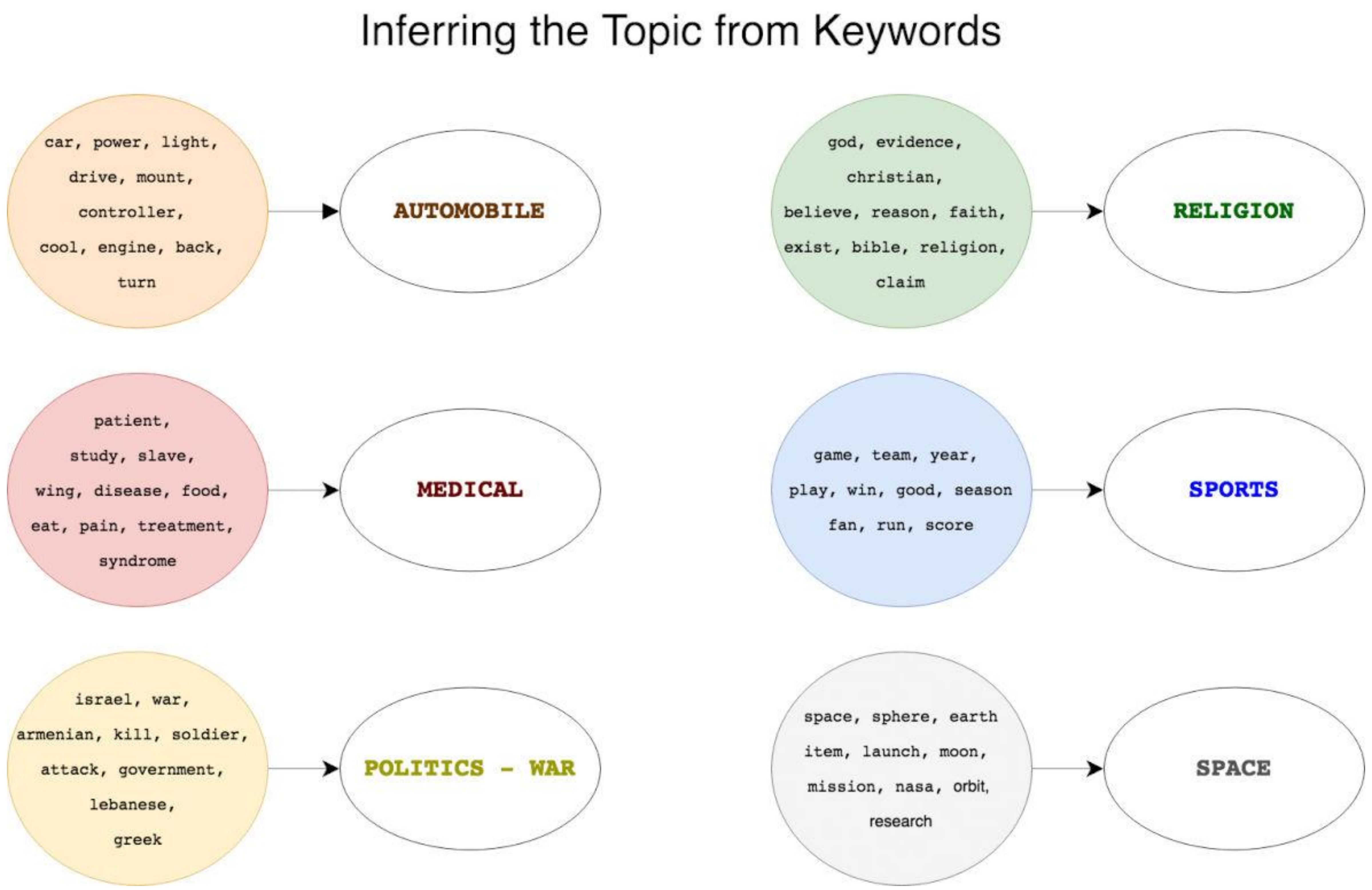
BERTopic with RAPIDS

Topic Modeling gives us an opportunity to perform analysis quickly while deriving valuable insights

