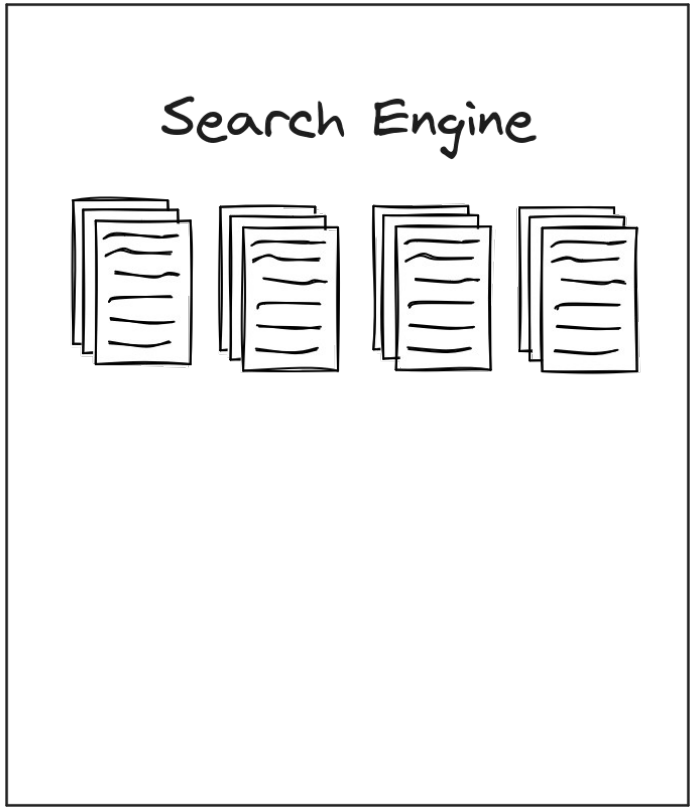
Web Clustering Algorithms for Selective Search

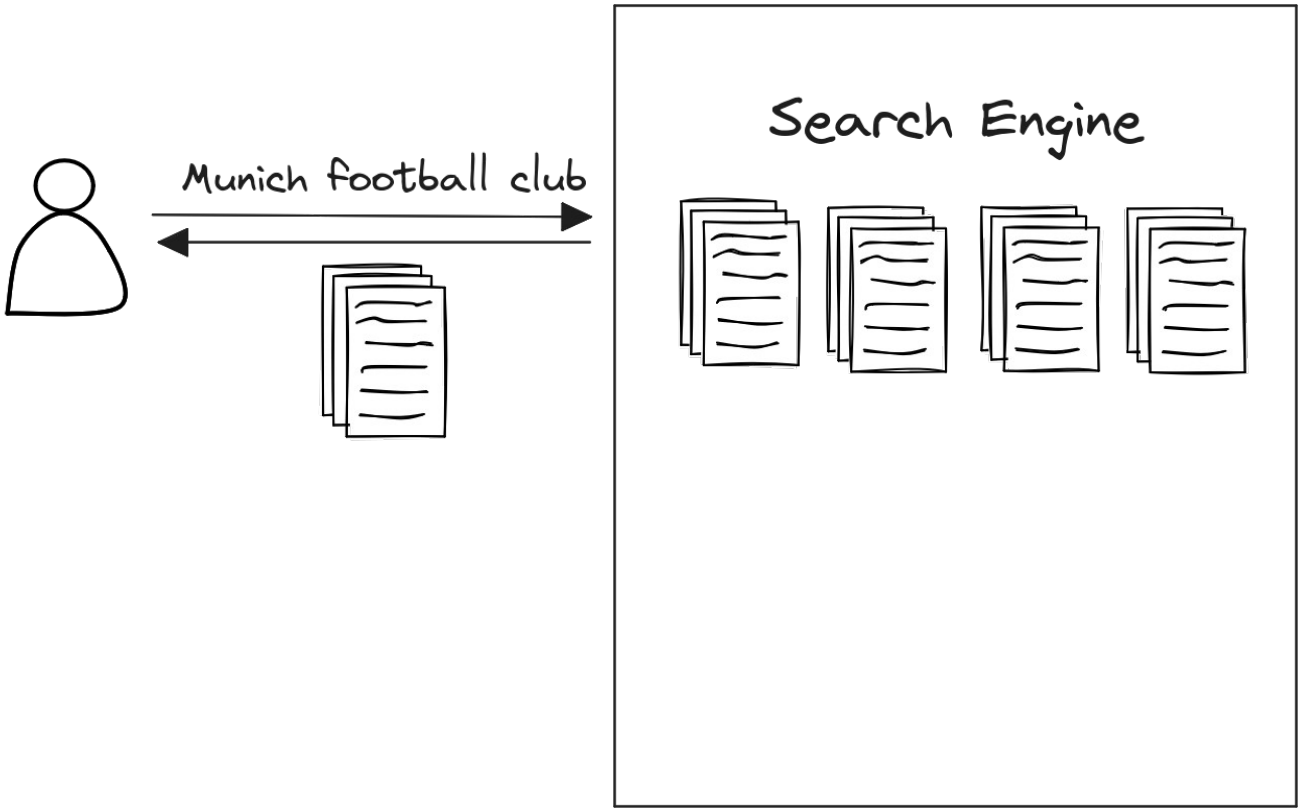
Gijs Hendriksen
Djoerd Hiemstra
Arjen P. de Vries

