Database: The Guthenberg Project (20 books from Children's Literature)

Started coding on a Jupyter Notebook to clean the dataset, split the books into paragraphs and sentences and play around with embeddings

First embedding : Sentence-BERT (cosine similarity = semantic distance)

Next step : Find the mathematical formulation for finding an ordering that minimizes the distance between neighbouring sentences (order 1) and implement it, consider higher orders ?