Tekstianalytiikka Helsingin kaupungilla

Mikko Niemi

Helsinki

Mitä tekstianalytiikka on

Tekstianalytiikka

- Luonnollisen kielen käsittely edistynein menetelmin (Natural Language Processing)
- Tekstien ryhmittely, luokittelu, tuottaminen, tunnesisältöjen havaitseminen
- Datapohjaiset menetelmät ja säännölliset kieliopit
- Datapohjaiset menetelmät voivat perustua neuroverkkoihin tai muihin tilastollisiin menetelmiin

- Yleensä tekstianalytiikan mallit eivät sisällä käsitteellistä ymmärrystä
- Mallit ovat yleisiä sääntöjä, eivät pureudu yksittäistapauksiin

Tekstianalytiikka Helsingin kaupungilla

Tekstianalytiikka kaupungilla

- Palautteiden ja kyselyiden jatkuva analyysi valmiin työkalun avulla
- Useita kokeiluja kokeilukiihdyttämössä (kaupungin sisäinen toiminto nopeille digitalisaatiokokeiluille)
- Pienkehitys digitalisaatioyksikössä

- Kuva: Kokeilukiihdyttämö - Tekoälyn hyödyntäminen Kerrokantasi vastausten luokittelussa ja analysoinnissa

Tekstianalytiikka ja arkkitehtuuri

- Tekstianalytiikan työkaluja käyttäen on koulutettu omia malleja (Etuma, SAS)
- Python-pohjaisia itse tehtyjä tekstianalytiikan malleja pari kappaletta
- Oma käyttöliittymä OpenAI:n generatiivisiin malleihin

- Nämä mallit eivät muodosta yhtenäistä kokonaisuutta

- Koneoppimisen ympäristöistä ja hallinnasta (MLops) on olemassa suunnitelma
- Tulevaisuudessa jäsentyneempi arkkitehtuuri mahdollinen myös NLP-malleille

