STIN3034 NATURAL LANGUAGE PROCESSING

Assignment 2: Creating Context Free Grammar and Parsing

(10% due 18 May 2023)

This assignment will contribute 10% of your overall grading for this course. The objective of this assignment is to enable students to analyze sentence structure by constructing their own Context Free Grammar (CFG) using NLTK.

Your tasks:

1. This assignment should be done in a group of 4/5 students.
2. Each group chooses **Three (3)** newspaper articles from any local English/Malay newspapers on the topic of your group's interest (e.g. air travel, road accidents, local sports, technology etc.). Ensure all 3 articles is from the same interest.
3. Extract the first 10 sentences from each of the chosen article. Create a text file named yourgroupname_myData.txt to store the extracted sentences. (Should have 30 sentences).
4. Construct CFG for the identified sentences. Create and edit your CFG in a text file and name it yourgroupname_myCFG.cfg.
5. Write a program in python to read each sentence from the text file and parse it using the CFG that you have created. The python program should be written in file named yourgroupname_myProgram.py.
6. For the parser, compare the performance of the top-down, bottom-up, and left-corner parsers available in NLTK.
7. The output of your program is the resulting parse tree for each of the 30 sentences for each of the 3 parsers.
8. (Bonus mark) Then, use *timeit* to log the amount of time each parser takes on the same sentence.
9. Your program should run all three parsers on all3 sentences, and (bonus mark: prints a 3-by-30 grid of times, as well as row and column for the totals.)
10. (Bonus mark) Add a discussion on your findings.
11. Submit the following softcopy - All the files (yourgroupname_myData.txt, yourgroupname_myCFG.cfg, yourgroupname_myProgram.py., a word file containing screenshot of your output and discussion on the *timeit* results) should be zipped into yourgroupname_assignment2 and uploaded in the UUM Online Learning system no later than 11.59pm Thursday 18th May 2023.
12. Late submissions will be penalized.