Observations:

1. All exact match sentences are removed by all other methods.
2. N-gram method removed the largest number of sentences. Many longer sentences are solely removed by this method.
3. Section titles/short phrases are likely to be solely removed by LCS.
4. Very short sentences are likely to be solely removed by sentence transformer (cosine similarity).
5. Sentences removed by Levinstein, in most cases, are also removed by sentence transformer or LCS or both.
6. There are some problems with sentence tokenization using Spacy. (bullet points like “1.”, “2.” will be considered as unique sentences/the end of the previous sentence. )

Suggestion:

N-gram method is not very informative.

Levinstein seems to be a conservative metric supported by other methods.

To do:

Accessibility: measure SMOG, KFGL, and GFI

Check the tool cTakes, for concept linking