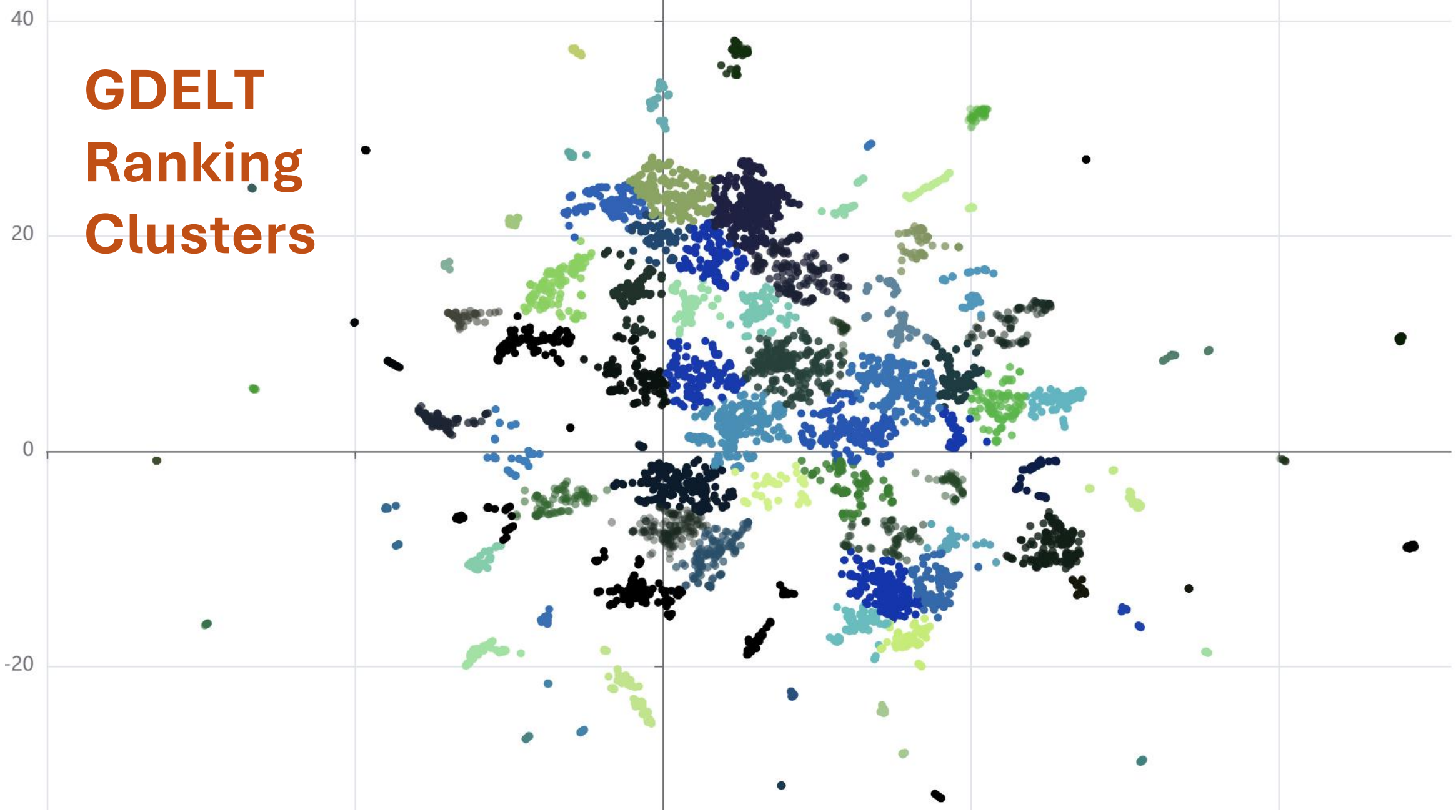
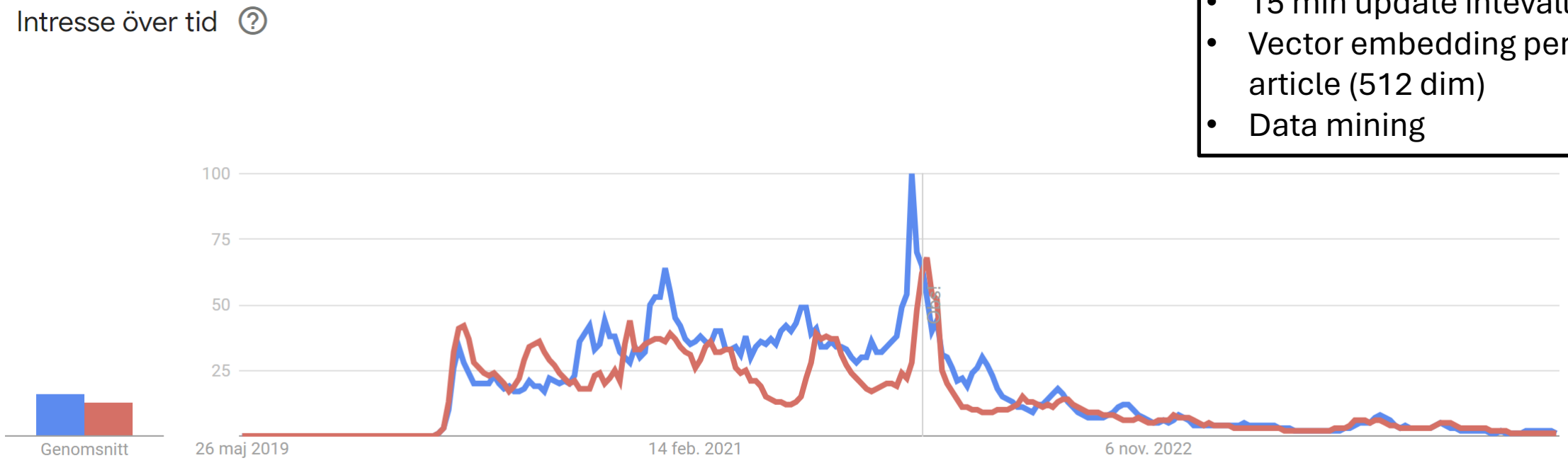

