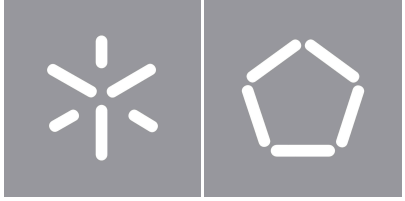
Universidade do Minho
Escola de Engenharia

Nome completo do autor

Título Título Título Título Título Título Título Título Título Título Título Título Título Título Título

setembro 2023



Universidade do Minho
Escola de Engenharia

Nome completo do autor

Título Título Título Título Título Título Título Título Título Título Título Título Título Título

Dissertação de Mestrado
Mestrado em Engenharia Informática

Trabalho efetuado sob a orientação de
Nome do Orientador
Nome do Coorientador

setembro 2023

Direitos de Autor e Condições de Utilização do Trabalho por Terceiros

Este é um trabalho académico que pode ser utilizado por terceiros desde que respeitadas as regras e boas práticas internacionalmente aceites, no que concerne aos direitos de autor e direitos conexos.

Assim, o presente trabalho pode ser utilizado nos termos previstos na licença abaixo indicada.

Caso o utilizador necessite de permissão para poder fazer um uso do trabalho em condições não previstas no licenciamento indicado, deverá contactar o autor, através do RepositóriUM da Universidade do Minho.

Licença concedida aos utilizadores deste trabalho:

[Caso o autor pretenda usar uma das licenças Creative Commons, deve escolher e deixar apenas um dos seguintes ícones e respetivo lettering e URL, eliminando o texto em itálico que se lhe segue. Contudo, é possível optar por outro tipo de licença, devendo, nesse caso, ser incluída a informação necessária adaptando devidamente esta minuta]

CC BY

https://creativecommons.org/licenses/by/4.0/ *[Esta licença permite que outros distribuam, remixem, adaptem e criem a partir do seu trabalho, mesmo para fins comerciais, desde que lhe atribuam o devido crédito pela criação original. É a licença mais flexível de todas as licenças disponíveis. É recomendada para maximizar a disseminação e uso dos materiais licenciados.]*

CC BY-SA

https://creativecommons.org/licenses/by-sa/4.0/ *[Esta licença permite que outros remisturem, adaptem e criem a partir do seu trabalho, mesmo para fins comerciais, desde que lhe atribuam o devido crédito e que licenciem as novas criações ao abrigo de termos idênticos. Esta licença costuma ser comparada com as licenças de software livre e de código aberto «copyleft». Todos os trabalhos novos baseados no seu terão a mesma licença, portanto quaisquer trabalhos derivados também permitirão o uso comercial. Esta é a licença usada pela Wikipédia e é recomendada para materiais que seriam beneficiados com a incorporação de conteúdos da Wikipédia e de outros projetos com licenciamento semelhante.]*

CC BY-ND

https://creativecommons.org/licenses/by-nd/4.0/ *[Esta licença permite que outras pessoas usem o seu trabalho para qualquer fim, incluindo para fins comerciais. Contudo, o trabalho, na forma adaptada, não poderá ser partilhado com outras pessoas e têm que lhe ser atribuídos os devidos créditos.]*

CC BY-NC

https://creativecommons.org/licenses/by-nc/4.0/ *[Esta licença permite que outros remisturem, adaptem e criem a partir do seu trabalho para fins não comerciais, e embora os novos trabalhos tenham de lhe atribuir o devido crédito e não possam ser usados para fins comerciais, eles não têm de licenciar esses trabalhos derivados ao abrigo dos mesmos termos.]*

CC BY-NC-SA

https://creativecommons.org/licenses/by-nc-sa/4.0/ *[Esta licença permite que outros remisturem, adaptem e criem a partir do seu trabalho para fins não comerciais, desde que lhe atribuam a si o devido crédito e que licenciem as novas criações ao abrigo de termos idênticos.]*

CC BY-NC-ND

https://creativecommons.org/licenses/by-nc-nd/4.0/ *[Esta é a mais restritiva das nossas seis licenças principais, só permitindo que outros façam download dos seus trabalhos e os comparti-*

Agradecimentos

Escreva aqui os seus agradecimentos. Não se esqueça de mencionar, caso seja esse o caso, os projetos e bolsas dos quais se beneficiou enquanto fazia a sua investigação. Pergunte ao seu orientador sobre o formato específico a ser usado. (As agências de financiamento são bastante rigorosas quanto a isso.)

Declaração de Integridade

Declaro ter atuado com integridade na elaboração do presente trabalho académico e confirmo que não recorri à prática de plágio nem a qualquer forma de utilização indevida ou falsificação de informações ou resultados em nenhuma das etapas conducente à sua elaboração.

Mais declaro que conheço e que respeitei o Código de Conduta Ética da Universidade do Minho.

Universidade do Minho, Braga, setembro 2023

Nome completo do autor

Resumo

Escrever aqui o resumo (pt)

Palavras-chave palavras, chave, aqui, separadas, por, vírgulas

Abstract

Write abstract here (en)

Keywords keywords, here, comma, separated

Conteúdo

Lista de Figuras

Lista de Tabelas

Parte I

Material Introdutório

Capítulo 1

Introdução

As tecnologias de reconhecimento óptico de caracteres (OCR), tem um papel fundamental na conservação, disponibilização e proliferação de documentos de épocas anteriores à digitalização, ou de origem física sem contrapartida digital.

