



**Universidade de Aveiro**

**Ano 2021/2022**

**António Jorge  
Ferreira Ramos**

**Validação da autenticidade de documentos  
impressos**



Universidade de Aveiro  
Ano 2021/2022

**António Jorge  
Ferreira Ramos**

## **Validação da autenticidade de documentos impressos**

Dissertação apresentada à Universidade de Aveiro para cumprimento dos requisitos necessários à obtenção do grau de Mestre em Engenharia Informática, realizada sob a orientação científica do Doutor André Zúquete, Professor Auxiliar do Departamento de Eletrónica, Telecomunicações e Informática da Universidade de Aveiro

## **o júri**

presidente

Prof. Doutor João Antunes da Silva  
professor associado da Universidade de Aveiro

Prof. Doutor João Antunes da Silva  
professor associado da Universidade de Aveiro

Prof. Doutor João Antunes da Silva  
professor associado da Universidade de Aveiro

Prof. Doutor João Antunes da Silva  
professor associado da Universidade de Aveiro

**agradecimentos /  
acknowledgements**

Agradeço a ajuda dos orientadores, amigos e familiares pelo incentivo de realizar a tese.

**palavras-chave**

Marca de água, códigos de barras 128, documentos impressos, documentos digitais, segurança, Qrcode, verificação de integridade

**resumo**

**keywords**

Watermark, barcodes, textual watermark, printed documents, security, image processing

**abstract**



# Índice

Índice.....	ii
Lista de tabelas.....	iii
Lista de Figuras.....	iv
1. Introdução.....	1
1.1. Descrição do sistema existente .....	1
1.2. Objetivos.....	1
2. Estado da Arte .....	3
2.1. Código de barras .....	3
2.1.1. Análise de código de barras .....	3
2.2. Texto com marca de água .....	5
2.2.1. Ataque zero com mudança de texto .....	5
2.2.1.1. Exemplo .....	6
2.2.1.1. Vantagens.....	6
2.2.1.2. Desvantagens.....	6
2.2.2. Ataque zero sem mudança no texto .....	6
2.2.2.1. Vantagens.....	7
2.2.2.2. Desvantagens.....	7
2.2.3. Documentos texto baseados em Eigenvalues.....	7
2.2.3.1. Testes.....	7
2.2.3.4. Vantagens.....	8
2.2.3.5. Desvantagens.....	8
2.2.4. Line-Shift Coding.....	8
2.2.4.1. Word-Shift Coding.....	9
2.2.4.2. Character Coding.....	9
2.2.4.3. Vantagens dos métodos .....	9
2.2.4.4. Desvantagens dos métodos.....	9
2.2.5. ECL zero-based watermarking .....	9
2.2.5.1. Vantagens.....	9
2.2.5.2. Desvantagens.....	9
3. Qrcode .....	11
4. Código de barras 128.....	13
5. Entropia documento.....	15
6. Sistema de verificação de documentos impressos ou digitais.....	17
6.1. Arquitetura.....	17
6.2. Estrutura do documento .....	18
6.3. Base de dados .....	18
6.4. Criação do código de barras.....	19
6.5. Operações desempenhar pelo utilizador .....	19
6.6. Processamento do documento.....	19
6.7. Retificação do documento .....	21
6.8. Verificação de integridade .....	21
6.9. Retificação Manual .....	<b>Error! Bookmark not defined.</b>
7. Caso de estudo .....	24
8. Conclusões.....	30
9. Futuro trabalho .....	32
10. Referências.....	34
Apêndice A.....	35
Processamento de ficheiro PDF.....	35
Iron Software.....	36



## Lista de tabelas

Tabela 1 - Estudo de código de barras lineares .....	3
Tabela 2 - Possíveis escolhas .....	5
Tabela 3 - Resultados Eigenvalues .....	8
Tabela 4 - Classificação de erros do Qrcode .....	11

## Lista de Figuras

Figura 1 - Exemplo de um ataque zero .....	6
Figura 2 - Esquema de geração e extração da marca de água.....	7
Figura 3 - Antes da Watermark .....	7
Figura 4 - Depois da Marca de água.....	7
Figura 5 - Comparação Marca de água.....	8
Figura 6 - Exemplo line-shift .....	8
Figura 7 - Exemplo word-shift .....	9
Figura 8 - Exemplo Character Coding.....	9
Figura 9 - Funcionamento do algoritmo .....	9
Figura 10 - Exemplo Qrcode.....	11
Figura 11 - Estrutura Qrcode.....	11
Figura 12 - Código de barra 128 .....	13
Figura 13 - Tabela ASCII.....	13
Figura 14 - Arquitetura solução.....	17
Figura 15 - Diagrama da Base de dados .....	18
Figura 16 - Processamento exemplo.....	19
Figura 17 - Erro ao clicar em aceitar ou rejeitar sem antes de processar o ficheiro .....	20
Figura 18 - Reconhecimento caracteres .....	20
Figura 19 - Finalização do processamento .....	20
Figura 20 - Finalização da marca de água .....	20
Figura 21 - Mensagem Documento Aceite.....	20
Figura 22 - Mensagem documento já aceite.....	20
Figura 23 - Abrir Documento .....	21
Figura 24 - Retificação documento .....	21
Figura 25 - Algoritmo que conta número de retas que se interseitam .....	22
Figura 26 - Algoritmo interseção retas .....	22
Figura 27 - Exemplo verificação integridade .....	22
Figura 28 – Posições .....	<b>Error! Bookmark not defined.</b>
Figura 29 - Original.....	<b>Error! Bookmark not defined.</b>
Figura 30 - Retificar Manual.....	<b>Error! Bookmark not defined.</b>
Figura 31 - Processamento diagrama de fluxo .....	24
Figura 32 - Documento exemplar sem marca de água .....	25
Figura 33 - Tempo de execução .....	25
Figura 34 - Documento com marca de água.....	25
Figura 35 - Documento .....	26
Figura 36 - Código de barras.....	26
Figura 37 - Marca de água documento .....	26
Figura 38 - Obtenção de caracteres .....	26
Figura 39 - Representação dos valores dos caracteres.....	26
Figura 40 - Pontos para verificação de integridade .....	26
Figura 41 - Fluxograma Retificar.....	26
Figura 42 - Resultado Retificação .....	27
Figura 43 - Resultado Análise Forense .....	27
Figura 44 - Scan normal.....	27
Figura 45 - Resultado scan normal.....	28
Figura 46 - Scan torto.....	28
Figura 47 - Resultado scan torto.....	28
Figura 48 - Código de extração dos caracteres num ficheiro PDF .....	35
Figura 49 - Obtenção dos valores.....	35
Figura 50 - Execução jar .....	35
Figura 51 – Iron Barcode.....	36
Figura 52 - Geração código de barras.....	36
Figura 53 - Leitura de código de barras.....	36



# 1. Introdução

No âmbito da tese de Mestrado de Engenharia Informática da Universidade de Aveiro, pretende-se que seja desenvolvido um módulo aplicacional a integrar com uma solução de gestão de informação classificada desenvolvida pela empresa iCreate Consulting.

Esta solução prevê a possibilidade de impressão de documentos classificados, devendo ser implementados mecanismos de marcação dos mesmos no sentido de garantir a sua autenticidade.

## 1.1. Descrição do sistema existente

Estes documentos vão ser tratados em áreas de segurança física destinadas para o efeito, onde para aceder o utilizador têm que ser pessoas idóneas (de confiança) e acesso a documentos com informação classificada deverá ter um certificado periódico.

Cada utilizador tem acesso a um determinado Posto de Trabalho, e aos documentos que foram destinados a esse mesmo Posto de Trabalho.

Por outro lado, esse utilizador só tem permissão para ver documentos até um nível de classificação para o qual está autorizado: por exemplo se só pode ver documentos até ao nível Confidencial, não terá acesso aos documentos Secretos ou de nível de classificação superior.

Apesar de ter acesso a documentos localizados num determinado posto, deve respeitar o princípio da “Necessidade de Conhecer”, não devendo consultar indiscriminadamente os documentos a que tem acesso.

## 1.2. Objetivos

Para assegurar a sua autenticidade propõe-se dois métodos de segurança no documento.

O primeiro é criar um código de barras que permita validar de uma forma explícita se o documento teve origem no sistema, que possível terá mais informações que caracterize o documento, nomeadamente marcas de segurança. A ideia deste primeiro é passar uma pistola que leia o código de barras e averiguar se o código de barras é válido e se a informação do código de barras corresponde ao documento.

O segundo é criar uma watermark segura, única, por documento, gerado por um conjunto de informações do documento, que permitam auxiliar em aspetos mais forenses.



## 2. Estado da Arte

Este capítulo tem como objetivo agregar informações relativas ao tema da tese, sendo dividido em dois subcapítulos, códigos de barras e *text watermarking*. O subcapítulo código de barras servirá para verificar se um documento é fidedigno, isto é comparar a assinatura do documento com a informação que estará no código de barras. O *text watermarking* verificará se a informação do documento foi modificada ou se é o original através de métodos de marca de água.

### 2.1. Código de barras

Código de barras é um modo de representar informação num estado visual, que possa ser visível sem de ter a necessidade de escrever texto. É classificado em duas categorias, linear (1D) e 2D.

Um dos requisitos da empresa era que o código de barras não ocupasse muito espaço no documento, tendo espaço livre no cabeçalho e no rodapé. Como a categoria linear ocupa menos espaço em altura, tendo um comprimento variável, dependendo do número de caracteres contidos na informação do mesmo. Existem várias subcategorias nos códigos de barras, para analisar as mesmas criou-se uma tabela (1), que compara os mesmos, escolhendo o ideal para inserir no documento. Esta tabela é constituída por colunas:

- Nome: designação do código de barras;
- Imagens exemplificativas;
- Comprimento - variável ou fixo;
- Uso;
- Necessidade de aparelho especial para ler.

Em alguns códigos de barras existe um espaço reservado que não se pode alterar que se chama *checksum* ou *check digit*, cujo objetivo é verificar se a informação do código de barras foi gerada corretamente.

