Plataforma de busca e indexação em nuvem privada

Renato dos Santos Ribeiro [1]; Prof. Dr. Andreiwid Sheffer Correa [1];

[1]- Instituto Federal de Educação, Ciência e Tecnologia de São Paulo – Campus Campinas

Objetivo

Desenvolver um sistema web como diretório virtual para indexação de conteúdo de texto a partir de documentos presentes nas redes corporativas.

Introdução

O armazenamento dos dados na nuvem permite reduzir os custos e centralizar informações na empresa. Além disso, permite a adição e remoção de servidores de forma muito mais rápida e a um custo de manutenção mais baixo.

Como consequência, o volume de arquivos se tornou imenso, o que pode tornar a busca por um determinado conteúdo custosa.

Desse modo, o objetivo deste trabalho é o desenvolvimento de um sistema web que possa servir de diretório virtual. Esse diretório fará a indexação de forma automática de conteúdo presente nos documentos de uma rede corporativa, com suporte para formatos populares como o PDF e os provenientes das suítes de escritório mas conhecidas, permitindo também a indexação de conteúdo de texto presente em arquivos de imagens.

Materiais e Métodos

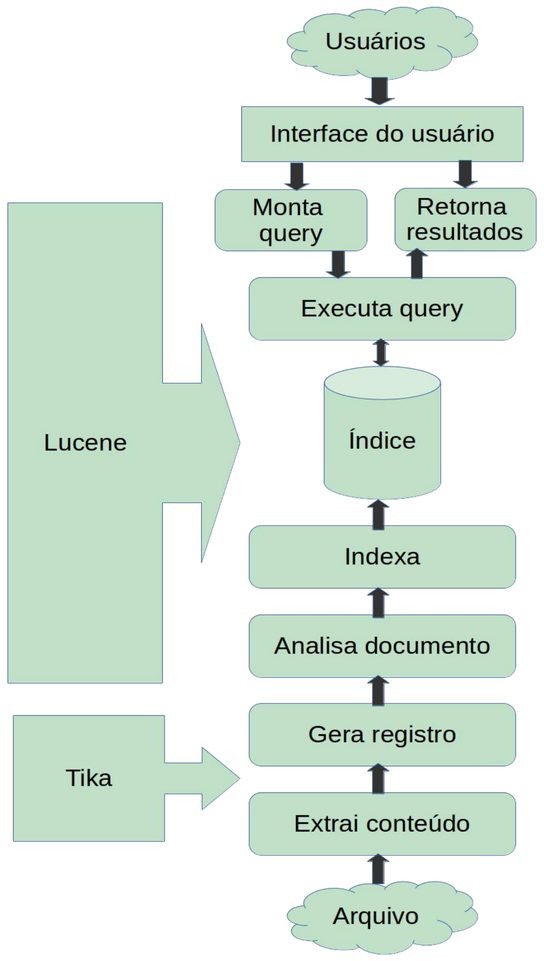
O sistema web utilizará como linguagem Groovy, com a interface web em React.

O servidor utilizará as seguintes tecnologias para realizar o processo de indexação e disponibilizar a busca:

- Tika
- Lucene
- Tesseract (ferramenta para *Optical Character Recognition,* OCR, para extração de conteúdo de texto em imagens)

A Figura 1 representa o fluxo de informações no contexto do sistema com a definição das responsabilidades de cada um dos módulos mencionados:

Figura 1: Responsabilidade dos módulos



Fonte: Produzido pelo autor (adaptado de *Lucene in Action (2)*)

Conforme mostrado na Figura 1, o Tika será utilizado para extrair o conteúdo e metadados utilizados para gerar o registro que será indexado.

O Tika possui licença Apache versão 2.0 (1), e possui suporte aos formatos mais populares de arquivos, o que permite que seja utilizado como única bilblioteca para a extração de conteúdo para todos os arquivos que serão suportados pelo sistema.

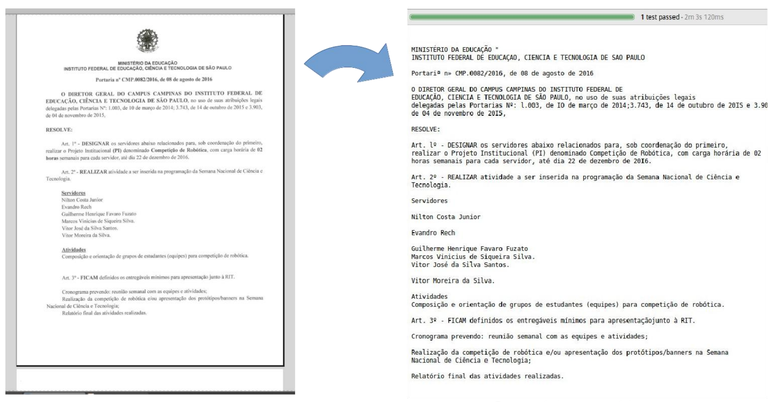
O Lucene possui licença Apache versão 2.0 (2), e será utilizado para gravar o conteúdo extraído pelo Tika. Também será utilizado para realizar as buscas nos conteúdos indexados. O Lucene, possui suporte a *highlights*, ou seja, ele traz o trecho em que o termo buscado foi encontrado, dando destaque ao termo. Esta é uma funcionalidade que traz bastante valor ao sistema.

O Tesseract possui licença licença Apache versão 2.0 (3) e servirá para suporte ao Tika, uma vez que ele possui integraçao que permite ao Tika extrair conteúdo de texto por OCR. Seu uso é um pré requisito, para extrair conteúdo de texto de imagens, e para extrair conteúdo de texto em imagens dentro de arquivos.

Resultados preliminares

Como resultado preliminar tem-se o sistema rodando em ambiente local, parcialmente implementado.O código do sistema realiza a extração de conteúdo de texto e imagem, além de imagens inseridas dentro de arquivos PDF. A figura 2 traz uma demonstração:

Figura 2: Demonstração de extração



PDF Digitalizado

Conteúdo extraído

Fonte: Produzido pelo autor

Conclusão

A partir dos resultados preliminares, o sistema extrai com sucesso arquivos com texto simples, imagens, arquivos PDF e arquivos produzidos pelos pacotes office mais populares. Dado ao resultado até agora alcançado, será desenvolvido a API para disponibilização dos dados, e a interface para consumo dos mesmos.

Referências

(1) Mattmann, C. A.; Zitting J. L. 2012. Tika in Action. Editora Manning Publications. 24 p. Ebook.

(2) McCandless, M; Hatcher E; Gospodnetic O. 2010. Lucene in Action. Editora Manning Publications: 6-10. Ebook.

(3) Zdenko Podobny. Documentação. Disponível em: <https://github.com/tesseract-ocr/tesseract> . Acesso em 09 set. 2017.