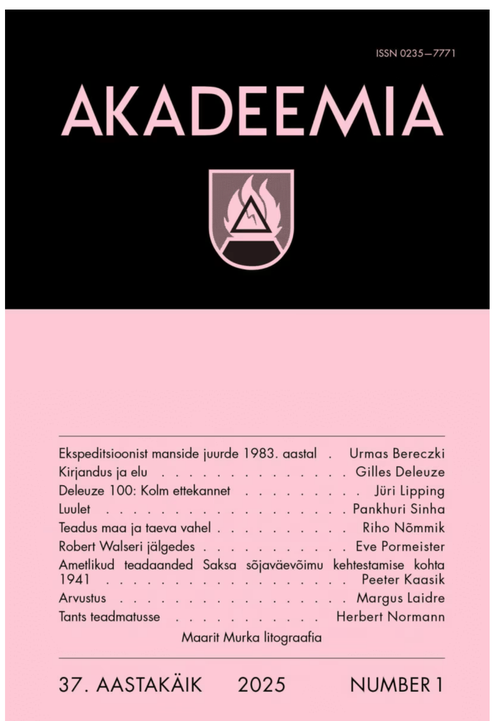
KULTUURIAJAKIRJA AKADEEMIA KAARDISTAMINE

Hendrik Matvejev, informaatika, BA III
Riki-Taavi Nurm, matemaatika, BA III

UNIVERSITY OF TARTU
1632

Sissejuhatus

Akadeemia on alates 1989. aasta aprillist igakuiselt ilmuv kultuuriajakiri, mille eesmärk on „vahendada eri teadusharude tänapäevast taset ja arengut“ [1]. Veebilehel Akad.ee on kätte saadavad kõigi numbrite sisukorrad tekstina ning nende sisu PDF-failidena. Veebilehe otsing toimub võtmesõnaliselt nende sisukordade peal, seega puudub semantilise otsingu võimalus. Selle probleemi lahendamiseks lõime veebilehe ja masinõppe andmetöötlusahela, mis muudab Akadeemia arhiivi semantiliselt otsitavaks ja interaktiivseks, kaardistades suurem osa ilmunud artiklitest. Veebilehele pääseb postril asuval QR-koodiga. Inspiratsiooniks oli sarnane projekt OpenSyllabus Galaxy [6].



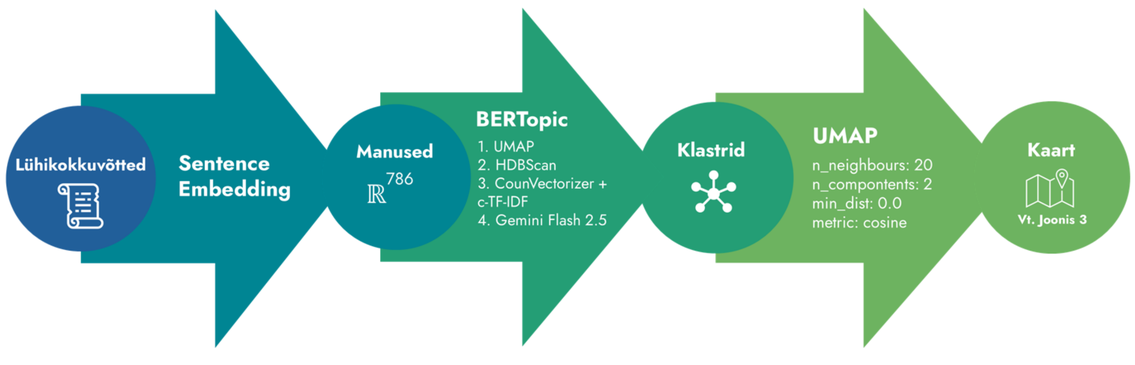
Joonis 1. Akadeemia 2025. aasta jaanuari numbri esikaas [1].

