

Date & Time

Monday, July 10th @ 11:30 AM Eastern

Zoom Link

<https://rowan.webex.com/meet/breitzman>

Thesis Title

Enhancing Inter-document Similarity for More Accurate Literature Reviews

Thesis Abstract

Document similarity, a core theme in Information Retrieval (IR), is a machine learning (ML) task associated with natural language processing (NLP). It is a measure of the distance between two documents given a set of rules. For the purpose of this thesis, two documents are similar if they are semantically alike, and describe similar concepts. While document similarity can be applied to multiple tasks, we focus our work on the accuracy of models in detecting referenced papers as similar documents. Multiple approaches have been used to determine the similarity of documents in regards to literature reviews. Some of such approaches use the number of similar citations, the similarity between the body of text, and the figures present in those documents. This researcher hypothesized that documents with sections of high similarity but a global low similarity are prone to being overlooked by existing models as the global score of the documents are used to measure similarity. In this study, we aim to detect, measure, and show the similarity of documents based on the maximum similarity of their subsections. By comparing subsections of the documents in our corpus, we were able to improve the performance of some models by over 100%.

Candidate Bio

Richard Igbiriki received his Bachelors of Science in Computer Science from Lincoln University, PA, USA in 2017. Following his graduation from Lincoln, he founded and ran a startup for two years, worked at Africa's best technology company, Andela, and led the engineering team of a well-funded YC company. At Rowan, he served as a Teaching Fellow for the Computer Science department. Presently, he works as a Software Engineer II at Microsoft, contributing to the development of OneDrive for Windows 10 and Windows 11.

Thesis Advisor

Dr. Anthony Breitzman