Integration of Deep Computer Vision Foundation Models for Document Interpretation and Anonymisation

Bachelor Thesis

BSc – Computer Science
Eastern Switzerland University of Applied Sciences
Campus Rapperswil-Jona

Fall Term 2023

Authors	Marc Havrilla
Version	2024-01-12
Advisors	Marco Lehmann
External Co-Examiner	Johanni Michael Brea
Internal Co-Examiner	Stefan F. Keller

Contents

CHAPTER 1

Abstract

Visual Document Understanding (VDU) models, combined with Optical Character Recognition (OCR) or OCR-free, offer businesses and institutions a great opportunity to digitalise their processes and improve workflows. The digitalisation is progressing. However, challenges like sufficient knowhow to integrate VDU models, compliance with data protection regulations and identifying the processes, where VDU models offer the most significant benefit, have to be resolved.
The main goal of the work is to analyse and evaluate the practicality and appropriateness of available VDU models for processing of documents (e.g. PDF of scanned documents) and to demonstrate these in a Proof-of-Concept (POC) application. Even though some regulatory aspects, especially regarding anonymisation, are discussed in the work, the developed application does not aspire to be regulatory compliant.

During this work, two areas have been identified, where a tool to extract text from an image, identify relevant entities of personal information and anonymise these, is beneficial. First, the anonymisation of medical documents makes more data available for research and educational purposes. A second application is data leakage prevention, where detecting client data from screenshots would lower the risk of data breaches.
Various tools exist to extract text from an image. In the scope of this project, three tools have been integrated i.e., Tesseract, Amazon Textract and OpenAI GPT-4V(ison). The application extracts the text of uploaded documents or images and provides the user with the resulting text from all three tools. The user will be able to select the text with the best quality. Afterwards, a Named Entity Recognition (NER) Transformer model (i.e., bert-base-NER model) is used to identify the names of persons in the extracted text. The last step is the pseudonymisation of the entities. A randomly generated unique string replaces the entities in the text, so that a person cannot be identified based on the name in the text.

Another feature of the application is the evaluation of the OCR accuracy. The user is able to upload an additional ground truth file, which will then be compared with the output of the uploaded images. To calculate the OCR accuracy the Jaro Similarity string comparison algorithm is used. Furthermore, the NER model can also be tested by uploading the expected entities of the document in a separate file. The test will then show how many of the provided entities have been found in the extracted text.

It is impressive how powerful today's text extraction and NER models have become. However, during the work, it was recognised that they are not yet off-the-shelf and just ready to use. Neither works each tool perfectly, so errors are propagated to subsequent processes nor are the outputs of each tool standardised. To overcome such limitations, the process of text extraction and entity recognition should be executed by one model, which is also fine-tuned on the specific document types.

CHAPTER 2

Management Summary

On of the goals of this project is to build a Proof-of-Concept application using Visual Document Understanding (VDU) to find names of persons in a PDF file or image containing text. Afterwards, the identified names will be anonymised or pseudonymised. This means, the application has to be able to extract text from an image, identify the name entities in the text, anonymising the entities and return a text file with the anonymised text. Difference between anonymisation and pseudonymisation is that pseudonymisation can be reverted, whereas anonymisation is irreversible.

The above mentioned process steps can be achieved by using VDU models, which often depend on Optical Character Recognition (OCR) but not always. VDU means to extract information from digital images or documents and to process these information. To make an image readable for a computer, tools like OCR engines are used. There exist various tools which are able to perform text extraction, entity recognition or anonymisation. However, there are not many that combine all this processes into one application like the Anonymiser app does. Anonymiser App is the name of the POC application. The second goal is to analyse and evaluate the practicality and appropriateness of currently available tools and techniques, which can be used to extract text, to identify name entities or to perform anonymisation.

What Is The Purpose?

VDU offer businesses and institutions a great opportunity to digitalise their processes and improve workflows. During this work, two areas have been identified, where a tool like the Anonymiser app is beneficial. First, the anonymisation of medical documents makes more data available for research and educational purposes. Secondly, in the area of data leakage prevention, where detecting client data from screenshots would lower the risk of data breaches.

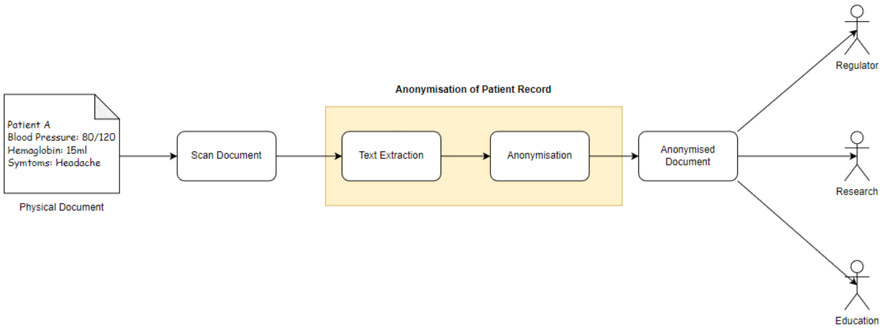


Figure 2.1: Anonymiser app used for the anonymisation of patient records

How Is Built and How Does It work?

Various tools exist to extract text from an image. To build the Anonymiser app, three text extraction tools have been included i.e., Tesseract, Amazon Textract and OpenAI GPT-4V(ision). In a first step, the application extracts the text of uploaded documents or images and provides the user with the resulting text from all three tools. The user will be able to select the text with the best quality. In a second step the tool will look for all names of persons in the text. This is done with a Named Entity Recognition (NER) Transformer model. The output of this phase is a list with all personal names. Afterwards, the anonymisation or pseudonymisation of the entities is done. Each name in the list will be replaced by a randomly generated unique string. In a last step, these random strings overwrite also the names in the uploaded text so that a person cannot be identified anymore based on the name.

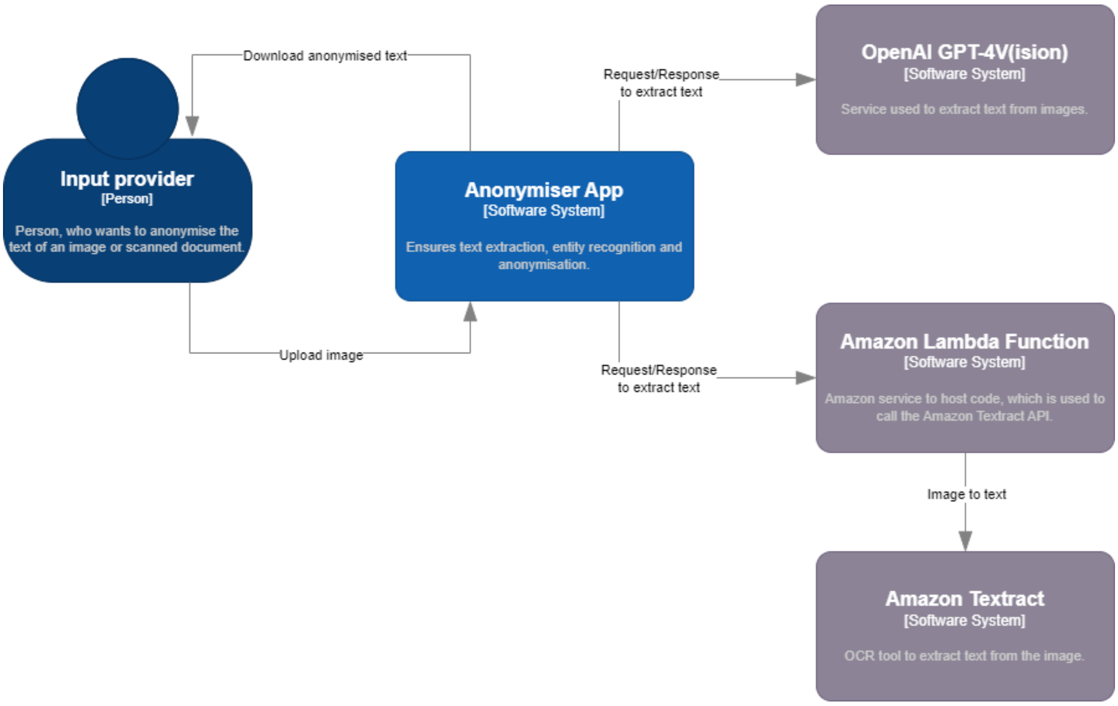

Download anonymised text
Request/Response to extract text
OpenAI GPT-4V(ision)
[Software System]
Service used to extract text from images.
Input provider
[Person]
Person, who wants to anonymise the text of an image or scanned document.
Anonymiser App
[Software System]
Ensures text extraction, entity recognition and anonymisation.
Upload image
Request/Response to extract text
Amazon Lambda Function
[Software System]
Amazon service to host code, which is used to call the Amazon Textract API.
Image to text
Amazon Textract
[Software System]
OCR tool to extract text from the image.

Figure 2.2: Highlevel illustration how the Anonymiser app works

How to Check If the Application Works Properly?

To check whether the application works properly, a feature to compare the extracted text with the original text is integrated. The user is able to upload the original text, also called ground truth, which will then be compared with the output of the uploaded images. The Anonymiser app will then compare the two text and provide a result how similar these text are. If the similarity is very low, adjustments on the Anonymiser app have to be done. Furthermore, the Anonymiser app is also able to test if all names in a text have been detected.

Conclusion and Outlook

Today's text extraction and NER models have become powerful. However, during the work, it was recognised that they are not just ready to use. There are a lot of additional tools and applications that have to be downloaded and installed. In addition, neither works each tool perfectly, so errors are propagated to subsequent processes nor are the outputs of each tool standardised. Meaning, after each step, a post processing has to be performed. To overcome such limitations going forward, the process of text extraction and entity recognition should be executed within the same phase and the tool used, has to fit perfectly to the specific task. Nevertheless, the POC works and valuable experiences have been gained.

Glossary

Bidirectional Encoder Representations from Transformers Transformers model pretrained on a large corpus of English data. 18

Consolidated Receipt Dataset A Consolidated Receipt Dataset for Post-OCR Parsing. 38

Data Definition Language Syntax for creating and modifying database objects. 23

Data Leakage Prevention Actions to prevent the unintentional leakage of sensitive data. 8

European Union Union of states within Europe. 20

European Union Agency for Cybersecurity ENISA works with the EU, its member states, the private sector and Europe's citizens to develop advice and recommendations on good practice in information security. 21

Federal Act on Data Protection Act to protect the personality and fundamental rights of natural persons whose personal data is processed. 20

General Data Protection Regulation Regulation on the protection of personal data within the European Union. 20

Named Entity Recognition Natural language processing method. 7, 18

Natural Language Processing Machine learning technique to reveal the structure and meaning of text. 30

Optical Character Recognition Converts an image of text into a machine-readable text format. 7, 35

Personally Identifying Information Information which make a person identfiable. 23

Proof-of-Concept Minimal application, to show that the idea works. 7

Universally Unique Identifier A UUID is a unique 128 bit long value. 23, 30

Visual Document Understanding End-to-end paradigm for visual document interpretation. 6

Visual Question Answering VQA is a new dataset containing open-ended questions about images. 39

Part I

Product Documentation

CHAPTER 3

The Current Problem

According to PWC *"Digitization is the process of translating analog information and data into digital form. For example, scanning a photo or document and storing it on a computer" [7].* The digitalisation of processes offers excellent opportunities and facilitates workflows.

