OPTICAL CHARACTER RECOGNITION (OCR) TECHNOLOGY APPLIED ON DOCUMENT CLASSIFICATION TOWARDS SMART DOCUMENT MANAGEMENT SYSTEM

Jenie L. Plender (1), Melbert R. Bonotan (2), Maria Alexis A. Barbosa (3), Donalyn A. Plaza (4)

College of Computing and Information Sciences, Caraga State University, Ampayon, Butuan City, Philippines
Email: jlplender@carsu.edu.ph; mrbonotan@carsu.edu.ph; malexisbarbosa@gmail.com; daplaza@gmail.com

Abstract: Document Classification is one important office operational process especially on program accreditation. Program accreditation is a significant milestone in educational institutions where a program is subjected for thorough assessment. Caraga State University (CSU) – Ampayon Campus, Butuan City Philippines, endeavors to maintain quality standards by submitting its curricular programs to accreditation. To assess programs' performance, necessary pieces of evidence and support documents are required to be compiled based on the AACCUP OBQA Survey Instrument for evaluation. Preparing and collecting sufficient documents is such tedious work and it takes ample time and manpower to complete the tasks on schedule despite the heavy workload of the faculty members and/or accreditation taskforce. To address this downside on the accreditation process, a Smart Document Management System is eminent. The system utilizes the OCR Technology to scan and read documents for classification pre-processing using the Fuzzywuzzy application. The documents in pdf format will be classified against the ten (10) areas of the AACCUP OBQA Survey Instrument using the pre-determined rules/keywords. Technically, raw data (pdf document) is converted to system object data for data matching. The dataset containing the pre-determined keywords extracted from the survey instrument are matched to the keywords extracted from the scanned documents. Upon successful data matching, the documents is automatically tagged to a specific area in the survey instrument. The system-generated document classification results are compared to that of a Human (Accreditation in-charge) and both machine and human output yielded a 100% match. This result indicates that this study contributes to the enhancement of the current document collection of CSU and a baseline structure towards smarter document management system.

Keywords: *Accreditation, Document Tagging, Optical Character Recognition (OCR), Smart Document Management System*