- Topic modeling extracts meaning from text by identifying recurring themes
- Generally, we either use a pre-trained BERT which isn't trained on the domain-specific task, or we perform a trivial task like sentimental analysis to train a model, resulting in sub-optimal situations
- We can leverage the widespread applications of topic modeling to generate actionable insights quickly



Inferring the Topic from Keywords

car, power, light, drive, mount, controller, cool, engine, back, turn

AUTOMOBILE

god, evidence, christian, believe, reason, faith, exist, bible, religion, claim

RELIGION

patient, study, slave, wing, disease, food, eat, pain, treatment, syndrome

MEDICAL

game, team, year, play, win, good, season fan, run, score

SPORTS

israel, war, armenian, kill, soldier, attack, government, lebanese, greek

POLITICS - WAR

space, sphere, earth item, launch, moon, mission, nasa, orbit, research

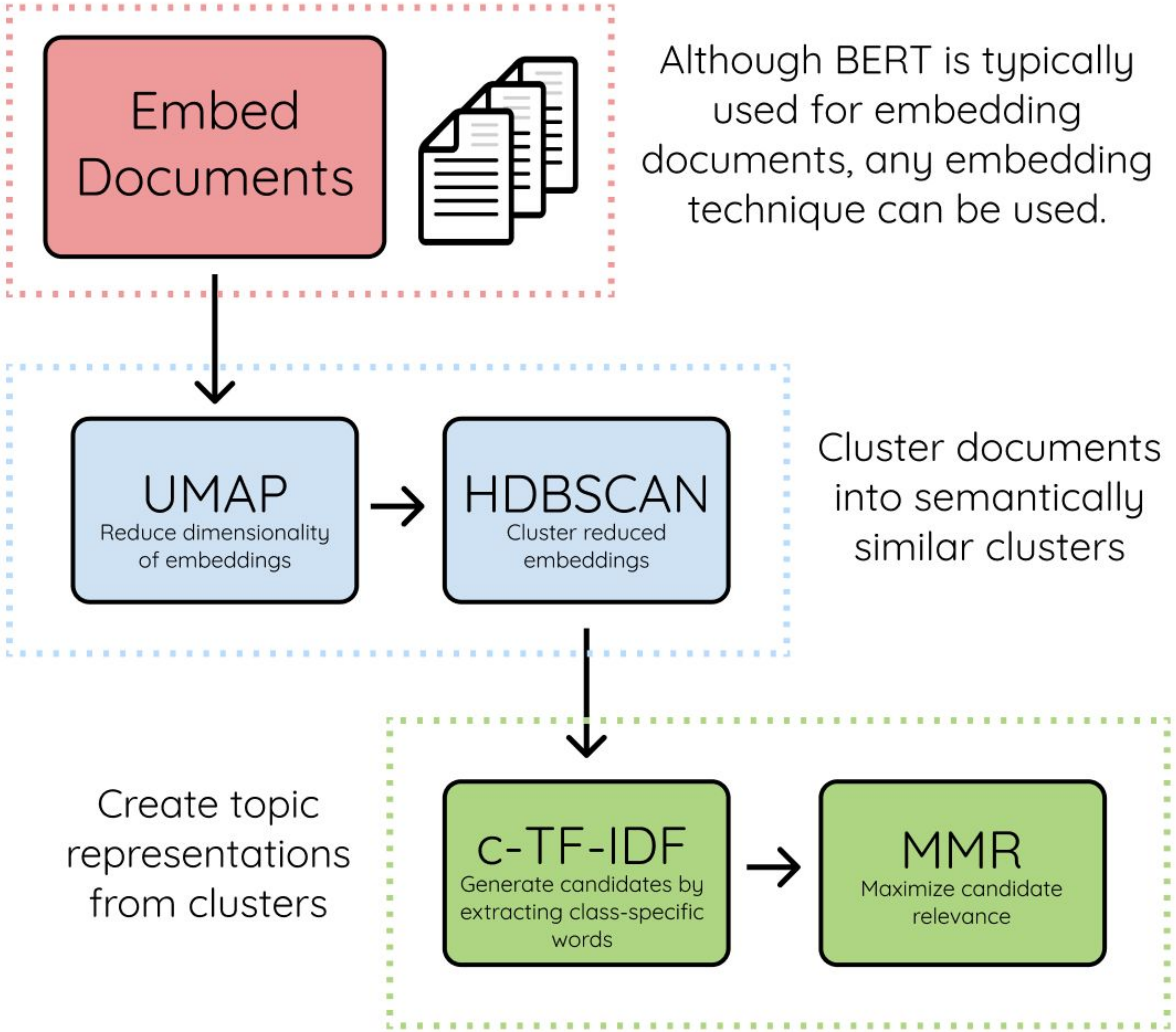
SPACE

BERTopic Solution

- BERTopic is a topic modeling technique that leverages 🤗 transformers and c-TF-IDF to create dense clusters allowing for easily interpretable topics whilst keeping important words in the topic descriptions.
- BERTopic supports guided, (semi-) supervised, and dynamic topic modeling. It even supports visualizations similar to LDAvis!
- Research Paper: https://arxiv.org/pdf/2008.09470.pdf
- Python Package: https://github.com/MaartenGr/BERTopic

BERTopic package stats:

- #stars - **2,924**
- #Issues - **604**
- #Downloads last week - **12,756**
- #Downloads last month - **60,093**



Embed Documents
UMAP
Reduce dimensionality of embeddings
HDBSCAN
Cluster reduced embeddings
c-TF-IDF
Generate candidates by extracting class-specific words
MMR
Maximize candidate relevance

Although BERT is typically used for embedding documents, any embedding technique can be used.

