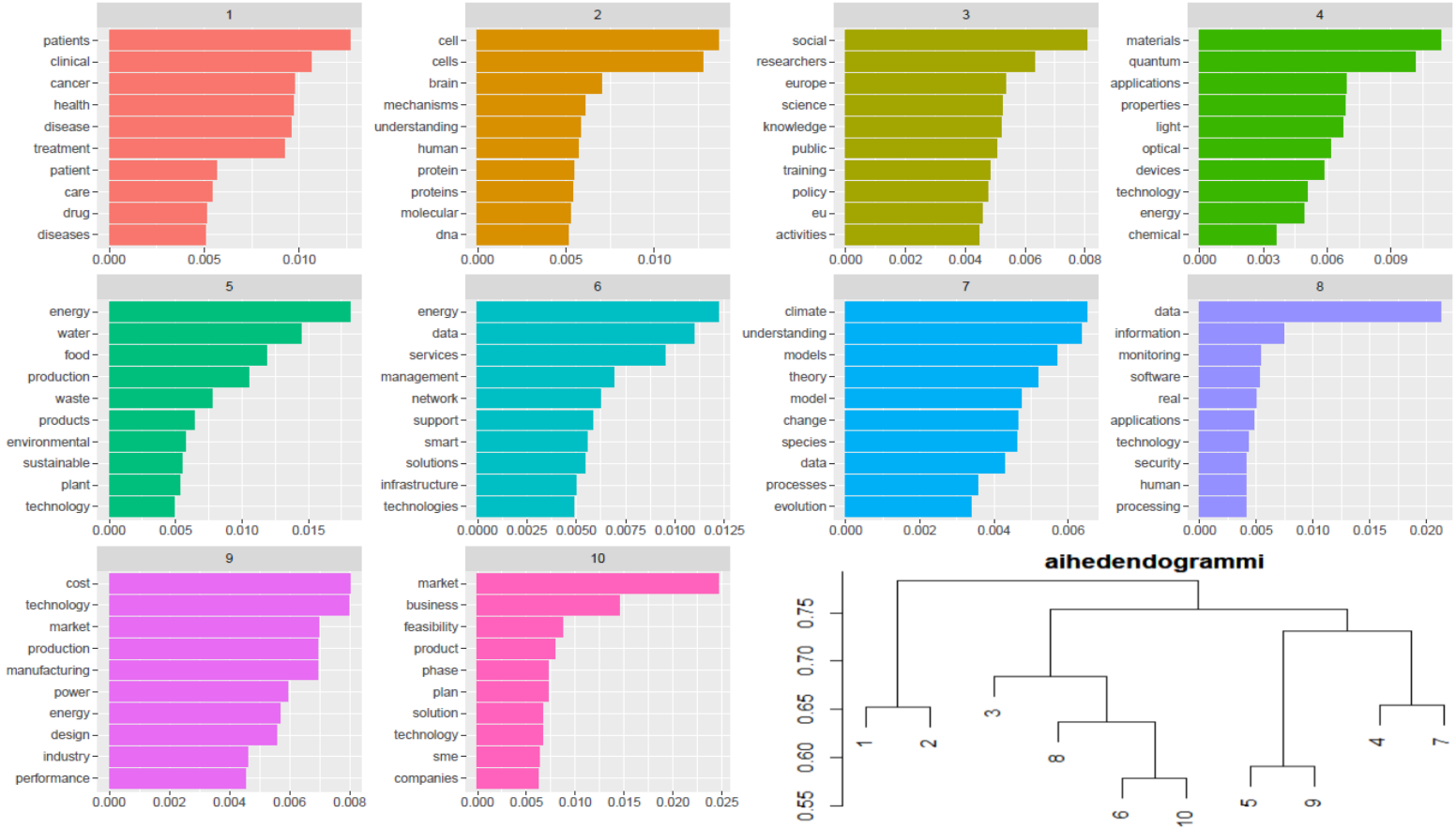
- Koherenssi on mikä tahansa kvantitatiivinen tapa arvioida aiheen sanojen yhteensopivuutta laskemalla, kuinka usein termit esiintyvät yhdessä verrattuna niiden yleisyyteen koko aineistossa.
- Laskemiseen monia tapoja, textmineR-paketissa aiheen {a, b, c, d} koherenssi lasketaan keskiarvona seuraavista erotuksista:
  - P(a | b) – p(b), P(a | c) – p(c), P(a | d) – p(d)
  - P(b | c) – p(c), P(b | d) – p(d)
  - P(c | d) – p(d)
- Suurempi koherenssi implikoi mallin parempaa sopivuutta. Arvot voivat olla positiivisia tai negatiivisia – rajat riippuvat sanojen kokonaismäärästä.
- Korkeakaan koherenssi ei takaa laadukasta mallia – on syytä käyttää myös inhimillistä järkeä aiheiden arvioinnissa.

