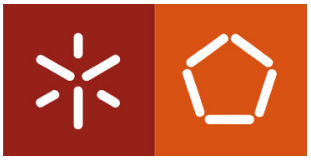
Universidade do Minho

Escola de Engenharia



Scripting no Processamento de Linguagem Natural

Mestrado em Engenharia Informática

Desenvolvimento de uma Ferramenta de Anonimização de Texto

Trabalho Prático 2

Gonçalo Ferreira - [PG50404]
Rui Braga - [PG50743]

Junho, 2023

1 Introdução e Objetivos

O presente documento serve de documentação ao desenvolvimento de uma ferramenta de anonimização de ficheiros de diversos formatos, que permite a anonimização de nomes, locais, números de telemóvel, entre outros dados privados que precisem de ser anonimizados.

A anonimização de dados é um processo que tem como objetivo remover ou ocultar informações pessoais identificáveis dos dados, de forma a garantir a privacidade dos indivíduos e a proteção de dados sensíveis. Instituições como Tribunais ou Hospitais possuem grandes quantidades de informação que pode ser utilizada para vários fins. Tratando-se de informações sensíveis, tais como acórdãos judiciais e relatórios médicos é necessário proceder à anonimização dos mesmos de modo a garantir a segurança e privacidade das pessoas mencionadas nesses documentos.

Com isto em mente, os principais objetivos estabelecidos para o desenvolvimento desta ferramenta são:

- O desenvolvimento de uma ferramenta instalável em Python, que permita a identificação de dados sensíveis, e os anonimize (podendo para isso, ocultar os mesmos ou substituir)
- A ferramenta deverá suportar vários tipos de formatos de texto, incluindo (mas não exclusivamente) .pdf, .txt, .html, .md.
- A ferramenta deverá estar funcional para textos de entrada em Português e Inglês.

Este projeto está incluído na componente prática da unidade curricular de Scripting no Processamento de Linguagem Natural (do perfil de Engenharia de Linguagens, parte integrante do Mestrado em Engenharia Informática), e o presente relatório será constituídos pelas seguintes secções: a **Arquitetura da Solução**, que descreverá o desenho do algoritmo e os objetivos e capacidades da ferramenta; a **Implementação**, onde de forma breve serão expostos alguns detalhes acerca da implementação, e por fim a **Demonstração**, onde serão apresentados alguns testes de forma a apresentar o funcionamento da ferramenta desenvolvida. Na conclusão, o grupo avaliará o trabalho desenvolvido, apresentando aspetos que desejaria melhorar no futuro da ferramenta.

2 Arquitetura

O processo de anonimização de um ficheiro de texto, deverá passar pelo seguinte processo:

1. A ferramenta deverá ser executada com argumentos, de forma a escolher o ficheiro de input e a estabelecer as flags para controlar o processamento do texto.
2. A ferramenta trata agora de ler o ficheiro de input, e passar o componente de reconhecimento de entidades, que com recurso a modelos de NLP, e expressões regulares deverá reconhecer entidades a anonimizar
3. O output da etapa anterior, deve então ser passado para um processador que associa as entidades encontradas, e trata da sua anonimização no texto original, através de ocultação ou substituição.
4. Por fim, um ficheiro de output será produzido, com as entidades encontradas anonimizadas.

O processo acima descrito será a base do processamento realizado pela a ferramenta. Contudo para cumprir o requisito de anonimização de PDFs serão adicionados alguns passos extra, nomeadamente:

- Para que a componente de reconhecimento de entidades seja capaz de encontrar as entidades, o PDF deverá ser transformado em ficheiros texto, um por página.

- Para que o output da ferramenta seja igual ao input, o PDF será também transformado numa sequência de imagens, que serão depois processadas com um módulo python de "computer vision", que reconhecerá nas imagens, o texto indicado como anonimizável.
- Para além de encontrar nas imagens, a ferramenta irá também desenhar uma forma retangular sobre o token a ser anonimizado.
- A ferramenta deverá no final da execução unir as imagens e transformar as mesmas nas páginas do PDF original, mantendo a sua ordem.