Andmete kogumine ja puhastamine

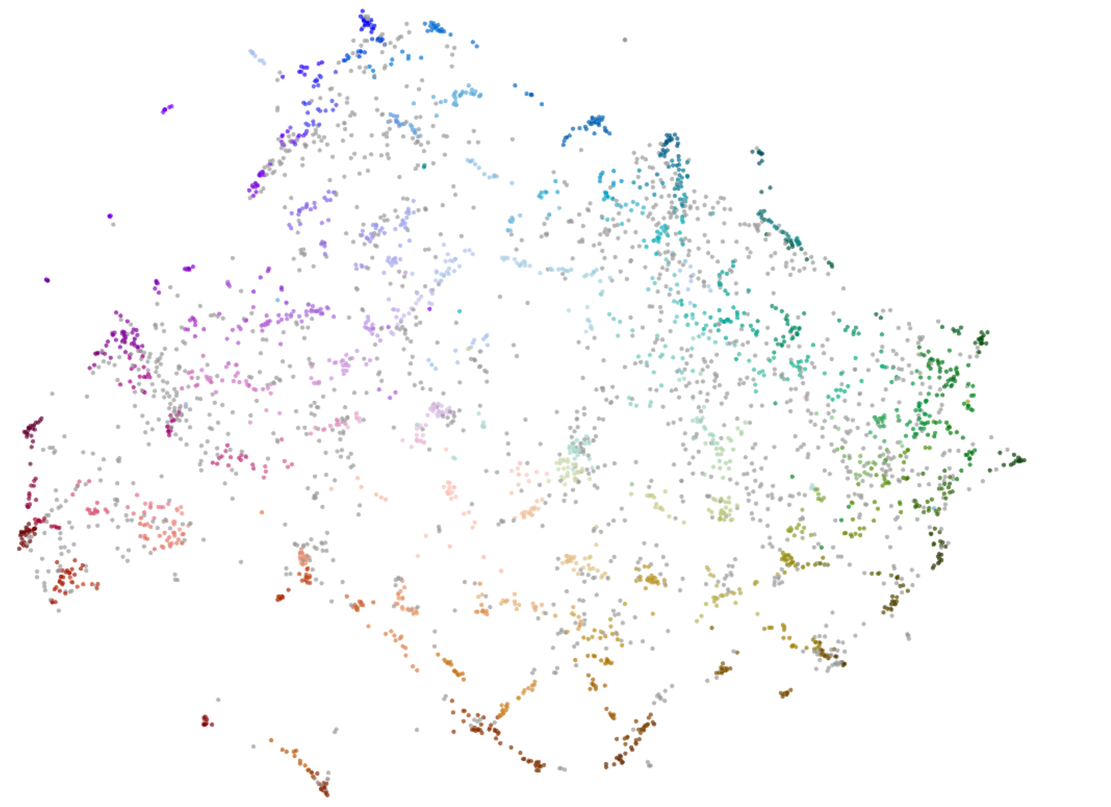
Ajakirja artiklitest ei eksisteerinud avalikku digitaalset andmestikku, mistõttu pidime selle ise looma. Andmete kogumiseks laadisime Akad.ee arhiiv veebilehelt alla numbrite HTML-sisukorrad, mis sisaldasid käsitsi sisestamisest tulenevaid ebakõlasid ja vormindusvigu. Seejärel laadisime alla vastavate numbrite PDF-failid Akadeemia veebilehelt ning Digari arhiivist [2], kusjuures mõne numbri skanneeringud olid puudu või ainult osaliselt olemas. Jagasime PDF-failid üksikuteks artikliteks sisukorra andmestiku järgi, kasutades optilist tekstituvastust (OCR), et leida õiged leheküljenumbrid. Viimaks eraldasime OCR-i abil artiklite inglisekeelsed sisukokkuvõtted ning viisime need vastavusse artiklitega. Terve protsessi jooksul puhastasime ja vajadusel kontrollisime andmeid käsitsi, mille tulemusena saime struktureeritud JSON-andmestiku, mis sisaldab artiklite metaandmeid ja inglisekeelseid lühikokkuvõtteid. Valminud andmestikus on olemas suurem osa artiklitest 1989. aasta aprillist kuni 2025. aasta veebruarini: kokku 5403 artiklit, millest 3946 on olemas inglisekeelne lühikokkuvõte.

Manuste loomine, klasterdamine ja kaardistamine

Kasutasime `all-mpnet-base-v2` tihedate lausevektorite mudelit [7], et luua inglisekeelsetest kokkuvõtetest 768-mõõtmelised vektormanused. Klastrite leidmiseks ning nende nimetamiseks kasutasime BERTopic [3] teeki, mis töötab neljas etapis: vähendab dimensioone UMAP algoritmiga; leiab klastrid HDBScaniga; leiab klastrite peamised märksõnad CountVectorizeriga ja c-TF-IDF-ga ning annab klastritele pealkirjad Gemini Flash 2.5ga. Viimaks kasutasime UMAP-algoritmi [5], et kujutada vektorid kahemõõtmelisele tasandile. Protsessi tulemusena saime kaardistatavad andmed: artiklite 2D koordinaadid ning kolm kihti klastreid vastavate pealkirjadega. Seda protsessi on kujutatud Joonisel 2.



