Assignment 2 (CSC3160)

Introduction

In this assignment, we will focus on regular expressions and Byte-Pair Encoding (BPE). The total score is 100 marks and it will be converted into **6%** for your final grades.

[30 marks] Regular Expressions

Task 1: HTML Cleaning (30 marks)

Write Python code using regular expressions to clean the HTML webpage.
Input: `input.txt`
Reference Output: `output.txt`

[70 marks] Byte-Pair Encoding

Task 2: Implement Byte-Pair Encoding (BPE) (30 marks)

1. Implement a Byte-Pair Encoding (BPE) algorithm to learn subword tokens. Validate your implementation with the following setup. In every iteration, track the tokens with the most frequency and the occurrence.
   - The text is "aaabdaaababc aa"
   - The number of merges $k = 2$
   - the expected tokenization output is "{aa}{ab}d{aa}{ab}{ab}c {aa}"', where '{' and '}' indicate the new tokens.
2. Apply BPE to the vocabulary of `output.txt`. In every iteration, track the tokens with the most frequency and the occurrence. Set the number of merges $k = 20$.
3. Among new tokens in Task 2.2, what are the three longest tokens. Do not list the component of the listed word. For example, if you have two tokens "speech" and "eech", please only put "speech". Discuss the reasons why those three tokens were selected.
4. Set the number of merges $k = 100$ and discuss the results.

You may use libraries such as NLTK for N-gram modeling. Refer to the lecture notes on BPE.