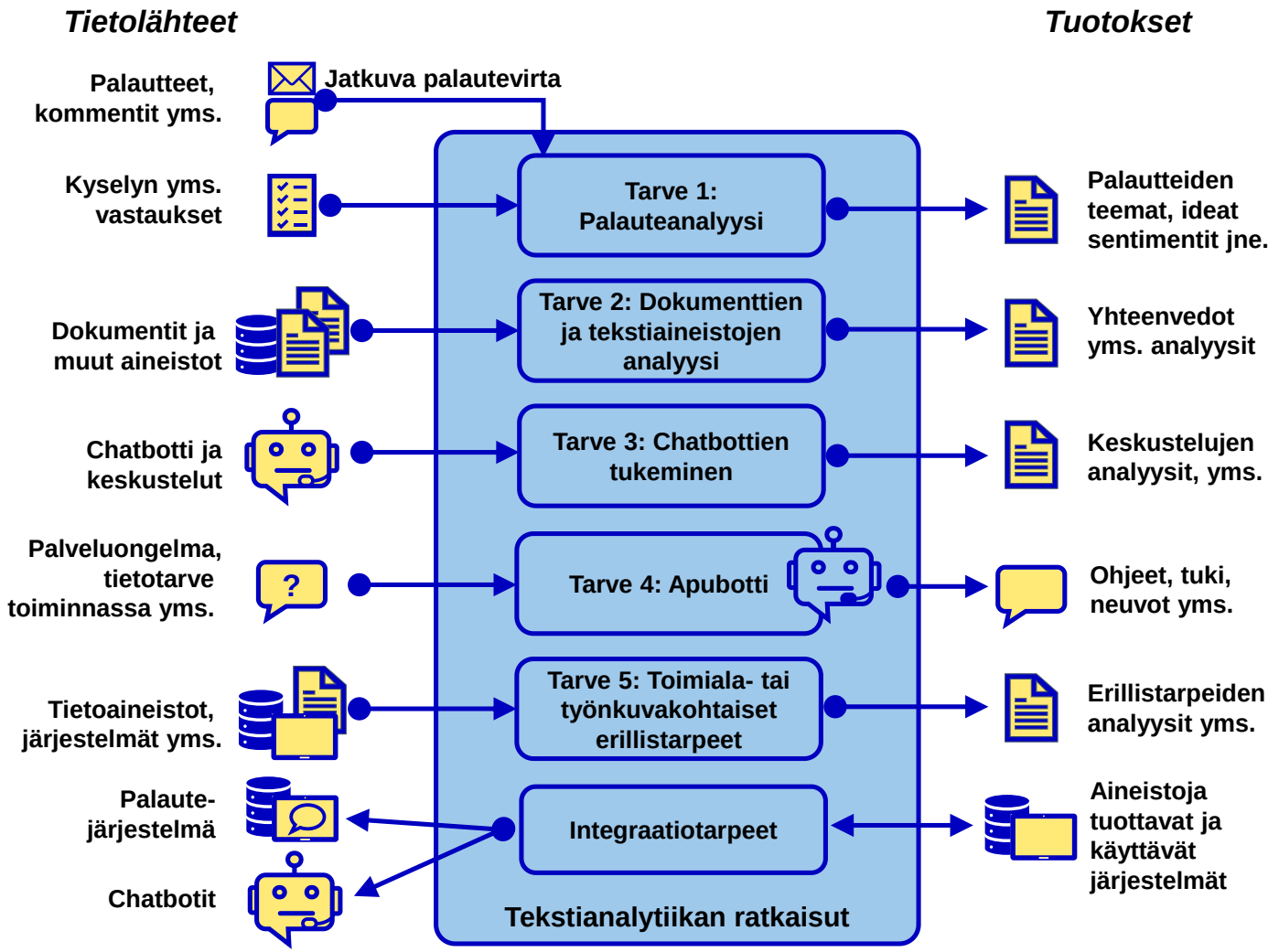
Kehittäminen

- Olemme selvittäneet tekstianalytiikan tarpeita kaupungin toimialoilla ja liikelaitoksissa vuonna 2023
- Keskityimme kartoittamaan tekstiaineistoja, analyysitarpeita ja saatavilla olevia välineitä