Käytännön toteutus

- Aineisto tokenisoidaan eli hajotetaan yksittäisten sanojen tai n-grammien (esim. climate change, big data solutions) tasolle.
- Termilistasta poistetaan liian yleiset ja liian harvinaiset sanat, sekä kohdekielen ns. hukkasanat (and, no, for yms.). Lisäksi poistetaan manuaalisesti muita geneerisiä sanoja (esim. objective, approach, significantly, include).
- Termeistä luodaan document-term-matrix (DTM), joka kertoo, kuinka monta kertaa kukin termi esiintyy kussakin dokumentissa – käytännössä DTM on siis hyvin harva matriisi.

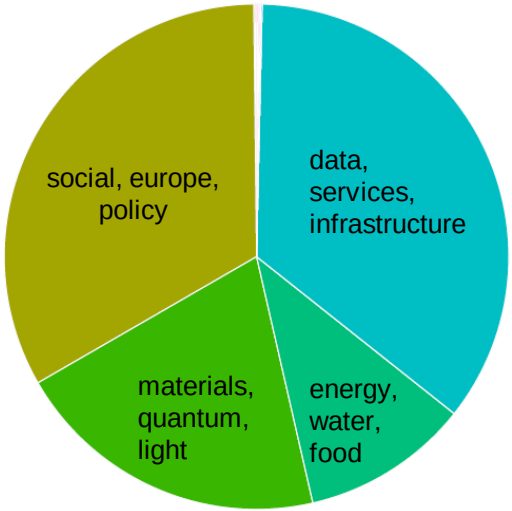

koherenssi ja aiheiden määrä
koherenssi
-0.0014
-0.0016
-0.0018
-0.0020
10
20
30
40
50
malli
bigrammit: matala alfa ja beta
unigrammit: korkea alfa, matala beta
unigrammit: matala alfa ja beta
unigrammit: matala alfa, korkea beta

Esimerkkimalli: 10 aihetta, topicmodels-paketin α-optimointi

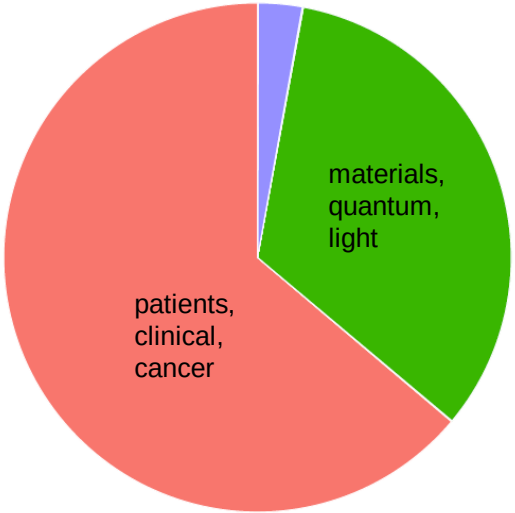


Dokumenttien aihejakaumia

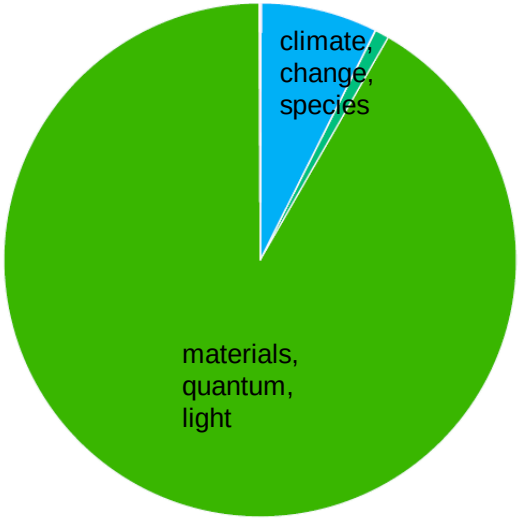
”An ambitious yet realistic roadmap to fusion electricity by…”



