ABSTRACT

The development of internet technology increases the number of scientific documents including scientific journals. The large quantity of these documents affects the time required to process information. By extracting keyword which represents the content of document observed, information processing on the entire document is no longer required. However, extracting keyword manually is both ineffective and inefficient as it is time and resource consuming. Therefore, automation of keyword extraction is applied to handle the drawbacks caused by manual keyword extraction.

In this final project, Conditional Random Field (CRF) model will be implemented to extract keywords from document by viewing the keyword extraction process as a sequence labelling process. This model requires training process to produce optimum feature supporting parameters. Testing process will be conducted to figure out the influences of factors such as number of features used, feature extraction approach used, the number of training documents, and the involvement of *stopwords* elimination on keyword extraction performance achieved by the system. The testing process will be conducted on two groups of document (medicine and public health documents) that have different rate of content homogeinity in order to figure out performance achieved on each document group.

The result shows that best performance of keyword extraction on medicine documents produces *precision* of 0.3892, *recall* of 0.714, and *f-measure* of 0.4648 while the best performance of keyword extraction on public health documents produces *precision* of 0.2833, *recall* of 0.692, and *f-measure* of 0.3901. The increase of the number of feature used and involvement of *stopwords* elimination produce lower performance of keyword extraction.

Keywords : Keyword Extraction, Keyword, *Conditional Random Field* (CRF) Model