PROCESSO SELETIVO 001/2020/PJC/MT

TESTE PRÁTICO

ADMINISTRADOR DE DADOS

3º ETAPA

PROPOSTA INDEXAÇÃO DE DOCUMENTOS

Autor: Ewerton Luiz Utrago Gonçalves

Soluções Open Source

Apache TIKA - O kit de ferramentas Apache Tika ™ detecta e extrai metadados e texto de mais de mil tipos de arquivos diferentes (como PPT, XLS e PDF). Todos esses tipos de arquivo podem ser analisados por meio de uma única interface, tornando o Tika útil para indexação de mecanismo de pesquisa, análise de conteúdo, tradução e muito mais.[TRADUDIZO AUTOMATICAMENTE][TIKA, 2021]

Postgres text search - é a técnica de indexação, pesquisa e relevância do PostgreSQL, que utiliza um conjunto de regras naturais para adicionar suporte a modo verbal (derivações de um verbo), através da utilização de dicionários e algoritmos específicos. Neste contexto, este artigo apresenta de forma prática o uso deste recurso no PostgreSQL.[DEVMEDIA 2021]

Ferramenta Similar

Search engines: Postgres Full Text Search - Trata-se de um plug-in para Moodle [...] *implementado usando indexação de texto completo Postgres com indexação de arquivo opcional usando Apache Tika*. Disponivel em: https://moodle.org/plugins/pluginversion.php?id=14948

Recursos necessários

- Servidor com SO linux instaldo
- PostgreSQL - Sistema de Banco de Dados
- PHP - Linguagem de programação
- Apache TIKA - Recurso para converter os binarios em texto.
- Nginx - Servidor HTTP
- *Postgres text search* - Recurso para indexação e busca dos textos convertidos pelo TIKA.
- Java - Ultizado para rodar o Apache TIKA.

Etapas da implentação

1. Montar servidor Web com os seguintes recursos instalados: PostgresSQL, PHP, Apache TIka, Nginx e Java

2. No banco de dados, na tabela do documento, criar uma coluna do tipo Text para armazenar o texto do documento e uma coluna do tipo tsvector para armazenar a indexação do texto. Criar também um indice para a coluna tsvector.

3. No bando de dados criar uma trigger que atualizar a coluna tsvector toda vez que um texto for inserido ou atualizado.

4. Criar um script PHP que utilize o Apache Tika para converter em texto os binarios dos arquivos e salve na coluna text criada no passo 2. (usar shell_exec para executar o jar)

5. Criar um modulo do sistema para Buscas, onde o usuario possa inserir palavras ou frases completas para busca. Utilizar o PostgreSQL Full Text para fazer essa busca. O sistema de ter a opção de busca simples(apenas em resumos e sumário) e busca completa(texto inteiro).

Pros

- Busca no texto inteiro
- Mais resultados de busca em uma só consulta
- Mais chances do usuario encontrar o que pesquisou
- Possibilidade de rankear documentos de acordo com o numero de referências .

Contras

- Maior processamento
- Maior volume de dados

Resultados Esperados

Esperesse com a indexação que os usuarios recebem resultados mais satisfatorios e que consigam encontrar o que buscam.

REFERENCIAS

DEVMEDIA, 2021 *PostgreSQL: Full Text Search na prática*, https://www.devmedia.com.br/postgresql-full-text-search-na-pratica/21362#:~:text=De%20que%20se%20trata%20o,de%20dicion%C3%A1rios%20e%20algoritmos%20espec%C3%ADficos.

TIKA, 2021 - *Apache TIKA* https://tika.apache.org/

MOODLE 2021, Search engines: Postgres Full Text Search , https://moodle.org/plugins/pluginversion.php?id=14948

Shahriar Shovon, PostgreSQL Full Text Search Tutorial,LINUXHINT 2021, https://linuxhint.com/postgresql-full-text-search-tutorial/

Full text search in milliseconds with PostgreSQL, Lateral 2021, https://blog.lateral.io/2015/05/full-text-search-in-milliseconds-with-postgresql/

Juliano Atanazio, Full Text Search, Slideshare 2013, https://pt.slideshare.net/spjuliano/fts-26392077