RESUMEN

La Secretaria Nacional de Educación Superior, Ciencia, Tecnología e Innovación (SENESCYT), concede becas a los estudiantes universitarios que desean estudiar en el extranjero. Para la aplicación de la mencionada beca, los estudiantes deben buscar por si solos información de las universidades con las cuales el estado ecuatoriano tiene convenio, pero al realizar esta búsqueda se encuentran con la gran pregunta ¿Qué voy a estudiar?, ¿Cuál es la carrera que debo elegir?, esta ha sido la razón para implementar un sistema que ayude a la selección, búsqueda de información y recomendación de carreras para los postulantes de becas del Senescyt. El presente artículo detalla la implementación de un sistema de Búsqueda Documental genérico, el cual puede ejecutar de forma independiente búsquedas de texto de páginas web y archivos PDF de los cuales su contenido se encuentra almacenado en una base de datos documental. El producto final desarrollado y que es la parte central de este proyecto de investigación es el desarrollo de un web crawler conjuntamente con un sistema de búsqueda e indexación de documentos que integran el uso de las librerías Apache Tika y Apache Lucene, dando como resultados las funcionalidades de recolección de información, extracción de contenido en texto plano, indexación de documentos y búsqueda para facilitar la integración con el sistema web AQUINO.

PALABRAS CLAVE:
WEB CRAWLER
BÚSQUEDA
TIKA
LUCENE
DOCUMENTAL
INDEXACIÓN
REPOSITORIOS
BASE DE DATOS NOSQL.

ABSTRACT

The Ministry of Higher Education, Science, Technology and Innovation (SENESCYT) award scholarships to college students who wish to study abroad. For the application of that scholarship , students must find their own information of the universities with which the Ecuadorian state has an agreement , but to perform this search are the big question What will I study? , What is the career I choose?, this was the reason for implementing a system to help the selection, information search and recommendation racing for applicants Senescyt scholarship. This article details the implementation of a generic document search system, which can run independently text searches of web pages and PDF files which content is stored in a document database. The final product developed and the central part of this research project is the development of a web crawler conjunction with a search system and indexing of documents that integrate the use of libraries Apache Lucene and Apache Tika, giving as a result the data collection capabilities, content extraction in plain text, indexing and searching documents to facilitate integration with web system AQUINO.

KEYWORDS:
WEB CRAWLER
SEARCH
TIKA
LUCENE
DOCUMENTARY
INDEXING
REPOSITORIES
NoSQL DATABASE.