However, during the Corona pandemic, it has become apparent that the digitalisation transformation process has not made everywhere the same progress. Especially in the health sector the digitalisation lags. This might be the reason why there are many efforts to change these circumstances. According to the Federal Office for Public Health (FOPH), there will be a significant investment in digitalisation.
Also, Baden's cantonal hospital aims to bring its medical data to the cloud in order to enhance data analytics capabilities. There is a massive potential in the available data for scientific purposes, but it is currently not optimally used.

Making all information digitally available also has drawbacks. The risk of data breaches rises as the incident from May 2023 shows, when data from around 425'000 Swiss citizens living abroad appeared in the dark net. The leaked data was sensitive.
The abovementioned aspects are examples, which show that bringing the digitalisation further is essential, but measures to mitigate data beaches are also needed.

For sure, many solutions exist to digitalise processes. Nevertheless, Visual Document Understanding (VDU) offers another great opportunity for businesses and institutions. Especially when physical documents have to be digitialised and further processed. However, current VDU models are not yet well known, and experiences in integrating them have not yet reached the masses.

CHAPTER 4

Aim and Scope

This project aims to analyse and evaluate the practicality and appropriateness of available VDU models for processing documents (e.g., PDF of scanned documents) and to demonstrate these in a Proof-of-Concept (POC) application. The developed application is called the Anonymiser app and is able to extract text from image files as well as from PDF documents. Furthermore, the app will identify personal names in the extracted text and anonymise or pseudonymise these names.

The app's implementation is based on literature research in Optical Character Recognition (OCR), Named Entity Recognition (NER) and anonymisation. There is a summary of different available tools for OCR and anonymisation.

Based on the insights gained, the Anonymiser app will be built and integrated with some of the available tools. The findings made with the available tools and models will be documented during the implementation.
Even though regulatory aspects regarding anonymisation are discussed, the developed application does not aspire to be regulatory compliant nor is the chapter in the documentation conclusive.

In the end, the Anonymiser app has to be able to anonymise a personal name within an image or PDF and the experiences made during the implementation are documented.

4.1 Fields of Application

In the chapter *"The Current Problem"* two areas have been mentioned, which would take advantage of a tool like the Anonymiser app.

First, the Anonymiser app can be used to anonymise scanned medical documents. There are already aspirations ongoing to make the health care sector more digital. Anonyimsed documents can be used for secondary purposes like research or educational purposes and to transmit information to the regulator without a fax machine.

Below figure provides an overview how the Anonymiser app could be integrated in the process.

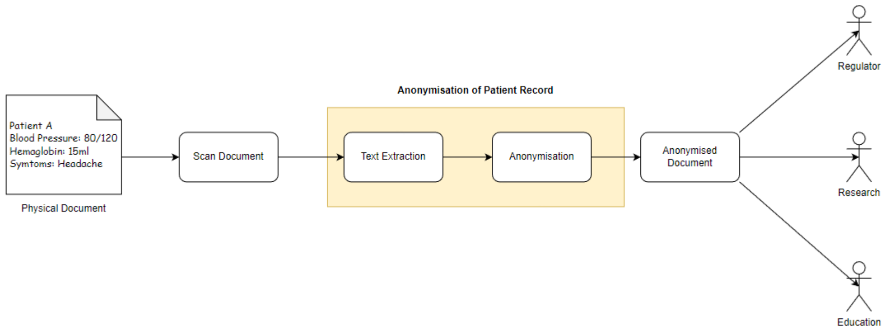


Figure 4.1: Anonymisation process for documents in the health sector

Another purpose for the Anonymiser app can be found in the Data Leakage Prevention (DLP). There exist already tools, which prevent the leakage of sensitive data. However, there exist loopholes, like embedding screenshots of sensitive data in a PDF file, which will then be send with an email outside the company.

The Anonymiser app, closes this loophole. Before an email with a PDF attachment will be sent outside the company, the Anonymiser app will extract the text in the PDF file and match the content against a database, containing all clients of a company. In case some clients appear in the email, it will be be flagged and pseudonymised. Now, the flagged email will be reviewed by a Information Security Office in order to assess the risk of a data leakage attempt. The content of the email is pseudoanonymised for this purpose in order to guarantee a high level of data privacy. In case the suspicion is confirmed, the pseudonymisation will be reverted and the email can be checked again.

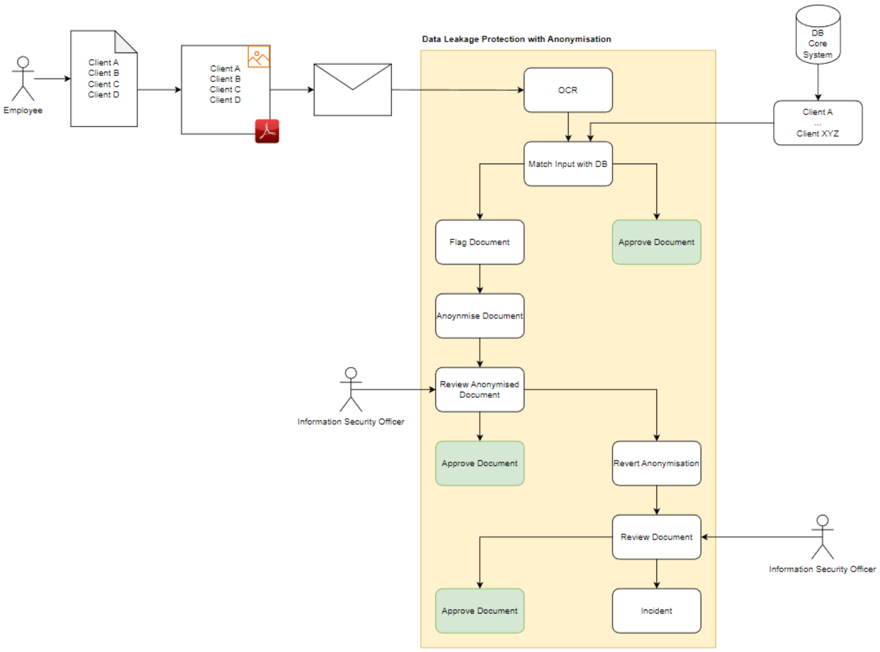


Figure 4.2: Anonymiser app used for data leakage prevention

4.2 Vision of the Product

Once a year, the big tech companies present their latest technological advances. Machine Learning (ML) and Natural Language Processing (NLP) are often essential parts of these presentations. When watching one of these presentations, the idea of the Anonymiser app was born. It is impressive how easy it became to translate menus in restaurants or articles by hoovering the mobile phone above the text and selecting the preferred language.

So, is there a possibility to use the latest technologies to make an app, which easily anonymises documents and images? The tool should offer an intuitive interface to upload documents and automatically anonymise or pseudonymise the whole document. Direct and indirect client identifying data is detected and replaced with the most suitable anoymisation method, so no relevant information is lost. Furthermore, the structure of the document will remain the same and can easily be shared.

The Anonymiser app should work on a mobile phone, where a picture of a document can be taken easily, anonymised and sent to a business partner. Ideally, the app is also integrated into email services, where the anonymisation of the attachments is done by simply clicking one button. This enables the sharing of relevant information, without breaching regulations.

CHAPTER 5

Requirements

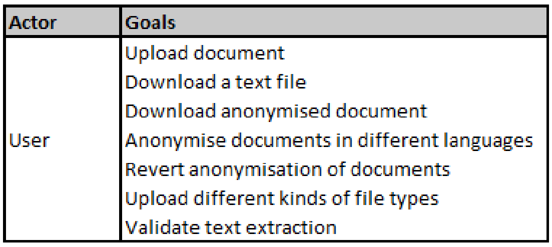
This chapter defines the use cases, the functional requirements as well as the non-functional requirements for the Anonymiser app.

5.1 Functional Requirements

The user goals lay the basis for the use cases. This section describes the user goals of the Anonymiser app, as well as the use cases.

5.1.1 List of User Goals

Below table provides an overview of the goals the actor has, while using the image to text Anonymiser app.



Actor	Goals
User	Upload document Download a text file Download anonymised document Anonymise documents in different languages Revert anonymisation of documents Upload different kinds of file types Validate text extraction

Table 5.1: Overview of actors and goals

5.1.2 Use Case Diagram

Based on on the user goals, following use cases have been derived. Some of the use cases have been flagged as optional and are not necessarily part of the POC.

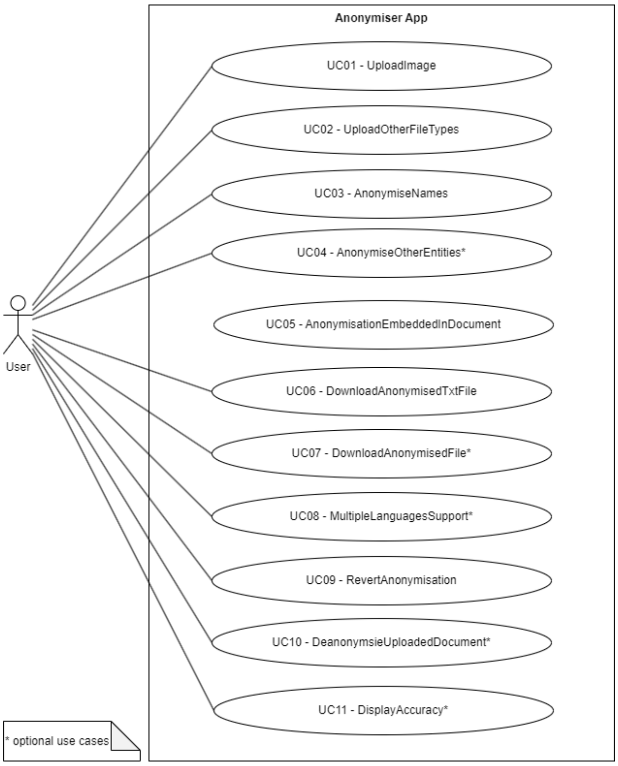
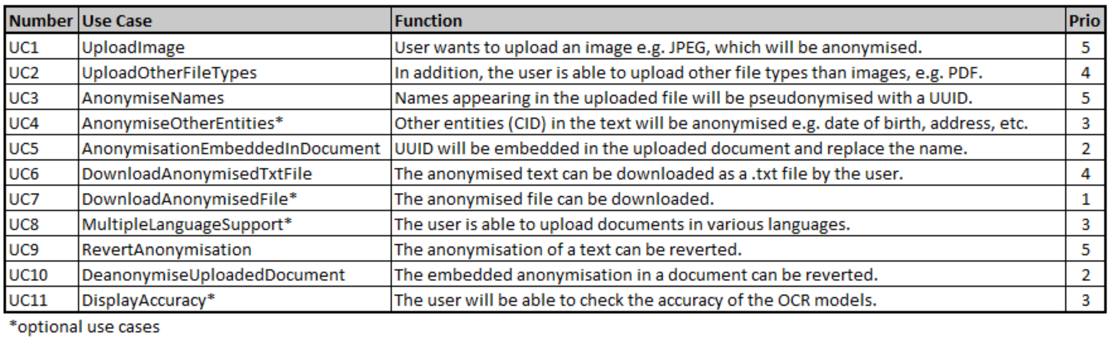

Anonymiser App
UC01 - UploadImage
UC02 - UploadOtherFileTypes
UC03 - AnonymiseNames
UC04 - AnonymiseOtherEntities*
UC05 - AnonymisationEmbeddedInDocument
UC06 - DownloadAnonymisedTxtFile
UC07 - DownloadAnonymisedFile*
UC08 - MultipleLanguagesSupport*
UC09 - RevertAnonymisation
UC10 - DeanonymsieUploadedDocument*
UC11 - DisplayAccuracy*
User
* optional use cases

Figure 5.1: Use case diagram

5.1.3 List of Use Cases

Below list provides an overview of the defined use cases with a short description.