Kehitteillä ja tuotannossa olevia malleja

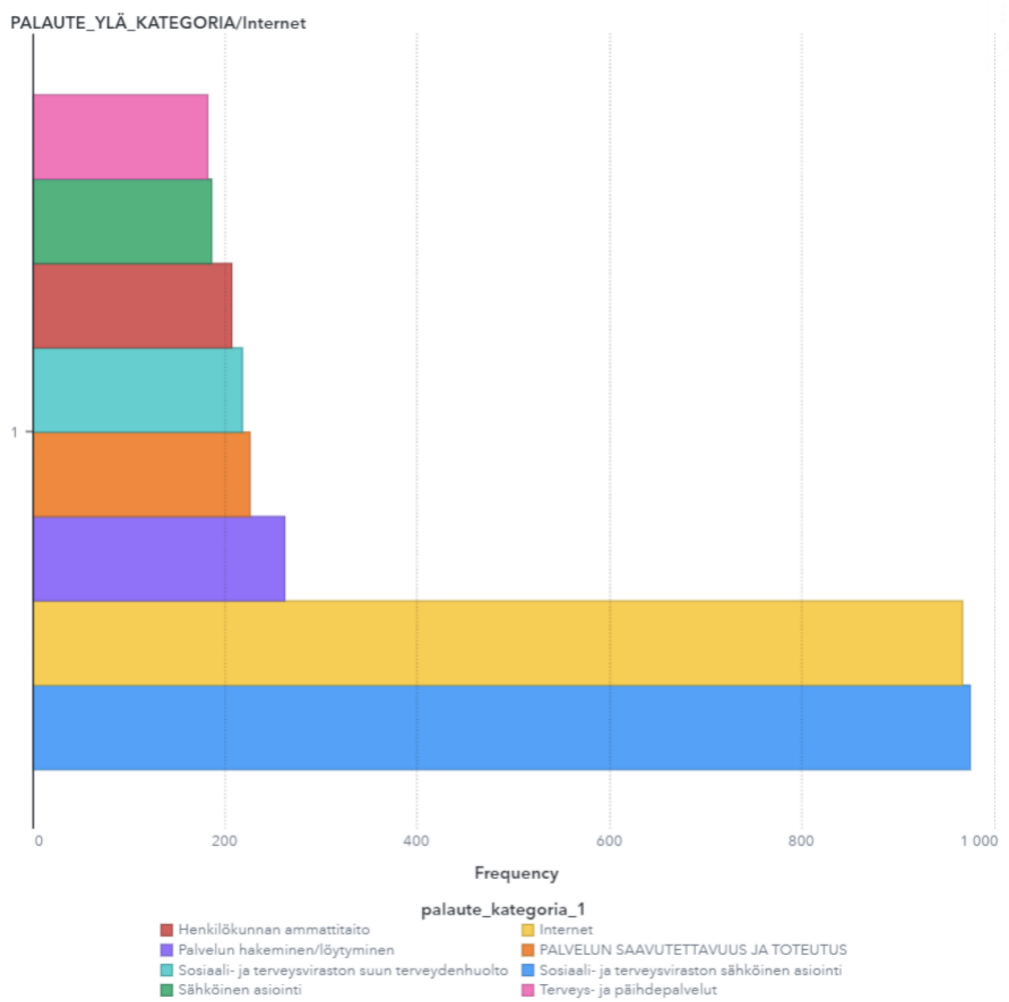
- Jatkuva palautteiden ja muun aineiston analysointi
- Generatiivisten mallien kokeilut: oma käyttöliittymä OpenAI:n malleille, Kokeilukiihdyttämön kokeilut
- Kaupunkilaispalautteiden analyysi

