Bibliobuild

Bhushan Milind Borole, Master of Science in Computer Science

University of Dublin, Trinity College, 2022

Supervisor: Yvette Graham

While doing some research, one tends to get lost in finding papers related to the idea or any other paper/article. Since everything is digital and connected, there are humongous amounts of research papers or academic articles available at your fingertip. They are a source of data that can have many wide applications, which people have not worked upon in the past, but searching through them is a dilemma. This dissertation aims to reduce that task by using ML algorithms to find similarity between papers moreover, by generating a graph that will help in easy visualization of the same. The significant challenges faced were: execution time and visualizing a massive graph. Checking whether two papers are similar or not is based on their *Title* and *Abstract*. ACL Anthology dataset was used that had 77,000 research papers, out of which approximately 30,000 were in English and had both *Title* and *Abstract* fields. We find semantic similarities between their titles and abstract's using a rich machine learning model (i.e., BERT) and then apply weights to merge them and get a single similarity metric value. Then if the value crosses the threshold, it can be said that the given two papers are similar or share some similar ideas. Based on the dataset, a prototype graph was created, which showed similar papers properly, although with some inaccuracies. The ML models were correlated with the STS-B dataset, which gave a 0.84 correlation and another 0.9 correlation. There was no closed-source tool/data used.