Priorities: 5 Highest, 1 Lowest



Number	Use Case	Function	Prio
UC1	UploadImage	User wants to upload an image e.g. JPEG, which will be anonymised.	5
UC2	UploadOtherFileTypes	In addition, the user is able to upload other file types than images, e.g. PDF.	4
UC3	AnonymiseNames	Names appearing in the uploaded file will be pseudonymised with a UUID.	5
UC4	AnonymiseOtherEntities*	Other entities (CID) in the text will be anonymised e.g. date of birth, address, etc.	3
UC5	AnonymisationEmbeddedInDocument	UUID will be embedded in the uploaded document and replace the name.	2
UC6	DownloadAnonymisedTxtFile	The anonymised text can be downloaded as a .txt file by the user.	4
UC7	DownloadAnonymisedFile*	The anonymised file can be downloaded.	1
UC8	MultipleLanguageSupport*	The user is able to upload documents in various languages.	3
UC9	RevertAnonymisation	The anonymisation of a text can be reverted.	5
UC10	DeanonymiseUploadedDocument	The embedded anonymisation in a document can be reverted.	2
UC11	DisplayAccuracy*	The user will be able to check the accuracy of the OCR models.	3

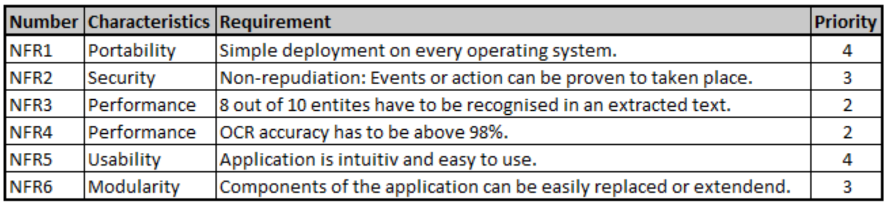
*optional use cases

Table 5.2: List containing the use cases

5.2 Non-Functional Requirements

Suitable non-functional requirements for the Anonymiser app have been defined according to the ISO/IEC 25010:2011 Standard.

Priorities: 5 Highest, 1 Lowest



Number	Characteristics	Requirement	Priority
NFR1	Portability	Simple deployment on every operating system.	4
NFR2	Security	Non-repudiation: Events or action can be proven to taken place.	3
NFR3	Performance	8 out of 10 entites have to be recognised in an extracted text.	2
NFR4	Performance	OCR accuracy has to be above 98%.	2
NFR5	Usability	Application is intuitiv and easy to use.	4
NFR6	Modularity	Components of the application can be easily replaced or extendend.	3

Table 5.3: List containing the non-functional requirements

Thresholds for NFR3 and NRR4

A OCR accuracy of 98% is considered an acceptable result [23, 52]. The recognition of the name entities in a text has to work in 8 out of 10 cases. This due to the fact that this project is a POC and the used NER model is not fine-tuned on the types of documents and the language.

CHAPTER 6

How to Extract Text from an Image?

This chapter addresses some theoretical aspects of techniques to extract text from an image. It provides an overview of how optical character recognition works. Afterwards, the chapter will discuss different providers of OCR tools.

6.1 Optical Character Recognition (OCR)

Optical character recognition exist almost for 50 years. It is a technology that changes text of a scanned document or an image into machine readable file [8]. Within the last years, the technology has become much more powerful and the accuracy has become almost perfect. Today, the technology is often used to automate document processing workflows [24].

6.1.1 How Does OCR work?

In order to extract text from a document, the OCR process is divided in pre-processing, segmentation, normalization, future extraction, classification and post processing [32].

Pre-processing

Aim of the pre-processing is to prepare the image for the subsequent phases by eliminating undesired characteristics or noise. Nevertheless, it has to be ensured that important information are not lost [32]. Following operations are part of the pre-processing:

- Binarization: Separates the text from the background. To do so, the image is converted to black and white [38].
- Noise Reduction: Due to technology advancements of the image acquisition devices, such as a camera [60], noise can be reduced [32].
- Skew Correction: Skewed documents have to be rotated in order to make the text lines horizontal or vertical [38].
- Morphological Operations: Comparing each pixel with its neighbors to add or remove pixels to characters [59].
- Thresholding: Separating information from its background [32].
- Thinning and Skeletonisation: Thinning a character until it reaches mostly one pixel width [32, 27].

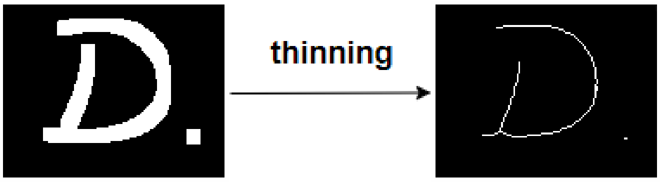


Figure 6.1: Illustration of thinning. Figure inspired by source [27]

Segmentation Phase

Segmentation is still considered to be a major part of the pre-processing phase [32]. During the segmentation phase, the image is split into different parts, which will then be further processed in order to isolate the text from the background [46]. There exist several methods to achieve a segmentation. Segmenting large regions into smaller sub regions is called **top-down approach**. The opposite is the **bottom-up approach**, which looks for interesting pixels and groups them together until the become a character, a word, lines and finally a text block. A combination of these two approaches is called **hybrid approach** [32].

Other approaches exist[32] and one of them is called **histogram projection method**[46]. The uploaded image is converted to a binary image and hence only consists of black and white pixels. Afterwards, the pixels reflecting the text are called foreground pixels, the other pixels are called background pixels.
The histogram projection counts all the foreground pixels in a row or column and plots them. By doing so, one can for example recognise the text lines, as shown in the below figure.

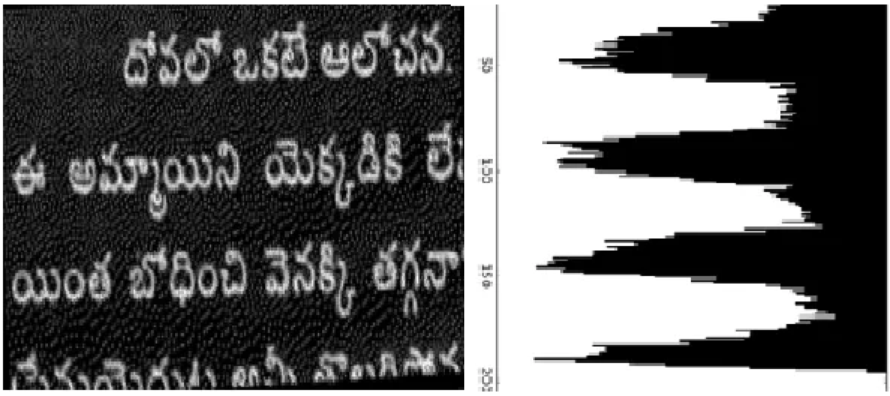


Figure 6.2: Example how segmentation works with histogram projection. Source figure: [46]

The histogram projection is also used to segment the document. First the lines will be segmented (Line Level Segmentation), then the line will be split into single words (Word Level Segmentation). Finally, the words are segmented into single characters (Character Level Segmentation)[46].

Normalization Phase

In the normalization phase, the isolated characters from the segmentation phase are minimized to a particular size. The selected size depends on the algorithm used for the feature extraction phase [32].

Feature Extraction Phase

Feature extraction is about the technique used to extract the most relevant information or properties of a character in order to classify the character during the classification phase. Secondly, the feature selection process looks for the most relevant features or properties of a character, which improve the classification accuracy. Generally, there are two kind of features called statistical feature and structural features.[42]. There are three major statistical features:

- Zoning: The character is split into different zones. These zones can be overlapping. Afterwards, the density of points in the zones will be used to find the correct representation [50].

- Crossings and Distances: Lines partition a character into different regions [42]. The feature value consists of the number of crossings from the contour and the lines [50].
- Projections: Characters, which are 2D elements, will be transformed to 1D elements. To do so, a projection histogram can be used. *See section Segmentation.* To represent a character, the number of foreground pixels in each column or row will be counted, which then results in a profile. These profiles describe the shape of a character [50].

Structural features describe the topological and geometrical properties of a character. Meaning, how long is a character, are there any loops, strokes or curves [50].

Classification Phase

In the classification phase the characters are recognised. The feature extraction provided the input for the classification phase which will then be compared with a database of many features [50]. The provided information are then used to identify the correct characters or words. There exist different methods to perform classification like template matching, statistical techniques and neural networks [42].

Post Processing

During the post processing, information of context and shape are incorporated into all OCR states to improve the recognition rate. [50].

6.2 OCR-Free Visual Document Understanding

This chapter will provide an overview of an alternative approach to OCR.

6.2.1 Transformers

In order to explain how OCR-free VDU works, a high-level overview about transformers is provided. A transformer model acts as the foundation of OCR-free VDU and also for the NER model implemented in the Anonymiser app.

Transformers a Brief Overview

In 2017, the paper *Attention Is All You Need* introduced the transformer model for language modeling and machine translation. It consists of a simpler architecture than the best performing models at this time and was also more performant [10].

Since the introduction in 2017, the transformer have become the basis for some well-known models like BERT and GPT-3. Furthermore, transformers are nowadays also used for computer vision and reinforcement learning tasks. The architecture of a transformer is deemed to be simpler, as it uses a encoder-decoder architecture, word embeddings, attention mechanisms, softmax and so on. However, it does not need recurrent neural networks or convolutional neural networks, which make things more complicated [53, 10].