Radboud University

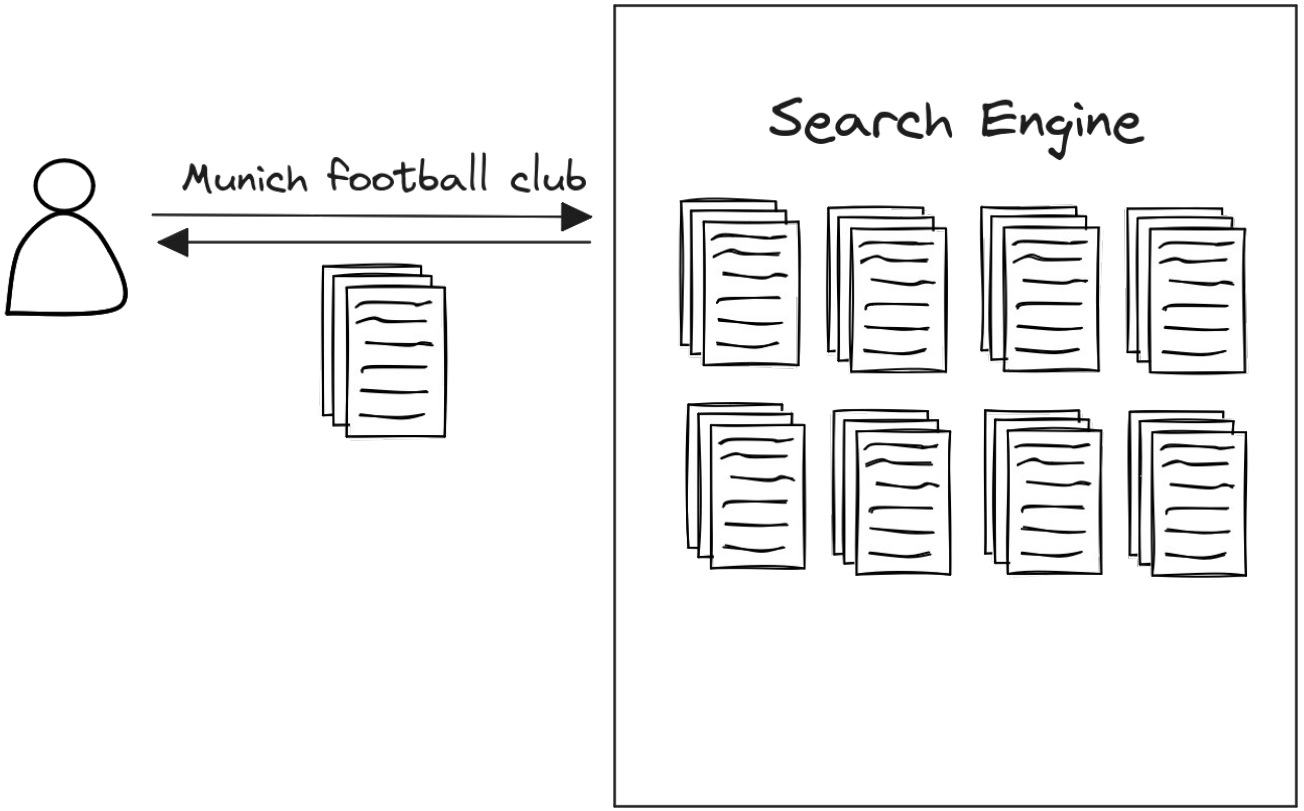
SCALABILITY OF A SEARCH ENGINE



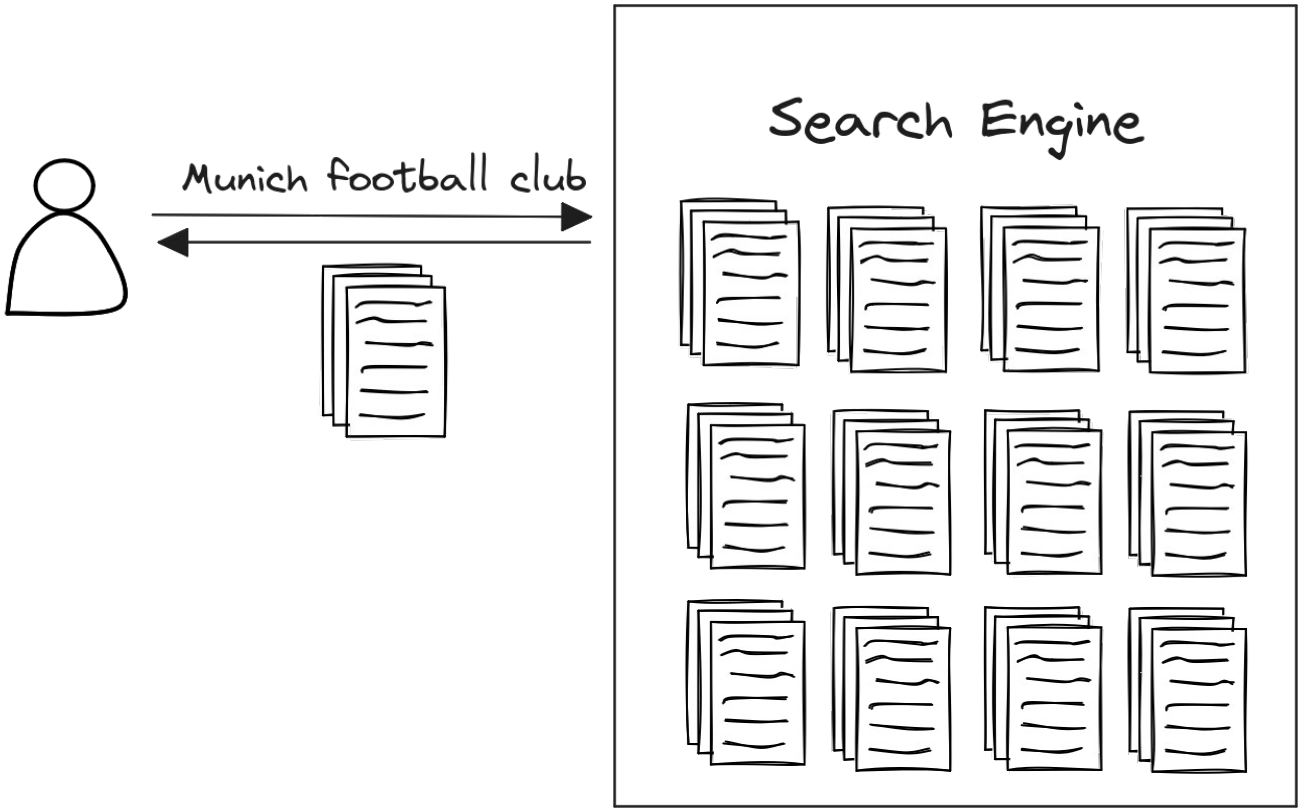
SCALABILITY OF A SEARCH ENGINE



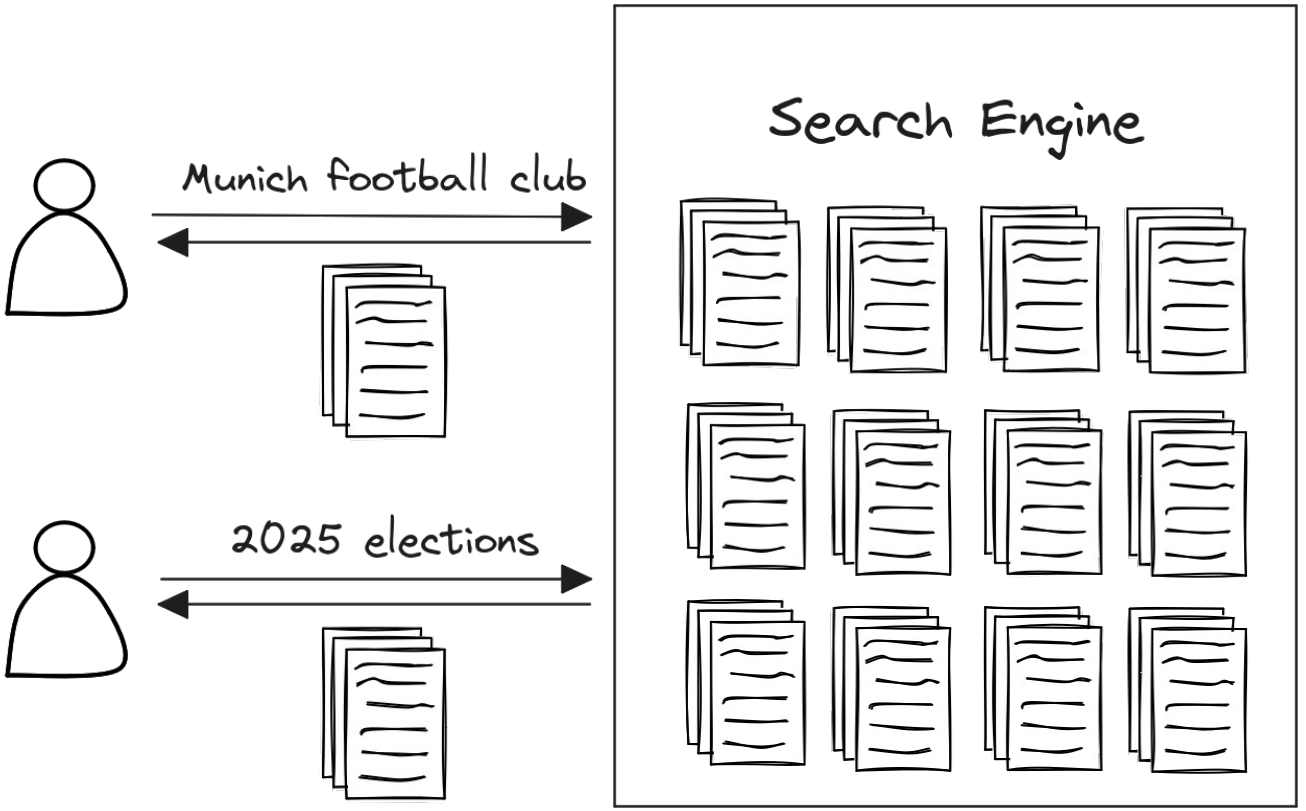
SCALABILITY OF A SEARCH ENGINE



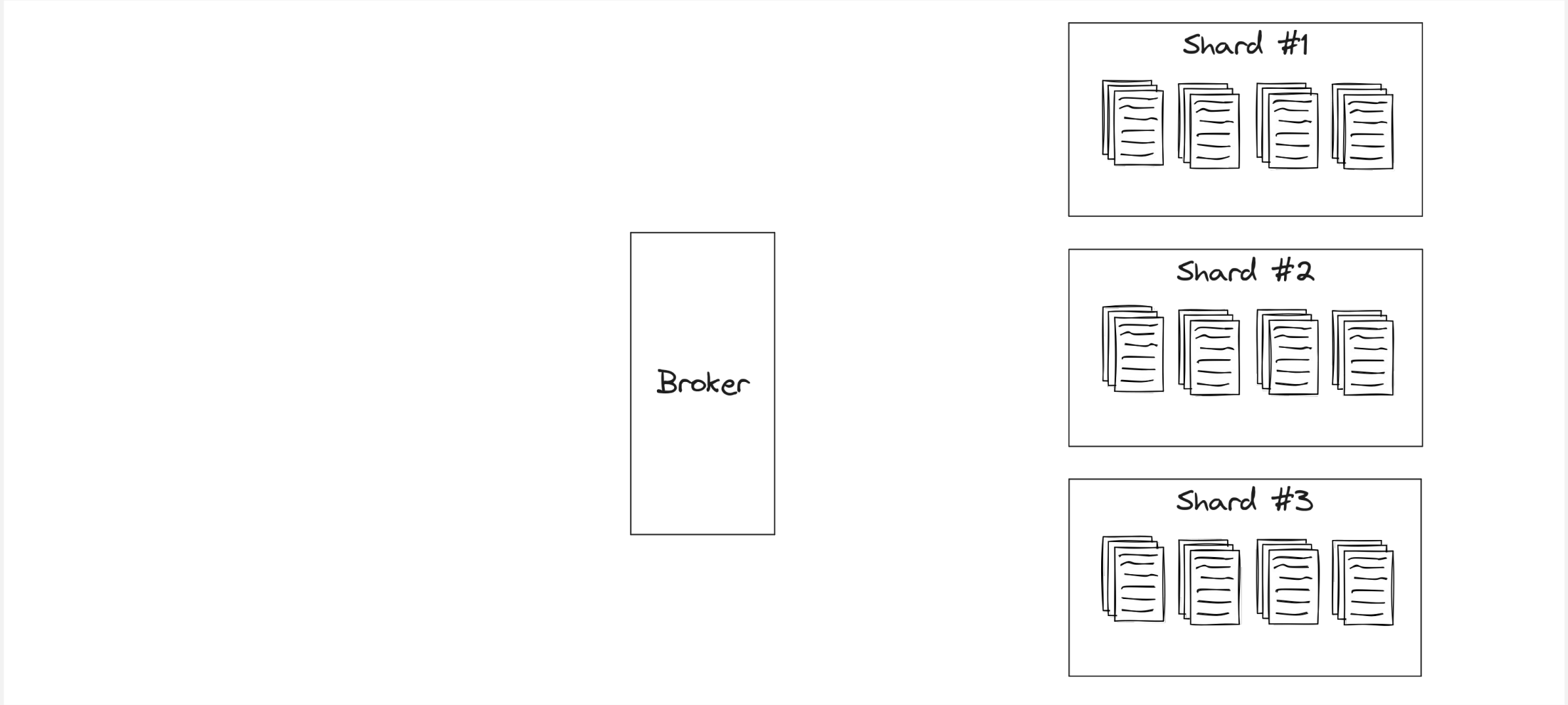
