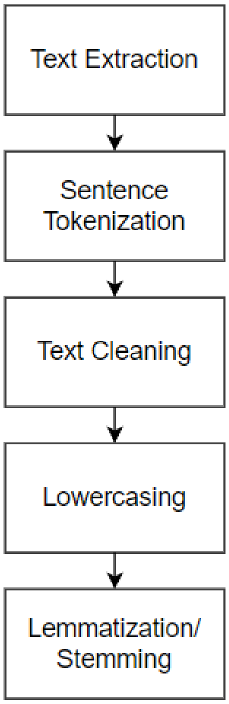
DEEP NEURAL NLP - ASSIGNMENT 2
RESEARCH PAPER SIMPLIFICATION PIPELINE DESIGN

Submitted by:
Jainisha Choksi (202211019)
Hinal Desai (202211035)
Man Desai (202211040)
Rutvik Prajapati (202211053)

1. Pre-Processing Module:



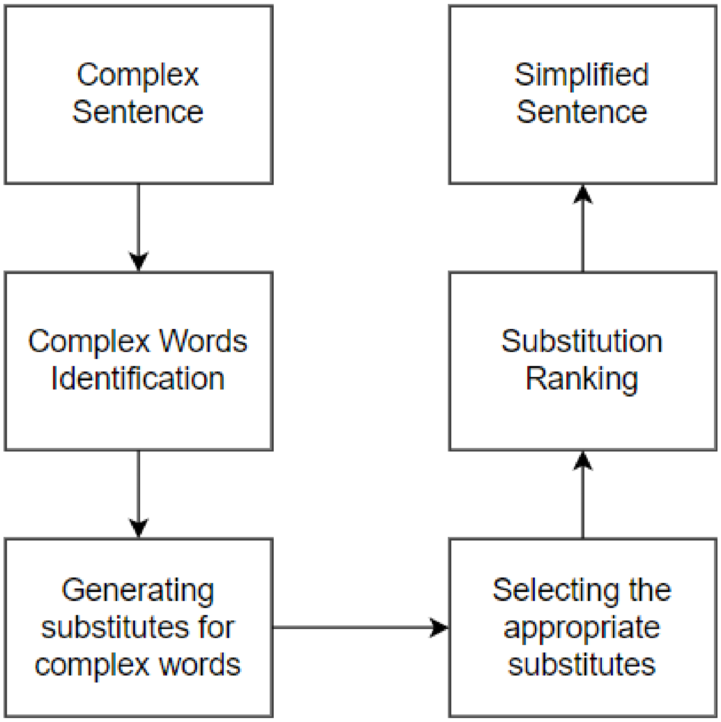
Description of the modules involved:

- **Text Extraction:** involves extracting all of the text content from the research paper.
- **Sentence Tokenization**: involves splitting/tokenizing the extracted text into sentences. The linguistic unit is sentence.
- **Text Cleaning**: involves removing the non-essential elements such as headers, footers, citations, urls, non-textual content and converting numbers to text etc.
- **Lowercasing:** involves lowercasing all the tokenized text for consistency.
- **Lemmatization/Stemming:** involves reducing the words into their base forms.

2. Definition of simplification:

Simplifying means to modify the content and structure of the text so as to make it easier to read and understand, but keeping the main idea and original meaning intact. It involves rephrasing the complex sentences into simpler sentences while retaining the original meaning.

Example:


Complex Sentence
Complex Words Identification
Generating substitutes for complex words
Selecting the appropriate substitutes
Substitution Ranking
Simplified Sentence

3. Scope of Simplification:

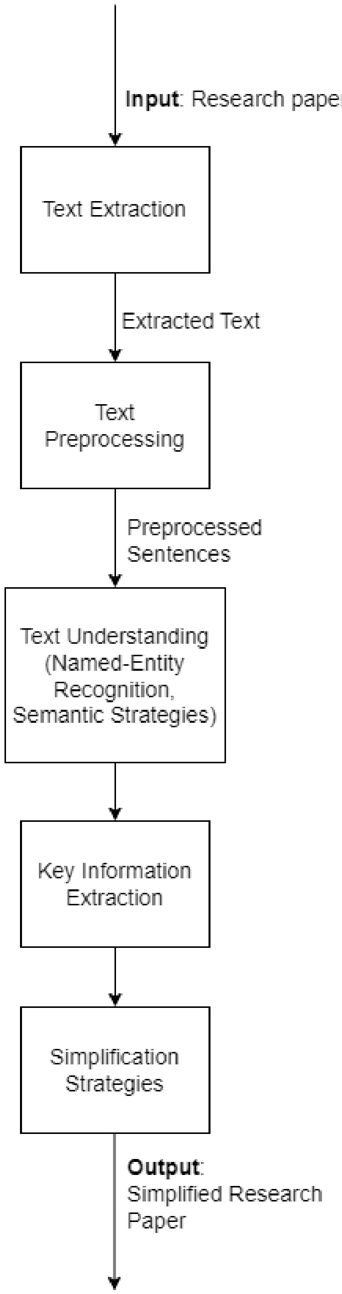
Given a research paper as an input, we aim to produce a simplified version of that research paper with a simpler vocabulary and sentence structure while preserving the main ideas in the original sentence. This is different from text summarization in which a very short summary of the input text is provided to capture only the main ideas of the text. Our focus is on **lexical simplification.**

Lexical simplification: replace complex words with their simpler synonyms without changing the grammatical meaning of the text.

4. Problem Formulation:

Input: Research paper or document in natural language text format.
Output: Simplified version of the input text, with complex sentences rephrased.
Loss function: Cross entropy loss function (negative log likelihood)

5. Methodology (at broad level):


Input: Research paper
Text Extraction
Extracted Text
Text Preprocessing
Preprocessed Sentences
Text Understanding (Named-Entity Recognition, Semantic Strategies)
Key Information Extraction
Simplification Strategies
Output: Simplified Research Paper

Model type: Sequence-to-Sequence models (based on RNNs and transformers)

6. Timeline for completion:

Week number	**Tasks**
1	Problem formulation, preparing blueprint, finding simplification strategies
2	Text extraction and preprocessing
3	Finalizing the simplification strategies, Text understanding (NLU, semantic strategies)
4	Model building and training
5	Tweaking the model, Evaluation

7. Task Delegation among members:

Team Member	**Tasks**
Jainisha Choksi (202211019)	● Text Extraction ● Text Understanding ● Model building and training
Hinal Desai (202211035)	● Key information extraction from preprocessed text ● Text Understanding ● Model building and training
Man Desai (202211040)	● Text preprocessing ● Simplification strategies
Rutvik Prajapati (202211053)	● Text preprocessing ● Simplification strategies

8. References:

- https://medium.com/nlplanet/two-minutes-nlp-quick-intro-to-text-simplification-f5f9d7be4a3c
- https://arxiv.org/abs/2110.05071
- https://arxiv.org/pdf/2302.11957.pdf