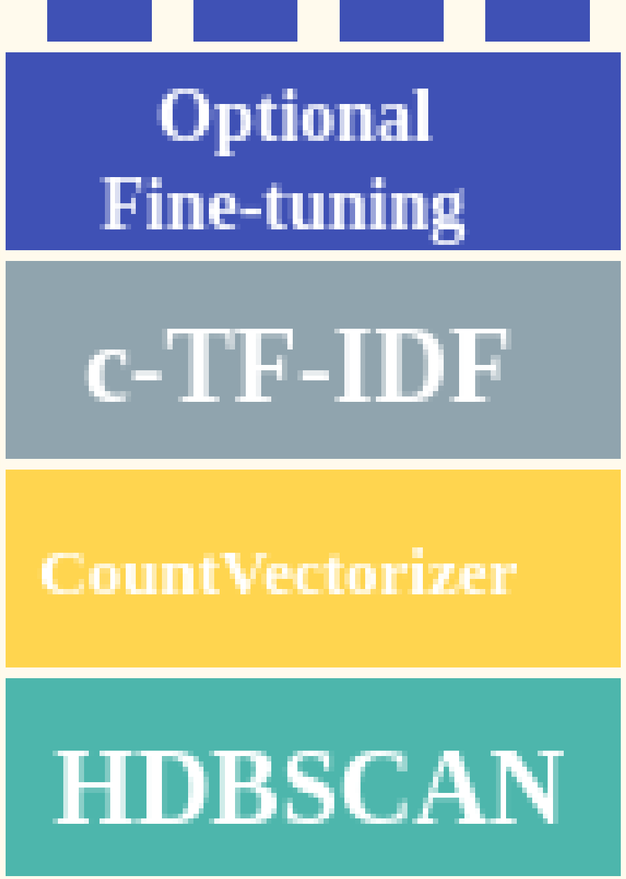
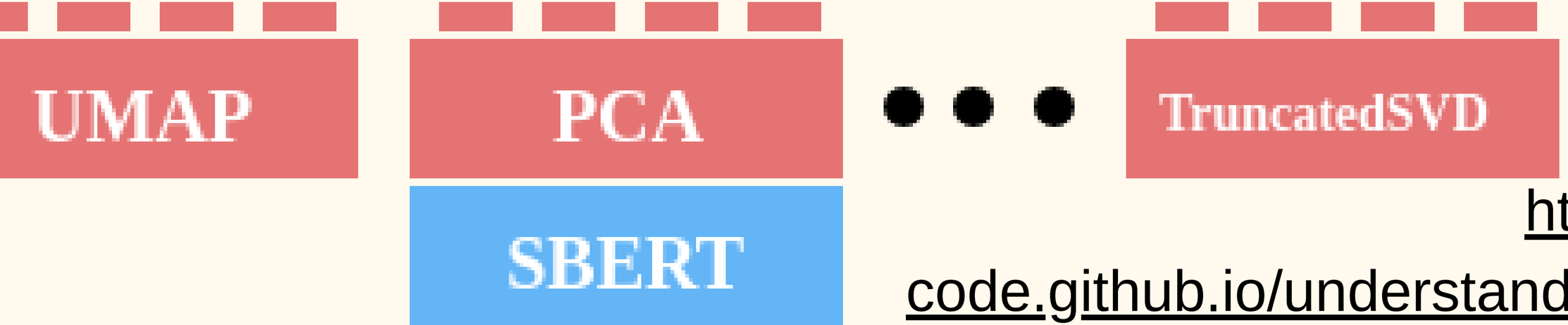
MASTERING TOPIC MODELLING & CLUSTER NAMING