SCALABILITY OF A SEARCH ENGINE



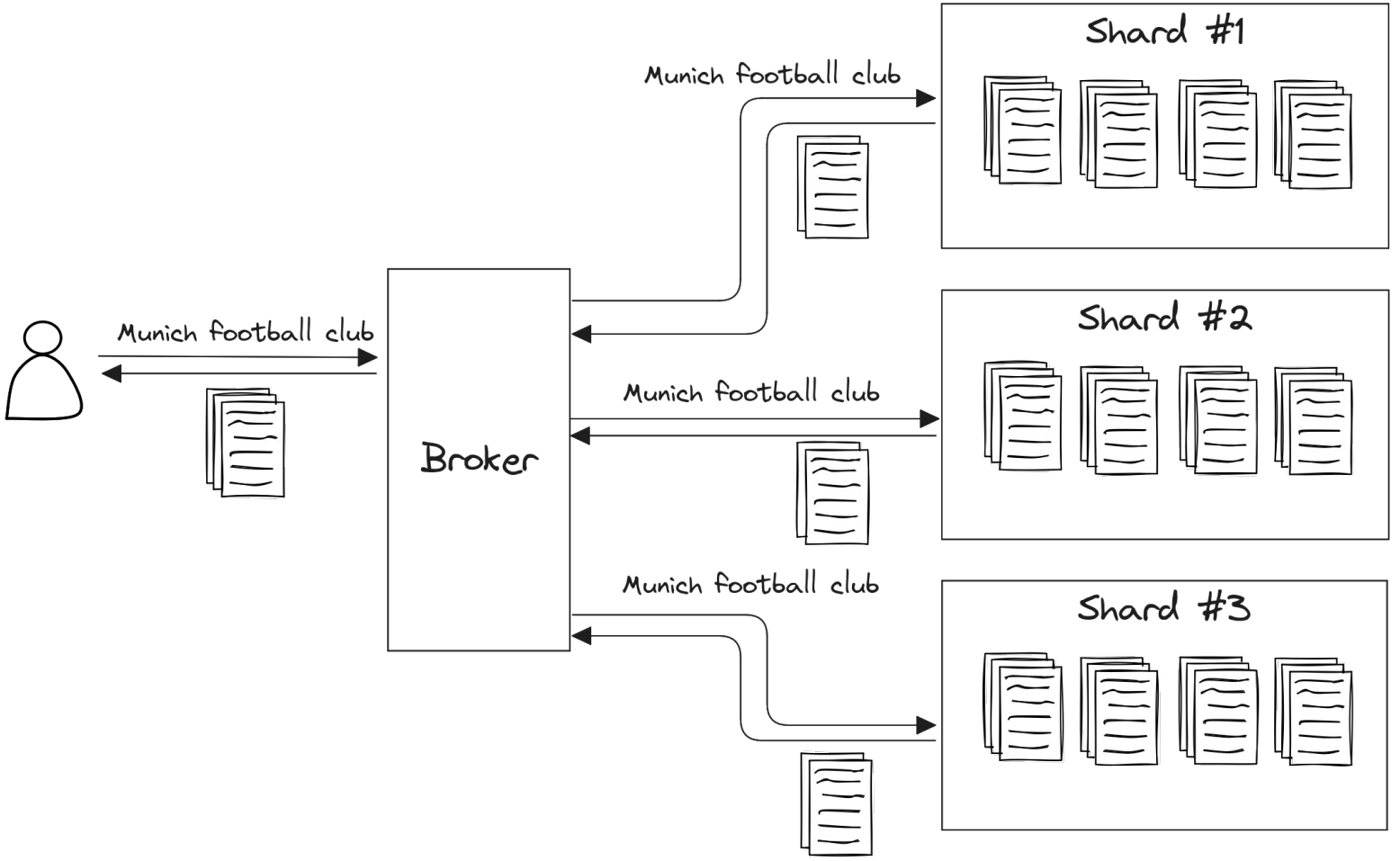
SCALABILITY OF A SEARCH ENGINE



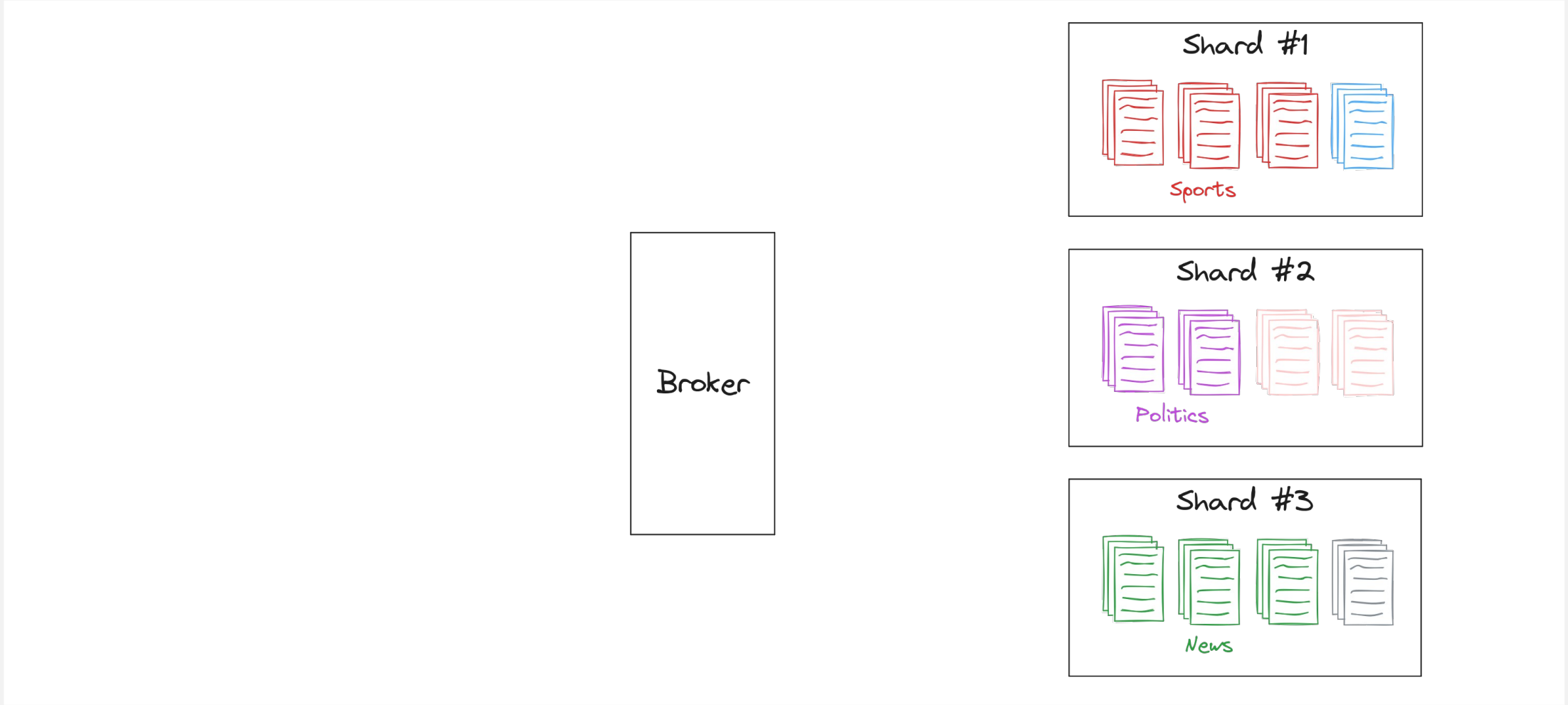
DISTRIBUTED SEARCH



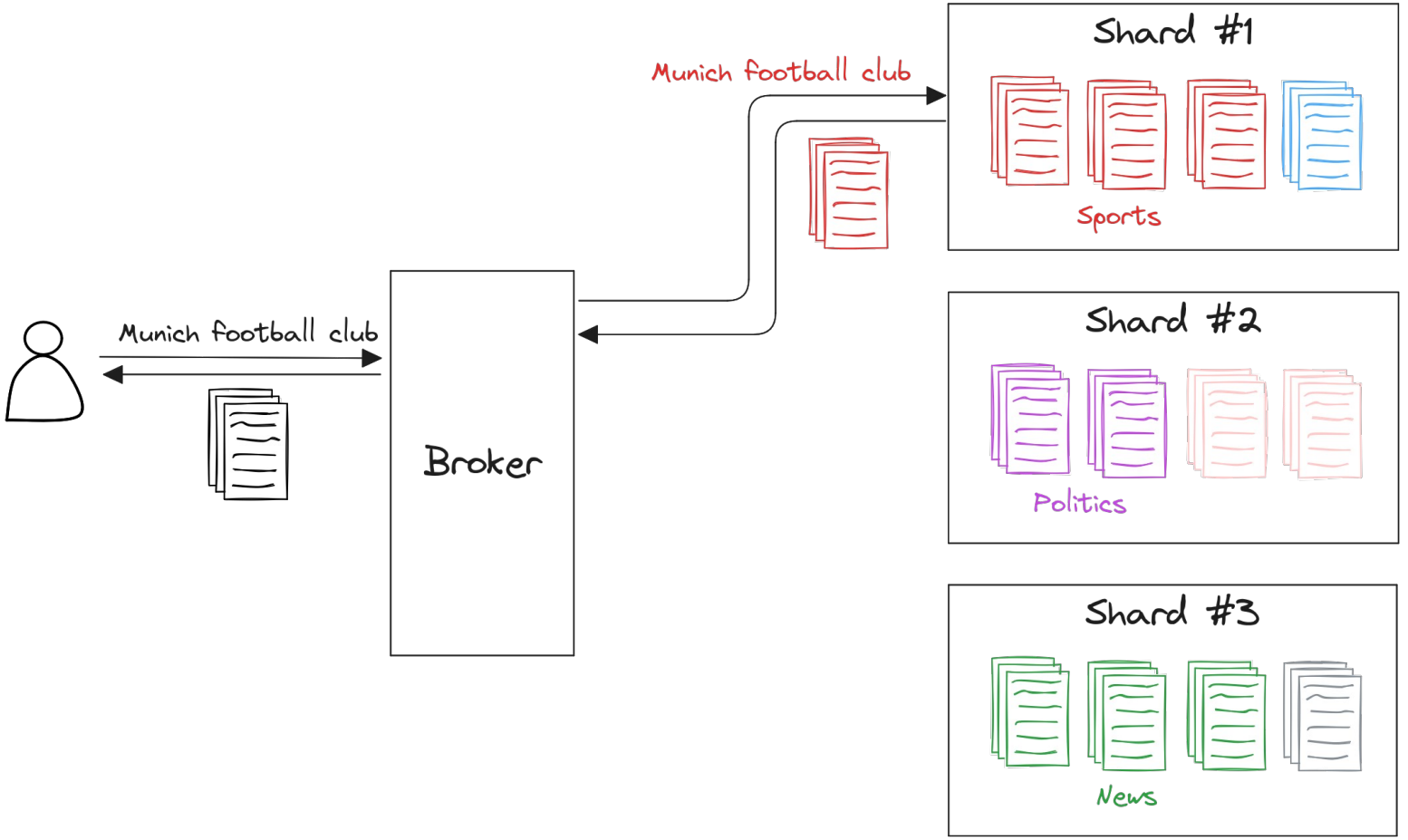
DISTRIBUTED SEARCH



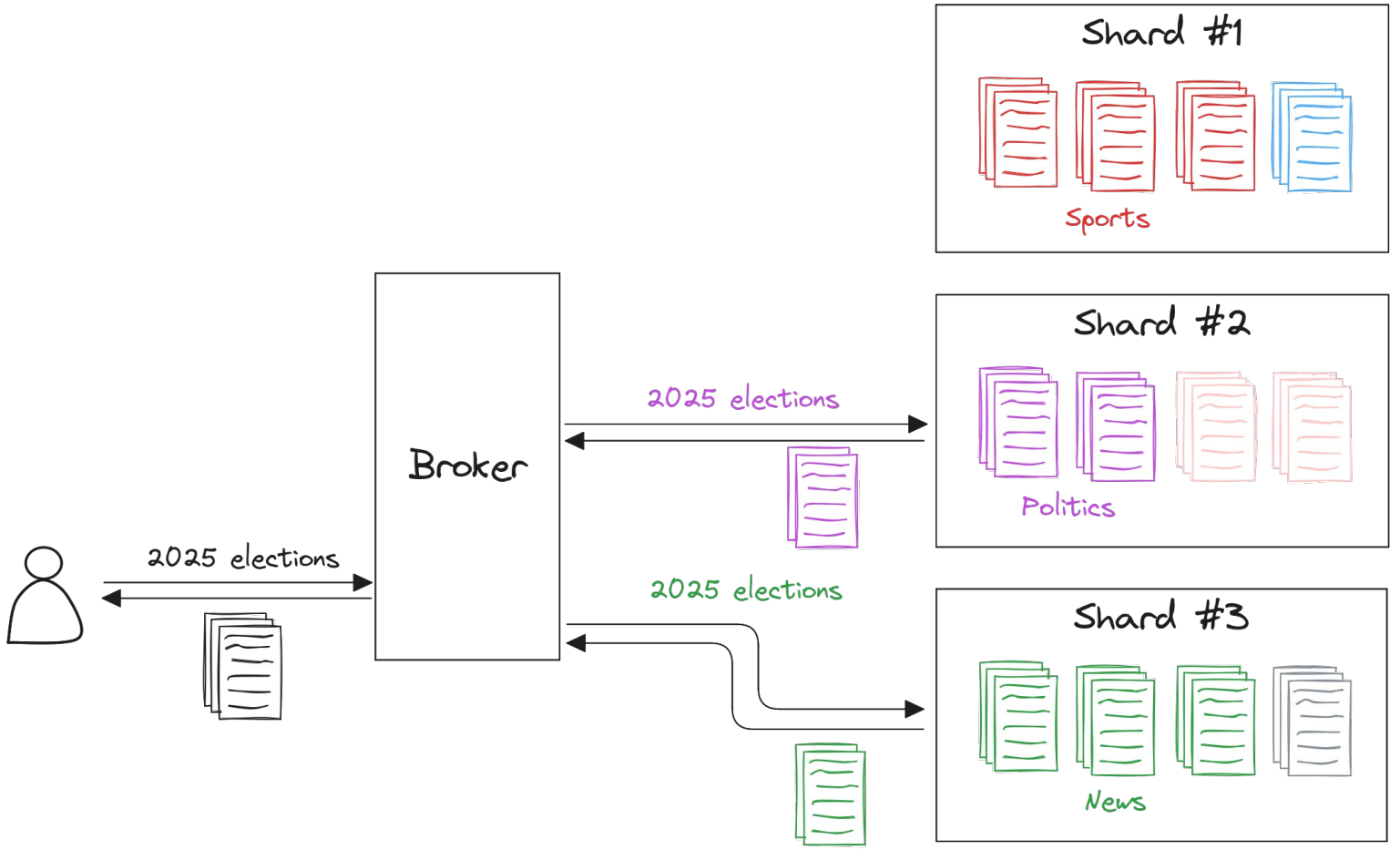
SELECTIVE SEARCH