Existem duas maneiras de ler um código de barras:

1. Máquina própria (leitor de barras de barras)
2. Smartphone (iOS ou android) com uma aplicação que permita a leitura.







Devido as questões de segurança, a empresa proíbe uso de telemóveis pessoais dentro do estabelecimento, o que levou descartar a hipótese de usar os smartphones como leitores. E leitores de barras especiais, já que têm um custo adicional para a empresa. Para solução optou-se por usar um package em C#, *Iron Bar Code* [1] que permite a leitura de códigos de barras em ficheiros PDF ou imagens (PNG), sendo apenas o reconhecimento em zonas sem distorção, ou seja, sem que apareça texto ou imagens por cima do código de barras.

#### 2.1.1. Análise de código de barras

Este subcapítulo tem como objetivo expor a análise de vários códigos de barras lineares, cujos parâmetros de escolha foram comprimento (variável ou fixo), tamanho das barras. A tabela 1 apresenta um conjunto de códigos de barras com as características previamente expostas.

Na tabela 2 encontram-se os possíveis códigos de barras que porventura se possam utilizar. Estes códigos foram selecionados com o critério tamanho máximo de caracteres que este pode guardar, podendo mudar no futuro de acordo com a escolha da empresa. A informação que se pretende guardar nos códigos de barras é a data do documento assinado, o local do posto de trabalho e um identificador único do documento, podendo ser mudado no futuro.

*Tabela 1 - Estudo de código de barras lineares*

Nome	Imagem	Fixo/ Variável	Tamanho das barras	Uso
Post Code Austrália		Fixo	4	Correios Austrália
CodaBar		Fixo	2	Librarias
Code 25		Variável	2	Librarias
Code 11		Fixo	2	Telefones, já não se usa muito devido à sua antiguidade
Code 32		Fixo	2	Farmácia
Code 39		Fixo	2	Vários

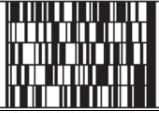















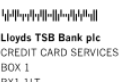






Code 49		Variável	Vários	Vários
Code 93	 CODE93	Variável	Vários	Vários
Code 128	 CODE128	Variável	Vários	Vários
EAN 2		Variável	Vários	Revistas
EAN 5		Variável	Vários	Livros
EAN-8		Variável	Vários	Retalho
GS1-128	 (05)95812345678901(310)0000123	Variável	Vários	Vários
GS1 Databar	 (01)00075678164125	Variável	Vários	Vários
Intelligent Mail barcode	 Wikimedia Foundation Inc. P.O. BOX 76350 SAN FRANCISCO CA 94107-8350	Fixo	4	Correios USA
ITF-14	 1 97 55432 10221 1 9	Variável	2	Encomendas
ITF-6	 123457	Variável	2	Vários
JAN	 5 901234 123457 >	Variável	Vários	Usado no Japão
Planet		Variável	Grande/Pequeno	Correios USA
Plessey	 4 3 2 1 6 Start 43216 Check Digit Stop	Variável	2	Catálogos, revistas, inventários
PostBar		Fixo	4	Correios Canada
PostNET		Fixo	Grande/Pequeno	Correios USA
RM4SCC/KIX	 Lloyds TSB Bank plc CREDIT CARD SERVICES BOX 1 BX1 1LT	Fixo	4	Correios
RM Mailmark C		Fixo	4	Correios
Universal Product Code	 9 118765432109 118	Variável	Vários	Retalho
Telepen	 ABC abc1234	Variável	2	Librarias da Inglaterra

Tabela 2 - Possíveis escolhas

Nome	Imagem	Tamanho máximo de dados	Tipos de dados	Vantagens	Desvantagens
Code 39		43	Letras e números	Ocupa menos espaço.	Não processa caracteres especiais tal como ç.
Code 128		48	Letras, números e caracteres especiais	Faz encoding de strings muito eficiente.  Tem <i>checksum</i>	Precisa de espaço reservado no início, fim e <i>check symbols</i> .  Tem pouco limite de informação que pode guardar.
Code 93		30	Letras, números e caracteres especiais	Open Source	Não tem checksum

## 2.2. Texto com marca de água

A *marca de água* é um método que permite a salvaguarda de documentos originais e verificação da autenticidade de documentos ou imagens. É usada nos dias de hoje como por exemplo nas notas europeias, para impedir a impressão de notas não autorizadas. Existem muitos métodos e ataques de marcação de água, neste documento irá ser abordado o método marca de água de texto e os métodos ataques zero devido às necessidades apresentadas pela empresa, contudo poderá mudar consoante o desenvolvimento da tese.

### 2.2.1. Ataque zero com mudança de texto

O objetivo dos ataques zero é aumentar a fragilidade da marca de água do documento contra ataques de pessoas não autorizadas, normalmente denominados por hackers. Estes ataques tendem a alterar o texto do documento sem alterar a marca de água. Estes ataques normalmente são de inserção, reordenação e remoção de texto.

Uma das soluções propostas pelos autores do artigo [2] foi ataque de substituição que explora a troca de ordem das palavras no texto. Esta solução é constituída por dois elementos:

1. Lista de palavras que se pretende substituir no documento.
2. Lista de palavras contidas no documento depois da substituição.

Em suma, escolhe-se um conjunto de palavras no documento que se quer substituir por novas. Quando uma palavra é selecionada para se substituir todas as ocorrências da mesma são substituídas pela palavra nova. A percentagem do ataque determina a quantidade de modificação do documento. Se a palavra escolhida tiver várias ocorrências no documento, a percentagem vai aumentar.

A percentagem da contagem de palavras no documento é dada pela equação 1. A percentagem do ataque de substituição é dada pela expressão da equação 2.

Existem duas maneiras de escolher a lista de palavras para a substituição, normal e avançada. A normal consiste em selecionar palavras aleatórias do documento. A avançada selecionar palavras com mesmo tamanho do que as palavras da lista de substituição.

Em suma, a marcação de água textual proposto por [2] consiste em 3 passos, selecionar palavras do documento para substituição, selecionar as novas palavras e ocorrência de implementação no texto do documento.

$$WOR(x) = \frac{\text{Number of Occurrence of } x \text{ in Document}}{\text{Total Number of Words in Document}}$$

Equação 1

$$WSR = \frac{\sum WOR(x)}{\text{Total Number of Words in Document}}$$

Equação 2



### 2.2.1.1.Exemplo

Conforme mencionado no artigo [2] a figura 1, demonstra um exemplo de tipos de substituição que podem ser usados para mascarar informações importantes em um texto. O exemplo demonstra a substituição da palavra "The", que é o pronome mais comum em inglês, por "close" na substituição normal e por "job" na substituição avançada.

A ideia por trás desse método é permitir que palavras-chave ou informações importantes não sejam perdidas, enquanto ainda são mascaradas para outras pessoas que não têm acesso às palavras originais. A substituição de palavras é realizada de tal forma que apenas quem possui o conhecimento necessário para substituir as palavras é capaz de entender o significado original do texto.

Este método pode ser útil em várias situações, como na proteção de informações sensíveis em documentos, comunicações confidenciais, entre outros. No entanto, é importante notar que este método não é infalível.

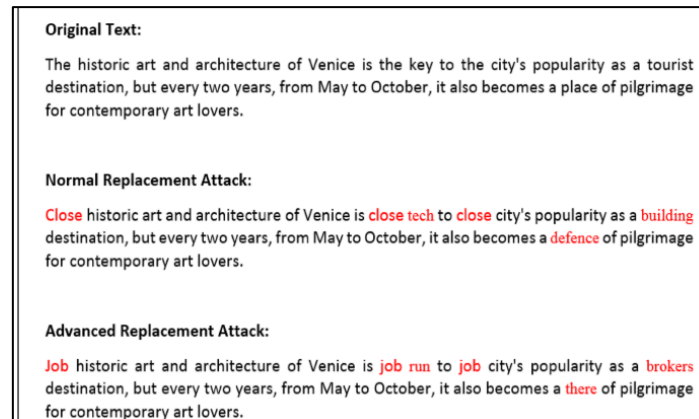


Figura 1 - Exemplo de um ataque zero

### 2.2.1.1.Vantagens

Uma das vantagens do método de substituição descrito é permitir a alteração de palavras em um texto por um utilizador fidedigno e, ao mesmo tempo, permitir que o algoritmo saiba em qual zona do documento as alterações foram feitas.

Isso pode ser útil em situações em que uma equipe de revisão precisa fazer alterações em um documento, como em um processo de edição de textos. Ao usar esse método, a equipe de revisão pode fazer as alterações necessárias sem perder informações importantes ou prejudicar a integridade do documento. Além disso, o algoritmo pode identificar facilmente as alterações realizadas e rastreá-las para fins de auditoria e verificação de alterações autorizadas.

### 2.2.1.2.Desvantagens

Uma das principais desvantagens é que as mudanças no texto podem não fazer sentido para pessoas que não estão familiarizadas com a substituição de palavras usada. Isso pode tornar a leitura do texto difícil e confusa, especialmente se muitas palavras forem substituídas. Se a substituição for realizada de forma excessiva, o texto pode se tornar ininteligível.

Além disso, o método de substituição pode não ser capaz de detectar métodos de cópia de texto, como copiar e colar o texto em um novo documento. Se o novo documento for criado com base no texto original antes da substituição, as informações confidenciais podem ser utilizadas sem que a substituição tenha qualquer efeito.

Outra limitação é que esse método de substituição não garante a segurança completa dos dados, pois é possível que alguém com conhecimento suficiente possa decifrar a substituição e utilizar as informações originais.

### 2.2.2. Ataque zero sem mudança no texto

A solução proposta no artigo [3] consiste em usar as características do documento para gerar a marca de água com o seguinte layout "autor:watermark:data:tempo" registrada pela uma autoridade certificada (CA). O objetivo principal deste método é usar as palavras que tenham um tamanho maior do que quatro caracteres para proteger contra possíveis ataques.

A escolha de palavras maiores do que quatro caracteres é devido a generalidade dos ataques serem direcionados a palavras com tamanho menor do que quatro caracteres.

Para gerar a marca de água, é necessário saber todas as frases de um documento, e assim percorrer cada palavra achando as palavras maiores do que quatro caracteres, quando acabar, guardar o primeiro caracter de cada dessas palavras e por fim gerar a marca de água.