Com este processamento de PDFs, existirá o problema de o PDF se tornar "read-only"porque, este será agora constituído por imagens, das quais não é possível extrair texto. O grupo pesou os prós e contras desta abordagem, e chegou a conclusão que esta seria a melhor alternativa no processamento de PDFs, que acaba por manter a estrutura (por vezes muito complexa e difícil de tratar) e permitindo a fácil leitura a um utilizador, ao custo de impossibilitar a extração de texto; de qualquer forma, será também possível transformar os PDFs em texto, e anonimizar estes textos.

Os argumentos passados ao programa são:

```
options:
  -h, --help            show this help message and exit
  -o OUTPUTFILE, --outputFile OUTPUTFILE
                        Name of the output file (without extension)
  -p PATTERNS, --patterns PATTERNS
                        Name of a file containing patterns to be detected
                        (without extension)
  -org, --organizations
                        Extends entity recognition to include organizations
  -m, --moreRestrictive
                        Extends entity recognition to include entities
                        that are highly restrictive
  -p2t, --pdf2text      Flag that indicates if pdf should be turned into .txt
  -r, --replace         Flag that indicates that the program should replace
                        entities found
  -l LANGUAGE, --language LANGUAGE
                        Language of the input file: en or pt
```

É possível, através da opção, $-p$ a utilização de um ficheiro que contenha algumas expressões regulares que serão utilizadas para identificar mais entidades a serem anonimizadas. Desta forma o utilizador poderá definir as suas próprias expressões regulares, e utilizar a nossa ferramenta para anonimizar de acordo com a sua preferência.

A ferramenta é capaz de anonimizar em duas línguas diferentes (definidas com a opção $-l$): português e inglês. É também capaz de, permitir que o utilizador defina o nível de restrição, $-org$ inclui a anonimização de organizações e $-m$ inclui a anonimização de todas as entidades que o *spacy* identificar.

3 Implementação

Para a implementação da ferramenta o grupo suportou se nos seguintes módulos python:

- **Spacy** - módulo Python de processamento avançado de linguagem natural; permitiu a identificação e labeling de entidades, como nomes, localizações etc.
- **Re** - biblioteca de métodos, que permite a utilização de expressões regulares, para executar pesquisas e substituições em texto; permitiu auxiliar o spacy na identificação de algumas entidades mais específicas, como moradas, números de telemóvel etc.

- **PyPDF, pdf2image & fpdf** - módulos utilizados para a transformação de pdf em texto, transformação de pdf em imagens e composição de imagens num único PDF.
- **CV2 & Pytesseract** - módulos de "computer vision", que permitiram identificar as entidades reconhecidas no texto, nas páginas do PDF

3.1 Identificação de Dados a Anonimizar

Para a identificação de dados, divide-se o texto por linhas. Com o texto dividido, procuram-se correspondências nos excertos de dados começando pelos padrões extra que o utilizador pode definir. De seguida, procuram-se correspondências com as expressões regulares de moradas e de datas que podem envolver espaços. O passo seguinte é dividir o excerto de texto em tokens com o método *split*() e procurar correspondências com as restantes expressões regulares.

Por fim, os excertos de texto são processados com o modelo carregado com o *spacy*, de forma a identificar as entidades presentes no texto, que não foram detetadas com as expressões regulares.

```
1. dividir o texto original em linhas
2. procurar correspondências com os padrões extra que o cliente definiu
3. procurar correspondências com os padrões de morada e data que podem
   envolver espaços
4. dividir uma linha em tokens
5. procurar expressões regulares por token
6. processar uma linha com um modelo do spacy
7. identificar restantes entidades
```

3.2 Anonimização do texto

Concluída a identificação das entidades anonimizáveis, resta agora à ferramenta dependendo do tipo de ficheiro, e de se o utilizador selecionou a opção de substituição ou de censura de entidades (para PDFs apenas está disponível a censura de entidades). No caso dos ficheiros de texto:

- Se foi escolhida a opção de **censura** de entidades, o programa utilizará o módulo re, para substituir as entidades encontradas no texto por uma string do mesmo tamanho com "*".
- Caso tenha sido escolhida a opção de **substituição**:
  - O programa armazenará as entidades num dicionário, organizado pelo tipo da entidade. Esta etapa irá permitir manter o contexto do texto, sendo que sempre que for mencionada uma entidade A, a mesma vai ser sempre substituída pela mesma marca textual;
  - Para entidades marcadas com o rótulo de "Pessoa", o nome será substituído por uma marca textual, composta pelas iniciais de cada nome, seguidas por um ponto e rodeadas por 2 #;
  - As restantes entidades, serão marcadas de acordo com o rótulo atribuído, resultando numa marca textual do seguinte tipo #[*label*][*id*]#; o id é apenas um escalar, que serve como identificador daquela entidade

