# Abstract

A Website on the Internet has shown a lot of information content in each block. Unlike conventional data or text, web pages not only have a main content but also typically contain a large amount of information that is not part of the main content of the pages, e.g., navigation bars, copyright, user guide, links, synopsis and also advertisement. The blocks information that is not the main content or irrelevant information in web pages is called web pages noise.

On this final project, feature weighting technique will be used to improve performance of classification with detection the noisy information in web pages. First, web pages will be modelled with structure tree Documents Object Model (DOM) tree and Compressed Structure Tree (CST) to capture the common structure and compare information block in a website. The Information that is captured will be used to measure and evaluate the importance of each node which is built from Compressed Structure Tree.

Based on the CST and the importance of weighting value, this method will put on a weight to each feature in each content block. The weighting result will be used to web mining process.

**Keyword:** *CST, DOM,* Noise Detection, Elimination, web mining