SELECTIVE SEARCH



SELECTIVE SEARCH



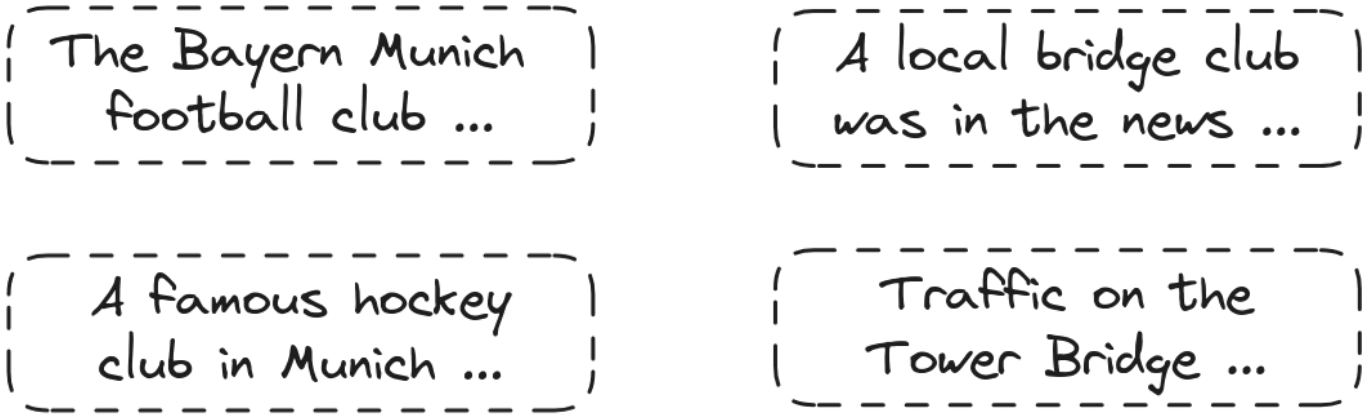
HOW TO CLUSTER THE DOCUMENTS?

HOW TO CLUSTER THE DOCUMENTS?

- K-means on document language models
  - Kullback-Leibler Divergence as distance metric

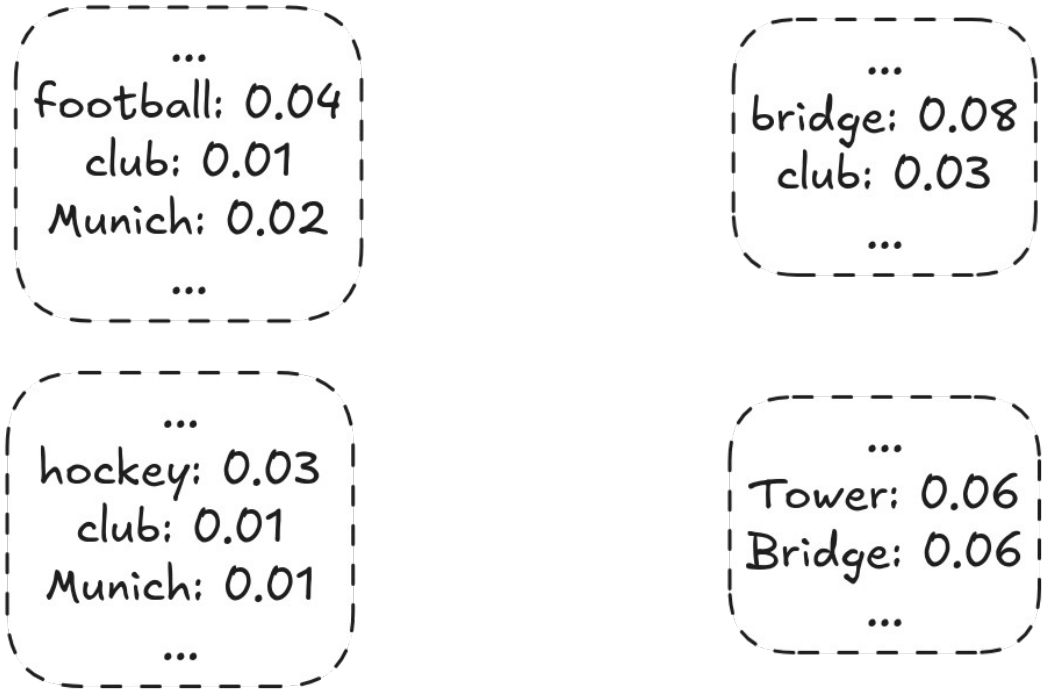
HOW TO CLUSTER THE DOCUMENTS?