Additionally, transformer models can also be used for transfer learning. Rather than training the own model from scratch, which requires a lot of compute power and causes a carbon footprint, a pretrained model can be used. A pretrained model is a model, which is already trained on a vast amount of data. In transfer learning, the pretrained model is used to initialize the weights of another model. The other model is then fine-tuned, which means that the model is trained with a task specific dataset. The advantages of transfer learning are, fine-tuning requires less data and hence less computation power and to get good results, much less resources are needed [33].

Transformer Architecture

The main components of a transformer model are the encoder and the decoder [33].

- Encoder: Takes an input and creates features. To do so, it uses a self-attention mechanism[54] to enrich the features with context from the whole sentence[53].

- Decoder: Optimized for generating output. The decoder takes the encoder's features along with other inputs to generate an output [33].

Depending on the task, a model can only use an encoder, or a decoder or the encoder and decoder together[25]. An encoder-decoder model for example is well suited for translation tasks. The encoder takes a sentence in English as input, extracts the features and the decoder afterwards uses the feature to output the sentence in French [53].

Another important aspect of a transformer is the attention layer. To illustrate this mechanism, an example from the area of NLP is used. In the natural language, every word has a meaning, which is deeply affected by the context. For a model to get a better understanding of the context, the attention layer highlights certain words, which have to be considered. A good example is again the translation of a English sentence to French. If the input to the model is "You like soccer!", the model has to pay attention to the word "you", because in French the word "like" needs to be conjugated to be correctly translated. [33]

Models using OCR for Visual Document Understanding vs. OCR-Free models

Current VDU models often use off-the-shelf OCR engines to extract text from documents and mainly focus on the understanding task [11]. However, recently more and more models like Donut or DUBLIN came up, which are end-to-end image to text models, which do not use OCR [12]. The OCR-free models use other methods, e.g. Swim Transformer to extract text from images [13]. Even though OCR has shown it usefulness, there are some limitation especially in non-Latin languages or for handwritten content. Another shortcoming of OCR is the lack of capturing the visual context in a document. The new OCR-free models try to address these challenges. Furthermore, they outperform OCR dependent VDU models in computational resources, performance and accuracy [11, 12].

6.3 Available Tools for Text Extraction in Images

There is a vast number of tools, which are able to extract text from an image or pdf file. The below list is far from comprehensive and only contains a selection of different tools and briefly summarises their key features:

6.3.1 Amazon Textract

Textraxt is a service provided by Amazon (Amazon Textraxt and only Textract are used interchangeable). Amazon describes Textract as follows: *"Amazon Textract is a machine learning (ML) service that automatically extracts text, handwriting, layout elements, and data from scanned documents. It goes beyond simple optical character recognition (OCR) to identify, understand, and extract specific data from documents[15]."* Various kinds of document can be uploaded. Some examples are patient registration form, an ID, or a simple letter [35]. To use Amazon Textract an Amazon Web Services account is required and furthermore, the service is also not for free as a certain amount will be charged per processed document [16].

Amazon says, Textract goes beyond simple OCR. What other techniques are used to get an optimal result are not known due to missing information.

6.3.2 OpenAI GPT-4V(ision)

GPT-4V(ision) (GPT-4V(ision) and GPT-4V are used interchangeable) is a service provided by Microsoft and belongs to the family of Large multimodal models (LMM). Compared to Large Language Model (LLM), a LMM is equipped with multi-sensory skills, such as visual understanding, and hence achieve a stronger generic intelligence. According to the latest paper, GPT-4V(ision) uses OCR to recognise characters in an image. Based on that, it provides several other services like image recognition, object localization, image captioning, visual question answering, visual dialogue, dense caption, and so on. GPT-4V(ision) is able to handle various input documents like receipts, research papers, handwritten notes and so on.[14].

A unique strenght of GPT-4V(ision) is the capability of understanding and following instructions. To get a desired output of an image text, the user simply has to input a textual instruction[14].

To be able to use GPT-4V(ision) a GPT-4 account is required, which charges a fee per processed token. As of December 2023 GPT-4V(ision) was only available as a preview version [36].

6.3.3 Tesseract

"Tesseract is an open source text recognition (OCR) Engine, available under the Apache 2.0 license [51]." There exist no graphical user interface, but the engine can be used via command line or by using an API[51]. According to the manual, Tesseract provides an overview of image processing operations, like binarisation, which are initially done to improve the output quality. Nevertheless, if the output does not meet the expectations, there are numerous recommendations, how the quality of the outcome could be improved [1].

In contrast to the services of Microsoft and Amazon, Tesseract is optimised to recognise solely word sentences. In case other documents like receipts or lists should be progressed, some adjustments have to be done. For example, an appropriate segmentation mode needs to be selected. Furthermore, it is a known issue that Tesseract has troubles to extract text from tables and therefore needs manual adjustments [1].

6.3.4 Donut Model

"Donut is an end-to-end (i.e., self-contained) VDU model for general understanding of document images." [11] The name Donut is an abbreviation and stands for Document Understanding Transformer. The architecture of the Donut model consists of a transformer-based visual encoder and textual decoder. The encoder takes an image as input and splits the images into patches and then creates the embeddings. This is done by using the a vision transformer called Swin Transformer [13]. The decoder takes the output from the encoder as input and provides a token sequence as result. Each token is represented as a one-hot vector and represents a word in the vocabulary. To perform this task, the pre-trained BART model is used. Finally, the output tokens are converted into a readable format like JSON [11].

The Donut model needs to be fine-tuned in order to perform a certain task accurately or to process certain types of images like letters or receipts. Nevertheless, the authors of the paper claim that the Donut needs less computational resources, is more flexible and is less prone to error propagation than models which use OCR [11].

CHAPTER 7

Indentifying Personal Information Using Huggingface

The Anonymiser app uses the bert-base-NER model available on Huggingface for the identification of the person names. Therefore, this chapter will summarize and highlight some of the theoretical aspects from the Huggingface Natural Language Processing course [4]. It starts with a brief introduction to NLP and then provides more insights about Named Entity Recognition and the tokenization method used for Bidirectional Encoder Representations from Transformers (BERT).

7.1 Natural Language Processing

With the help of machine learning, NLP focuses on the human language and tries to understand the context of words, rather than just understanding a single words individually. The below list shows some examples of NLP tasks:

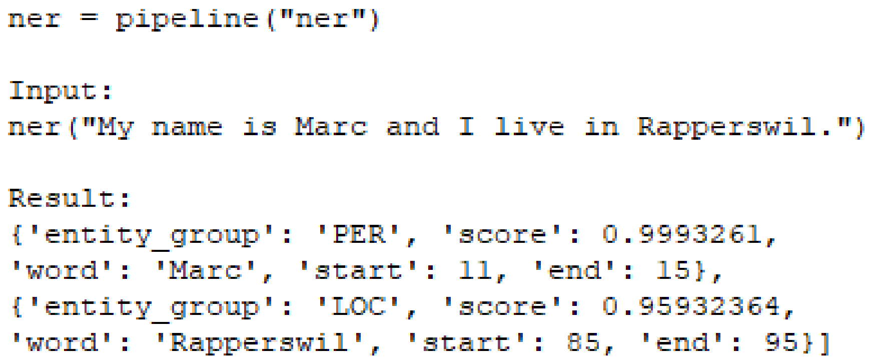
- Classification of whole sentences: Used to detect spam or to check if the grammar of a sentence is correct.
- Classification of each word in a sentences: Identify words in a sentence like noun, verb, adjective or NER (Person, Location, Organisation).
- Generating text content: Auto-generates text based on a prompt.
- Extracting an answer from a text: Answering a question from the user.
- Generation of a new sentence based on an input text: Summarizing a text or translation tasks.

To perform such NLP task a pipeline is used. The pipeline consists of two main components, the model and the tokenizer. A pipeline executes three main steps:

- The raw text has to be converted into numbers. This task is done by the tokenizer. The tokenizer splits the input into words, sub words, or symbols, which are called tokens. Each token is mapped to a number and additional input is added.
- The tokens are then passed to the model. The different layers in the model will manipulate the vectors and create an output.
- The results from the model cannot necessarily be interpreted, hence a post processing will make the result human readable.

7.2 Named Entity Recognition

Token classification is a task, where a label is assigned to each word. Hence, this can be used for Named Entity Recognition, where labels like persons, location or organisation is assigned to an entity. Below code fragment shows the example sentence, which is analysed by a Named Entity Recognition model:



```
ner = pipeline("ner")

Input:
ner("My name is Marc and I live in Rapperswil.")

Result:
{'entity_group': 'PER', 'score': 0.9993261,
'word': 'Marc', 'start': 11, 'end': 15},
{'entity_group': 'LOC', 'score': 0.95932364,
'word': 'Rapperswil', 'start': 85, 'end': 95}]
```

Figure 7.1: Example input and output of a NER pipeline

7.3 WordPiece Tokenization Used in BERT

There exist several kind of tokenizers (e.g. Word-based, Character based, Subword tokenization, and so on). As the Anonymiser app the bert-base-NER model uses, this section explains how the WordPiece tokenization works. As the WordPiece algorithm is not open source, not all information are correct.
The WordPiece tokenization algorithm works as follows:

- Starting point is the pretrained vocabulary. In the case of the ber-base-NER model it consists of almost 30'000 entries.
- To create the tokens, each word of the input is compared with the vocabulary
- Starting at the beginning of each word, the longest subword is looked up in the vocabulary. The longest subword becomes a token.
- This will be repeated until the whole word or text is tokenized.

Below examples illustrates how the WordPiece tokenization algorithm works. The example is from the Huggingface WordPiece Tokenization Youtube video. The tokenization is done for the word "huggingface".

The algorithm starts and looks for the longest subword found in the provided vocabulary. In this case, it is "huggi". Next, the algorithm starts to lookup the longest subword of "ngface", which is "##n". The added prefix "##" flaggs all the characters inside a word. Afterwards, the longest subword for "gface" is looked up. In the vocabulary "##gfac" is found. Now, only "e" is remaining, which will become the token "##e".

Vocabulary	**Tokens**
##a, ##c, ##e, ##f, ##g, ##i, ##n, ##r, ##s, ##u, f, h, l, hu, ##fa, ##fac, fa, fac, hug, ##gfac, hugg, huggi	huggi ##n ##gfac ##e

Table 7.1: Tokenization example for the word "Huggingface" and a given vocabulary

CHAPTER 8

Anonymisation

This chapter provides an overview of some theoretical and regulatory aspects in regard to anonymisation and pseudonymisation. It will also give an overview of some tools to perform anonymisation or pseudonymisation.

8.1 Difference Between Anonymisation and Pseudonymisation

With the introduction of General Data Protection Regulation it has become especially important to distinguish the two methods anonymisation and pseudonymisation. Anonymisation means that sensitive data is replaced by unrelated characters and hence a person cannot be re-identified. Or in other words, it is impossible to bring the data in relation with a person[34, 49, 9]. With pseudonymisation, the sensitive data is replaced as well and it is not possible to link the data to a person anymore. However, with the help of additional information the pesudonymisation can be reverted. This means, pseudonymisation substitutes sensitive values with reversible data[34, 9].