Por exemplo na seguinte frase: "O José gosta muito de ler" a marca de água ficava JGM (José Gosta Muito). Depois de percorrida todas as frases juntasse as marcas de água de cada frase, dando a marca de água final.

A figura 2 demonstra como é que a watermark é gerada e extraída. Deteta-se um ataque quando o padrão da marca de água tem pelo menos ser 70% igual ao documento comparativo.

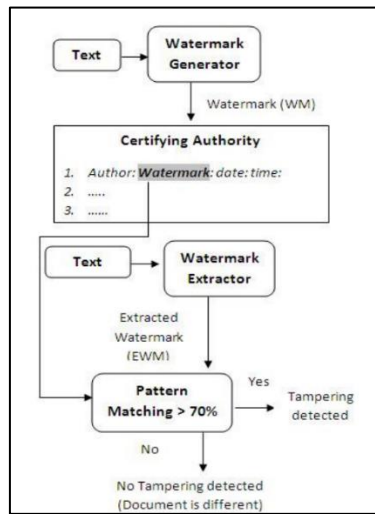


Figura 2 - Esquema de geração e extração da marca de água

### 2.2.2.1. Vantagens

Tem uma entidade certificadora que vai possuir a marca de água original, que é a única pessoa que pode comparar os documentos, permitindo assim que o documento esteja só na posse de uma pessoa, diminuindo os ataques ou distribuição do documento. A composição da entidade certificadora é similar ao pretendido no código de barras.

### 2.2.2.2. Desvantagens

Tem testes realizados com ataques aleatórios, como inserir, apagar, reordenar e alterar, que para textos reais os resultados poderão ser diferenciados, levando com que o comportamento do algoritmo seja diferenciado entre documentos, dificultando a sua implementação para a retificação da integridade do documento.

### 2.2.3. Documentos texto baseados em Eigenvalues

O artigo [4] propõe um algoritmo que guarda as posições de todas as palavras do documento numa matriz e o peso das mesmas em ASCII. Um documento normalmente é constituído por palavras, espaçamento, números e pontuações. Os autores consideram cada contagem para calcular o peso ASCII para gerar o esquema da marca de água com base numa chave privada, que um utilizador fidedigno retifica o documento recebido.

#### 2.2.3.1. Testes

No artigo [4] existe testes realizados a um documento com 47 linhas, não tendo disponibilizado o mesmo, apenas prints de pequenas secções do mesmo.

A figura 3 demonstra um excerto do texto antes de ter a marca de água, e na figura 4 demonstra o texto depois da geração da marca de água. Para simular um ataque foi alterada a palavra “OFF” para “ON” numa zona do texto e pelos resultados binários na figura 5, sendo que em primeiro está o documento original e de seguida o alterado, retira-se que os números são diferentes. Uma pequena alteração da palavra levou à mudança de 11 bits.

Os autores realizaram testes na mudança de vogais, consoantes, caracteres especiais, palavras, números pontuações, e alterações aleatórias no texto, obtendo os resultados da tabela 3, podendo retificar que na coluna “tamper detection” está tudo a 100% que leva a retificar que o algoritmo detetou alterações em todos os casos.

point in time receives a spe  
rally speaking, the device 1  
made it available. Unauthori  
de the tampered version av  
istance, a hacker operating  
re image 111 as it is made a  
nper with the image 111 as li  
a tampered version of the  
a variety of means.  
140 extracts the verificatio

(a) Before watermarking

Figura 3 - Antes da Watermark

point in time receives a spe  
rally speaking, the device 1  
made it available. Unauthori  
de the tampered version av  
istance, a hacker operating  
re image 111 as it is made a  
nper with the image 111 as li  
a tampered version of the  
a variety of means.  
140 extracts the verificatio

(b) After watermarking

Figura 4 - Depois da Marca de água

Before Tampering Secret key generated is,  
0000000000001001011100100000000000001000000101101  
0000000000001011101100010100000000001111001111100  
0100000000001110010001001111110110001101000110010  
1110000000  
After Tampering Secret key generated is,  
0000000000001001011100011000000000001000000101101  
0000000000001011101100010010000000001111001111100  
01000000000011100100010100110110110001101000110010  
1011000000

Figura 5 - Comparação Marca de água

Tabela 3 - Resultados Eigenvalues

Type of alteration (a single character)	Average eigen value shift	% bit change in secret key	Tamper Detection
Vowels	68.89	12.29	100%
Consonants	66.83	11.67	100%
Words	95.22	14.38	100%
Numerals	169.28	13.96	100%
Punctuations	53.94	10	100%
Random Alterations	127.95	15.63	100%

#### 2.2.3.4. Vantagens

Guarda a posição das palavras do texto numa matriz. Tem uma identidade certificadora que guarda a marca de água e é a única pessoa que pode verificar se o documento é original ou falsificado.

#### 2.2.3.5. Desvantagens

Difícil de compreensão, a execução do algoritmo leva muito tempo de execução e processamento para textos grandes.

#### 2.2.4. Line-Shift Coding

No artigo [5] expuseram outra forma de fazer marcação de água denominado, *Line-Shift Coding*, que consiste em mover as linhas de texto de um documento para cima ou para baixo, enquanto as linhas adjacentes não são movidas. Na figura 6, demonstra-se um exemplo. Neste exemplo a linha do meio começada por “Effects...”, foi movida para baixo 1/300 inches, que equivale a 0.00846666667 cm. Esta mudança não é perceptível ao olho humano, só através de um OCR, Optical Character Recognition, é que se conseguiria comparar ficheiros para verificar se existe ou não mudança nas linhas.

the Internet aggregates traffic flows from many end systems. Understanding effects of the packet train phenomena on router and IP switch behavior will be essential to optimizing end-to-end efficiency. A range of interesting

the Internet aggregates traffic flows from many end systems. Understanding effects of the packet train phenomena on router and IP switch behavior will be essential to optimizing end-to-end efficiency. A range of interesting

Figura 6 - Exemplo line-shift

#### 2.2.4.1. Word-Shift Coding

No mesmo artigo [5], previamente exposto, expuseram outra forma de fazer marcação de água denominado *Word-Shift Coding*, que consiste em mover as palavras para a esquerda ou para a direita, enquanto as palavras adjacentes não são alteradas. A figura 7, tem-se um exemplo. A segunda linha, contém quatro palavras movidas com espaçamento de 1/150 inch, que equivale a 0.0169333333 cm, enquanto na primeira linha não se altera, a terceira é uma junção das duas anteriores. Como referido no ponto anterior esta mudança não é perceptível a olho humano, e para retificar a marca de água recorrer a um OCR.

the Internet aggregates different sessions from many end systems. Understanding  
the Internet aggregates different sessions from many end systems. Understanding  
the Internet aggregates different sessions from many end systems. Understanding

Figura 7 - Exemplo word-shift

#### 2.2.4.2. Character Coding

No mesmo artigo [5], previamente exposto, deram outra forma similar, referidas anteriormente, mas desta vez consiste em mover o carácter para cima ou baixo, enquanto adjacentes não se alteram. A figura 8, apresenta um exemplo. A primeira letra “e” da palavra “internet” foi movida para baixo 1/600 inches que equivale a 0.0042333333 cm. Como foi referido anteriormente noutros capítulos é necessário um OCR, para retificar as mudanças feitas no texto, isto é dispendioso em termos de tempo, e difícil de percepção.

the Internet aggregates  
**Internet**

Figura 8 - Exemplo Character Coding

#### 2.2.4.3. Vantagens dos métodos

Tem marca de água invisível, ou seja, não é perceptível para os olhos de um humano.

#### 2.2.4.4. Desvantagens dos métodos

Uso de OCR para verificar documento, e ter acesso ao texto para inserir o espaçamento. Erros de impressão, por exemplo faltar letras no documento, pode levar à mal classificação de um documento, por exemplo ser falso, quando se tem a certeza de que o documento é original e não foi modificado.

#### 2.2.5. ECL zero-based watermarking

O artigo [6], propõe um tipo de marcação de água denominado *ECL zero-based watermarking* que consiste num algoritmo que mantém o conteúdo original do texto do documento e constrói a *marca de água* pela seleção de caracteres no documento. A marca de água é guardada num sítio de confiança que se denomina *Certifying Authority*, para quando houver a necessidade de comparar o documento com a *marca de água* para verificação da sua autenticidade, perceber que se o documento é original ou falsificado. Os autores forneceram uma imagem (figura 9) em que se demonstra como funciona o algoritmo.

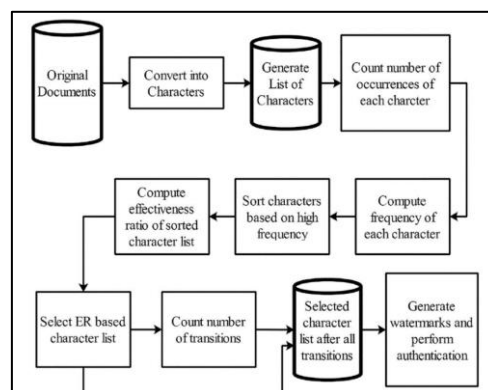


Figura 9 - Funcionamento do algoritmo

#### 2.2.5.1. Vantagens

Tem uma identidade certificadora que garante que o documento original esteja seguro com uma pessoa confiável na empresa.

#### 2.2.5.2. Desvantagens

Falta de testes em cópias de documentos. Pouco efetivo com documentos que tenham um *effectiveness ratio* (ER) menor que 0.1. Effectiveness Ratio determina quais os caracteres para a geração da marca de água.



### 3. Qrcode

Atualmente vemos qrcode, nas faturas, em cartões de empresas e até mesmo na rua, maior parte das pessoas já viu, mas maior parte pode não saber o processo de leitura ou como surgiu o mesmo, para isso dediquei este capítulo ao qrcode.

O Qrcode (figura 10) surgiu em 1994 pela empresa Denso Wave, onde originalmente foi criado para categorizar peças de automóvel. Os Qrcode podem ter links para páginas web, texto, um endereço geográfico, uma imagem, um vídeo ou contacto telefónico.