- K-means on document language models
  - Kullback-Leibler Divergence as distance metric



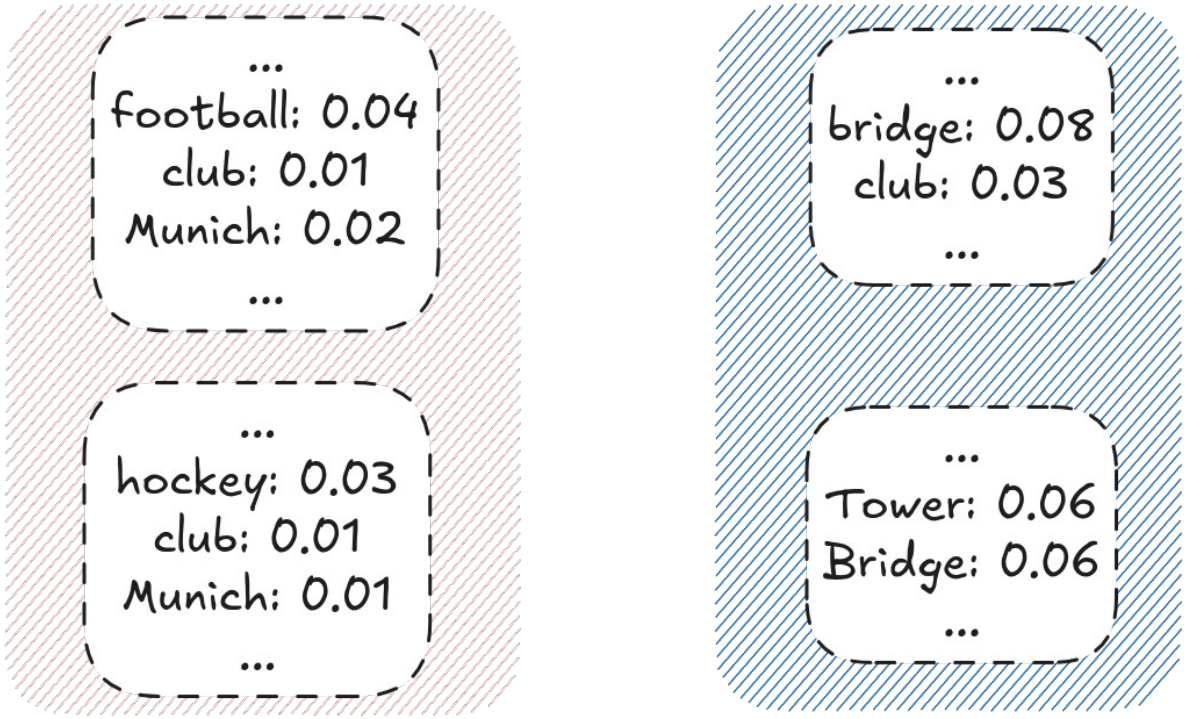
HOW TO CLUSTER THE DOCUMENTS?

- K-means on document language models
  - Kullback-Leibler Divergence as distance metric



HOW TO CLUSTER THE DOCUMENTS?

- K-means on document language models
  - Kullback-Leibler Divergence as distance metric



QKLD: A QUERY-BIASED DISTANCE METRIC

- What if the clusters don’t align with user interests?
- Solution: use a query log to determine important terms
  - New distance metric: QKLD

[1] Dai, Z. et al. 2016. Query-Biased Partitioning for Selective Search. CIKM 2016.

QKLD: A QUERY-BIASED DISTANCE METRIC

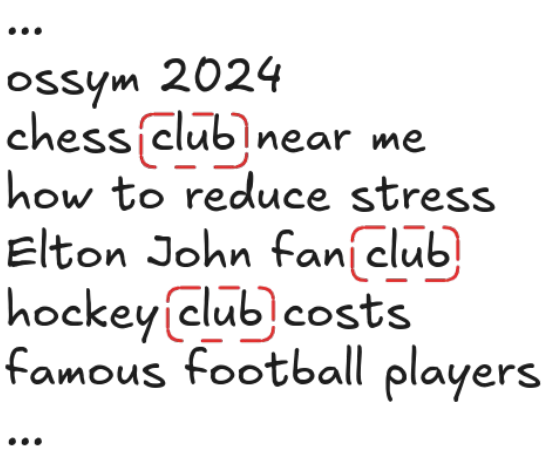
- What if the clusters don’t align with user interests?
- Solution: use a query log to determine important terms
  - New distance metric: QKLD

...
ossym 2024
chess club near me
how to reduce stress
Elton John fan club
hockey club costs
famous football players
...

[1] Dai, Z. et al. 2016. Query-Biased Partitioning for Selective Search. CIKM 2016.

QKLD: A QUERY-BIASED DISTANCE METRIC

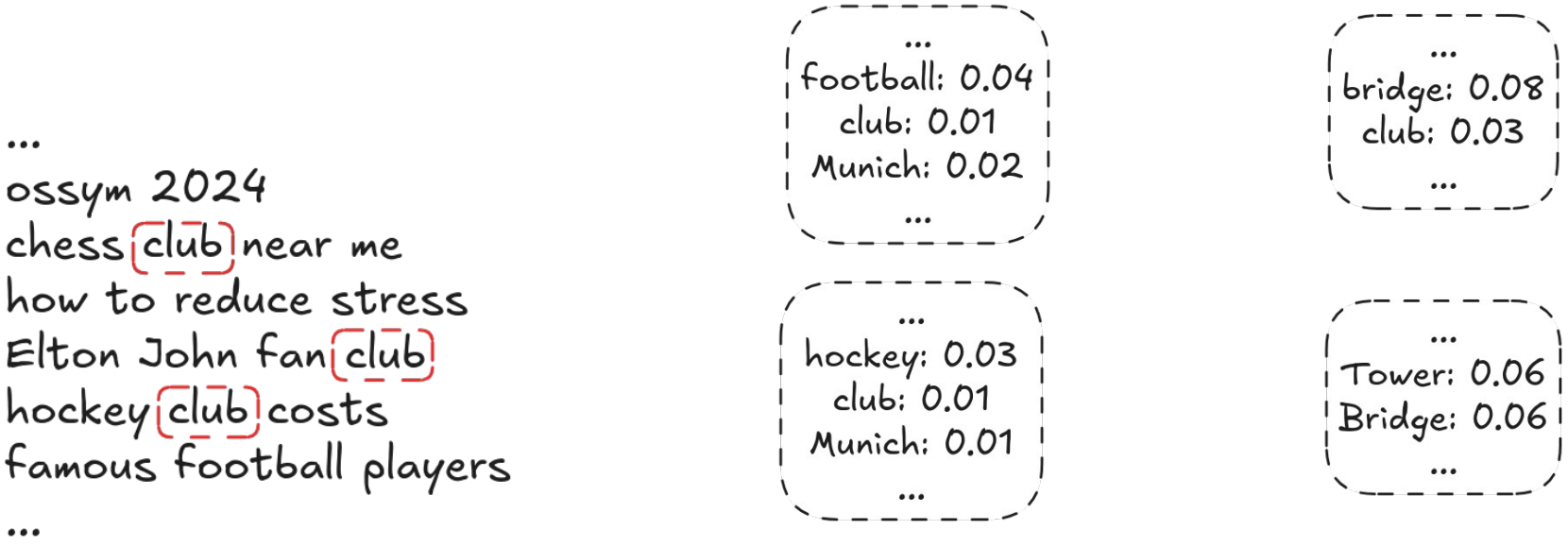
- What if the clusters don’t align with user interests?
- Solution: use a query log to determine important terms
  - New distance metric: QKLD



[1] Dai, Z. et al. 2016. Query-Biased Partitioning for Selective Search. CIKM 2016.

QKLD: A QUERY-BIASED DISTANCE METRIC

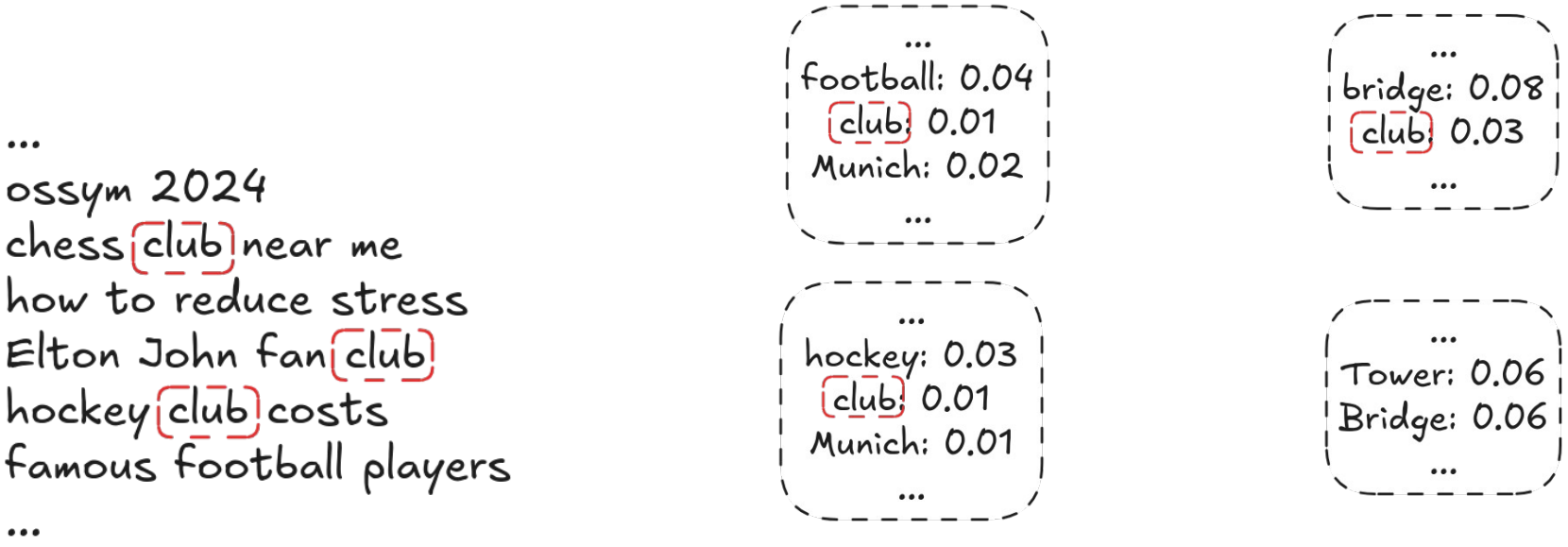
- What if the clusters don't align with user interests?
- Solution: use a query log to determine important terms
  - New distance metric: QKLD



[1] Dai, Z. et al. 2016. Query-Biased Partitioning for Selective Search. CIKM 2016.

