# Projeto de Formatura – Turmas 2016

## PCS - Departamento de Engenharia de Computação e Sistemas Digitais

Engenharia Elétrica – Ênfase Computação

**Tema:** Particle: Ferramenta de Busca Bibliográfica Baseada em Conceitos de Mineração de Texto

## Objetivo

O objetivo do trabalho realizado foi desenvolver uma ferramenta de busca bibliográfica de artigos científicos que permitisse extrair informações e indexar documentos do gênero no formato PDF.

## Projeto

O sistema construído foi baseado em alguns conceitos empregados por ferramentas de busca de páginas web, tais como mineração de texto e algoritmos de ordenação de resultados de busca. As técnicas de mineração de texto foram aplicadas na extração do conteúdo textual de documentos PDF para a obtenção de palavras-chave, por meio da análise do conteúdo do tópico de resumo do artigo.

## Arquitetura

A figura 1 mostra a arquitetura geral do sistema, cujo os componentes são descrito a seguir.

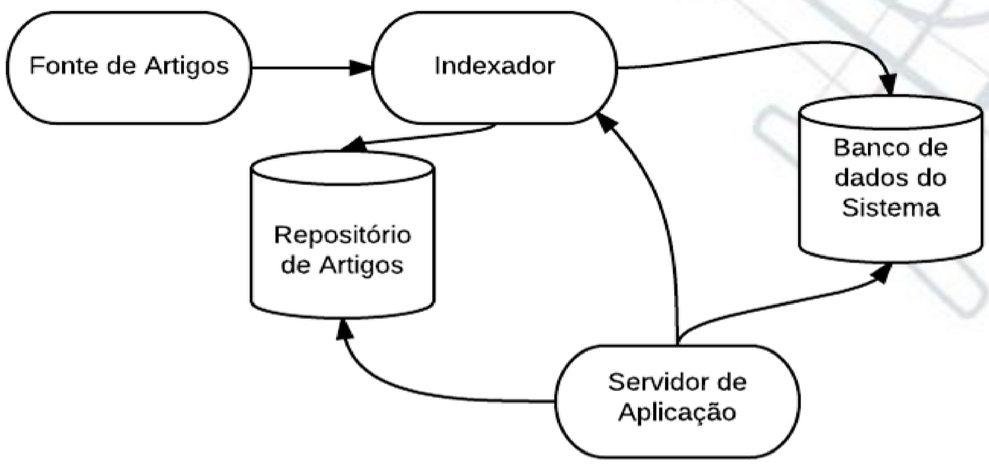


Figura 1 - Arquitetura do Sistema

### Fonte de Artigos

Os artigos podem ser extraídos de páginas web ou repositórios que disponibilizam seus artigos para indexação na ferramenta. A título de ilustração, criou-se um *bot* para indexação dos artigos disponibilizados no site do LTA (Laboratório de Linguagens e Técnicas Adaptativas).

### Indexador

O indexador é o componente responsável por extrair informações, por meio de técnicas de mineração de texto, contidas nos documentos dos artigos científicos.

### Servidor de Aplicação

Este componente tem a função de processar as requisições feitas pelas páginas da ferramenta acessadas pelo usuário.

### Banco de dados e Repositório de Artigos

No banco de dados são armazenadas e indexadas as informações extraídas pelo indexador, e o repositório de artigos é responsável por arquivar os documentos submetidos ao sistema.

## Extração de dados de arquivos PDF

Devido à estrutura dos arquivos PDF, o conteúdo textual desse tipo de documento nem sempre disponibiliza seu conteúdo textual na mesma sequência em que se espera que o artigo deva ser lido. Para solucionar esse problema e assim extrair informações sobre o artigo, foi desenvolvido um algoritmo que busca extrair o texto do arquivo na ordem correta.
Além disso, foi desenvolvida uma funcionalidade que permite ao usuário descrever a particular estrutura de seu artigo, para que seja possível uma extração mais precisa dos dados nele contido.

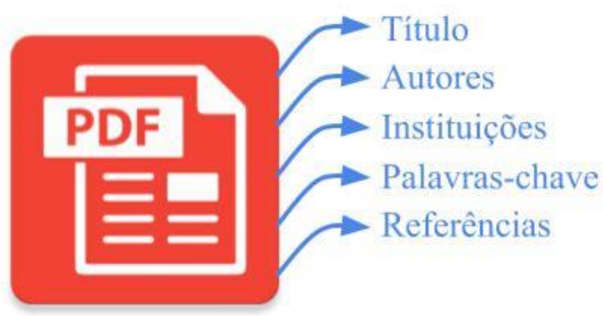


Figura 2 - Alguns metadados extraídos de um artigo típico

## Extração de palavras-chave

Devido à ausência de informações referentes às palavras-chave dos artigos em algumas normas, foram implementados dois algoritmos clássicos, baseados em processamento de linguagem natural, para determinar tais palavras-chave: RAKE e TextRank.

Integrante: Luís Arthur Bighetti Valini

Professor Orientador: João José Neto