Tarpeiden yleiskuva



Tekstianalytiikan tarpeita

- Aineistojen analysointi
- Tekstin tuottaminen



- Kuva: Kokeilukiihdyttämö - Sote-toimialan avoimen palautteen tekstianalysointi ja luokittelu

Keskeisiä aineistoja

- Kaupunkilaispalautteet
- Palautekyselyt
- Toimintaan liittyvät asiakirjat kuten asemakaavat tai palotarkastuspöytäkirjat
- Operatiivisten järjestelmien tekstidata
- Päätöstekstit
- Talouden ja hankinnan asiakirjat

Metsä- ja luontoaiheisten palautteiden analyysi

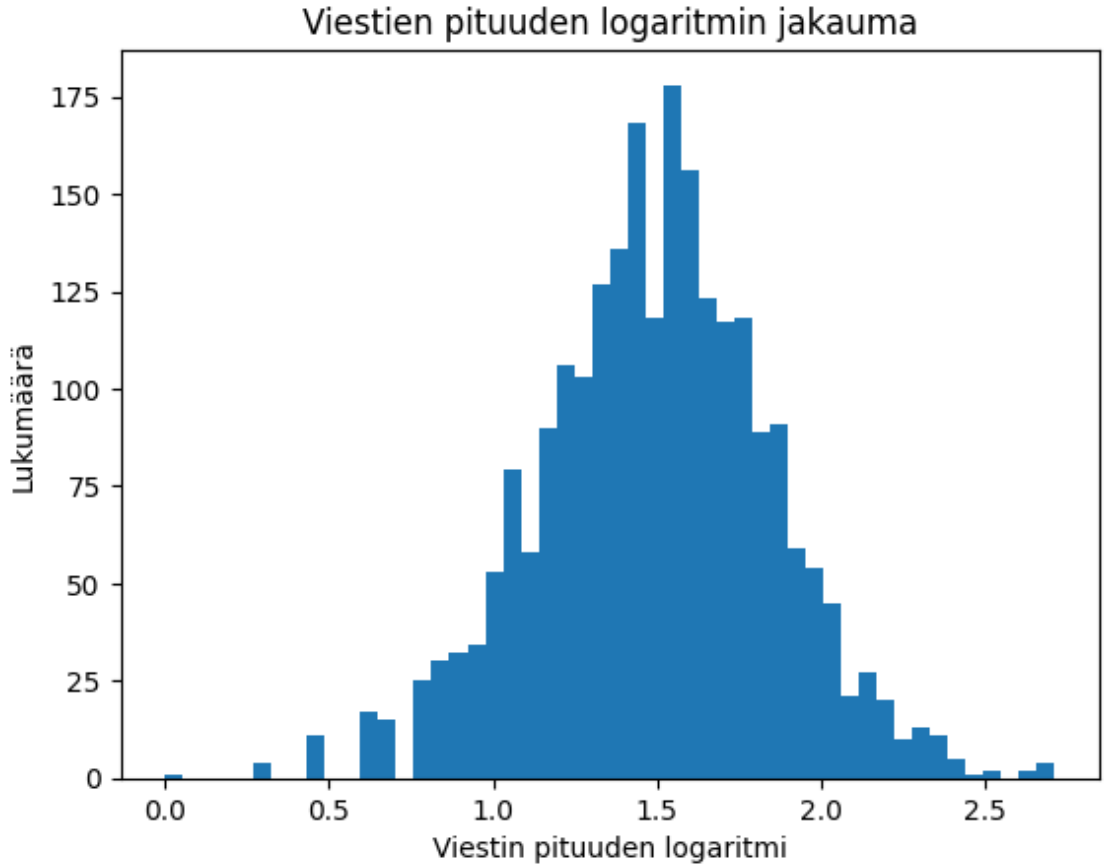
Tutkimuskysymys

- Miten lähiluonto tulee esille kaupunkilaispalautteissa? Mitä kaupunkilaiset halusivat sanoa metsiin liittyen?

- Aineistona vanhat anonymisoidut kaupunkilaispalautteet
- Kiinnostavaa on ainakin klusterointi ja aihemallinnus
- Analyysi on keskeneräinen

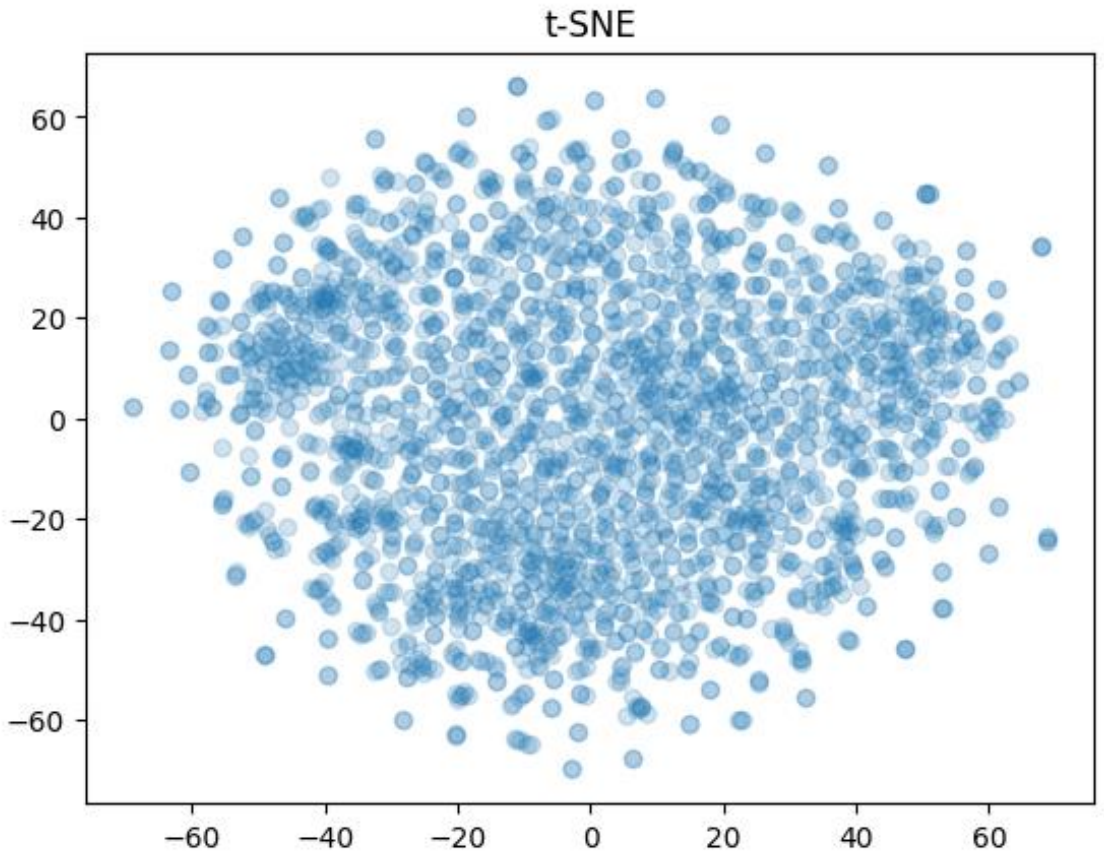
Aineisto

- Kaupunkilaispalautteita 2020-luvulta
- Kaupunkiympäristön toimiala ja muutama muu organisaatio
- Esikäsitelty viesti sisältää sanan ”metsä”, ”luonto” tai ”luontoarvo”
- Noin 2350 viestiä