GDELT
Ranking
Clusters
40
20
0
-20

Choose a date for our demo

	manual labelled:
• gsg_2020-11-16_2020-11-19_20_perc.parquet.br	Covid
• gsg_2022-02-23_2022-02-26_20_perc.parquet.br	Ukrainian War
• gsg_2024-04-13_2024-04-16_20_perc.parquet.br	Iran- Israel
• gsg_2024-04-30_2024-05-03_20_perc.parquet.br	Palestian protest

Trending news, GDELT data, motiviation intruduction.

- Trending news
- GDELT Data
- (Global Database of Event Language and Tone)
- 15 min update intevall
- Vector embedding per article (512 dim)
- Data mining



Covid: UK(blue), USA(red)

Assumptions

- General base

  The news articles with similar topics will be clustered together by the USE embedding vector, and the denser (more articles in it) the cluster the trending the topic (of those articles) is at a certain timestamp.

- Topic

  For each group of articles (cluster) there exist a topic that describes it perfectly

- Dimensionality reduction

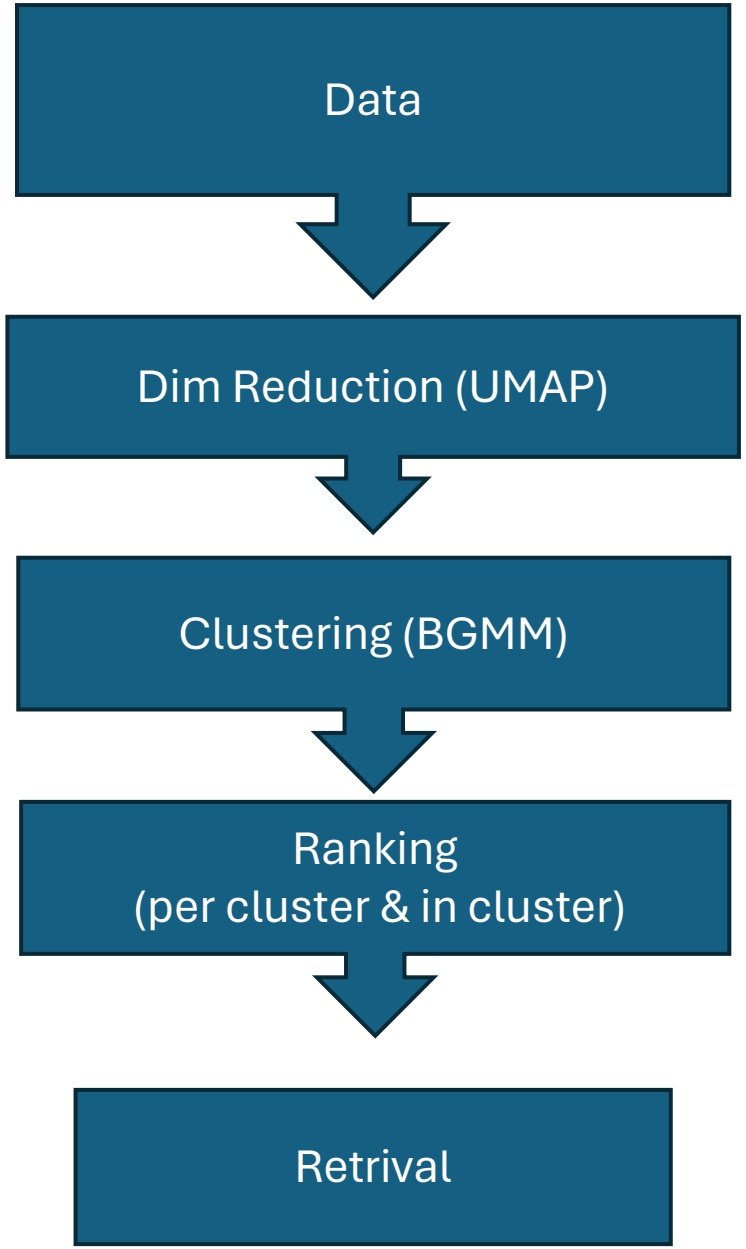
  All articles within the same topic will be pushed together close enough to be clustered into a single cluser

