Universidade do Minho
Escola de Engenharia

# Plano de Trabalho de Dissertação

Ano Letivo 2023/2024

| | |
|---|---|
| **Nome Estudante** | Gonçalo Braz Afonso |
| **N.º Estudante** | A93178 |
| **Curso** | MIEI |
| **Título da Dissertação** (em Português) | Extração de informação de documentos estruturados antigos utilizando OCR |
| **Título da Dissertação** (em Inglês) | Extraction of information from old, structured documents using OCR |

**Enquadramento e Motivação** (150 - 200 palavras)

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 e 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 notórias 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.

**Objetivos e Resultados Esperados** (150 - 200 palavras)

Espera-se então com este trabalho construir uma solução de modo a mitigar os problemas apresentados pelos documentos supramencionados e assim tornar os resultados de OCR nestes mais satisfatórios.
Para isto, espera-se que sejam conseguidos os seguintes objetivos/tarefas:

- Investigação do estado da arte em OCR
- Criação/Seleção de casos de estudo – de modo a identificar e exemplificar os problemas alvo
- Criação de métricas de validação dos resultados de OCR
- Criação de um toolkit de OCR que apresente ferramentas úteis e independentes para o tratamento destes documentos – ex.: limpeza de blocos de Tesseract; identificação de zonas do template (como cabeçalho ou rodapé); marcador de blocos (como delimitadores ou potenciais títulos); analisador de informação do documento (como tamanho de letra normal ou quantidade de colunas); etc.

- Exploração de métodos de extração de linguagem contemporânea aos documentos de modo a permitir uma modernização do texto
    - Exemplo: utilização de documentos do projeto Gutenberg com as suas contapartidas modernizadas para a criação de um dicionário de linguagem datada e moderna
- Criação de uma ferramenta (produto final) que proporciona um pipeline dos módulos criados de forma a conseguir extrair a informação do documento com maior fidelidade e em diferentes formatos (ex.: markdown com os artigos extraídos; html que mantém o formato original do documento; pdf com camadas sobrepostas mais eficiente – por ter menos blocos de texto vazios)

No decorrer destas tarefas, espera-se a exploração de tecnologias de scripting, NLP, OCR e machine learning.

**Calendarização**

| | Set | Out | Nov | Dez | Jan | Fev | Mar | Abr | Mai | Jun |
|---|---|---|---|---|---|---|---|---|---|---|
| Investigação e estado da arte | ● | ● | | | | | | | | |
| Casos de estudo | | ● | ● | | | | | | | |
| Métricas de validação | | | ● | ● | | | | | | |
| Toolkit | | | | ● | ● | ● | ● | ● | | |
| Ferramenta Final | | | | | ● | ● | ● | ● | ● | |
| Relatório | | | | | | | | | ● | ● |

**Referências Bibliográficas** (5 - 10 referências)


- R. Smith, "An Overview of the Tesseract OCR Engine," Ninth International Conference on Document Analysis and Recognition (ICDAR 2007), Curitiba, Brazil, 2007, pp. 629-633, doi: 10.1109/ICDAR.2007.4376991.
- T. -T. -H. Nguyen, A. Jatowt, M. Coustaty, N. -V. Nguyen and A. Doucet, "Deep Statistical Analysis of OCR Errors for Effective Post-OCR Processing," 2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL), Champaign, IL, USA, 2019, pp. 29-38, doi: 10.1109/JCDL.2019.00015.
- W. Bieniecki, S. Grabowski and W. Rozenberg, "Image Preprocessing for Improving OCR Accuracy," 2007 International Conference on Perspective Technologies and Methods in MEMS Design, Lviv, UKraine, 2007, pp. 75-80, doi: 10.1109/MEMSTECH.2007.4283429.
- M. A. Souibgui and Y. Kessentini, "DE-GAN: A Conditional Generative Adversarial Network for Document Enhancement," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 3, pp. 1180-1191, 1 March 2022, doi: 10.1109/TPAMI.2020.3022406.
- V. N. Sai Rakesh Kamisetty, B. Sohan Chidvilas, S. Revathy, P. Jeyanthi, V. M. Anu and L. Mary Gladence, "Digitization of Data from Invoice using OCR," 2022 6th International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2022, pp. 1-10, doi: 10.1109/ICCMC53470.2022.9754117.
- A. B. Salah, J. p. Moreux, N. Ragot and T. Paquet, "OCR performance prediction using cross-OCR alignment," 2015 13th International Conference on Document Analysis and Recognition (ICDAR), Tunis, Tunisia, 2015, pp. 556-560, doi: 10.1109/ICDAR.2015.7333823.
- W. B. Lund and E. K. Ringger, "Error Correction with In-domain Training across Multiple OCR System Outputs," 2011 International Conference on Document Analysis and Recognition, Beijing, China, 2011, pp. 658-662, doi: 10.1109/ICDAR.2011.138.
- F. Shafait, D. Keysers and T. Breuel, "Performance Evaluation and Benchmarking of Six-Page Segmentation Algorithms," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 30, no. 6, pp. 941-954, June 2008, doi: 10.1109/TPAMI.2007.70837.


**Justificação de Coorientação** (se aplicável)

**Assinaturas**

**Estudante**

Assinado por: **Gonçalo Braz Afonso**
Num. de Identificação: 14368138
Data: 2023.10.11 11:42:00+01'00'

CHAVE MÓVEL

**Orientador** (tal como previsto no ponto 1 do Artigo 169.º do

Assinado por: **JOSÉ JOÃO ANTUNES GUIMARÃES DIAS DE ALMEIDA**
Num. de Identificação: 03676378
Data: 2023.10.11 13:52:50 +0100

**Diretor do Ciclo de Estudos**

**Orientador** (tal como previsto no ponto 3 do Artigo 169.º do RAUM. Neste caso, é obrigatório existir um Orientador pelo ponto 1 do Artigo 169.º do RAUM)

Assinatura digital qualificada com Cartão de Cidadão ou Chave Móvel Digital. Para os estudantes, nos casos em que tal não seja possível, os mesmos deverão imprimir este plano, assinar manualmente e, após digitalização, os restantes intervenientes usam a assinatura digital qualificada.

O plano de trabalho deve ser preenchido *offline* e realizado o *upload* do mesmo, depois de assinado, no formulário do requerimento de pedido de admissão à dissertação, disponível em http://dissertacao.eng.uminho.pt