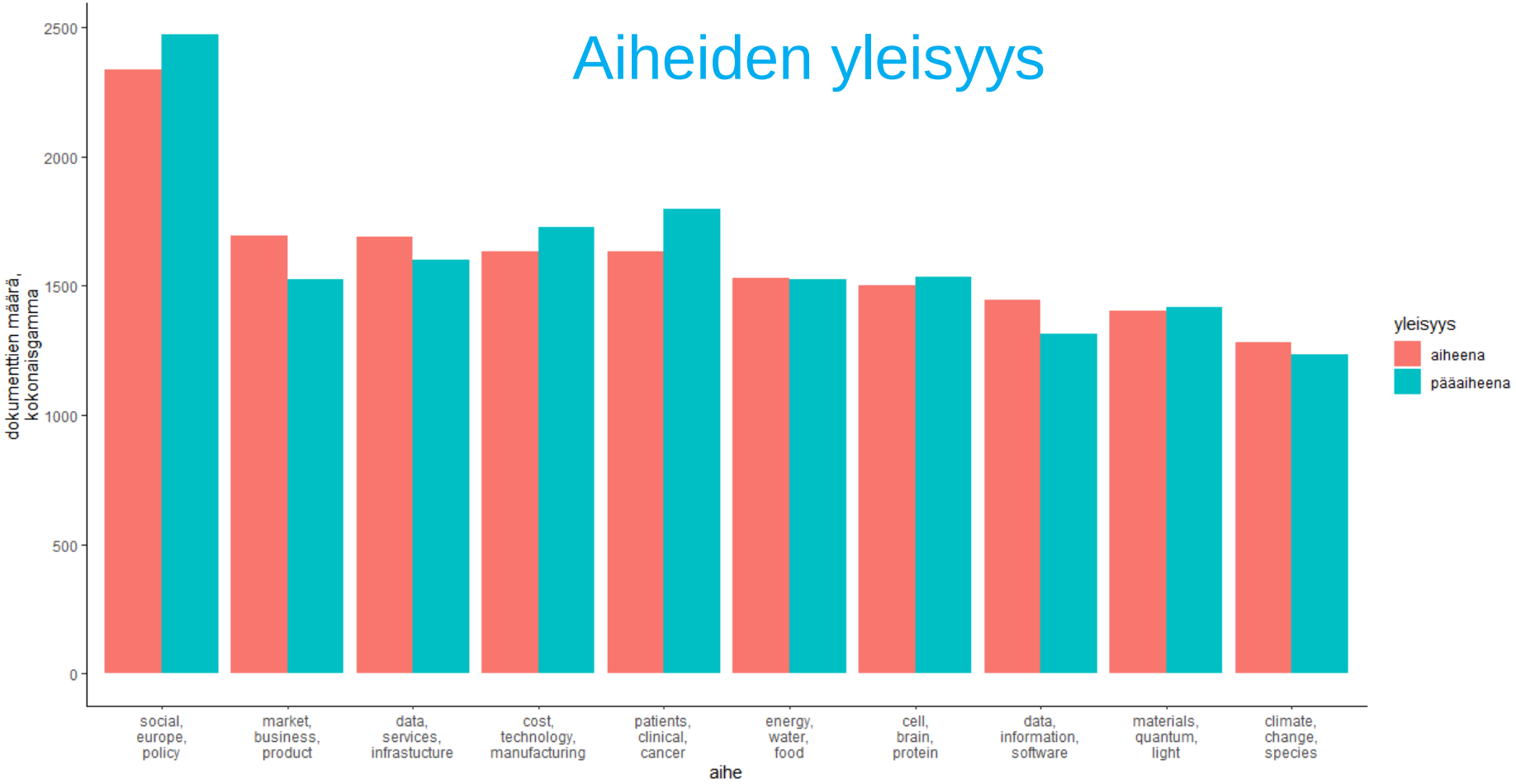
“HYPMED will develop and evaluate a novel PT-RF insert…”



”Molecular multifunctional switching materials attractive…”



Aiheiden yleisyys



Havaintoja

- Aineiston luokittelu geneerisiin tieteenaloihin (esim. luonnontieteet, IT-alat) onnistuu, ja dokumenttien mallinnetut aiheet vaikuttavat vastaavan todellisia aiheita.
- Kuitenkin näiden geneeristen alojen sisällä pienempien aiheiden erittely vaikeaa: fysiikka ei erotu kemiasta eikä tietoturva 5G-verkoista. Tämä voi johtua aineiston pienuudesta.
- Ideoita ja tavoitteita jatkoanalyysiin:
  - Tarkempien aiheiden erittely käyttäen mahdollisesti apuna myös tutkimusten otsikoita sekä sofistikoituneempia datansiistimiskeinoja ja erilaisia LDA-toteutuksia
  - Mallin laadun arviointi aineiston project ID –sarakkeen avulla
  - H2020-aineistolla treenatun mallin sovittaminen vanhempaan aineistoon
  - Tutkimustrendien tarkastelu aikasarja-analyysillä
  - Aiheiden läheisyyden tutkiminen: Θ-matriisin sarakkeiden korrelaatiot?

OPETUS- JA KULTTUURI-MINISTERIÖ

UNDERVISNINGS- OCH KULTUR-MINISTERIET

MINISTRY OF EDUCATION AND CULTURE