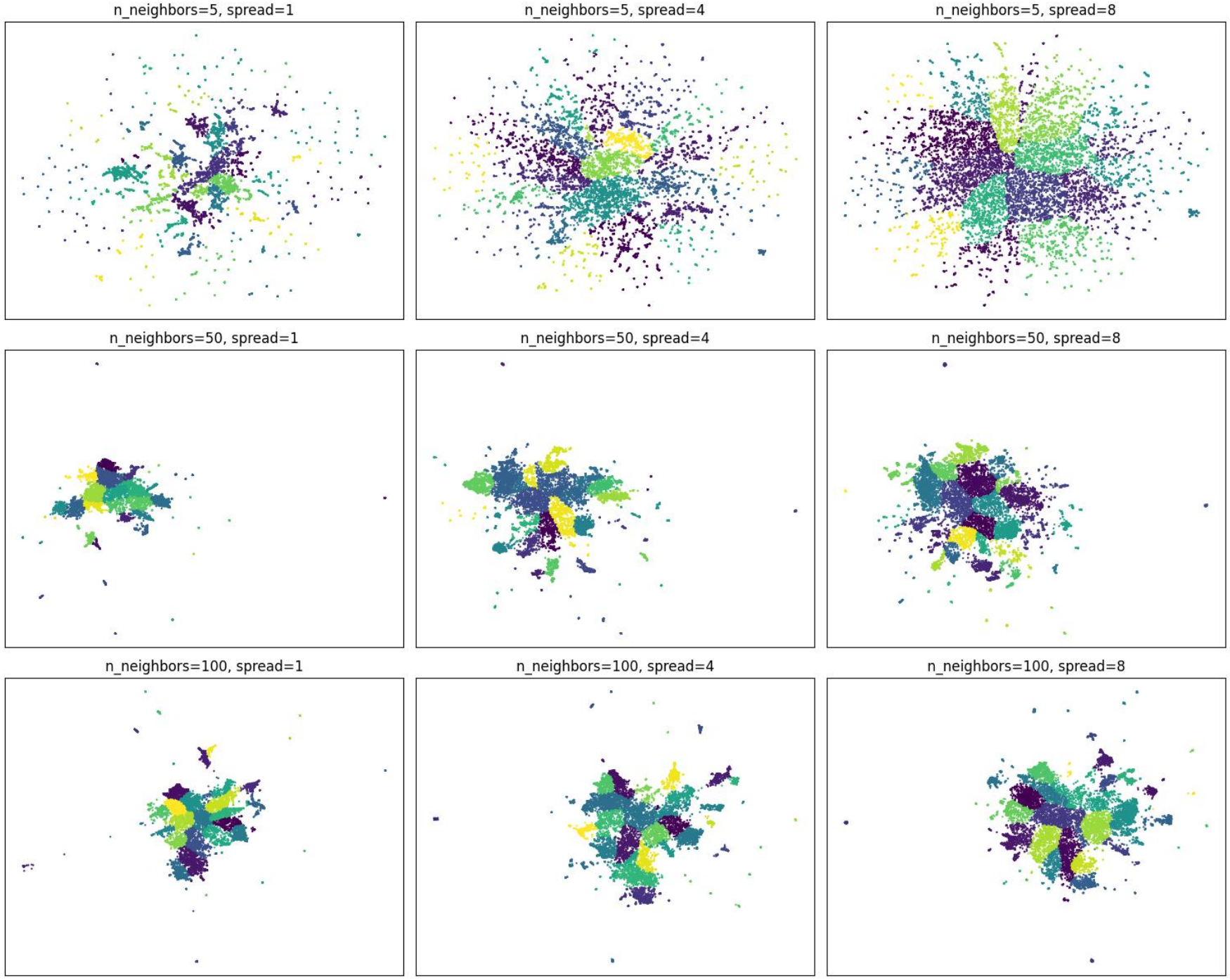
  All articles with different topics will be separated far enough, to build multiple clusters