GROOTENDORST, MAARTEN}

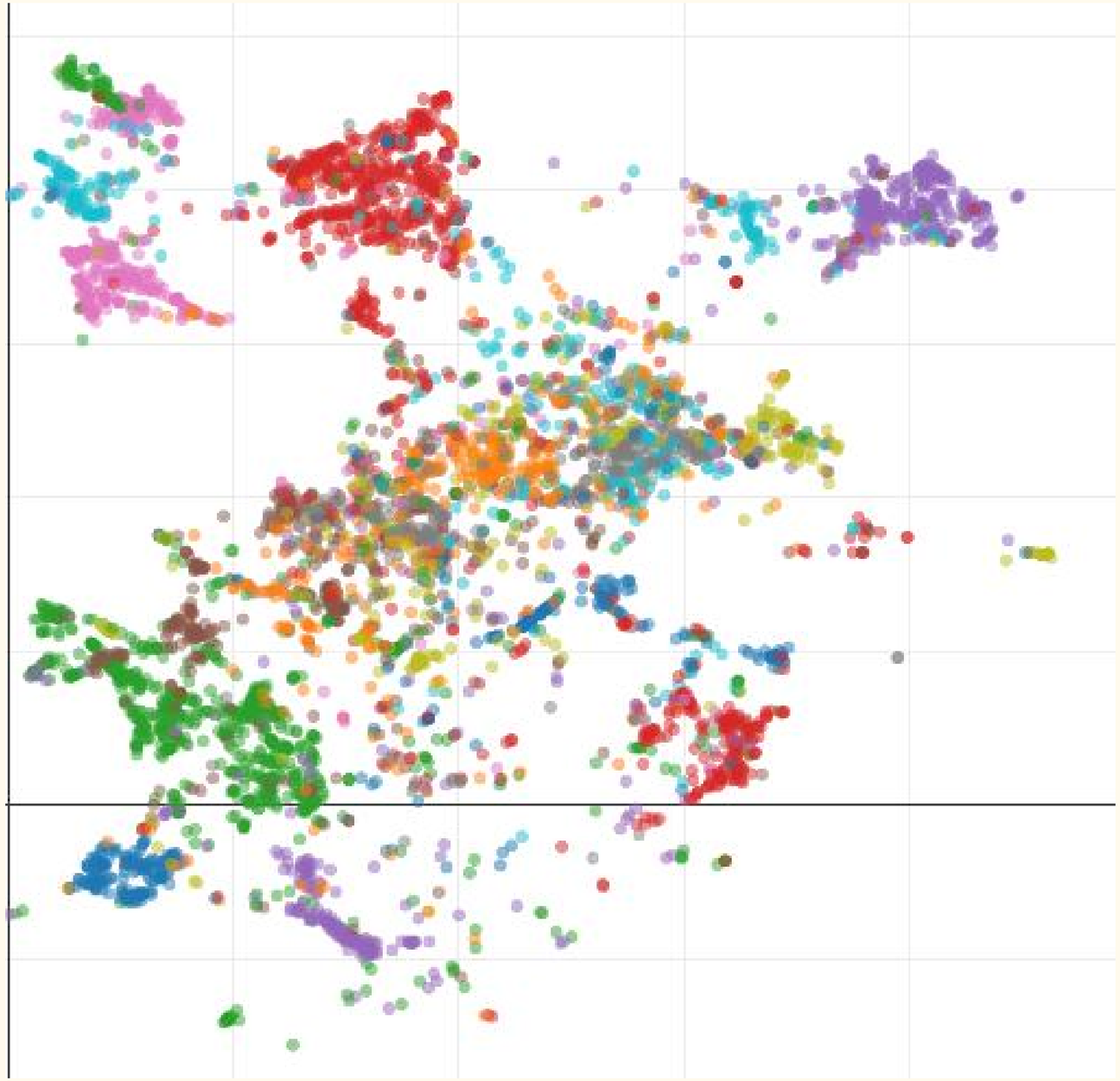
AUTOMATING TOPIC IDENTIFICATION BASED ON THE DOCUMENTS



https://pair-code.github.io/understanding-umap

CHALLENGE SOLVED: NAMING THE CLUSTERS

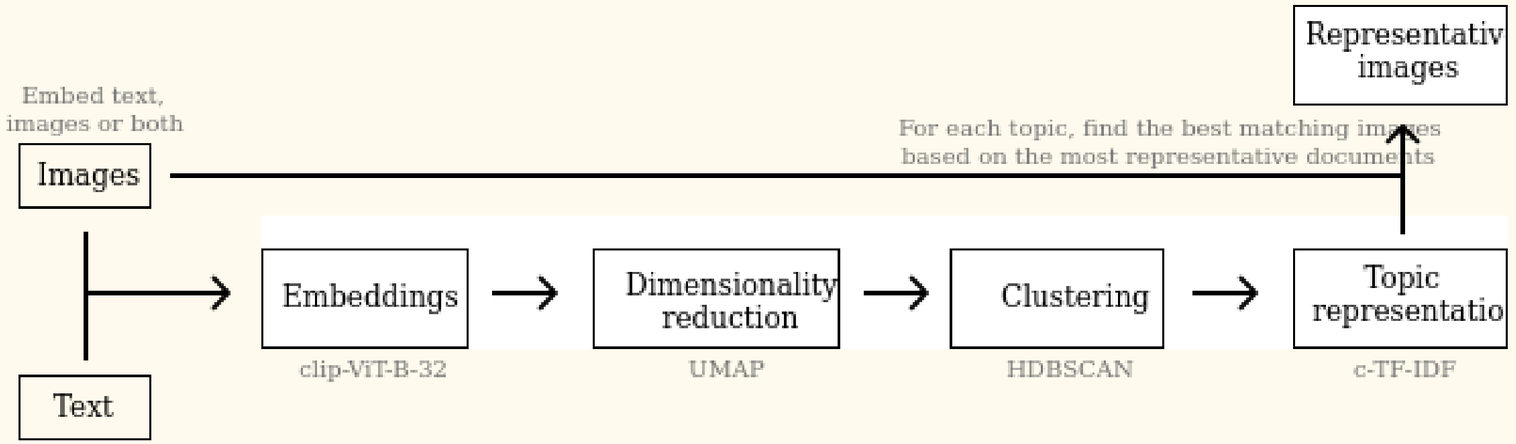
- TEXT DATA CAN BE CLUSTERED BASED ON THE EMBEDDING NUMBERS, AND DIMENSION REDUCTION ALGORITHMS. THESE CLUSTERS STILL NEED A NAME
- CLASS BASED TF-IDF (C-TF-IDF) ALGORITHM IS USED FOR ARRIVING AT THE NAME OF THE CLUSTER
- VISUALISING THE CLUSTERS IN 3D SPACE WILL REQUIRE COORDINATES FOR THE TEXT. UMAP CAN BE USED FOR THE SAME
- UNDERSTANDING HOW THE REPRESENTATION OF TOPICS CAN BE DIFFERENT USING VARIOUS STARTEGIES LIKE KEY WORDS, PARTS OF SPEECH OR USING THE LLMS



HTTPS://GITHUB.COM/INSIGHTBUILDER

CODE, PRACTICE AND FURTHER EXPLORATION

- THE DATA CAN BE OF ANY KIND, IF THE FEATURE CAN BE EXTRACTED THEN TOPIC CAN BE EXTRACTED.
- MORE THE DATA POINTS LONGER IT WILL TAKE TO EMBED AND GENERATE TOPICS
- SAVING THE MODEL TO HUGGING FACE HUB, USING PYTORCH AND SAFETENSORS IS THE CORRECT WAY TO
- HOW ABOUT MULTI-MODAL, DYNAMIC DATA, ?



HTTPS://GITHUB.COM/INSIGHTBUILDER

THANKS FOR WATCHING

REMEMBER TO PRACTICE WITH EXAMPLES

HTTPS://GITHUB.COM/INSIGHTBUILDER