Crawler de Clientes em PDF

Este projeto implementa um crawler em Python para buscar nomes de clientes de um arquivo Excel em documentos PDF, utilizando uma tolerância de similaridade de 80%.

Requisitos

Para executar este crawler, você precisará das seguintes bibliotecas Python:

- `pandas`
- `openpyxl`
- `PyPDF2`
- `fuzzywuzzy`
- `python-Levenshtein`

Você pode instalá-las usando pip:

```
pip install pandas openpyxl PyPDF2 fuzzywuzzy python-Levenshtein
```

Estrutura do Projeto

```
crawler_clientes/
├── input/
│   ├── clientes.xlsx       # Arquivo Excel com a lista de
clientes
│   └── documento.pdf       # Arquivo PDF para busca
├── output/
│   └── resultados_busca.xlsx # Arquivo de saída com os
resultados da busca
└── crawler.py              # Script principal do crawler
```

Como Usar

1. **Prepare seus arquivos:**
    - Coloque seu arquivo Excel (`.xlsx`) contendo a lista de clientes na pasta `input/` . A primeira coluna da primeira aba será usada para extrair os nomes dos clientes.

    - Coloque seus arquivos PDF (`.pdf`) na pasta `input/` .

2. **Execute o crawler:** Abra o terminal na raiz do projeto (`crawler_clientes/`) e execute o script `crawler.py` :

   `bash python3.11 crawler.py`

3. **Verifique os resultados:** Um novo arquivo Excel chamado `resultados_busca.xlsx` será gerado na pasta `output/` . Este arquivo conterá a lista de clientes do seu Excel original, indicando se cada cliente foi encontrado no PDF e o nível de similaridade.

Configuração de Similaridade

A tolerância de similaridade padrão é de 80%. Você pode ajustar este valor na função `find_matches` dentro do arquivo `crawler.py` :

```
def find_matches(client_list, pdf_text, threshold=80):
    # ...
```

Exemplo de Uso

Para testar o crawler, você pode usar os arquivos de exemplo gerados durante o desenvolvimento. Eles estão localizados na pasta `input/` .