QKLD: A QUERY-BIASED DISTANCE METRIC

- What if the clusters don't align with user interests?
- Solution: use a query log to determine important terms
  - New distance metric: QKLD



[1] Dai, Z. et al. 2016. Query-Biased Partitioning for Selective Search. CIKM 2016.

QKLD: A QUERY-BIASED DISTANCE METRIC

- What if the clusters don’t align with user interests?
- Solution: use a query log to determine important terms
  - New distance metric: QKLD

[1] Dai, Z. et al. 2016. Query-Biased Partitioning for Selective Search. CIKM 2016.

QINIT: QUERY-BIASED CENTROID INITIALIZATION

- Extract important terms from query log
- Cluster word embeddings of these terms
  - E.g. word2vec, GloVe
- Use clusters as initial seed “documents”
  - New initialization algorithm: `QInit`

[1] Dai, Z. et al. 2016. Query-Biased Partitioning for Selective Search. CIKM 2016.

QINIT: QUERY-BIASED CENTROID INITIALIZATION

- Extract important terms from query log
- Cluster word embeddings of these terms
  - E.g. word2vec, GloVe
- Use clusters as initial seed “documents”
  - New initialization algorithm: `QInit`

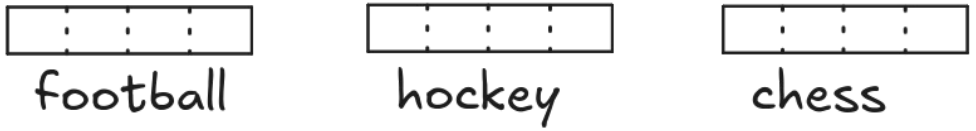
…
ossym 2024
chess club near me
how to reduce stress
Elton John fan club
hockey club costs
famous football players
…

[1] Dai, Z. et al. 2016. Query-Biased Partitioning for Selective Search. CIKM 2016.

QINIT: QUERY-BIASED CENTROID INITIALIZATION

- Extract important terms from query log
- Cluster word embeddings of these terms
  - E.g. word2vec, GloVe
- Use clusters as initial seed “documents”
  - New initialization algorithm: `QInit`

…
ossym 2024
chess club near me
how to reduce stress
Elton John fan club
hockey club costs
famous football players
…

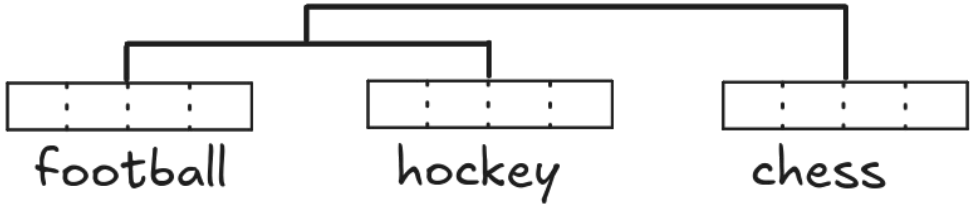


[1] Dai, Z. et al. 2016. Query-Biased Partitioning for Selective Search. CIKM 2016.

QINIT: QUERY-BIASED CENTROID INITIALIZATION

- Extract important terms from query log
- Cluster word embeddings of these terms
  - E.g. word2vec, GloVe
- Use clusters as initial seed “documents”
  - New initialization algorithm: `QInit`

...
ossym 2024
chess club near me
how to reduce stress
Elton John fan club
hockey club costs
famous football players
...

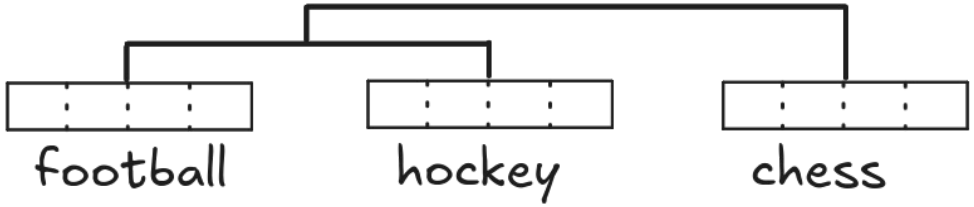


[1] Dai, Z. et al. 2016. Query-Biased Partitioning for Selective Search. CIKM 2016.

QINIT: QUERY-BIASED CENTROID INITIALIZATION

- Extract important terms from query log
- Cluster word embeddings of these terms
  - E.g. word2vec, GloVe
- Use clusters as initial seed “documents”
  - New initialization algorithm: `QInit`

...
ossym 2024
chess club near me
how to reduce stress
Elton John fan club
hockey club costs
famous football players
...



football: 0.02
hockey: 0.01

chess: 0.01

[1] Dai, Z. et al. 2016. Query-Biased Partitioning for Selective Search. CIKM 2016.

PROBLEMS WITH CLUSTERING

PROBLEMS WITH CLUSTERING

- How to efficiently cluster a large dataset?

PROBLEMS WITH CLUSTERING

- How to efficiently cluster a large dataset?
  - Sample-based clustering
  - Cluster only a subset
  - Map remaining documents to nearest centroid

PROBLEMS WITH CLUSTERING

- How to efficiently cluster a large dataset?
  - Sample-based clustering
  - Cluster only a subset
  - Map remaining documents to nearest centroid

- How to prevent large skew in shard sizes?

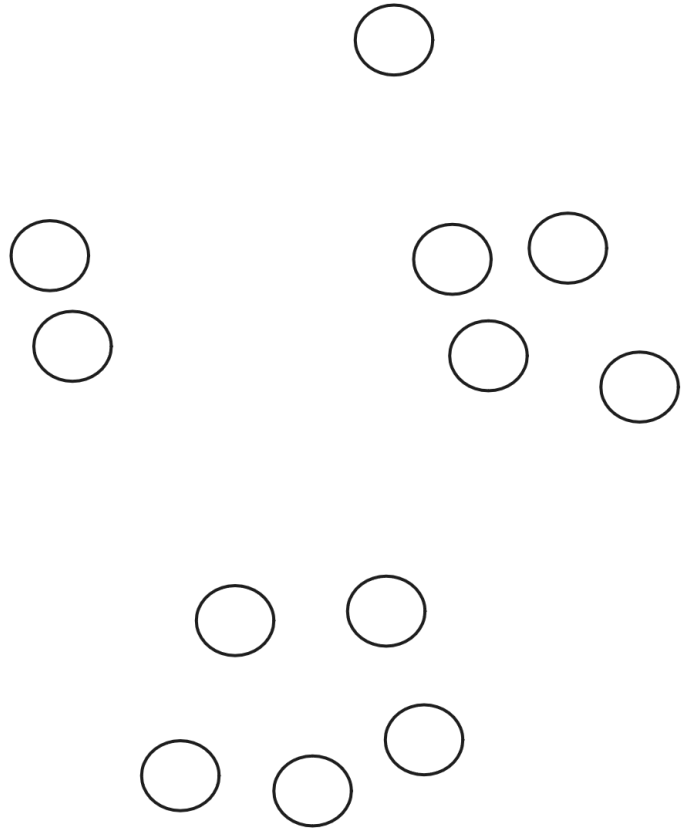
PROBLEMS WITH CLUSTERING

- How to efficiently cluster a large dataset?
  - Sample-based clustering
  - Cluster only a subset
  - Map remaining documents to nearest centroid

- How to prevent large skew in shard sizes?
  - Size-bounded clustering
  - Split large shards
  - Merge small shards

SIZE-BOUNDED SAMPLE-BASED CLUSTERING (SB² K-MEANS)

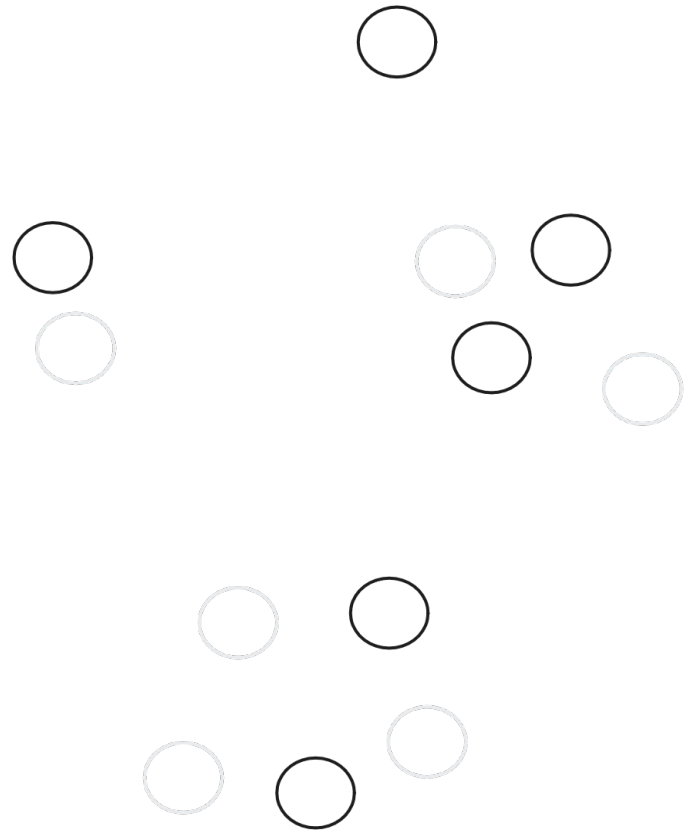
- Initial phase
  - Sample and cluster



[2] Kulkarni, A. 2013. Efficient and Effective Large-scale Search. Carnegie Mellon University.