Cluster documents into semantically similar clusters

Create topic representations from clusters

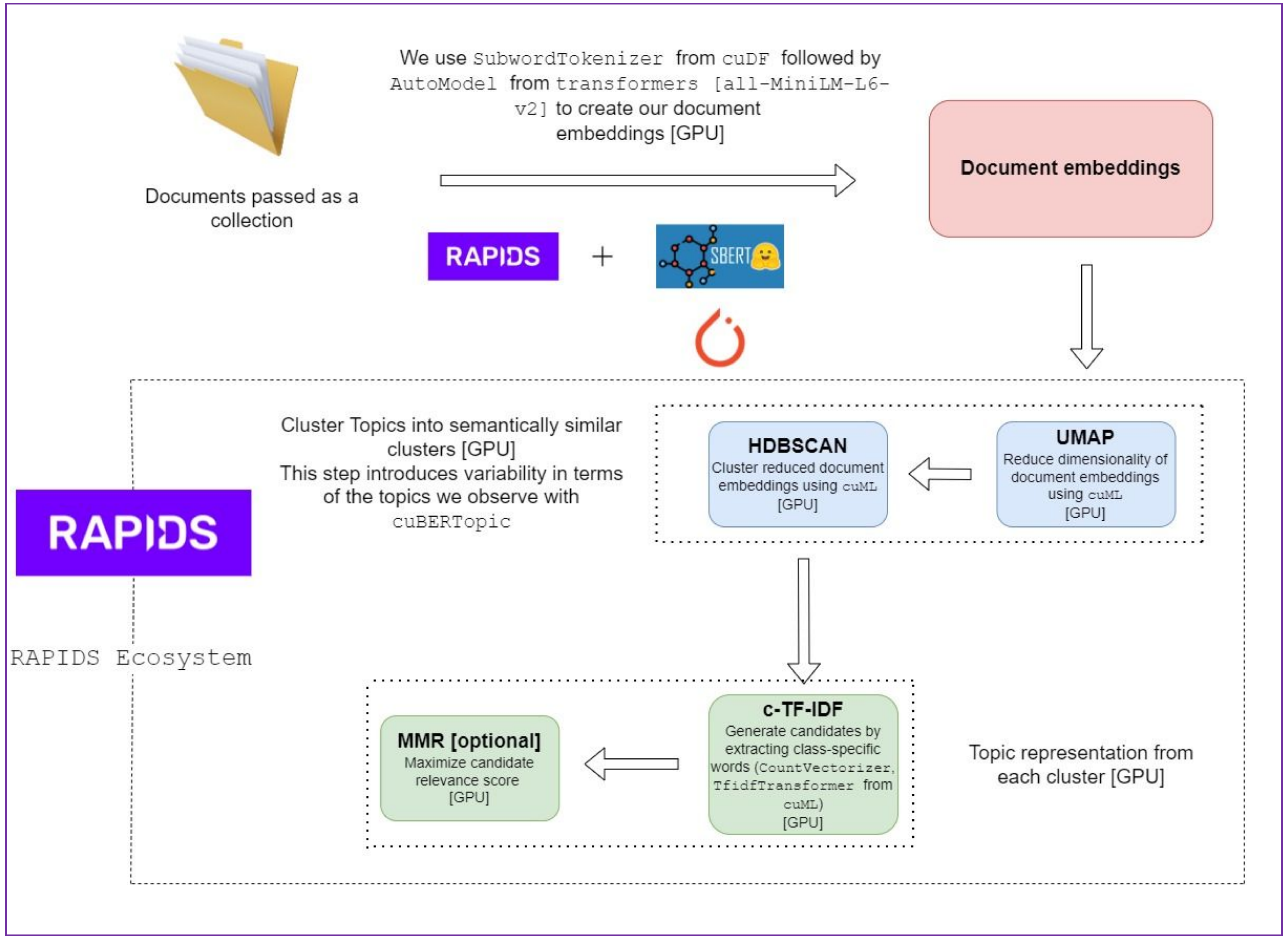
BERTopic
RAPIDS acceleration

Advantages of integrating BERTopic with RAPIDS

- Allows us to increase library interoperability by integrating ML framework of BERTopic with RAPIDS
- Due to the compartmentalized nature of BERTopic, we can use UMAP and HDBSCAN in isolation for various ML applications. So, any speedup in these individual pieces can be applied to a host of dimensionality reduction and clustering applications

Repo Link:
https://github.com/rapidsai/rapids-examples/tree/main/cuBERT_topic_modelling

Example Notebook:
https://github.com/rapidsai/rapids-examples/blob/main/cuBERT_topic_modelling/berttopic_example.ipynb


Documents passed as a collection
We use SubwordTokenizer from cuDF followed by AutoModel from transformers [all-MiniLM-L6-v2] to create our document embeddings [GPU]
RAPIDS
+
SBERT
Document embeddings
RAPIDS
RAPIDS Ecosystem
Cluster Topics into semantically similar clusters [GPU]
This step introduces variability in terms of the topics we observe with cuBERTopic
HDBSCAN
Cluster reduced document embeddings using cuML [GPU]
UMAP
Reduce dimensionality of document embeddings using cuML [GPU]
MMR [optional]
Maximize candidate relevance score [GPU]
c-TF-IDF
Generate candidates by extracting class-specific words (CountVectorizer, TfidfTransformer from cuML) [GPU]
Topic representation from each cluster [GPU]

RAPIDS integration benchmarks

Wikidata dataset

Time on 500K rows of Wikidata Dataset (in s)					
Stage	CPU	GPU	GPU with RAPIDS	Speedup (vs CPU)	RAPIDS Speedup (vs GPU)
UMAP	N/A	790	71	**N/A**	**11.1**
HDBSCAN	N/A	65	17	**N/A**	**3.8**

News dataset

Time on News Dataset (in s)				Speedup due to RAPIDS	
Stage	CPU	GPU	GPU with RAPIDS	Speedup (vs CPU)	Speedup (vs GPU)
UMAP	21	10	1	**21.0**	**10.0**
HDBSCAN	3	3	1	**3.0**	**3.0**