Já a anonimização dos PDFs requiriu um processamento um pouco mais complexo:

1. A imagem de cada página deverá ser carregada, e o texto nela contido terá de ser extraído. Esta extração é levada a cabo pelo módulo *pytesseract*, que é capaz de extrair todo o texto encontrado, em conjunção com as *"bounding boxes"*, que contêm cada palavra. Para obtenção de melhores resultados as cores da imagem foram transformadas para escala de cinzentos, para melhorar a deteção de texto.

2. Após obter todo o texto da imagem, é necessário percorrer todas as entidades identificadas no formato texto; assim, para cada entidade, é corrido o texto encontrado até que seja encontrada a entidade na sua totalidade

3. Após encontrar uma entidade na imagem, as *"bounding boxes"* das palavras são unidas e é utilizado o módulo cv2, para desenhar um retângulo azul sobre a *bounding box* que delimita a entidade, censurando desta forma a entidade.

4. Por fim a imagem é escrita num ficheiro, e será mais tarde junta com as restantes páginas para que um PDF seja criado.

4 Demonstração

4.1 Limitações Conhecidas

- Os modelos do Spacy de processamento de linguagem natural em português, são relativamente pequenos e por isso, a sua precisão na identificação de entidades poderá ser demasiado permissiva, reconhecendo entidades incorretamente.
- Os PDFs gerados após o processamento são uma coleção de imagens, o que os torna read-only, e impossibilita o processamento do seu texto
- A ferramenta não permite a substituição de entidades em PDFs

4.2 Testes

Ficheiro de teste:

```
Prezado João,

Para confirmar a sua presença neste evento exclusivo que irá decorrer
na Avenida da Liberdade, pedimos a gentileza de responder a este e-mail
até 30 de junho de 2023. Caso tenha alguma dúvida ou necessite de mais
informações, não hesite em entrar em contato conosco pelos meios indicados abaixo.

Atenciosamente,

Tony Stark
Stark Industries
im_stark@gmail.com
```

Ficheiro de teste depois de ter sido aplicada a ferramenta construída e com uma execução simples (sem opções extra):

```
**********,

Para confirmar a sua presença neste evento exclusivo que irá decorrer
na ******************, pedimos a gentileza de responder a este e-mail
até ******************. Caso tenha alguma dúvida ou necessite de mais
informações, não hesite em entrar em contato conosco pelos meios indicados abaixo.

Atenciosamente,

**********
```

```
***************
*****************
```

Com a opção de substituição ativada:

```
#P. J.#,

Para confirmar a sua presença neste evento exclusivo que irá decorrer
na #morada...0#, pedimos a gentileza de responder a este e-mail
até #data...0#. Caso tenha alguma dúvida ou necessite de mais
informações, não hesite em entrar em contato conosco pelos meios indicados abaixo.

Atenciosamente,

#T. S.#
#S. I.#
#email...0#
```

5 Conclusões

Neste relatório foi apresentada a arquitetura e implementação de uma ferramenta de anonimização em Python, capaz de esconder e/ou substituir diversos dados privados, de uma grande variedade de ficheiros de texto. A esta ferramenta o grupo adicionou diversas funcionalidades que melhoram a experiência do utilizador como a possibilidade de anonimização de expressões regulares customizadas, definidas através de um ficheiro de configuração.

Desta forma, e apesar de algumas limitações conhecidas da ferramenta (que o grupo gostaria de ter melhorado), o grupo faz uma avaliação muito positiva do projeto desenvolvida, apresentando uma ferramenta que cumpre todos os objetivos definidos, acrescentando novas funcionalidades, que o grupo considerou úteis para qualquer possível utilizador que deseje anonimizar os seus documentos.

Nota:
A ferramenta é facilmente instalável recorrendo ao módulo Python flit, e estará disponível no github `https://github.com/GoncaloPereiraFigueiredoFerreira/blue-pencil`