SIZE-BOUNDED SAMPLE-BASED CLUSTERING (SB² K-MEANS)

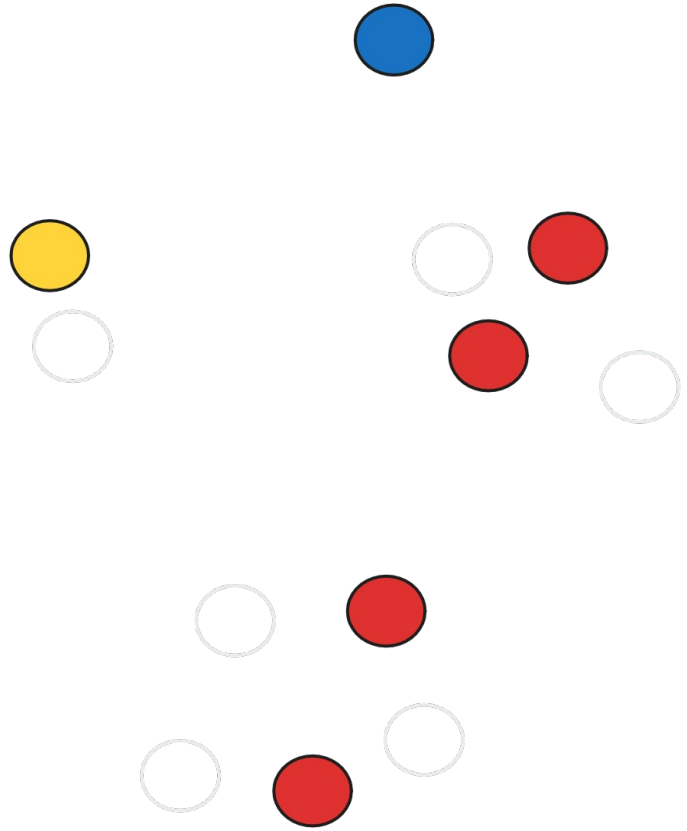
- Initial phase
  - Sample and cluster



[2] Kulkarni, A. 2013. Efficient and Effective Large-scale Search. Carnegie Mellon University.

SIZE-BOUNDED SAMPLE-BASED CLUSTERING (SB² K-MEANS)

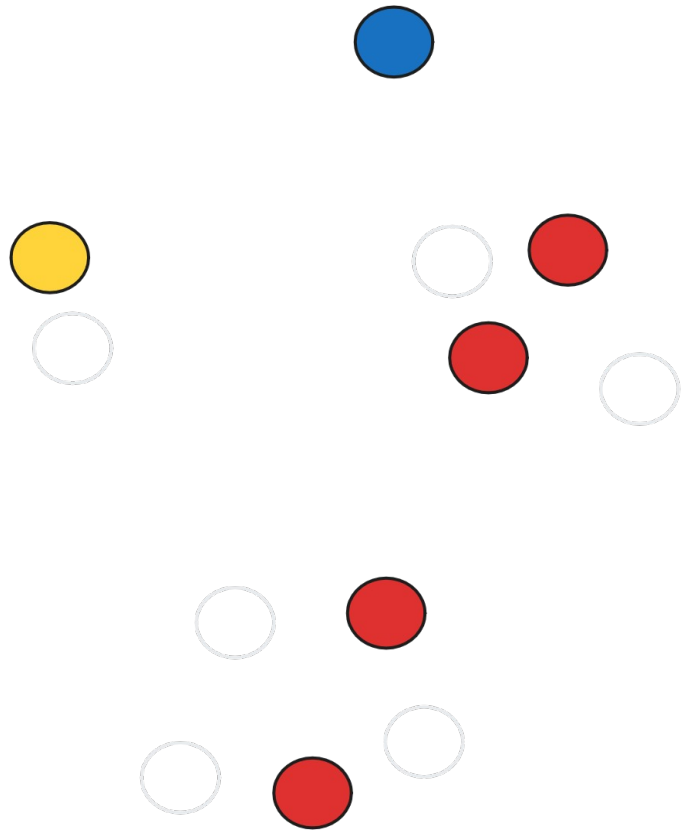
- Initial phase
  - Sample and cluster



[2] Kulkarni, A. 2013. Efficient and Effective Large-scale Search. Carnegie Mellon University.

SIZE-BOUNDED SAMPLE-BASED CLUSTERING (SB² K-MEANS)

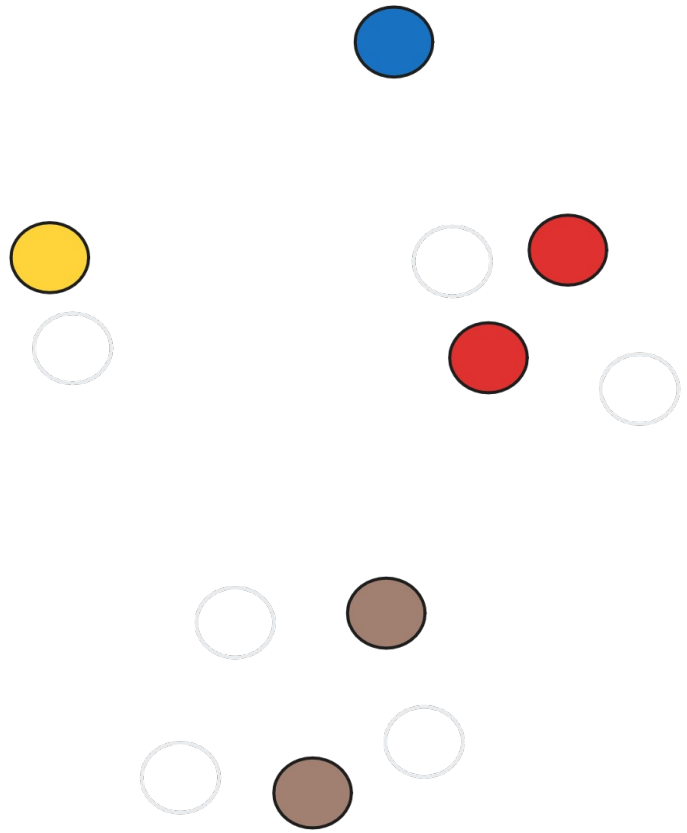
- Initial phase
  - Sample and cluster
- Split phase
  - Re-cluster large shards



[2] Kulkarni, A. 2013. Efficient and Effective Large-scale Search. Carnegie Mellon University.

SIZE-BOUNDED SAMPLE-BASED CLUSTERING (SB² K-MEANS)

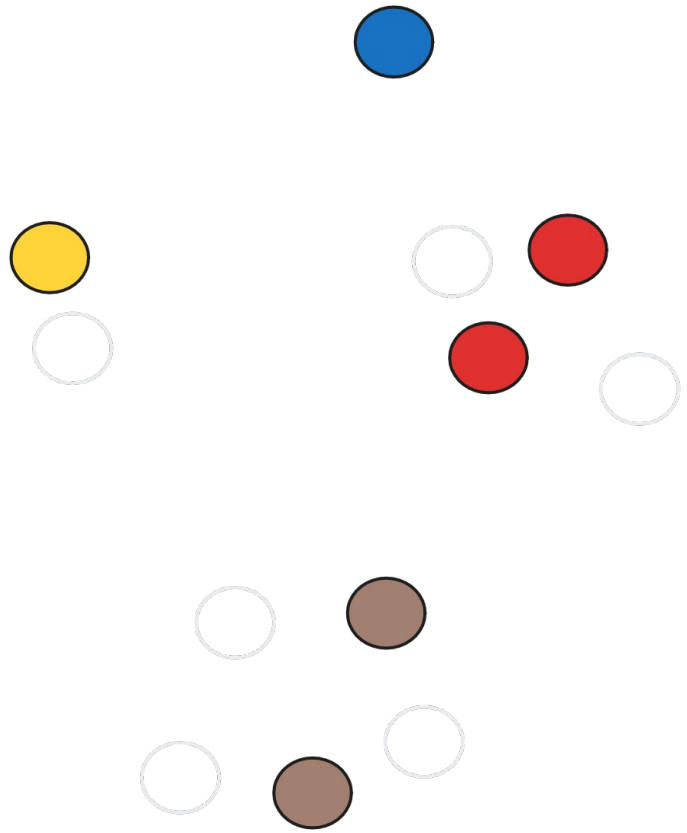
- Initial phase
  - Sample and cluster
- Split phase
  - Re-cluster large shards



[2] Kulkarni, A. 2013. Efficient and Effective Large-scale Search. Carnegie Mellon University.

