Progress Report

Group: NLP
Member: Mingwei Huang (mingwei6)
Nov 14th, 2021

Completed:

- Text collection and text preprocessing
- Topic modeling for the dataset
- A word2vec model for word similarity (may not be used)
- Algorithm for segregating paragraphs based on the topic distribution of each sentence
- Also uncommented code is pushed to the repo

Ongoing:

- model evaluation (may need to manually label 10-20 lectures)
  - Evaluate how the model performs on segregating paragraphs
  - (Maybe) evaluate topic modeling since I have multiple models (which needs word2vec)
- May also incorporate the data here

Future work:

- Try paragraph segregation by measuring sentence similarities with Sci-BERT
- (Maybe) Summarize each paragraph
  - baseline - BM25
  - evaluation (may need manual labeling too)
  - still Sci-BERT on keyword extraction

Problem:

- The bag-of-word performs too much worse than an average human, and the result is barely useful.
- I suspect the usage of a transformer will not improve the performance a lot even though it can take the context into consideration. Paragraph segregation requires an extremely fine understanding of the text which is even hard for an average human, and this level of difficulty is too hard for supervised learning, let alone the unsupervised one.