1. **Introdução**

Podemos citar como exemplo de utilização e aplicação de OCR atualmente, uma aplicação chamada Indexador e processador de evidencias digitais, mais conhecida como IPED. Essa aplicação foi desenvolvida por peritos da polícia federal afim de processamentos dos arquivos apreendidos durante a operação lava-jato.

Um dos recursos do IPED, é a OCR de imagens e PDFs e detecção de imagens contento textos como digitalizações. Ferramenta desenvolvida através da utilização do metadado OcrCharCount, Tesseract OCR, e com a linguagem de programação JAVA.

1. **Recursos necessários**

Levando em consideração a utilização do OCR em JAVA, considerando esse exemplo, os recursos necessários seriam: A codificação em JAVA, que iria envolver desenvolvedores dessa linguagem, o Tesseract ORC, conforme recomendado no edital, que é open source, e disponível no GIT, e claro, o banco de dados onde estaríamos criando os índices.

1. **Etapas da Implementação**
2. A possível hipótese de implementação consistiria em uma aplicação JAVA (ou poderia ser também outra linguagem que aceitaria o Tesseract OCR), simplificada a apenas receber um arquivo, podendo ser através de ‘upload’ (escolhendo o arquivo) ou de maneira ‘arrastar e soltar’ o arquivo no EXE. A aplicação com a utilização do OCR iria analisar o arquivo escolhido, podendo ser arquivo texto, PDF ou imagem, e com o mapeamento dos caracteres criar palavras chaves com base em quantas vezes elas se repetiriam no documento.
3. Com base na análise e no arquivo escolhido, a aplicação já iria fazer um INSERT no banco de dados com o arquivo escolhido (um exemplo de banco de arquivos), e com as palavras chaves mais recorrentes já seria feito a criação dos índices (tipo B-tree ou tipo GIST, ou outra denominação de índice), juntamente a qual linha do banco é pertencente a esse index, cujo qual, nos levaria ao documento que fora anexado na aplicação.
4. **Prós e Contras**

Por parte do contra, teria que ser prototipada uma ferramenta possivelmente do zero que atenta os recursos do Tesseract OCR, juntamente com seu desenvolvimento.

Por parte do Prós, teríamos uma aplicação de indexação e conversão de OCR já dinamicamente criando (INSETS) registros no banco de dados conforme fosse sendo escolhidos os arquivos que forem pertinentes a serem salvos no banco de dados.

1. **Resultados Esperados**

Por parte dos resultados esperados, teríamos basicamente o ponto forte citado do item 4, que seria uma indexação e conversão de OCR já dinamicamente no banco de dados, o que propiciaria além da agilidade da criação do registro, seria também o controle sobre os índices criados, que por exemplo poderia ser criado conforme a contagem das palavras chaves do documento.