SIZE-BOUNDED SAMPLE-BASED CLUSTERING (SB² K-MEANS)

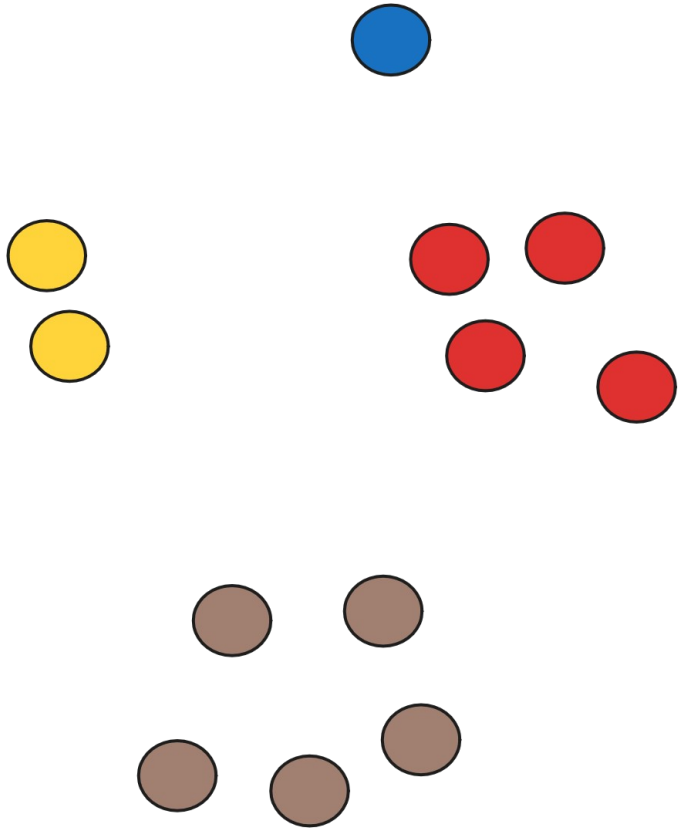
- Initial phase
  - Sample and cluster
- Split phase
  - Re-cluster large shards
- Project phase
  - Assign remaining documents



[2] Kulkarni, A. 2013. Efficient and Effective Large-scale Search. Carnegie Mellon University.

SIZE-BOUNDED SAMPLE-BASED CLUSTERING (SB² K-MEANS)

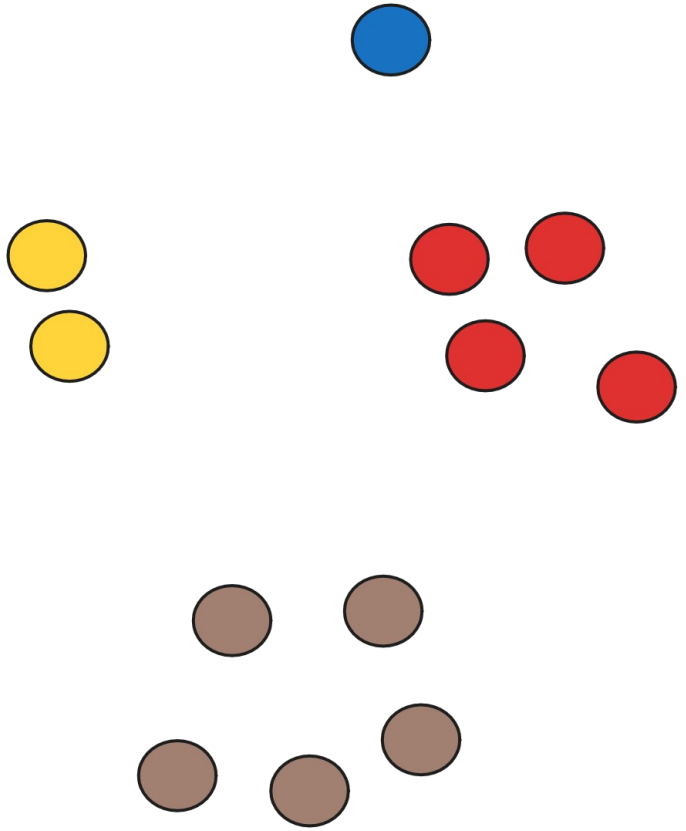
- Initial phase
  - Sample and cluster
- Split phase
  - Re-cluster large shards
- Project phase
  - Assign remaining documents



[2] Kulkarni, A. 2013. Efficient and Effective Large-scale Search. Carnegie Mellon University.

SIZE-BOUNDED SAMPLE-BASED CLUSTERING (SB² K-MEANS)

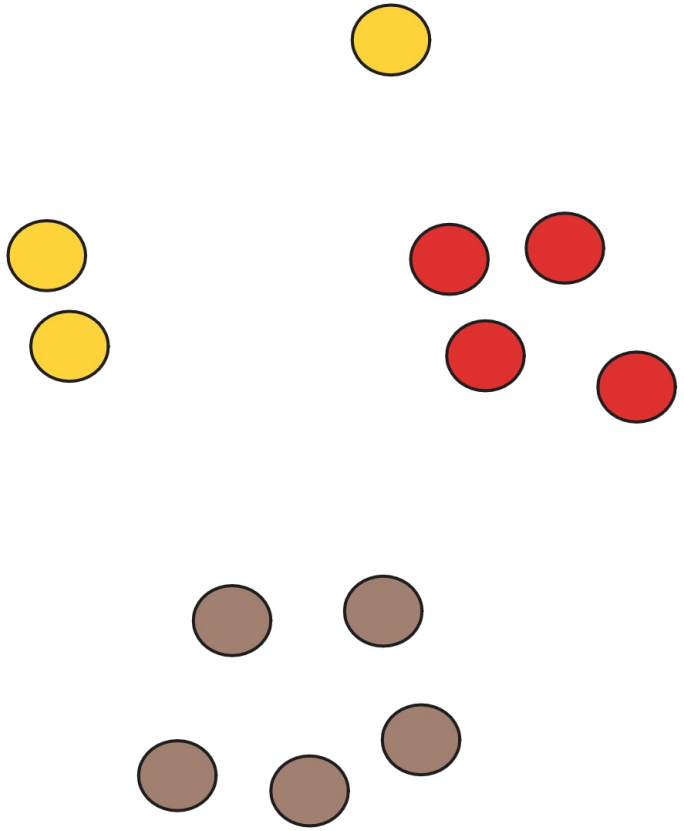
- Initial phase
  - Sample and cluster
- Split phase
  - Re-cluster large shards
- Project phase
  - Assign remaining documents
- Merge phase
  - Combine small shards



[2] Kulkarni, A. 2013. Efficient and Effective Large-scale Search. Carnegie Mellon University.

SIZE-BOUNDED SAMPLE-BASED CLUSTERING (SB² K-MEANS)

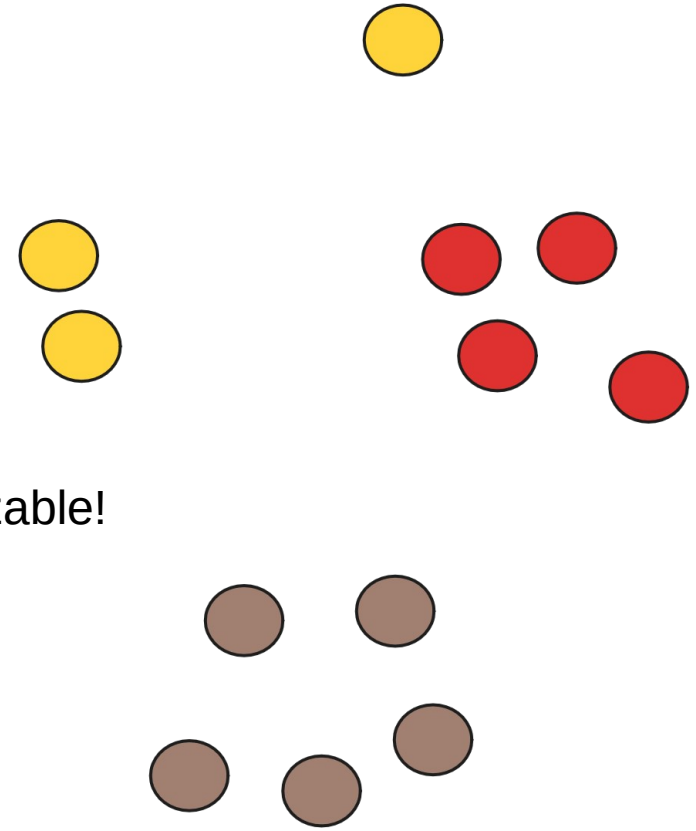
- Initial phase
  - Sample and cluster
- Split phase
  - Re-cluster large shards
- Project phase
  - Assign remaining documents
- Merge phase
  - Combine small shards



[2] Kulkarni, A. 2013. Efficient and Effective Large-scale Search. Carnegie Mellon University.

SIZE-BOUNDED SAMPLE-BASED CLUSTERING (SB² K-MEANS)

- Initial phase
  - Sample and cluster
- Split phase
  - Re-cluster large shards
- Project phase
  - Assign remaining documents
- Merge phase
  - Combine small shards



This step is parallelizable!

[2] Kulkarni, A. 2013. Efficient and Effective Large-scale Search. Carnegie Mellon University.

OUR CONTRIBUTIONS

- An open source implementation of SB² K-means

https://gitlab.science.ru.nl/informagus/document-clustering/

OUR CONTRIBUTIONS

- An open source implementation of SB² K-means
  - Including `QKLD` and `QInit`
  - Following the scikit-learn API
  - Written in Cython
  - Parallelization for Projection step

https://gitlab.science.ru.nl/informagus/document-clustering/

OUR CONTRIBUTIONS

- An open source implementation of SB² K-means
  - Including `QKLD` and `QInit`
  - Following the scikit-learn API
  - Written in Cython
  - Parallelization for Projection step

https://gitlab.science.ru.nl/informagus/document-clustering/

- Use cases
  - Verify and improve reproducibility of selective search papers
  - Allow other parties to cluster documents for research or (search) applications