Relativamente à estrutura do qrcode é constituído por 3 quadrados de deteção de posição (4.1. Figura 11) que permite a leitura em várias posições do scanner ou de um smartphone com câmara.

Por um padrão alinhamento (4.2. Figura 11) que corrige a distorção do qrcode em superfícies curvadas, o seu número varia consoante a informação contida.

Por padrões de temporização (4.3. Figura 11) que permite obter o tamanho da matriz de dados. Por informações da versão, que indica que versão do qrcode está a ser utilizada (1. Figura 11).

Por informações de formato, que contém informações sobre a tolerância de erros e o padrão da máscara de dados. Por códigos de dados e erros que podem ser do tipo L, M, Q, H, cujos estão apresentados na tabela 6.



Figura 10 - Exemplo Qrcode

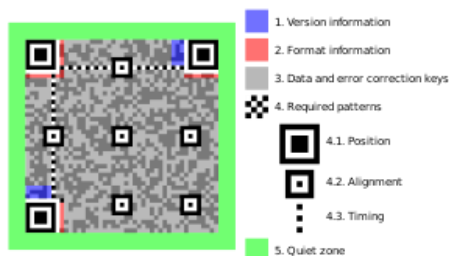


Figura 11 - Estrutura Qrcode

Tabela 4 - Classificação de erros do Qrcode

Nível de correção de erros	Percentagem de área danificada (%)
L (Low) – Baixo	7
M (Medium) – Médio	15
Q (Quartile) - Quartil	25
H (High) – Alto	30



#### 4. Código de barras 128

O código de barras 128 (figura 12) [7] é um tipo de código de barras linear que pode ser usado para codificar uma grande variedade de dados alfanuméricos, incluindo letras, números e caracteres especiais. Ele é chamado de 128 é capaz de codificar 128 caracteres ASCII (figura 13).

Este é muito utilizado em aplicações de logística, como no controle de estoques e na identificação de produtos em supermercados e lojas. Isso ocorre porque ele é capaz de codificar informações como o nome do produto, seu número de série, o código de barras do fabricante e outras informações relevantes em um único código.

É constituído por barras largas e compactas que representam cada caractere, juntamente com barras de início e fim que indicam onde começa e termina a sequência de caracteres. Ele é capaz de codificar uma grande quantidade de informação em um espaço relativamente pequeno, o que o torna uma ferramenta eficiente e econômica para o gerenciamento de inventário e outras aplicações similares.

É um código de barras universalmente reconhecido e amplamente utilizado em todo o mundo, o que leva ser uma opção confiável para empresas que necessitam de um sistema de identificação e rastreamento de produtos rápido e preciso.

Como referido inicialmente, o intuito da dissertação é validação de documentos impressos, para isso é necessário guardar informações acerca do documento em algum lado, para a retificação do original. Como o código de barras consegue agregar informação, como por exemplo um id referenciador para um documento onde por fora esteja uma base de dados que contenha informação do documento, optou-se por utilizar como medida de retificação inserindo-o no rodapé do documento.

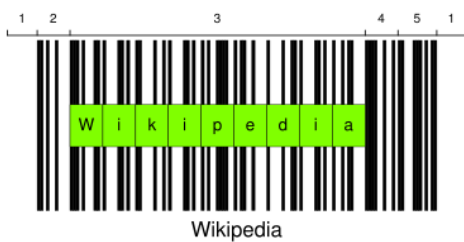


Figura 12 - Código de barra 128

# ASCII Table

Dec	Hex	Oct	Char	Dec	Hex	Oct	Char	Dec	Hex	Oct	Char	Dec	Hex	Oct	Char
0	0	0		32	20	40	[space]	64	40	100	@	96	60	140	`
1	1	1		33	21	41	!	65	41	101	A	97	61	141	a
2	2	2		34	22	42	"	66	42	102	B	98	62	142	b
3	3	3		35	23	43	#	67	43	103	C	99	63	143	c
4	4	4		36	24	44	\$	68	44	104	D	100	64	144	d
5	5	5		37	25	45	%	69	45	105	E	101	65	145	e
6	6	6		38	26	46	&	70	46	106	F	102	66	146	f
7	7	7		39	27	47	'	71	47	107	G	103	67	147	g
8	8	10		40	28	50	()	72	48	110	H	104	68	150	h
9	9	11		41	29	51	!	73	49	111	I	105	69	151	i
10	A	12		42	2A	52	*	74	4A	112	J	106	6A	152	j
11	B	13		43	2B	53	+	75	4B	113	K	107	6B	153	k
12	C	14		44	2C	54	,	76	4C	114	L	108	6C	154	l
13	D	15		45	2D	55	-	77	4D	115	M	109	6D	155	m
14	E	16		46	2E	56	.	78	4E	116	N	110	6E	156	n
15	F	17		47	2F	57	/	79	4F	117	O	111	6F	157	o
16	10	20		48	30	60	0	80	50	120	P	112	70	160	p
17	11	21		49	31	61	1	81	51	121	Q	113	71	161	q
18	12	22		50	32	62	2	82	52	122	R	114	72	162	r
19	13	23		51	33	63	3	83	53	123	S	115	73	163	s
20	14	24		52	34	64	4	84	54	124	T	116	74	164	t
21	15	25		53	35	65	5	85	55	125	U	117	75	165	u
22	16	26		54	36	66	6	86	56	126	V	118	76	166	v
23	17	27		55	37	67	7	87	57	127	W	119	77	167	w
24	18	30		56	38	70	8	88	58	130	X	120	78	170	x
25	19	31		57	39	71	9	89	59	131	Y	121	79	171	y
26	1A	32		58	3A	72	:	90	5A	132	Z	122	7A	172	z
27	1B	33		59	3B	73	;	91	5B	133	[	123	7B	173	{
28	1C	34		60	3C	74	<	92	5C	134	\	124	7C	174	
29	1D	35		61	3D	75	=	93	5D	135	]	125	7D	175	}
30	1E	36		62	3E	76	>	94	5E	136	^	126	7E	176	~
31	1F	37		63	3F	77	?	95	5F	137	_	127	7F	177	

*Figura 13 - Tabela ASCII*





## 5. Entropia documento

A presente dissertação analisa estudos de segurança em documentos digitais, é necessário introduzir um conceito fundamental denominada entropia num documento.

A entropia de um documento é uma mediada da incerteza ou desordem da informação contida nele, ou seja, serve para avaliar a segurança de um algoritmo ou sistema criptográfico.

Quanto maior for a entropia maior é dificuldade para a pessoa/hacker descobrir como um algoritmo funciona, o que significa que no desenvolvimento da tese pretendesse que a entropia seja mais pequena possível.

Para atingir esse objetivo é necessário ocultar informações sobre a verificação da integridade e o modo como se gera a marca de água para um documento e desenvolver métodos aleatórios.



## 6. Sistema de verificação de documentos impressos ou digitais

A primeira intenção para a criação da solução era usar Qrcode, contudo retificou-se que em scans, podiam desaparecer (desvanecer, ou seja, perder a cor) e influenciava na visualização do documento.

Para corrigir estes problemas pensou-se em sugestões como:

- processamento de imagem, que reconhece os Qrcode no texto (através de substituição de cores, por exemplo preto no branco dentro do Qrcode é uma zona crítica, para corrigir é necessário alterar o preto para branco), descartada devido ao elevado tempo de processamento da mudança de cores;
- posições aleatórias no documento, descartada devido a porventura o qrcode ser colocado num sítio não permitido, como por exemplo debaixo de uma imagem, contudo sempre há oportunidade de gerar de novo a marca de água do ficheiro;
- quantidade de Qrcode a colocar no documento;
- dimensão do Qrcode;

Com o desenvolver do algoritmo e o objetivo de traçar retas para a retificação de letras no documento, nos quadrados de posição do Qrcode, optou-se por substituir por uma imagem reduzida com um círculo que contivesse um X, o que levou ao surgimento da questão “se as imagens são visíveis com base nos pontos que se usa para a criação da retificação das interseções, o hacker não poderá decifrar como se calcula e obtém as mesmas?”, a resposta é sim, consegue. Para corrigir e melhorar a solução removeu-se completamente a imagem, colocando assim os pontos de origem da criação das retas incógnitos ao hacker.

A solução criada envolve duas ações distintas, processamento e retificação do documento.

O processamento do ficheiro envolve utilizar códigos de barras, para a retificação rápida de um documento que seja aceite na base de dados através do utilizador, e posições aleatórias no documento para guardar letras que resultam de interseções de retas.

A retificação do documento/verificação de integridade/análise forense consiste em determinar se a informação do código de barras corresponder ao documento, se não procedesse a uma análise mais profunda através das interseções de retas para verificar se as letras que aparecem no sistema são iguais ou não aquelas que aparecem no documento.

### 6.1. Arquitetura

De acordo com o descrito anteriormente é criada a arquitetura da figura 14, onde surge 3 camadas.

- Servidor destinado à base de dados que irá conter dados armazenados acerca dos ficheiros, códigos de barras, segmentos de reta traçados, posição dos caracteres no ficheiro de input e a geração da marca de água.
- Programa onde acontece o desenvolvimento da criação do código de barras, do processamento do ficheiro para obtenção das posições dos caracteres no documento, da retificação do documento, e verificação de integridade, esta também é responsável pelas conexões entre camadas, desencadeadas pelo utilizador.
- Utilizador é responsável por agregar as ações que o utilizador pode fazer como por exemplo processar o documento ou retificar.