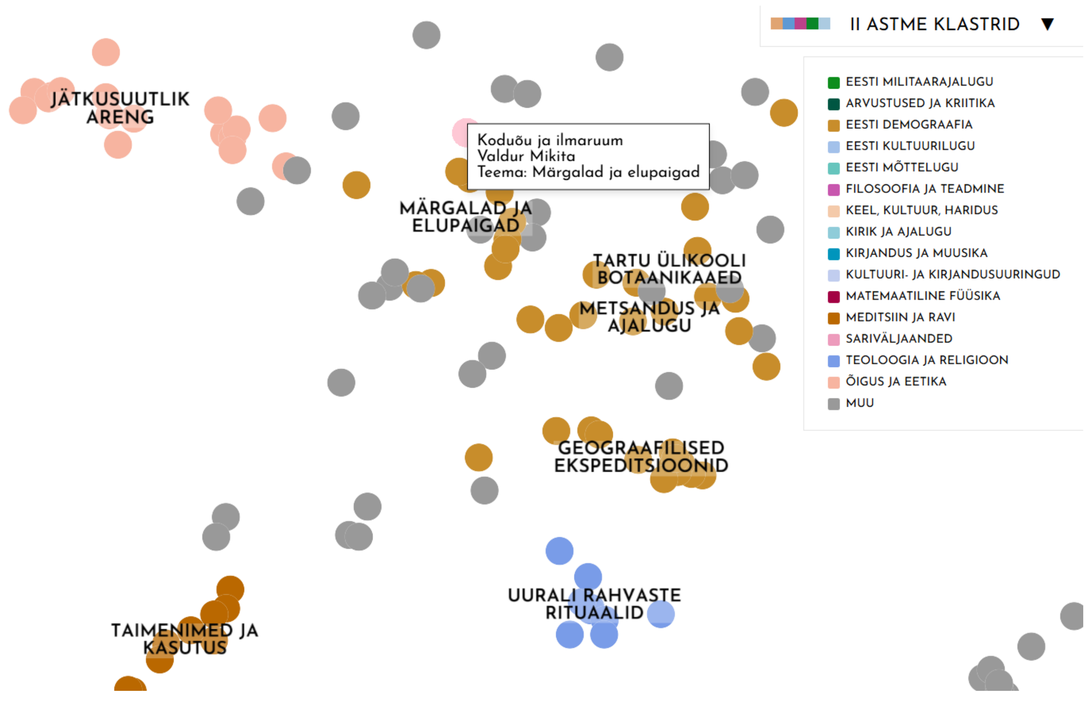
Joonis 2. Semantiliste klastritega kaardi loomise ahel.



Joonis 3. Akadeemia artiklite kaart värvitud madalaima ehk III astme klastrite järgi.

Arhiivileht ja interaktiivne kaart

Veebilehe disaini inspiratsiooniks on 2025. aasta esikaane disain (vt. Joonis 1). Veebirakendus põhineb Flaskil. Kasutajaliides rakendab 2D-projektsiooni ja loodud klastrite kuvamiseks tugevalt kohandatud DataMapPlot teeki [4]. Kasutajad saavad uurida arhiivi kaardivaate kaudu (vt. Joonis 3 ja 4), teha vabas vormis semantilisi otsinguid ja saada koosinussarnasuse skooridel põhinevaid lähimate–ehk semantiliselt sarnaseimate–artiklite soovitusi. Meie poolt loodud semantiline otsing ja kaart lihtsustavad Akadeemia artiklite leidmist ning võimaldavad näha nendevahelisi, varasemalt märkamata jäänud seosed.



Joonis 4. Klastrid ja nende sildid.

Viited

[1] Akadeemia. *Akadeemia*. 2026. URL: https://www.akad.ee/ (vaadatud 26.05.2026).
[2] Eesti Rahvusraamatukogu. *DIGAR*. 2005. URL: https://www.digar.ee/arhiiv (vaadatud 26.05.2026).
[3] Maarten P. Grootendorst. *BERTopic*. 2024. URL: https://maartengr.github.io/BERTopic/index.html (vaadatud 25.05.2026).
[4] Leland McInnes. *DataMapPlot*. 2023. URL: https://datamapplot.readthedocs.io/en/latest/ (vaadatud 26.05.2026).
[5] Leland McInnes, John Healy ja James Melville. *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*. September 2020. DOI: 10.48550/arXiv.1802.03426. (Vaadatud 25.05.2026).
[6] Open Syllabus. *Open Syllabus: Galaxy*. 2021. URL: https://galaxy.opensyllabus.org/ (vaadatud 26.05.2026).
[7] UKP Lab ja Hugging Face. *sentence-transformers/all-mpnet-base-v2*. Jaanuar 2024. URL: https://huggingface.co/sentence-transformers/all-mpnet-base-v2 (vaadatud 26.05.2026).