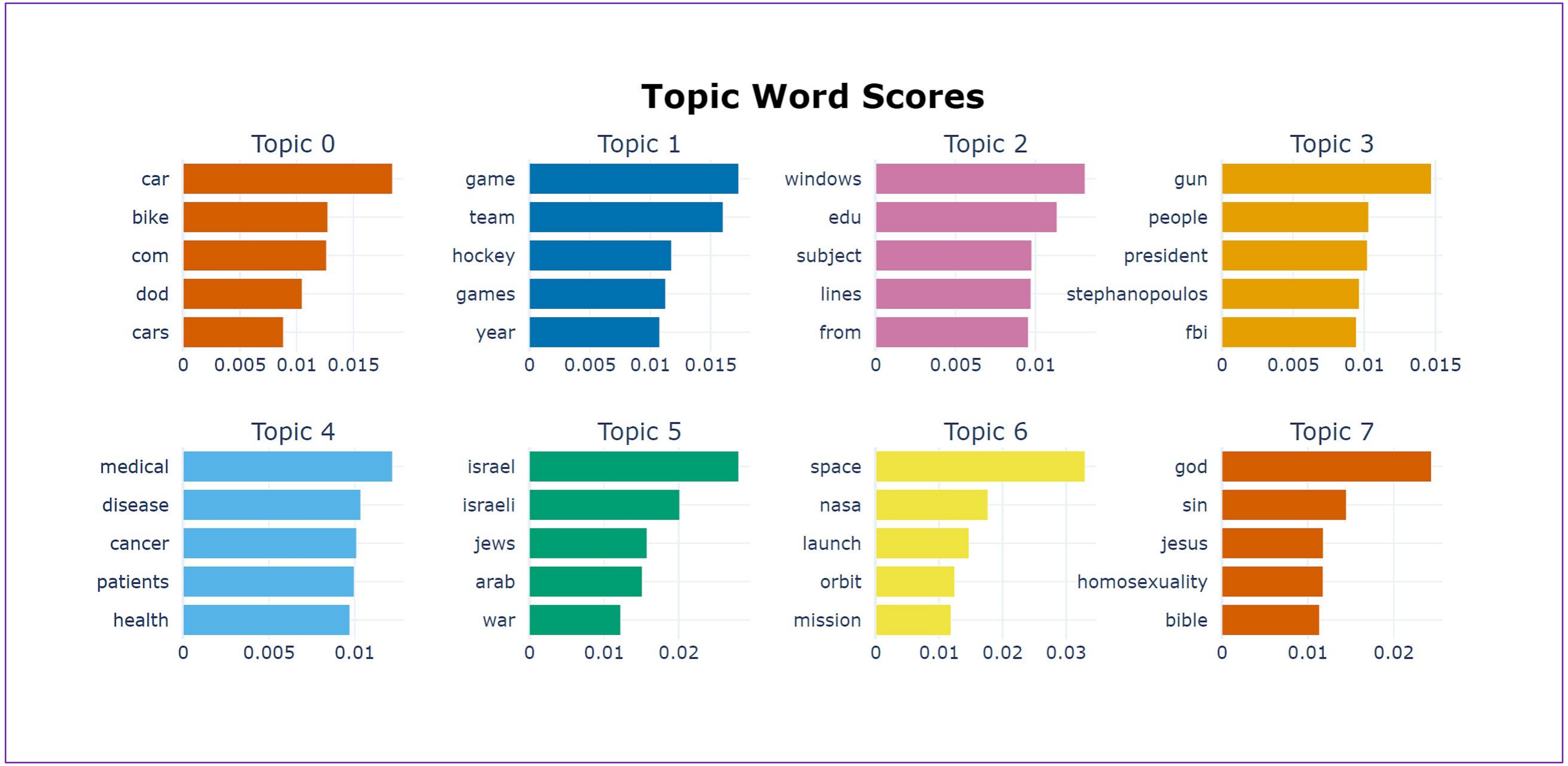
AN4 dataset

Time on Amazon Dataset (in s)				Speedup due to RAPIDS	
Stage	CPU	GPU	GPU with RAPIDS	Speedup (vs CPU)	Speedup (vs GPU)
UMAP	2350	1660	19	**123.7**	**87.4**
HDBSCAN	65	57.8	7	**9.3**	**8.3**

***We used A100 GPU and m6id.16xlarge CPU for the benchmarks**

***We have also released a stand alone cuBERT package**

Topic modeling results



*We used A100 GPU and m6id.16xlarge CPU for the benchmarks

*We have a stand alone cuBERT package which gives a 3x speedup in UMAP phase against cuML integrated BERTopic