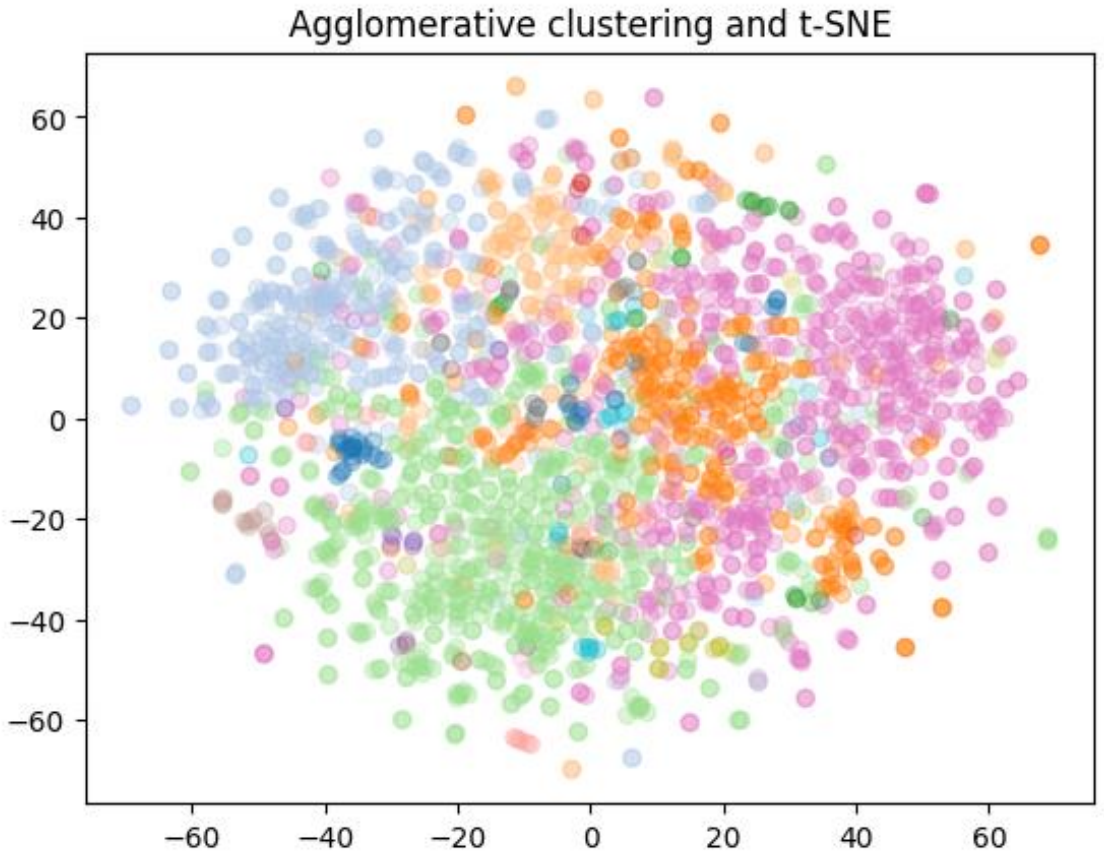
Tekstistä luvuiksi

- Turhien sanojen ja ilmausten poisto
- Perusmuotoistaminen
- Bag of Words (1,2,3-grams)
- TF-IDF-transformaatio
- Singulaariarvohajotelma (8200 ulottuvuudesta 800 ulottuvuuteen)
- Dimensionaliteetin vähentäminen t-SNE-menetelmällä (800 ulottuvuudesta kahteen)