8.2 Data Protection in Switzerland

In 1992, the first Federal Act on Data Protection (FADP) has become effective. It aimed to protect persons and regulates the processing of personal data by companies or the government [40]. As of 1. September 2023 the new FADP has become effective in Switzerland. The revised regulation strengthens the rights of persons and their data further. In addition, the regulation enhances the transparency. Persons should know what happens with their personal data [26].

The revision of the FADP in 2023 was required, as since 1992 the digitalisation progressed and the usage of personal data has increased significantly. Therefore, it was necessary to have a regulation, which correspond to current standards, social media platforms and consumer behavior [26].

Furthermore, especially in the European Union (EU) the regulations regarding data protection have become much more restrictive with the introduction of the General Data Protection Regulation (GDPR) as of 25th May 2018. This standard is not only applicable in the EU, but also in other territories as soon as personal data of European citizens is processed. The revised FADP therefore also aims to align with the regulations of the EU [26]. Consequently, besides the FADP also GDPR is applicable in Switzerland.
GDPR is applicable for organizations, which process personal data of EU citizens, sell products or services within the EU, or the behavior of persons within the EU are tracked [39].
The new FADP is for every person applicable, who processes data. It affects companies and their employees. The regulation is world wide applicable as soon as Swiss citizens are impacted [47].

8.3 Implications of Data Protection Regulations for Businesses

Even though it is not aim of this work to perform an analysis from a regulatory perspective, the below examples should highlight the additional effort, which is required from organization, when they use personal data.
With the introduction of the new FADP several new duties have been introduced [26]. Following a few examples of duties:

- In case of a data breach, the regulator and the affected persons have to be informed. This is the case for GDPR as well as FADP [26].
- Removal or anonymisation of personal data, if they are not required anymore and no retention period is applicable [26].
- Ensuring data protection with technical means. Furthermore, the processing of personal data should be restricted to a minimum [26]

8.4 Advantages of Using Anonymised Data

Is the anonymisation properly done, the data will not contain sensitive information anymore and hence GDPR is not applicable. This will enable a company to use the data less restricted and does not need to comply with GDPR rules [9, 56]. However, GDPR treats pseudonymised data differently. The European Union Agency for Cybersecurity (ENISA) mentioned in their guideline for pseudonymisation techniques and best practices that pseudonymisation can lead to relaxations of legal obligations, provided it is properly applied [41].

8.5 Techniques to Mask Data

In order to perform anonymisation or pseudonymisation, there are several techniques available. Below list provides an overview.

8.5.1 Pseudonymisation Techniques

The below list is based on the ENISA report for techniques and best practices.

- **Counter:** Replaces the sensitive number with a number. The number starts at 0 and will be incremented each time data is pseudonymised. Advantage of this technique is its simplicity, however a mapping table needs to be stored, which can lead to scalability issues.
- **Random number generator:** Takes a number from a set of numbers with equal probability to be selected. Otherwise, the approach is almost as the counter. It replaces the sensitive data with a random number. As random numbers are used, it provides a strong data protections. On the other hand, there is the risk of collision as the same random number could have been drawn multiple times. Furthermore, a mapping table needs to be stored, which makes it hard to scale for large datasets.
- **Cryptographic hash function:** The cryptographic hash function takes a string as input and maps it to a fixed length output. The hash function fulfill the properties of collision free and one-way. One-way means that is hard to find the input string which will map to a pre-defined output. The hash function is prone to brute force and dictionary attacks and hence it is deemed as a weak pseudonymisation technique.
- **Message authentication code (MAC):** Message authentication code is almost similar as the cryptographic hash function except that a secret key is used to generate a pseudonym. The secret key makes it impossible to reverse the pseudonymisation, provided the secret key is not compromised. The message authentication code is considered to be a robust pseudonymisation technique.
- **Encryption:** Encryption has several properties similar to MAC and hence also considered to be a robust pseudonymisation technique. A block cipher is used to encrypt sensitive data with a secret key. The key is used for the pseudonymisation as well as for the recovery.

The above list is not exhaustive. Besides the above mentioned pseudonymisation techniques it also important to define how the pseudonymisation is implemented, the so called policy (or mode). It is distinguished between deterministic pseudonymisation, document-randomized pseudonymisation and fully-randomized pseudonymisation. Using determinstic pseudonymisation, the same values will be mapped to the same pseudonym. The opposite is document-randomized pseudonymisation, whereby the same value is mapped to different pseudonyms. Fully-randomize pseudonymisation is an extension to document-randomized pseudonymisation.

8.5.2 Anonymisation Techniques

There are two main approaches to irreversibly anonymise data, which are called randomization and generalization [56].

Below lists are a summary of techniques provided in the opinion 05/2014 on Anonymisation Techniques[55]:

Randomisation

- **Noise addition:** Adding noise causes the data to be less accurate, but the overall distribution is still retained. For example will a certain amount be added or subtracted from the correct salary. If a third-party looks at the data, it will consider it as correct but it will not be able to refer the data to an individual. However, noise addition by itself is often not sufficient to anonymise the data and other techniques in combination are required.

- **Permutation:** Attributes within the dataset will be shuffled and hence linked to other individuals or subjects. The advantage of this technique is that the values are still correct and hence the distribution of the values remains the same. However, there are some pitfalls which have to be considered. The relation of the attributes has to be considered. If there is a strong relationship between attributes it still might possible to identify an individual. Assuming the attributes job and income, there is a strong link between these attributes and there is a high risk that an attacker is able to revert the permutation. Furthermore, permutation by itself might not be sufficient to completely anonymsie the data. A combination of other techniques might be required.

- **Differential privacy:** Differential privacy is also part of randomization techniques, but it follows a different approach. Rather than anonymise the data directly, it will only be anonymised if a third-party queries certain data. Differential privacy will tell how the data should be anonymised and how much of the data has to be anonymised so that an individual is not identifiable. One advantage of this approach is that a dataset is not just released, as it is only provided to authorised third parties upon a specific request. Nevertheless, it is important to monitor the issued queries in order to ensure that information gained do not lead to the identification of a subject.

Generalisation

Is the second approach, which belongs to the family of anonymisation techniques. Generalization aims to generalise or diluting sensitive data in order to make it hard to singling out an individual. For example, rather than providing the exact weight of 78kg, one provides a range 75kg - 85kg [55].

- **Aggregation and K-anonymity:** By grouping sensitive data from one individual with K other individuals, the goal is to avoid the singling out one single individual. All K individuals share the same values belong to the same equivalent class. To achieve this, the granularity of location is changed. Rather than providing a city name, the name of the region is captured. This would grouping several individuals together and make it harder to singling out one specific person. Furthermore, ranges could be provided for numerical values e.g. range for the salary, weight or height. Is aggregation and K-anonymity used, the K has to be carefully selected. If an attacker has some background information about a subject in the dataset, it is still possible to retrieve further information about that person e.g., the attacker knows that the person is born in 1950 and in the dataset all persons with a heart attack were born in 1950. The attacker can be sure that the victim had an heart attack as well.

- **L-diversity/T-closeness:** L-diversity is an extension of the K-anonymity. It aims to have at least L different values for each attribute in one equivalent class with K individuals. To achieve this, the number of groups is limited, if the variability of the attributes within this group is poor. In addition, it needs to be considered that the distribution of each attribute is of importance. Otherwise, an attacker could still derive with a high probability information about a subject. T-closeness is an extension of L-diversity. With T-closeness it is not only required to have L different attributes for an equivalent class, it is also required that the number of values reflect the original distribution of the data set. Consequently, T-closeness aims to be as close to the initial distribution of the attributes in the data set as possible.

8.6 Tools to Anonymise Client Data

There exist numerous tools to anonymise data. Some of them are open source, others are proprietary services like Amazon Macie. In this section a selection of the available tools have been analysed. Focus is on open source tools and it has been distinguished between tools to anonymise the data in the code or afterwards, when the sensitive data is already stored in the data base.

8.6.1 ARX

ARX is an open source tool, which is used to anonymise personal data. The tool is able to process structured data, which means the data has to be in tabular form [21]. ARX supports various techniques to anonymise data. For example K-anonymity, L-diversity and differential privacy and much more [22]. The tool offers a graphical user interface as well as a Java software library [21]. The anonymisation process is divided in four steps called **configuration perspective**, **exploration perspective**, **utility analysis perspective** and **risk analysis perspective**. In the configuration perspective, the data is imported. Afterwards, potential transformation are recommended, which is called exploration perspective. The utility analysis is used to assess the a specific transformation. Finally the risk analysis perspective, provides various metrics, which show the associated risks, like re-identification [20]

8.6.2 Amnesia

Like ARX, Amnesia is as well a popular open source tool to simplify the anonymisation process[37]. Amnesia guarantees GDPR compliance for provided datasets [17]. The anonymisation process consists of five steps. First, the data has to be uploaded as a delimited text file. Secondly, the user has to decide, which attributes should be anonymised. In a next step, the user has to select an available technique like K-anonymity, masking or pseudonymisation to anonymise the data. Afterwards, the tools provides different statistics (e.g. distribution) of possible solution, which then the user can choose from. Finally, the data can be stored. Amnesia can be used with a library or graphical user interface [18].

8.6.3 PostgreSQL Anonymiser

PostgreSQL anonymiser is an extension for PostgreSQL database. The extension is used to anonymise Personally Identifying Information (PII) or sensitive data. There exist 8 techniques, which can be used with the PostgreSQL Data Definition Language (DDL). The anonymised data can then be exported in another SQL file, stored permanently in the dataset or made available by views or roles [19].

8.6.4 Universally Unique Identifier (UUID)

A Universally Unique Identifier (UUID) is a 128 bit long value, which is unique and does not need a central registration process. Therefore, the UUID can be easily created and used for various purposes.

There exist five different version of UUID. To identify which version is used, the four most significant bits of the time stamp have to be considered [44].

- **UUID Version 1 (Time- and node-based):** Created out of the system time, the local clock sequence and the MAC address. Drawback of the UUID Version 1 is the missing anonymity, as the MAC address is encoded[31, 30].

- **UUID Version 2 (Security Version):** Version 2 is not often used as there exist several problems with the format and collisions can occur [57, 31].
- **UUID Version 3 (name based, MD5):** Version 3 is almost similar to the version 5 except that a different Hashing algorithm is used. Version 3 uses the MD5 algorithm. A namespace and a name has to be provided as input. The UUID for the same name in the same namespace created at different times have to be the same. Using another namespace or a different name has to result in different values [44].
- **UUID Version 4 (random):** Based on random numbers to create unique identifiers. The UUID is completely random and anonymous [31].
- **UUID Version 5 (name-based, SHA-1):** Version 5 uses the SHA-1 hashing algorithm. Due to the size of 160 bits of the SHA-1 value, it has to be truncated. The rest works analogous Version 3 [31, 57].