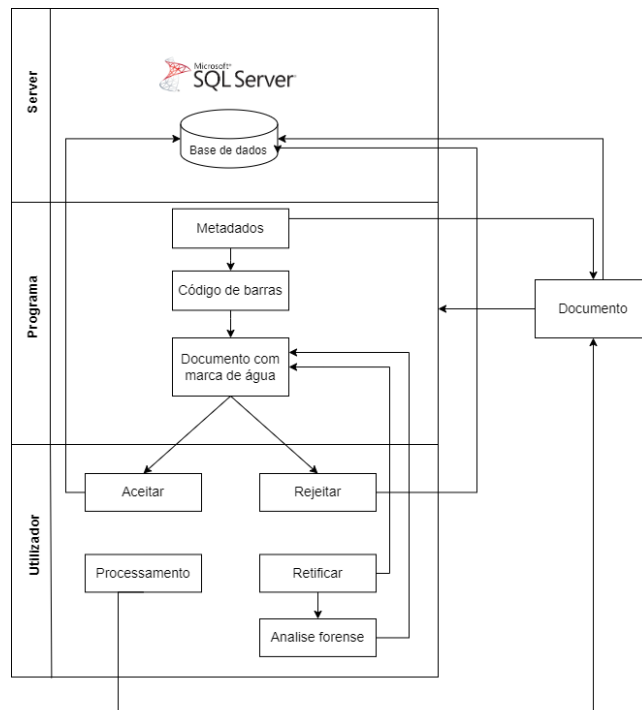


Figura 14 - Arquitetura solução

## 6.2. Estrutura do documento

Para a utilização de testes do programa foi disponibilizado pela empresa 4 documentos diferentes em formato PDF e digitais. Cabe à empresa inserir manualmente os ficheiros na diretoria correspondente na inicialização do programa e os metadados no programa para tratar os mesmos. É de salientar que os documentos disponibilizados têm uma classificação de segurança mínima pelo que se pode partilhar os resultados do mesmo.

O algoritmo lida com três tipos de documentos:

- Documento sem marca de água;
- Documento com marca de água;
- Documento (scan) com marca de água;

## 6.3. Base de dados

O intuito de usar base de dados no sistema é garantir segurança sobre as informações do documento através do uso de código de barras, como se referiu anteriormente, o código de barras tem um id aleatório representante das características do documento cujo é inserido na base de dados aquando do seu processamento do documento para a criação da marca de água.

Existem dois tipos de base de dados as relacionais e não relacionais (NoSQL). [8]

As bases de dados relacionais guardam os dados nas tabelas, tendo algumas delas partilha de informação, causando uma relação entre tabelas.

Cada tabela contém colunas que definem a informação que se pode guardar, e linhas que contém a informação.

Normalmente a tabela contém um identificador único que referencia cada linha chamado de chave primária (primary key), caso se queira referenciar os valores a outra tabela utiliza-se a chave estrangeira (foreign key), que obrigatoriamente tem de existir previamente.

A linguagem que se usa para tratar base de dados relacionais é SQL, existem vários programas que permitem correr SQL, como por exemplo mySQL [9], Oracle SQL Developer [10] e Microsoft SQL Server Management Studio 2018 [11].

As bases de dados não relacionais, não usam tabelas relacionais, em vez disso faz cria grupos em se que guarda a informação em informações diferentes.

Como o objetivo do trabalho é sempre perceber que documento é qual, é necessário haver relações entre a marca de água e o documento, então optou-se por usar uma base de dados relacional com o auxílio da ferramenta Microsoft SQL Server Management Studio 2018.

A base de dados SQL foi criada localmente, tendo um utilizador e base de dados dedicado apenas à consulta de informações pelo programa. As informações que se guardam são as características do documento, o código de barras, segmentos de reta traçados entre pontos, posições dos caracteres no documento de entrada (processamento apenas), e a criação da marca de água (aceitação ou rejeição).

Na figura abaixo, está presente um diagrama da base de dados que contém as tabelas usadas e respetivas conexões.

O diagrama é constituído por as seguintes tabelas:

- “document”: guarda características do documento (metadados);
- “barcode”: guarda informações relativas ao documento;
- “watermark”: guarda as confirmações do documento com marca de água, se foi aceite ou não, para efeitos de rastreamento;
- “forense\_analises”: guarda os segmentos de reta traçados entre dois pontos, o ponto de interseção e a letra que aparece no ponto de interseção para efeitos de verificação de integridade do documento;
- “position\_char\_file”: guarda as posições dos caracteres no documento.

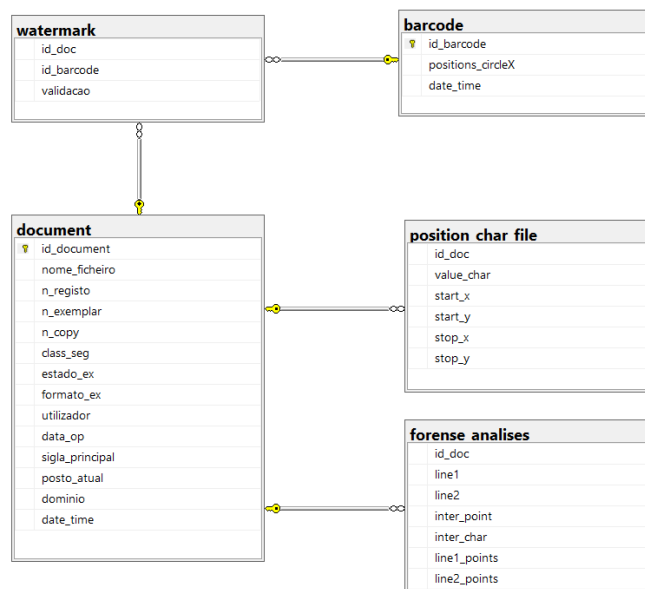


Figura 15 - Diagrama da Base de dados

## 6.4. Criação do código de barras

Para a criação do código de barras 128 foi utilizado um package para C# denominado Iron Barcode, abordado mais detalhadamente no Anexo.

Como referido anteriormente o código de barras irá ter um id referenciador para o documento e para as posições utilizadas para traçar os segmentos de reta no documento.

O código de barras 128 irá ser colocado sempre no cabeçalho da primeira página devido a ser uma zona livre do documento, e permitir maior parte das vezes a sua leitura.

## 6.5. Operações desempenhar pelo utilizador

Um utilizador exerce 4 tipos de funções.

- Escolher entre
  - Processamento: que processa o documento sem a marca de água para originar a marca de água;
  - Retificar: que retifica o documento de input com a marca de água para verificar se é o documento que se apresenta ser.
- Aceitar ou rejeitar o documento com marca de água;
- Comparar a informação do documento com a marca de água no submenu retificar, para determinar a sua autenticidade;
- Verificação de integridade, para determinar a zona da alteração do documento, se por venturar o documento não ser autêntico, através da comparação de informações que aparecem na aplicação versus o documento.

## 6.6. Processamento do documento

Conforme mencionado anteriormente o utilizador escolhe o ficheiro a processar, o algoritmo só aceita caso o ficheiro não tiver na base de dados e se o nome do mesmo não conter “watermark” que se adiciona ao nome do ficheiro já existente para diferenciar os ficheiros sem e com marca de água, também se adicionou o dia e a hora que se executou o processamento do ficheiro levando ao nome final do ficheiro ser “nome\_ficheiro\_watermark\_dd\_mm\_yy\_hh\_m\_ss”, sendo dd:dia, mm:mês, yy:ano, hh:hora, m: min, ss:segundos. Antes de processar, caso o utilizador queira visualizar se o ficheiro selecionado é pretendido, o algoritmo mostra o ficheiro numa janela nova com os botões processar, aceitar e rejeitar (figura 16). Caso o utilizador clique em aceitar ou rejeitar o ficheiro sem antes de clicar “processar” aparece o erro da figura 17.

Consoante o clique “processar” o programa vai abrir uma consola que vai executar um ficheiro jar, fechando automaticamente depois da execução, que vai retirar os caracteres do ficheiro e as respetivas posições no documento. Para retirar as posições o algoritmo foi realizado em Java porque em C# não foi encontrado packages que retirassem a posição com certeza dos caracteres no documento (figura 18).

O processamento demora à volta de 2 min, colocando a janela do processamento bloqueada até que as ações de calcular os pontos, obtenção dos pontos de interseção e letras, inserções na base de dados sejam concluídas.

Quando o processamento acaba o utilizador é abordado com uma mensagem da figura 19 para aceitar ou rejeitar o documento que se pode pré visualizar figura 20. Caso o documento seja aceite o utilizador recebe a mensagem “Documento Aceite!” figura 21, por ventura se tentar rejeitar ou aceitar de novo o documento é informado de que o documento já foi aceite (figura 22). Caso rejeite o documento é guardado, e o processamento é feito novamente. Caso feche a aplicação sempre poderá consultar o documento que foi gerado para a diretoria que abriu o documento na extensão previamente exposta. A escolha é feita no menu principal quando se escolhe a opção pretendida. É aberto a diretoria predefinida da execução (figura 23) e pode-se mudar consoante o necessário, tendo a noção que ficheiros fora da pasta do algoritmo podem dar origem a erros.



Figura 16 - Processamento exemplo

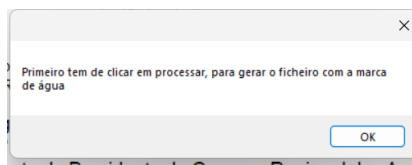


Figura 17 - Erro ao clicar em aceitar ou rejeitar sem antes de processar o ficheiro