A eficácia desta tecnologia é no entanto dependente de vários fatores: a qualidade das imagens alvo, como a resolução, estado do documento, coloração, qualidade/tipo de escrita; a estrutura dos documentos, quanto mais complexo, mais difícil é obter a informação de forma congruente de forma automática; linguagem do texto, sendo que por vezes diferentes tecnologias, como por exemplo o tesseract, procuram na procura de texto verificar a sua confiança na deteção com o vocabulário conhecido, o qual pode no entanto não coincidir com a época de produção do documento; etc.

Estas dependências são especialmente notórios quando se envolvem documentos menos recentes, os quais podem, além de apresentarem envelhecimento causado pelo tempo e danos pelas condições de armazenamento, devido às limitações tecnológicas, assim como por vezes à falta de convenções de formatação dos documentos, não disporem de uma consistência no formato e texto (template, alinhamento, dimensões dos caracteres,etc), usual nos documentos atuais. Estes fatores, resultam então num reconhecimento de texto não tão satisfatórios como se esperaria.

Estes documentos, são mais comummente, mais não exclusivamente, reconhecidos como anteriores à era da digitalização, sendo que o foco de trabalho será maioritaremente dirigido a documentos desta época, como jornais, revistas e outros, do século passado ou anteriores.

O seguinte documento pretende então servir como um estudo dos desafios apresentados por estes tipos de documentos perante OCR, assim como a procura de soluções para a melhoria dos resultados na deteção de texto e assim criar uma ferramenta que torne o processo de extração de informação destes tipos de documentos mais simples e confiável.

O trabalho realizado seguirá então por um processo de investigação do estado da arte, onde serão aprofundados teoricamente e na prática diferentes motores de OCR, de modo a permitir entender a sua

capacidade em tratar deste tipo de documentos, e procurar isolar diferentes dificuldades que os documentos lhes apresentam; seguido da utilização do estudo realizado para a criação de uma ferramenta ou conjunto de ferramentas que, fazendo uso destas tecnologias e mitigando ou resolvendo os problemas que elas demonstram na situação descrita, melhore o processo de extração de informação. Esta componente prática, entende então que torne possível a extração de informação dos documentos, isolando artigos ou outras peças contínuas de texto; detetar problemas nas imagens dos documento e aplicar/sugerir diferentes soluções para os corrigir; possibilitar uma modernização automática do texto detetado; habilitar a conversão do documento original (em pdf ou imagem) para um novo tipo de ficheiro com a informação extraída para text (por exemplo: html que mantenha a mesma estrutura do ficheiro original; markdown com os diferentes artigos extraídos isolados; pdf com camadas de texto sobre a imagem, limpo de blocos dispensáveis, com a sua performance de navegação melhorada).

Capítulo 2

Estado da arte

Estado da arte revisto; trabalho relacionado.

2.1 Citações

Exemplo de uma citação: **?**, cf. esta entrada em `dissertation.bib`. Outra forma de citar [**?**].

2.2 Expressões matemáticas

A equivalência massa-energia é descrita pela famosa equação

$$E = mc^2 \tag{2.1}$$

descoberta em 1905 por Albert Einstein. Em unidades naturais ($c = 1$), a fórmula expressa a identidade

$$E = m$$

2.3 Notas de rodapé

Este é um exemplo de uma nota de rodapé[1].

2.4 Acrónimos e Glossário

Dado um conjunto de números, existem métodos elementares para calcular seu **Máximo Divisor Comum**, que é abreviado como **MDC**. Este processo é semelhante ao usado para o **Mínimo múltiplo comum (MMC)**.

[1] The quick brown fox jumps over the lazy dog.

O **Latex** é especialmente adequado para documentos que incluam **matemática**. **Fórmulas** são corretamente e facilmente renderizados a partir do momento que nos habituamos aos comandos.

2.5 Índice

Neste exemplo, várias palavras-chave importantes serão usadas pelo que merecem aparecer no Índice.

Os termos no índice também podem ser aninhados .

Cf. o ficheiro `dissertation.bib` para ver algumas definições como UMinho .

Capítulo 3

O problema e os seus desafios

O problema e os seus desafios

3.1 Imagens

Exemplo de inserção de uma imagem como texto exibido,

— dentro no texto, bla-bla

— ou em formato flutuante

Figura 1: Legenda

Parte II

Core da Dissertação

Capítulo 4

Contribuição

Principais resultados e as suas evidências cientificas.

4.1 Introdução

4.2 Sumário

Capítulo 5

Aplicações

Aplicação do resultado principal (exemplos e casos de estudo)

5.1 Introdução

5.2 Sumário

Capítulo 6

Conclusões e trabalho futuro

Conclusões e trabalho futuro.

6.1 Conclusões

6.2 Perspetiva de trabalho futuro

Capítulo 7

Planeamento

7.1 Atividades

Tarefa	Out	Nov	Dez	Jan	Fev	Mar	Abr	Mai	Jun	Jul
Background e **EA**	•	•	•							
Preparação do **RPD**		•	•	•						
Contribuição				•	•	•	•	•	•	
Escrita							•	•	•	•

Tabela 1: Plano de atividades.

Índice

Parte III
Apêndices

Apêndice A
Trabalho de apoio

Resultados auxiliares.

Apêndice B
Detalhes dos resultados

Detalhes de resultados cuja extensão comprometeria a legibilidade do texto principal.

Apêndice C
Listings

Se for o caso.

Apêndice D
Ferramentas

(Se for o caso)

Utilizadores de LaTeX devem consultar TUG , o grupo de utilizadores TeX .

Coloque aqui informação sobre financiamento, projeto FCT, etc. em que o trabalho se enquadra. Deixe em branco caso contrário.