It exists a native library to use the UUID in Python [30].

8.6.5 Hashing

Another opportunity to randomize sensitive data are hashes. A hash function takes an input and produces an output called hash value. The hash value is an alphanumeric string and usually unique. Furthermore, a hash function is deterministic, which means that the same input will always produce the same output [48]. There exist a Python library called hashlib, which provides the opportunity to use various hashing algorithms [29]. Drawback of using hash functions to anonymise sensitive data is the fact that the hash function is deterministic. With dictionary or brute force attacks the hash values could be guessed [28, 45, 43].

CHAPTER 9

Architecture and Implementation

This chapter provides an overview how the Anonymiser app was built, what technologies are used and what design decisions were made.

9.1 Visualisation of the Software Architecture

To describe the architecture of the Anonymiser app, the C4 model has been selected [6].

9.1.1 Level 1 - Context

The context diagram below provides an overview of the landscape, in which the Anonymiser app is embedded. A user provides input to the app and the app either performs the anonymisation process on the server side, or some external services, like Amazon Textract or OpenAI GPT-4V(ision) are called. These services are tools to extract text from an image or PDF documents. The NER and the pseudonymisation are performed on the server.

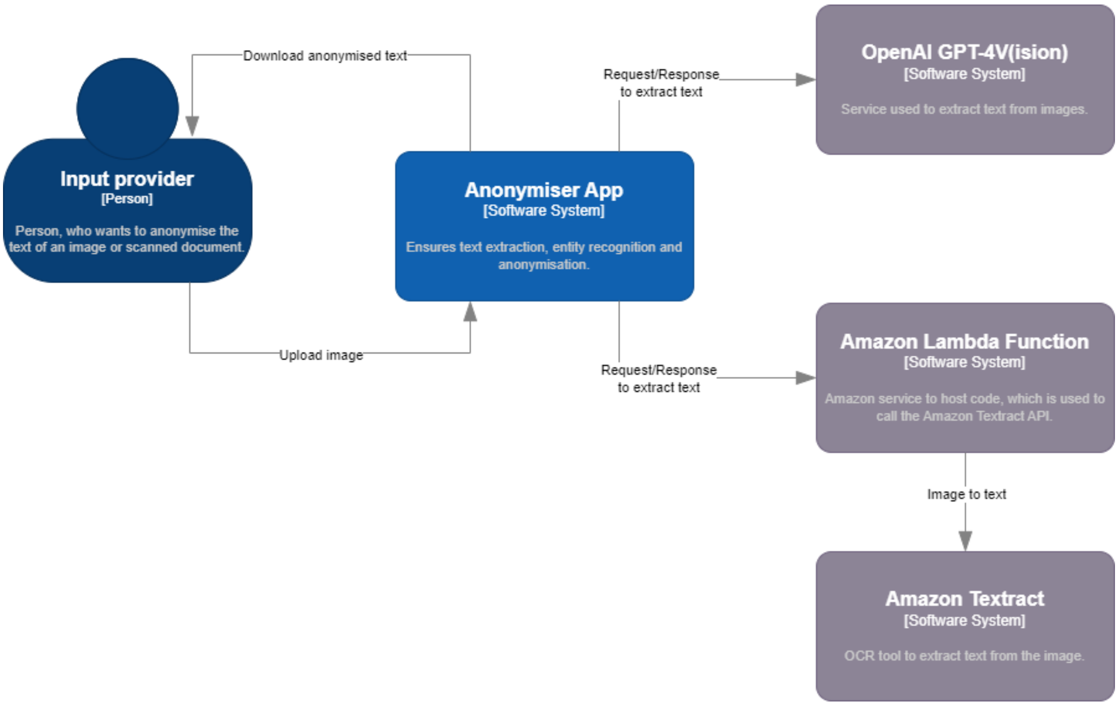

Download anonymised text
Input provider
[Person]
Person, who wants to anonymise the text of an image or scanned document.
Upload image
Anonymiser App
[Software System]
Ensures text extraction, entity recognition and anonymisation.
Request/Response to extract text
OpenAI GPT-4V(ision)
[Software System]
Service used to extract text from images.
Request/Response to extract text
Amazon Lambda Function
[Software System]
Amazon service to host code, which is used to call the Amazon Textract API.
Image to text
Amazon Textract
[Software System]
OCR tool to extract text from the image.

Figure 9.1: C4 context diagram

9.1.2 Level 2 - Containers

Zooming in one level results in the container diagram. The Anonymiser app consists of three containers called web application, API Service and database. The user accesses the web application and provides some input and selects the desired feature. The request will then be sent to the API Service, which handles the request. The API Service ensures the extraction of the text, triggers the NER model and performs the anonymisation. In addition, the API Service executes the accuracy calculations and the deanonymisation.

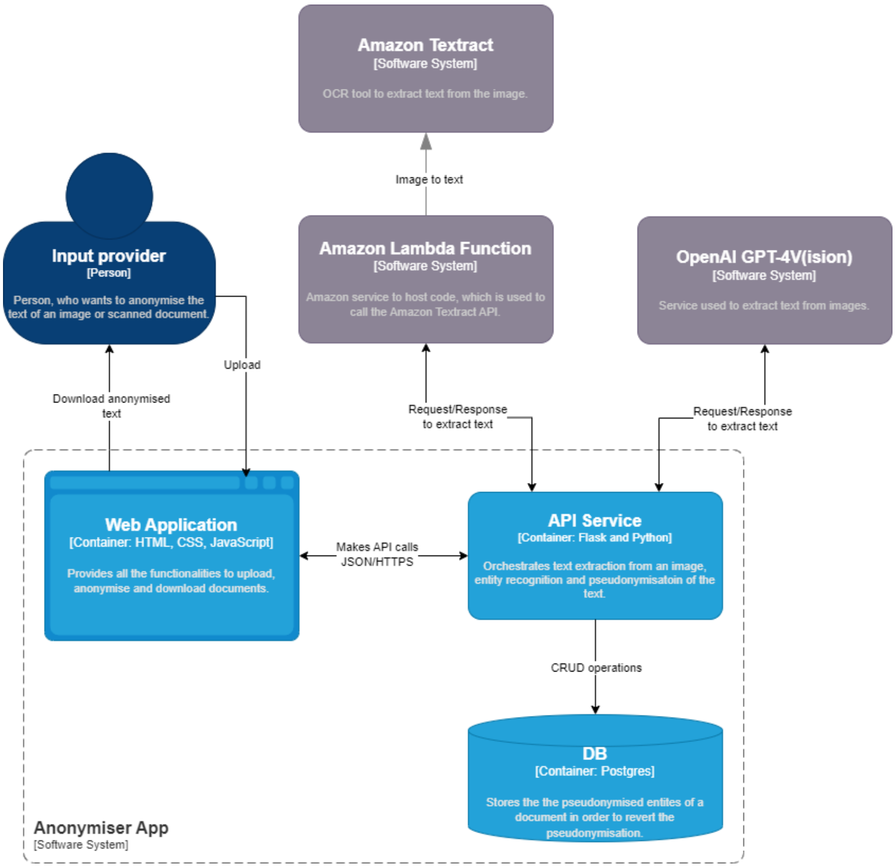

Amazon Textract
[Software System]
OCR tool to extract text from the image.
Image to text
Input provider
[Person]
Person, who wants to anonymise the text of an image or scanned document.
Amazon Lambda Function
[Software System]
Amazon service to host code, which is used to call the Amazon Textract API.
OpenAI GPT-4V(ision)
[Software System]
Service used to extract text from images.
Upload
Download anonymised text
Request/Response to extract text
Request/Response to extract text
Web Application
[Container: HTML, CSS, JavaScript]
Provides all the functionalities to upload, anonymise and download documents.
Makes API calls JSON/HTTPS
API Service
[Container: Flask and Python]
Orchestrates text extraction from an image, entity recognition and pseudonymisatoin of the text.
CRUD operations
DB
[Container: Postgres]
Stores the the pseudonymised entites of a document in order to revert the pseudonymisation.
Anonymiser App
[Software System]

Figure 9.2: C4 container diagram

9.1.3 Level 3 - Components

In the third level the containers are further expanded into components.

Web Application Components

The input provider accesses the web client. Some parts of the website are static components and the other parts are dynamic. When the user uploads PDF documents or images, the request will be passed to the API Service via the dynamic component.