- Clustering

  The clustering algorithm will perfectly cluster all articles within the same topic

Pipeline


Data
Dim Reduction (UMAP)
Clustering (BGMM)
Ranking
(per cluster & in cluster)
Retrival



Tuning Parameters

- n_neighbors

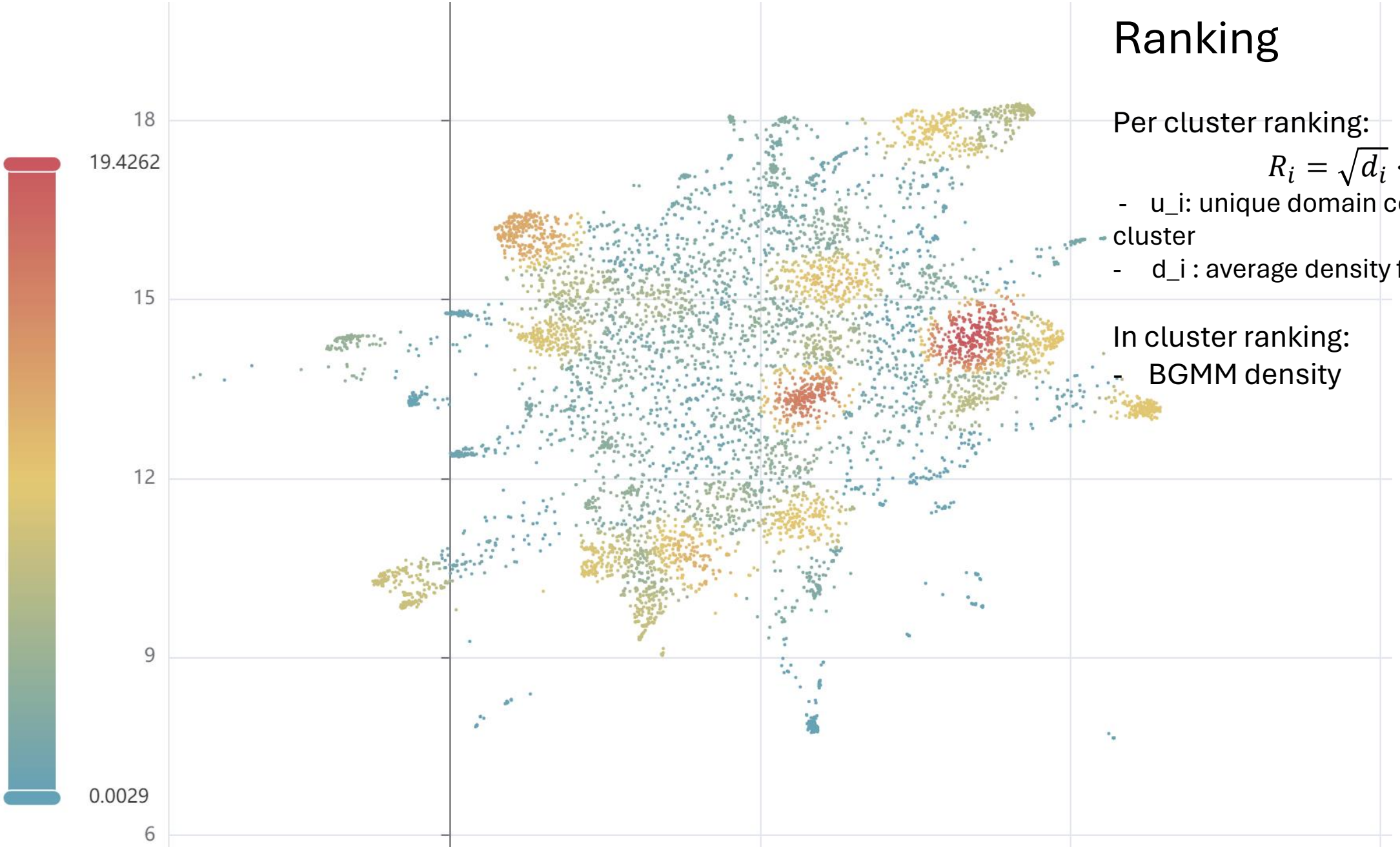
Local and global structure

- Spread

Clustering of datapoints

- Trade-off

50–100, 4



Ranking

Per cluster ranking:

$$R_i = \sqrt{d_i} \cdot u_i$$

- u_i: unique domain counter per cluster
- d_i : average density for cluster i

In cluster ranking:

- BGMM density

DEMO