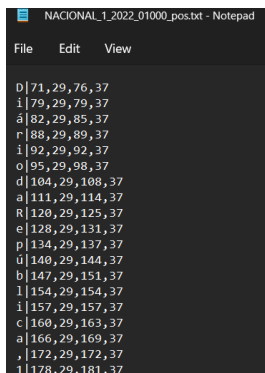


Figura 18 - Reconhecimento caracteres

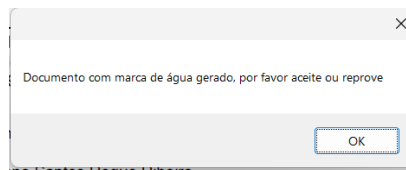


Figura 19 - Finalização do processamento

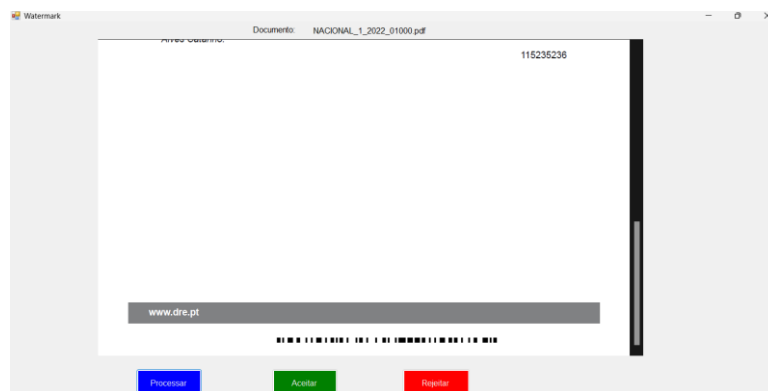


Figura 20 - Finalização da marca de água

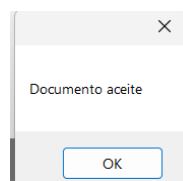


Figura 21 - Mensagem Documento Aceite

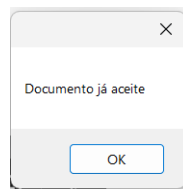


Figura 22 - Mensagem documento já aceite

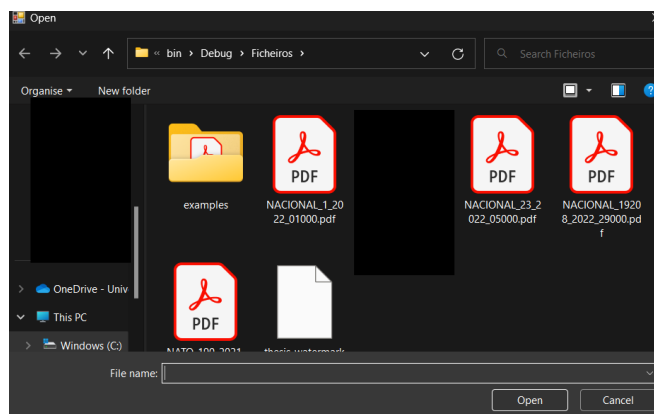


Figura 23 - Abrir Documento

## 6.7. Retificação do documento

Para retificar um documento é necessário que o documento escolhido para retificação contenha a marca de água e cuja informação esteja na base de dados. Caso esteja, o programa, lê o código de barras 128 e mostra as informações contidas para o utilizador numa janela onde é possível ter a visualização do documento com a marca de água escolhido e algumas informações sobre as características do documento (metadados) (figura 24). Cabe ao utilizador averiguar se o documento e as características demonstradas são diferentes ou iguais, se determinar que são diferentes o utilizador pode proceder à verificação de integridade do documento, referido no ponto 6, determinando as zonas que foram alteradas no documento.



Figura 24 - Retificação documento

## 6.8. Verificação de integridade

A verificação de integridade ou análise forense, serve para retificar se o documento foi alterado, e em que zonas.

A solução para retificar um documento, é criar 3 segmentos de retas calculados através dos pontos criados nas extremidades da folha 9 (3 × 3) pontos, totalizando 81(9 × 9) pontos.

Para saber o número total de retas usou-se a expressão somatória do Gauss [12] (equação 1). O resultado da equação dá igual ao valor no código através do sublinhado a preto na figura 25.

$$n \times \frac{n(n-1)}{2}, n = 9 \leftrightarrow 9 \times \frac{9(9-1)}{2} = 9 \times \frac{9 \times 8}{2} = 9 \times \frac{72}{2} = 9 \times 36 = 324$$

Equação 1 - Gauss

Aquando das retas traçadas vai se calcular o ponto de interseção da mesma [13], com a expressão da figura 26, de seguida com base nas coordenadas dos caracteres obtidos no processamento do ficheiro verifica-se se o ponto pertence ao subconjunto das coordenadas, dando como output a letra correspondida.

Na figura 27 mostra um possível output da verificação da integridade, sendo o círculo amarelo o ponto de interseção e a azul o carácter que está na base de dados, extraído no processamento.







## 7. Caso de estudo

Este capítulo é destinado a demonstrar o funcionamento do algoritmo criado num documento exemplar fornecido pela empresa.

Como referido anteriormente existem dois tipos de ações que se pretende desenvolver no algoritmo, processamento de um documento sem marca de água para originar a marca de água e retificação de um documento com marca de água, sendo ele em formato digital ou pelo scan.

O algoritmo para o processamento segue a lógica do fluxograma da figura 31.

O documento escolhido para efeitos de teste, segue-se na figura 32 tendo uma página, denominado “NACIONAL\_1\_2022\_01000”. Depois de 1 min e 25 segundos (figura 33) de processamento descrito no diagrama de fluxo segue-se o documento com a marca de água presente na figura 34.

Na figura 35 encontra-se o resultado do documento inserido na base de dados ao fim da execução do programa.

Na figura 36 tem-se a informação contida no código de barras que é colocado no documento.

Na figura 37 representa a ação do utilizador caso aceite ou rejeite a marca de água de um documento.

Na figura 38 representa um pequeno input da inserção de posições do carácter na base de dados. As colunas start\_x, start\_y, stop\_x, stop\_y são os valores representados na figura 39.

Na figura 40 representa uma parte das retas traçadas com base nos pontos dados através do código de barras.

A retificação segue a lógica do fluxograma da figura 41.

O resultado da retificação apresenta-se na figura 42, e caso a informação não seja consistente, realiza-se a verificação de integridade ou análise forense presente na figura 43. O tempo de processamento quer na retificação quer na análise forense é pouco, tendo só o tempo de processamento de mostrar as respetivas janelas com a pre-visualização do documento.

Existindo a possibilidade de os documentos chegarem em formato scaneados, tendo em conta ao documento digitalizado estar contido na base de dados, criou-se dois scans diferentes um com a folha direita (figura 44) e outro com a folha torta (figura 46), sendo que o algoritmo consegue detetar o código de barras num scan normal (figura 44) e realizar a verificação de integridade (figura 45), contudo para o scan torto não consegue, dando um erro de insucesso de leitura (figura 47).

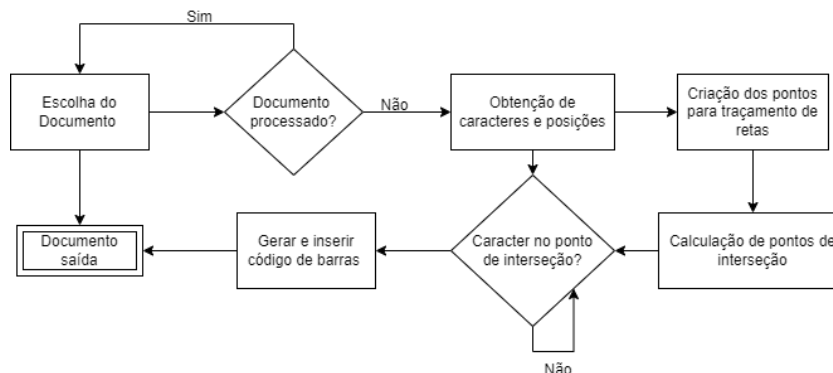


Figura 28 - Processamento diagrama de fluxo

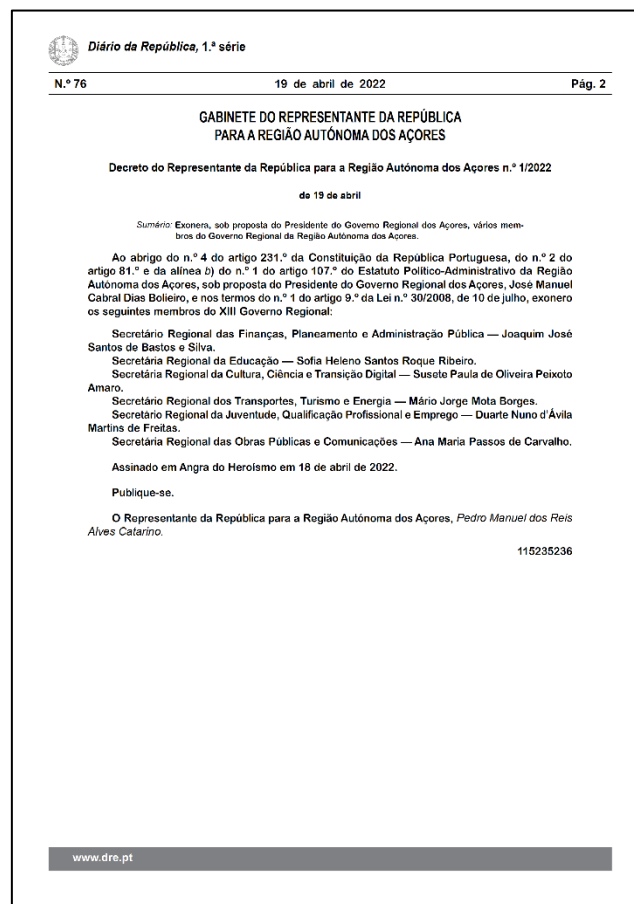


Figura 29 - Documento exemplar sem marca de água

Execution Time 00:01:25.2440000

Figura 30 - Tempo de execução

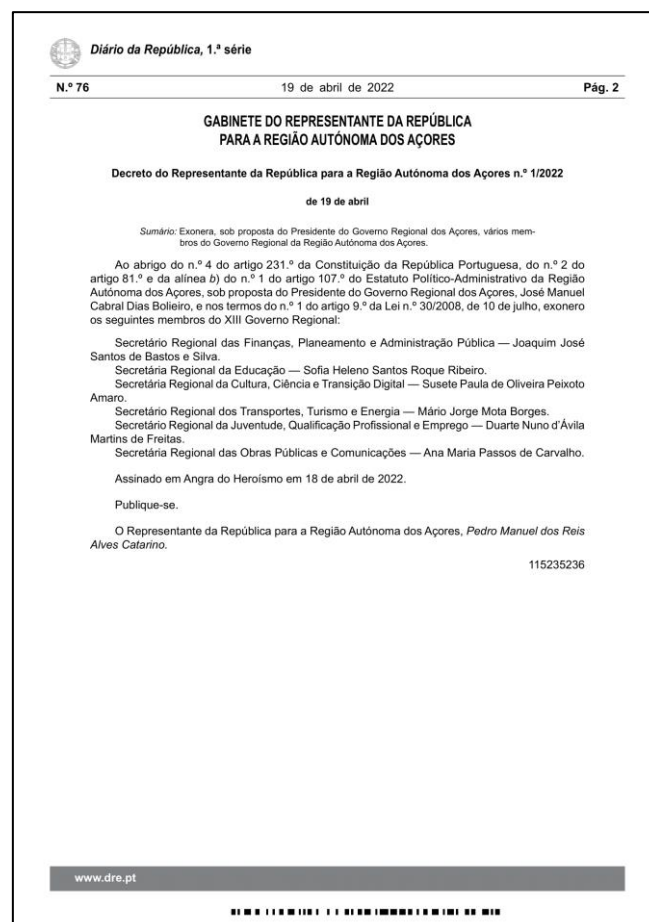


Figura 31 - Documento com marca de água

id_document	nome_ficheiro	n_registo	n_exemplar	n_copy	class_seg	estado_ex	formato_ex	utilizador	data_op	sigla_principal	posto_atual	dominio
1770408244	NACIONAL_1_2022_01000	1/2022/01000	1	0	S	Arquivado	Eletronico	João Francisco	31/01/2022 15:01:35	Decreto do Representante da República para os Aç.	Registo Central	NACIONAL

