Vector representation of scientific text for document retrieval.

By Santosh Gupta, Srihari Humbarwadi, Akul Vohra, and Liam Croteau

In our ongoing research, we are attempting to develop transformer-based models that can create vector representations of queries and research papers, with the goal of having the query vectors be similar (in terms of cos similarity) to document vectors. An area of friction in biomedical information retrieval is the large variation in keywords and phrasing, and often completely different terms are used to describe similar materials or processes, and this often makes searching with traditional term-based methods difficult. We aim to develop models that can create accurate vector representations of biomedical text. We are also aiming to create classifiers that can associate these vector representations of biomedical text, for cases where the texts themselves do not represent similar items/ideas, but the items/ideas tend to often appear together in biomedical literature.