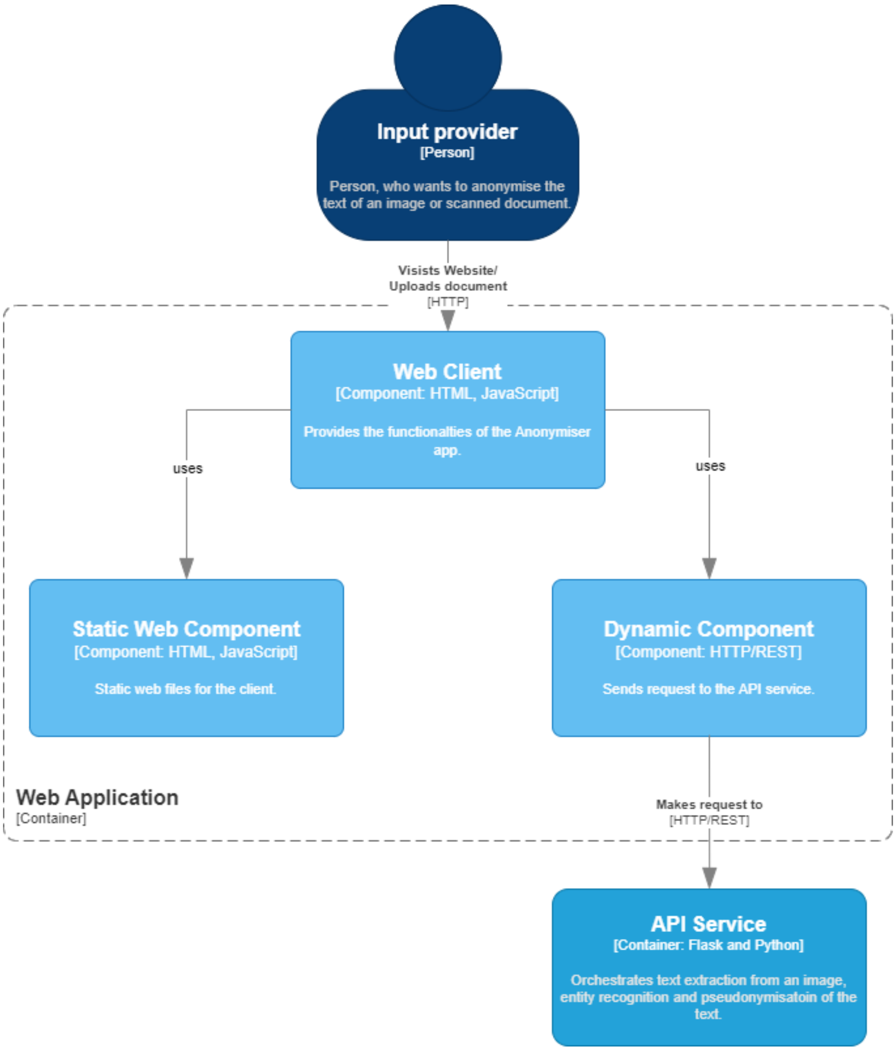
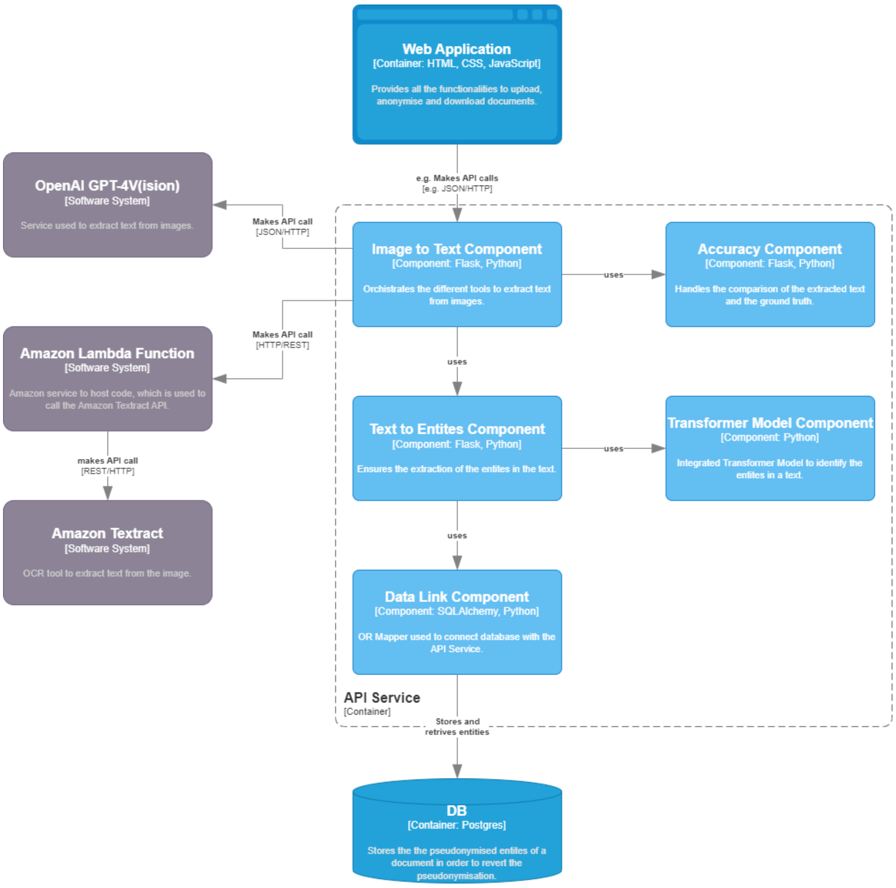

Input provider
[Person]
Person, who wants to anonymise the text of an image or scanned document.
Visists Website/
Uploads document
[HTTP]
Web Client
[Component: HTML, JavaScript]
Provides the functionalties of the Anonymiser app.
uses
uses
Static Web Component
[Component: HTML, JavaScript]
Static web files for the client.
Dynamic Component
[Component: HTTP/REST]
Sends request to the API service.
Web Application
[Container]
Makes request to
[HTTP/REST]
API Service
[Container: Flask and Python]
Orchestrates text extraction from an image, entity recognition and pseudonymisatoin of the text.

Figure 9.3: C4 component diagram for the web application

API Service Components

The uploaded image or document from the input provider is passed via the web application to the image to text component. To anonymise the document, it is passed to the Amazon Lambda Function and to the OpenAPI interface. Tesseract is directly executed in the image to text component. The results of the text extraction are either passed to the accuracy component for the string comparison, or forwarded to the text to entities component, where the entity extraction is managed. Once the transformer model component extracted the name entities, they are passed via the text to entities component and data link component to the database. The pseudonymisation of the entities is done in the text to entities component.



Web Application
[Container: HTML, CSS, JavaScript]
Provides all the functionalities to upload, anonymise and download documents.
e.g. Makes API calls
[e.g. JSON/HTTP]
OpenAI GPT-4V(ision)
[Software System]
Service used to extract text from images.
Makes API call
[JSON/HTTP]
Image to Text Component
[Component: Flask, Python]
Orchistrates the different tools to extract text from images.
uses
Accuracy Component
[Component: Flask, Python]
Handles the comparison of the extracted text and the ground truth.
Makes API call
[HTTP/REST]
Amazon Lambda Function
[Software System]
Amazon service to host code, which is used to call the Amazon Textract API.
uses
Text to Entites Component
[Component: Flask, Python]
Ensures the extraction of the entites in the text.
uses
Transformer Model Component
[Component: Python]
Integrated Transformer Model to identify the entites in a text.
makes API call
[REST/HTTP]
Amazon Textract
[Software System]
OCR tool to extract text from the image.
uses
Data Link Component
[Component: SQLAlchemy, Python]
OR Mapper used to connect database with the API Service.
API Service
[Container]
Stores and retrives entities
DB
[Container: Postgres]
Stores the the pseudonymised entites of a document in order to revert the pseudonymisation.

Figure 9.4: C4 component diagram for the API service

9.2 Technology Stack

The following table provides an overview of tools, which are used for the implementation of the Anonymiser app.

Domain	Technology
Environment for containers	Docker Containers
Frontend	HTML, CSS, JavaScript
Backend	Flask, Python
OR Mapper	SQLAlchemy
Database	PostgreSQL
OCR	Tesseract (local), Amazon Textract and OpenAI GPT-4V API calls
NER	bert-base-NER
OCR accuracy	Jellyfish library
PDF management	pdf2image library

Table 9.1: Technology stack of the Anonymiser app

9.3 Implementation of Main Features

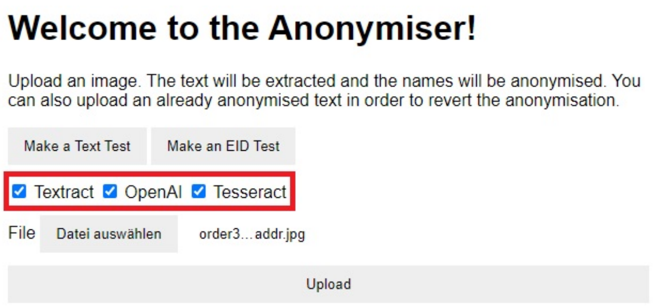
Following sections provide insights how some of the selected main features have been implemented.

9.3.1 Upload Images and PDF files

The Anonymiser app offers the opportunity to upload images (i.e. PNG and JPEG) as well as PDF files. When a PDF document is uploaded, it needs to be considered that a PDF file may consist of multiple pages. The pdf2image library is used to convert the PDF file in multiple PIL images, which will be stored in an list. The application will iterate over the list and extract the text of each image. The extracted text of each page will then be joined and displayed. Example is available in section 18.2.

9.3.2 Option Between Three Different OCR Engines

The user will get the opportunity to choose between different OCR engines or select all of them to provide an output. Once the output is available, the user can select the text with the best quality. So, it can be ensured, that the user is not dependent on only one tool, which might not always provide the perfect results for the users type of documents.



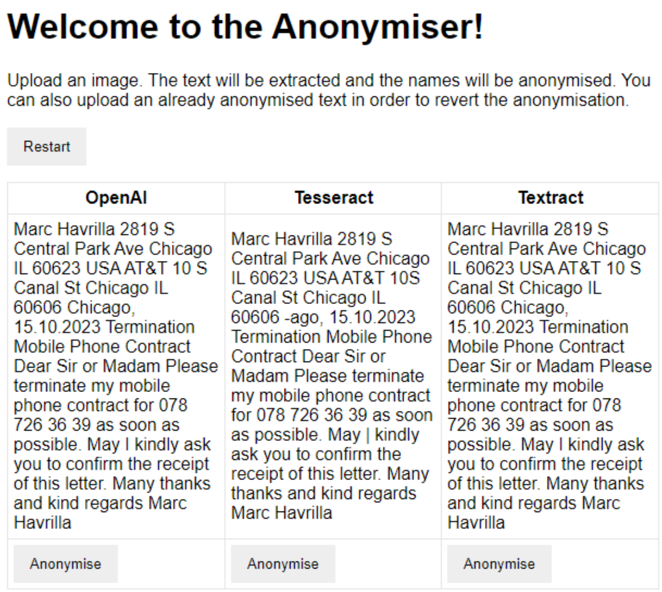
Welcome to the Anonymiser!

Upload an image. The text will be extracted and the names will be anonymised. You can also upload an already anonymised text in order to revert the anonymisation.

☑ Textract ☑ OpenAI ☑ Tesseract

File Datei auswählen order3...addr.jpg

Figure 9.5: Screenshot of the Anonymiser app, showing the opportunity to select the three OCR engines



Welcome to the Anonymiser!

Upload an image. The text will be extracted and the names will be anonymised. You can also upload an already anonymised text in order to revert the anonymisation.

OpenAI	Tesseract	Textract
Marc Havrilla 2819 S Central Park Ave Chicago IL 60623 USA AT&T 10 S Canal St Chicago IL 60606 Chicago, 15.10.2023 Termination Mobile Phone Contract Dear Sir or Madam Please terminate my mobile phone contract for 078 726 36 39 as soon as possible. May I kindly ask you to confirm the receipt of this letter. Many thanks and kind regards Marc Havrilla	Marc Havrilla 2819 S Central Park Ave Chicago IL 60623 USA AT&T 10S Canal St Chicago IL 60606 -ago, 15.10.2023 Termination Mobile Phone Contract Dear Sir or Madam Please terminate my mobile phone contract for 078 726 36 39 as soon as possible. May | kindly ask you to confirm the receipt of this letter. Many thanks and kind regards Marc Havrilla	Marc Havrilla 2819 S Central Park Ave Chicago IL 60623 USA AT&T 10 S Canal St Chicago IL 60606 Chicago, 15.10.2023 Termination Mobile Phone Contract Dear Sir or Madam Please terminate my mobile phone contract for 078 726 36 39 as soon as possible. May I kindly ask you to confirm the receipt of this letter. Many thanks and kind regards Marc Havrilla
Anonymise	Anonymise	Anonymise

Figure 9.6: Output of the three OCR engines

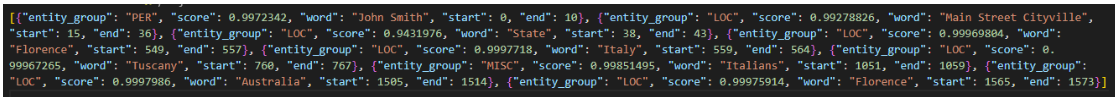
Implementation of the OCR Tools

- Tesseract: The implementation is done with the pytesseract package for Python, enabling simple method calls to extract the text from an image.