Figura 32 - Documento

id_barcode	positions_circleX
1	50,50 297,50 525,50 50,421 297,421 525,421 50,817 297,817 525,817

Figura 33 - Código de barras

id_doc	id_barcode	validacao
1770408244	1	1

Figura 34 - Marca de água documento

	id_doc	value_char	start_x	start_y	stop_x	stop_y
1	1770408244	D	71	29	76	37
2	1770408244	i	79	29	79	37
3	1770408244	á	82	29	85	37
4	1770408244	r	88	29	89	37
5	1770408244	i	92	29	92	37
6	1770408244	o	95	29	98	37
7	1770408244	d	104	29	108	37
8	1770408244	a	111	29	114	37
9	1770408244	R	120	29	125	37
10	1770408244	e	128	29	131	37
11	1770408244	p	134	29	137	37
12	1770408244	ú	140	29	144	37
13	1770408244	b	147	29	151	37
14	1770408244	l	154	29	154	37
15	1770408244	i	157	29	157	37
16	1770408244	c	160	29	163	37

Figura 35 - Obtenção de caracteres

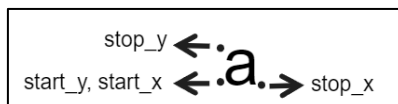


Figura 36 - Representação dos valores dos caracteres

	id_doc	line1	line2	inter_point	inter_char	line1_points	line2_points
1	1770408244	point1_l:point5_l	point2_l:point4_b	767,874	G	268,123:1297,1669	1297,123:148,1754
2	1770408244	point1_l:point5_l	point2_l:point7_r	936,1126	s	268,123:1297,1669	1297,123:148,3318
3	1770408244	point1_l:point5_l	point2_l:point8_r	1242,1586	ã	268,123:1297,1669	1297,123:1177,3318
4	1770408244	point1_l:point5_l	point2_l:point8_b	1243,1588	ã	268,123:1297,1669	1297,123:1177,3403
5	1770408244	point1_l:point5_l	point2_b:point7_r	893,1063	p	268,123:1297,1669	1177,208:148,3318
6	1770408244	point1_l:point5_l	point2_b:point8_l	1229,1567	ç	268,123:1297,1669	1177,208:1297,3318
7	1770408244	point1_l:point5_l	point3_r:point4_l	930,1118	s	268,123:1297,1669	2127,123:268,1669
8	1770408244	point1_l:point5_l	point3_r:point7_b	1243,1587	ã	268,123:1297,1669	2127,123:148,3403
9	1770408244	point1_l:point5_r	point2_b:point7_l	891,1183	G	268,123:1177,1669	1177,208:268,3318
10	1770408244	point1_l:point5_r	point3_l:point4_l	890,1182	G	268,123:1177,1669	2247,123:268,1669
11	1770408244	point1_l:point5_r	point3_l:point4_b	888,1178	G	268,123:1177,1669	2247,123:148,1754
12	1770408244	point1_l:point5_r	point3_r:point4_r	853,1118	r	268,123:1177,1669	2127,123:148,1669
13	1770408244	point1_l:point5_r	point3_b:point4_l	889,1180	G	268,123:1177,1669	2127,208:268,1669
14	1770408244	point1_l:point5_r	point3_b:point4_b	887,1176	G	268,123:1177,1669	2127,208:148,1754
15	1770408244	point1_l:point5_b	point2_l:point4_l	736,964	4	268,123:1177,1754	1297,123:268,1669
16	1770408244	point1_l:point5_b	point2_l:point7_r	893,1245	F	268,123:1177,1754	1297,123:148,3318

