Assignment

Sayandip Bhattacharyya

3rd Year, B. Tech, CSE

2021-2025 Batch, B. P. Poddar Institute Of Management and Technology

My Approach for this task:

1. Data Reading and Preprocessing:

- I began by reading the citation network dataset (`cit-HepTh.txt`) and the abstracts dataset (`cit-HepTh-abstracts.tar.gz`).

- The citation network dataset provides information about which papers cite other papers.

- In the abstracts dataset, I found abstracts of papers organized by year.

2. Loading Pretrained Sci-BERT Model:

- I loaded the pretrained Sci-BERT model (`allenai/scibert\_scivocab\_uncased`) using the Sentence Transformers library.

- This model, trained specifically for scientific text, allows me to generate embeddings for sentences or paragraphs.

3. Iterating Over Seed Papers:

- I selected a few seed papers (e.g., `'9201001', '9203201', '119203001'`) for which I wanted to calculate similarity scores.

- For each seed paper:

- If the paper exists in the citation network dataset, I proceeded to find its references and their abstracts.

4. Embedding Abstracts and Calculating Similarity:

- For each reference paper cited by the seed paper:

- I embedded the abstracts of both the seed paper and the reference paper using the Sci-BERT model.

- Then, I calculated the cosine similarity between the embeddings of the seed paper's abstract and the reference paper's abstract.

- Cosine similarity serves as a similarity metric here, indicating how similar the abstracts are.

- Higher similarity scores suggest greater thematic similarity between papers.

5. Output:

- I printed the reference paper IDs along with their similarity scores for each seed paper.

- This information helps in understanding the thematic similarity between papers based on their abstract content.