Agglomeratiivinen klusterointi

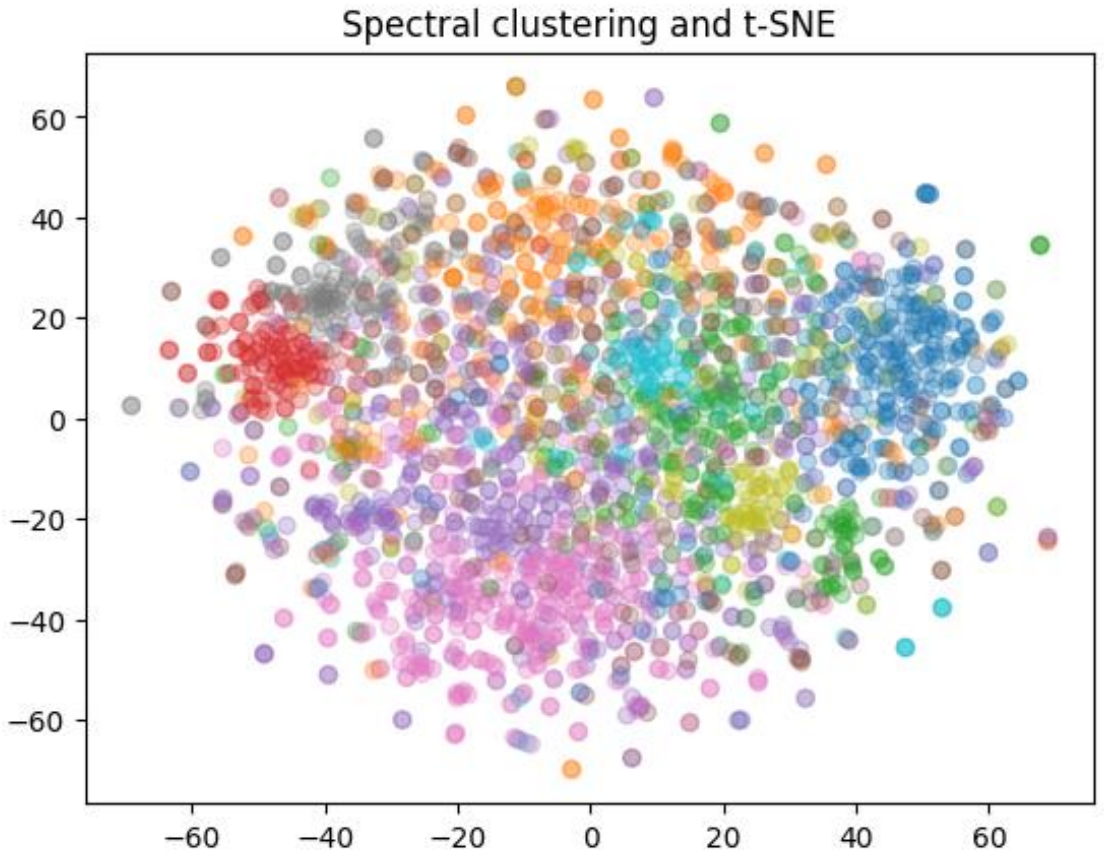
- 20 klusteria
- Kosinietäisyys
- Visualisoinnin pohjana t-SNE



Spektraalinen klusterointi

- 10 klusteria
- Kosinisimilariteetti
- Visualisoinnin pohjana t-SNE

- Sekä klusterointi että t-SNE laittavat samoja viestejä yhteen, mikä on lupaavaa
- Numeerisen siluetti-mittarin perusteella nämä klusterimallit eivät ole erityisen hyviä



Klusterien tulkintaa

- Sininen: puu, kävelytie, lenkkipolku
- Oranssi: ?
- Vihreä: ?
- Punainen: roskikset
- Violetti: ?
- Ruskea: ?
- Pinkki: ?
- Harmaa: roskia luonnossa
- Harmaanvihreä: ?
- Syaani: metsään jätetyt autot

Aihemallinnus

- Kokeilimme menetelmiä Latent Semantic Analysis ja Latent Dirichlet Allocation
- Sopiva aiheiden määrä arvioitava
- Seuraavaksi tutkimme ja kokeilemme erityisesti lyhyille teksteille soveltuvia aihemallinnuksen ja klusteroinnin menetelmiä

Yhteenveto

Yhteenveto

- Kaupunki on soveltanut tekstianalytiikkaa erityisesti palautteiden ja kyselyiden analysointiin
- Jatkuvia analyyseja ja yksittäisiä kokeiluja
- Kehitämme edelleen edellytyksiä tekstianalytiikan käyttöön