Summary:

Creating contextual embeddings for BERTopic

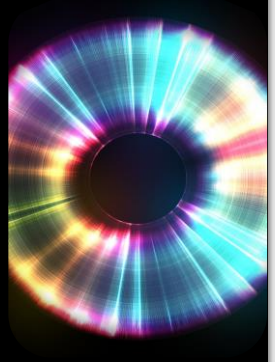
BERTopic is a Machine Learning framework that leverages transformers, whose architecture is effective at **capturing contextual relationships in large volumes of unstructured text**. One of the powerful applications of transformer-based models is topic modeling — unsupervised clustering of textual data into learned, semantically meaningful topics. BERTopic combines **contextual embeddings, dimensionality reduction, and clustering into a cohesive pipeline.** As such, BERTopic is particularly well-suited for clinical text, which is often complex (a pinnacle of professional jargon!).

In this example, the first step — generating fixed contextual embeddings for tokens across 2,000 clinical notes on the pre-trained bio_clinicalBERT model — has been completed and saved for future analysis in a .csv file. These embeddings form the foundation for downstream semantic analysis.

The results can be used for topic modeling and text summarizations for:

1. **Discharge planning.**
2. **Care quality assurance and compliance.**
3. **Population Health initiatives.**

Repository on GitHub: https://github.com/olga12kz-DS/BERT-Embeddings



Applying

bio_clinical_BERT model

1. Prepare the text by removing common stopwords and expanding clinical abbreviations.
2. Convert text to input IDs
   Generate attention masks
   Add special tokens ([CLS], [SEP])
   Pad and truncate
3. Mask and average token embeddings to produce a single 768-dimensional vector per clinical note.
4. Save results – 768-fixed vector length embeddings per note – in .csv