Figura 37 - Pontos para verificação de integridade

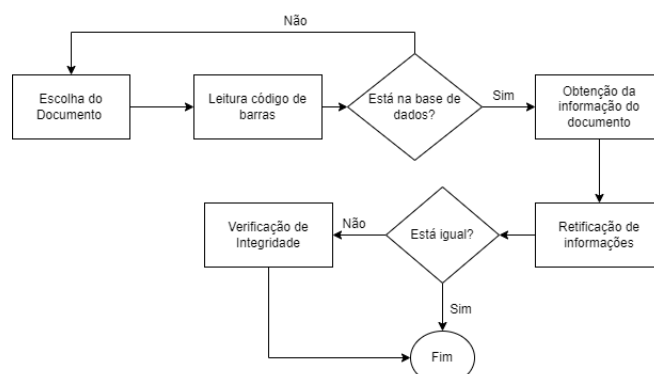


Figura 38 - Fluxograma Retificar



Figura 39 - Resultado Retificação

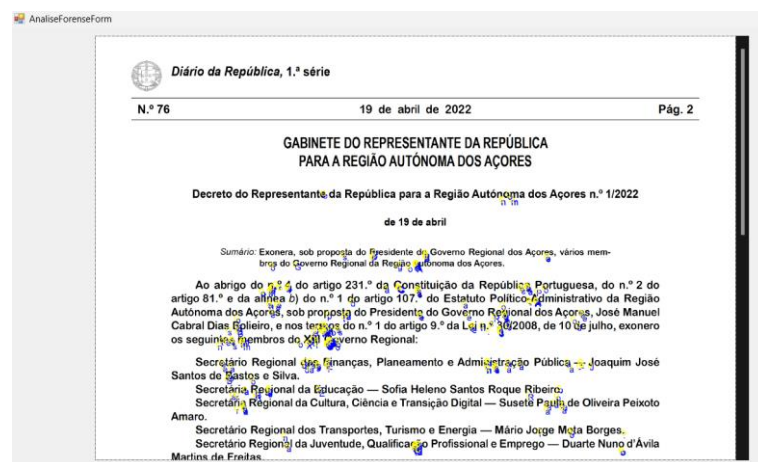


Figura 40 - Resultado Análise Forense



Figura 41 - Scan normal

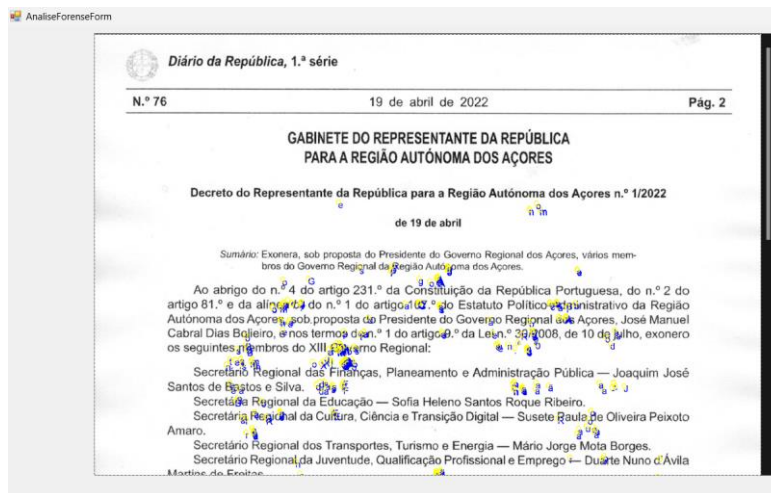


Figura 42 - Resultado scan normal

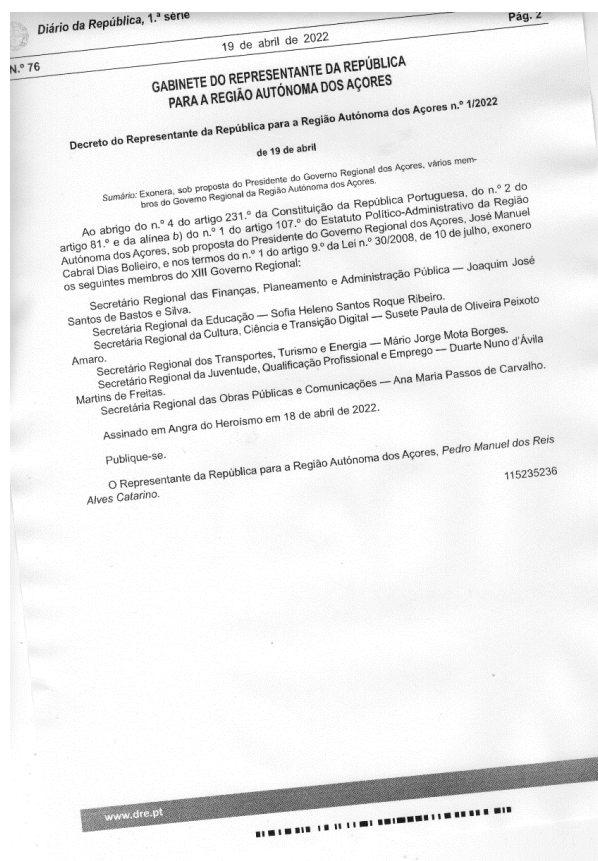


Figura 43 - Scan torto

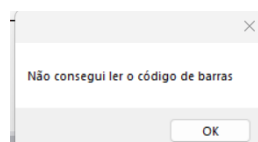


Figura 44 - Resultado scan torto





## 8. Conclusões



## 9. Futuro trabalho



## 10.Referências

- [1] “C# .NET Barcode Quickstart Guide, Code Examples | Iron Barcode.” <https://ironsoftware.com/csharp/barcode/examples/barcode-quickstart/> (accessed Jul. 26, 2022).
- [2] M. B. Mohd, S. Mohd, R. Tanzila, and S. A. Rehman, “Replacement Attack: A New Zero Text Watermarking Attack,” *3D Res.*, vol. 8, 2017, doi: 10.1007/s13319-017-0118-y.
- [3] Z. Jalil, A. M. Mirza, and H. Jabeen, “Word length based zero-watermarking algorithm for tamper detection in text documents,” *ICCET 2010 - 2010 Int. Conf. Comput. Eng. Technol. Proc.*, vol. 6, no. May, 2010, doi: 10.1109/ICCET.2010.5486185.
- [4] T. Rethika, I. Prathap, R. Anitha, and S. V. Raghavan, “A novel approach to watermark text documents based on eigen values,” *2009 Int. Conf. Netw. Serv. Secur. N2S 2009*, no. c, pp. 1–5, 2009.
- [5] J. T. Brassil, S. Low, and N. F. Maxemchuk, “Copyright protection for the electronic distribution of text documents,” *Proc. IEEE*, vol. 87, no. 7, pp. 1181–1196, 1999, doi: 10.1109/5.771071.
- [6] T. Saba, M. Bashardoost, H. Kolivand, M. S. M. Rahim, A. Rehman, and M. A. Khan, “Enhancing fragility of zero-based text watermarking utilizing effective characters list,” *Multimed. Tools Appl.*, vol. 79, no. 1–2, pp. 341–354, Jan. 2020, doi: 10.1007/S11042-019-08084-0/TABLES/5.
- [7] “Code 128 - Wikipedia.” [https://it.wikipedia.org/wiki/Code\\_128](https://it.wikipedia.org/wiki/Code_128) (accessed Jan. 19, 2023).
- [8] “Relational Vs. Non-Relational Databases | MongoDB | MongoDB.” <https://www.mongodb.com/compare/relational-vs-non-relational-databases> (accessed Jan. 19, 2023).
- [9] “MySQL.” <https://www.mysql.com/> (accessed Jan. 19, 2023).
- [10] “Oracle SQL Developer Downloads.” <https://www.oracle.com/database/sqldeveloper/technologies/download/> (accessed Jan. 19, 2023).
- [11] “SQL Server Management Studio (SSMS) - SQL Server Management Studio (SSMS) | Microsoft Learn.” <https://learn.microsoft.com/en-us/sql/ssms/sql-server-management-studio-ssms?view=sql-server-ver16> (accessed Jan. 19, 2023).
- [12] “The Story of Gauss - National Council of Teachers of Mathematics.” <https://www.nctm.org/Publications/TCM-blog/Blog/The-Story-of-Gauss/> (accessed Feb. 08, 2023).
- [13] “Line–line intersection - Wikipedia.” [https://en.wikipedia.org/wiki/Line–line\\_intersection](https://en.wikipedia.org/wiki/Line–line_intersection) (accessed Oct. 16, 2022).
- [14] “Apache PDFBox | A Java PDF Library.” <https://pdfbox.apache.org/> (accessed Set. 17, 2022).
- [15] “About Us | IronSoftware.com.” <https://ironsoftware.com/about-us/> (accessed Jul. 26, 2022).
- [16] “C# Tesseract OCR In 1 Line of Code | Iron OCR.” <https://ironsoftware.com/csharp/ocr/examples/simple-csharp-ocr-tesseract/> (accessed Jul. 26, 2022).

# Apêndice A

## Processamento de ficheiro PDF

Apesar de existirem alguns packages em c# retirassem as posições dos caracteres bem como os mesmos, não era suficiente, já que se precisava de um intervalo de valores donde começa e acaba a letra para determinar se o ponto de interseção pertence ou não à letra para ser usado na verificação de integridade, optou-se por utilizar o package Apache PDFBox [14] desenvolvido em java que conseguia, dar as posições do começo e fim de cada letra bem como a letra respetiva.

Para diminuir o tempo de processamento e para efeitos de teste apenas se lê a primeira página do documento (figura 48), sendo possível depois alterar para todas as páginas, contudo o tempo de processamento aumenta também. Para os valores lidos serem acedidos em C# criou-se um ficheiro temporário que vai guardar as seguintes características “character|start\_x,start\_y,stop\_x,stop\_y” como se representa na figura 49.

Contudo é necessário compilar o código em java e criar um ficheiro jar com os respetivos packages dependentes que permita a execução em c#, para isso abriu-se uma consola e executou-se um comando presente na figura 50.

```
PDDocument document = null;
file_name = args[0];
String[] f = file_name.split( regex: ".pdf");

File fi = new File( pathname: f[0]+"_pos.txt");

if(fi.exists())
{
    fi.delete();
}
else {
    fi.createNewFile();
}

String fileName = args[0];
try {
    document = PDDocument.Load( new File(fileName) );
    PDTextStripper stripper = new PositionCharacter();
    stripper.setSortByPosition( true );
    stripper.setStartPage( 0 );
    stripper.setEndPage( 1 );

    Writer dummy = new OutputStreamWriter(new ByteArrayOutputStream());
    stripper.writeText(document, dummy);
}
finally {
    if( document != null ) {
        document.close();
    }
}
```

Figura 45 - Código de extração dos caracteres num ficheiro PDF

```
protected void writeString(String string, List<TextPosition> textPositions) throws IOException {
    try{
        String[] f = file_name.split( regex: ".pdf");
        FileWriter file = new FileWriter( fileName: f[0]+"_pos.txt", append: true);
        String ch;

        for (TextPosition text : textPositions) {
            ch = text.getUnicode();

            //remove -
            if (ch.equals("-"))
                ch = " ";
            else if (ch.equals(""))
                ch = " ";

            if(!ch.isBlank() && !ch.isEmpty()) // remove spaces
                file.write( ch + "|" + Math.round(text.getX()) + "." + Math.round(Math.abs(text.getHeight() - text.getY()))
                    + "." + Math.round(text.getEndX() - text.getWidthOfSpace()) + "." + Math.round(text.getY()) + "\n");
        }
        file.close();
    } catch (IOException e){
        System.out.println("An error occurred");
        e.printStackTrace();
    }
}
```

Figura 46 - Obtenção dos valores

```
private void Get_PositionChar()
{
    Process process_file = new Process();
    process_file.StartInfo.UseShellExecute = false;
    process_file.StartInfo.RedirectStandardOutput = true;
    process_file.StartInfo.FileName = "java";
    process_file.StartInfo.Arguments = "-jar " + '"' + jar_file + '"' + " " + '"' + file_name + '"';
    process_file.Start();
    process_file.WaitForExit();
}
```

Figura 47 - Execução jar

## Iron Software

Para desenvolver o programa foi necessário usar tecnologias de criação e leitura de código de barras. O que levou a usar um software open-Source, que tem a opção de comprar a licença para efeito de segurança e apoio, denominado Iron Software [15].

Iron software é usado por mais de 6.9 milhões de utilizadores tendo vários produtos como o Iron PDF, Iron OCR, Iron Barcode, Iron XL, Iron Updates e Iron suite, listados na figura 51. Sendo que neste projeto utilizou-se Iron barcode para gerar e leitura do código de barras.

A geração do código de barras está presente na figura 52, e a leitura na figura 53.

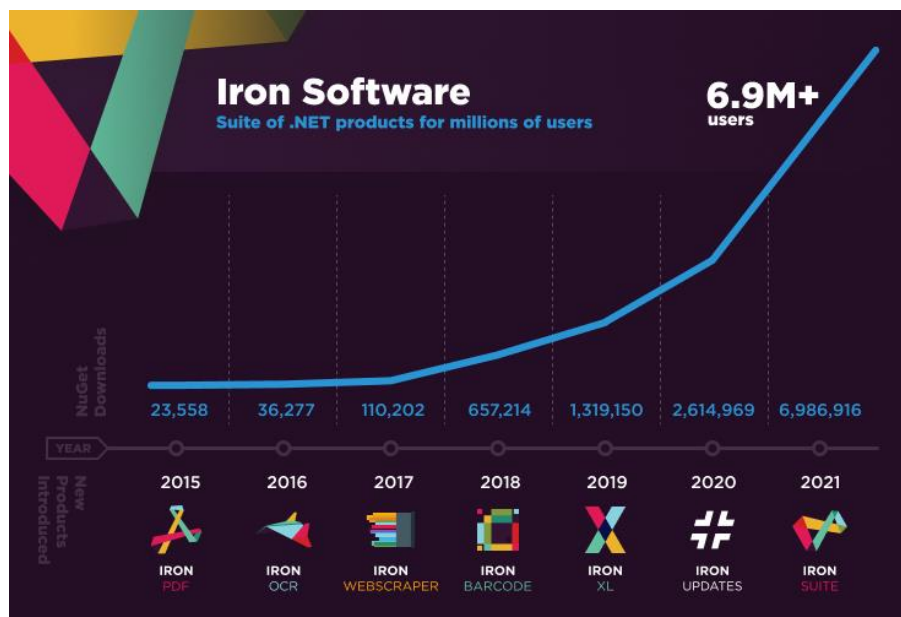


Figura 48 – Iron Barcode

```
public void Generate_barcode(int id_barcode)
{
    string data_barcode = id_doc.ToString() + ";" + id_barcode.ToString();
    GeneratedBarcode MyBarCode = BarcodeWriter.CreateBarcode(data_barcode, BarcodeWriterEncoding.Code128);
    MyBarCode.ResizeTo(MyBarCode.Width, resizedBarcode); // pixels
    MyBarCode.ChangeBarCodeColor(Color.Black);
    MyBarCode.SaveAsPng(filename + "_barcode.png");
}
```

Figura 49 - Geração código de barras

```
public string Read_barcode(string file_name)
{
    BarcodeResult BarCodeResult = BarcodeReader.QuicklyReadOneBarcode(file_name, BarcodeEncoding.Code128, true);
    if (BarCodeResult != null)
        return BarCodeResult.Value;
    return "insucesso";
}
```

Figura 50 - Leitura de código de barras