Discourse-based Text Summarization using RST

You may use any programming language or tool(s) to do this assignment. Please submit the following documents:

1. A PDF report answering the questions mentioned below.
2. Your code and README file in a tarball archive.

References:

- Mann & Thompson (1988). *Rhetorical Structure Theory: Toward a Functional Theory of Text Organization.*
- Pretrained RST Parser: isanlp_rst

Task: Generate extractive summaries of paragraphs using Rhetorical Structure Theory (RST). You will use a pretrained RST parser to obtain discourse structures and design simple rule-based methods to select key sentences for summarization.

1. (10 points) Run the pretrained `isanlp_rst` parser on the ten English paragraphs provided in the following file: `paragraph.txt`. For each paragraph, obtain and display the discourse tree and the relation labels generated by the parser.

2. (10 points) From the parsed output, extract the nucleus/satellite labels and relation types for each paragraph. Prepare this structured information for use in the summarization step.

3. (10 points) Using the extracted RST information, design a set of rules to select important sentences for summarization. You may, for instance, prefer nucleus sentences, give higher weight to top-level nodes, or retain sentences linked by relations such as *Result* or *Summary*. Generate one- to two-sentence summaries for each paragraph and include both the original paragraph and your generated summary in the report.

4. (10 points) For a modern baseline, generate a two to three-sentence summary for each paragraph using ChatGPT. Compare your rule-based summaries with these LLM-generated summaries. Report the Accuracy and F1-score of your summaries with respect to reference sentences or manually identified key content.

5. (5 points) Provide a brief discussion of your results. Summarize your observations about the summarization quality, note which rule patterns worked well or failed, and comment on possible ways the system could be improved.

Submission format: `rollno_a5.tar`