ThoughtRiver - Document Headers

Legal agreements are typically quite structured documents. They usually open with a title and a declaration of the parties to the agreement. This is often followed by a declaration section where key terms are defined to disambiguate later clauses.

The rest of the document is the body of the contract (which is often broken into sections) followed by a brief signature (“execution”) block.

The goal of this task is to produce an algorithm that will extract headers from a document. The implementation should take the provided JSON format as input. Please provide your solution with an interface (with suitable methods or functions) that takes the document as input and returns a list of the indices of paragraphs containing headers. The case where no headers are present should be handled appropriately. Your solution should be built such that it is easy to add additional logic and demonstrate a high code quality.

The actual logic to determine whether a paragraph is a header does not need to be 100% accurate, but we are interested in your ideas and the implementation of them.

Please provide a supporting document detailing your solution, how you arrived at it, and any known limitations. Given a larger dataset, how would you assess the quality of your solution, and what other inference methods might be appropriate?

For reference, I have included two documents in a JSON format and the source file that they were derived from. Each row in the JSON corresponds to a paragraph in the document. Additionally, I have included nda.json in an HTML format where the desired headers are highlighted.