- Amazon Textract: The code to extract the text from an image is hosted on a Amazon Lambda Function. To get the image text extracted, the Anonymiser app will send a HTTP request to the Amazon Lambda Function. Then, the code hosted on Amazon Lambda Function will call the Textract API. The response of the this API call is the text as JSON, which will be sent back to the Anonymiser app as JSON response.
- OpenAI GPT-4V(ision): OpenAI offers a package for Python to make requests to the OpenAI API. The request to the API needs to contain some instructions in natural language and the image, so GPT-4V knows what it has to do.

9.3.3 NER Transformer Model

After the OCR tools did the text extraction, the NER model identifies the name entities in the text. In order to do that, the bert-base-NER model is used. For the implementation the transformer package is downloaded, which provides the tokenizer and the models. Afterwards, the application initialises the pipeline with the selected model, tokenizer, type of Natural Language Processing task and the grouped entities parameter. The next step is to pass the text to the pipeline, which will result in a JSON response containing all the entities.



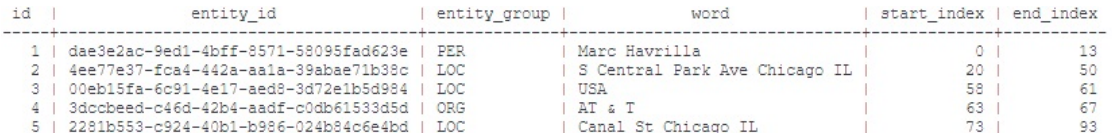
```
[{"entity_group": "PER", "score": 0.9972342, "word": "John Smith", "start": 0, "end": 10}, {"entity_group": "LOC", "score": 0.99278826, "word": "Main Street Cityville",
"start": 15, "end": 36}, {"entity_group": "LOC", "score": 0.9431976, "word": "State", "start": 38, "end": 43}, {"entity_group": "LOC", "score": 0.99969804, "word":
"Florence", "start": 549, "end": 557}, {"entity_group": "LOC", "score": 0.9997718, "word": "Italy", "start": 559, "end": 564}, {"entity_group": "LOC", "score": 0.
99967265, "word": "Tuscany", "start": 760, "end": 767}, {"entity_group": "MISC", "score": 0.99851495, "word": "Italians", "start": 1051, "end": 1059}, {"entity_group":
"LOC", "score": 0.9997986, "word": "Australia", "start": 1505, "end": 1514}, {"entity_group": "LOC", "score": 0.99975914, "word": "Florence", "start": 1565, "end": 1573}]
```

Figure 9.7: Example output of the bert-base-NER model

Due to the scope of this project, only the name entities, called "PER" are selected and further processed. Lastly, the found entities are stored in the PostgreSQL database.

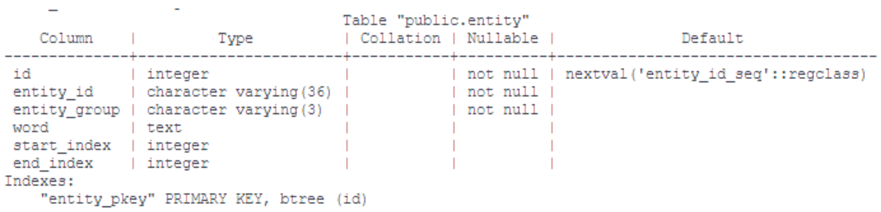
9.3.4 Pseudonymisation

The pseudonymisation of the found entities is done with a Universally Unique Identifier. The Python module UUID offers the usage of UUID objects. To pseudonymise the entities the UUID version 4 is used. Generally, the replacement of the entity with a UUID would constitute an anonymisation. However, as the entities and the UUID are stored in a PostgreSQL database, it can be reverted and hence is considered as pseudonymisation. To facilitate the communication between the API Service and the PostgreSQL database, SQLAlchemy is used as OR mapper. The database stores the attributes as follows:



```
 id  |              entity_id               | entity_group |             word              | start_index | end_index
-----+--------------------------------------+--------------+-------------------------------+-------------+-----------
   1 | dae3e2ac-9ed1-4bff-8571-58095fad623e | PER          | Marc Havrilla                 |           0 |        13
   2 | 4ee77e37-fca4-442a-aa1a-39abae71b38c | LOC          | S Central Park Ave Chicago IL |          20 |        50
   3 | 00eb15fa-6c91-4e17-aed8-3d72e1b5d984 | LOC          | USA                           |          58 |        61
   4 | 3dccbeed-c46d-42b4-aadf-c0db61533d5d | ORG          | AT & T                        |          63 |        67
   5 | 2281b553-c924-40b1-b986-024b84c6e4bd | LOC          | Canal St Chicago IL           |          73 |        93
```

Figure 9.8: Example of database entries



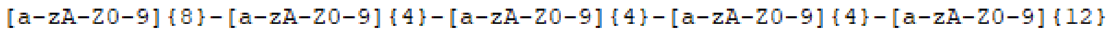
```
                                    Table "public.entity"
    Column    |         Type          | Collation | Nullable |                Default
--------------+-----------------------+-----------+----------+----------------------------------------
 id           | integer               |           | not null | nextval('entity_id_seq'::regclass)
 entity_id    | character varying(36) |           | not null |
 entity_group | character varying(3)  |           | not null |
 word         | text                  |           |          |
 start_index  | integer               |           |          |
 end_index    | integer               |           |          |
Indexes:
    "entity_pkey" PRIMARY KEY, btree (id)
```

Figure 9.9: Database schema used for the Anonymiser app extracted from the database container

After the creation of the UUID and the storage of the entity in the database, the name entity will be replaced by the UUID, so that the document is now pseudonymised.

Revert Psuedonymisation

The Anonymiser app offers the opportunity to revert the pseudonymisation. To achieve this feature, a Regex expression is used to find UUIDs in a text. This check is done after the text is extracted and before the text is passed to the NER model. If at least one UUID is found, the data link component queries the database, in order to get the entities belonging to the UUID. Then the UUID will be replaced with the names of the entities and the pseudonymisation is reverted.



```
[a-zA-Z0-9]{8}-[a-zA-Z0-9]{4}-[a-zA-Z0-9]{4}-[a-zA-Z0-9]{4}-[a-zA-Z0-9]{12}
```

Figure 9.10: Regex expression to find UUIDs in a text

In case there exists no UUID in the text, the text will automatically be passed to the pipeline for the entity recognition.

9.4 Design Decisions

To document some of the architectural decisions, the (WH)Y approach is used [58].

9.4.1 Amazon Lambda Function

To interact with the Amazon Textract service, an AWS Lambda Function has been created. Within the Lambda Function Python code is deployed, which calls the Amazon Textract API. However, this could have also been implemented with the Boto3 library, which could have been simply imported and installed. The Lambda Function would not have been required. Below (WH)Y statement provides the reasoning for this decision.

In the context of text extraction from an image, facing the need to reach 98% of OCR accuracy, I decided to set up the Amazon Textract service with an Amazon Lambda Function (and against the usage of the Boto3 library) to enhance my knowhow about AWS Lambda Function, accepting that the implementation could have been done much simpler and avoiding the effort needed to setup the AWS Lambda Functions.

9.4.2 Selecting the Tools For Text Extraction

There exist various tools, which extract text from images or documents. The Anonymiser app provides the user the opportunity to use three different tools. Following tools have been implemented:

- Tesseract
- Amazon Textract
- OpenAI GPT-4V(ision)

The three tools have been selected for the following reason:

In the context of text extraction from an image, facing the need to reach 98% of OCR accuracy, I selected Tesseract, Amazon Textract and OpenAI GPT-4V(ision) as text extraction tools (and against all the others like Donut, Google Cloud Vision API, etc.) to achieve a diverse range of tools (cloud service vs. local machine, stand alone OCR vs. end-to-end VDU model, paid service vs. open source, established vs. preview function), to extract the complete text of the image or document (not only NLP responses) and ensuring the fulfillment of the POC within the time frame, accepting that an OCR-free model is missing.

9.4.3 bert-base-NER Model

To recognise the name entities from the extracted text, the bert-base-NER models was selected based on the below decision:

In the context of name entity recognition of an extracted text, facing the need to recognise 8 out of 10 entities, I selected the bert-base-NER model as the NER model (and against all other NER models available on Huggingface), to achieve the identification of personal names without fine-tuning the model, accepting that the model is not perfectly suited for the kind of documents.

9.4.4 Using UUID For Anonymisation and Pseudonymisation

Anonymisation and the pseudonymisation are done by replacing the name entities with a UUID. This approach has been chosen based on the following decision:

In the context of anonymisation and pseudonymisation, facing the need to properly mask a personal name, I selected the UUID as anonymisation and pseudonymisation technique (and against hashing), to achieve a secure anonymisation and pseudonymisation, which is simple to implement, accepting that also a database is required to revert the pseudonymisation.

CHAPTER 10

Ensuring Quality

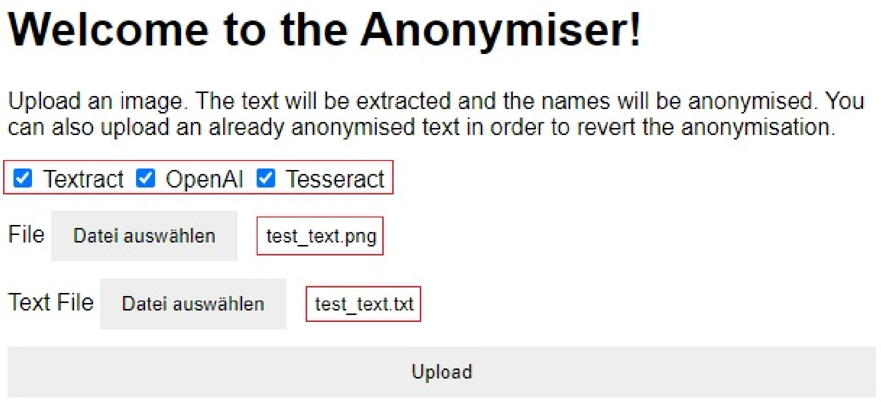
To ensure the quality of the Anonymiser app, two features have been implemented. These features enable the users to validate the quality of the extracted text for their specific documents.

10.1 Validate OCR Accuracy

Per default, the user has directly from the beginning the opportunity to select between three OCR engines. It is also possible to select all three of them, to get a first glance, which tool performs best for a certain type of document.

However, the output of the three OCR engines only provides a first visual impression of the quality. To get a more comprehensive overview, a dedicated test can be executed. Rather than just uploading the image containing the text, the user will also be able to upload a text file containing the ground truth of the text. The Anonymsier app will compare the text string extracted from the image with the ground truth. This provides the OCR accuracy, a percentage value that shows how well each tool performs on certain document types.

The screenshot below shows the upload of the different files the opportunity to select the OCR engines.



Welcome to the Anonymiser!

Upload an image. The text will be extracted and the names will be anonymised. You can also upload an already anonymised text in order to revert the anonymisation.

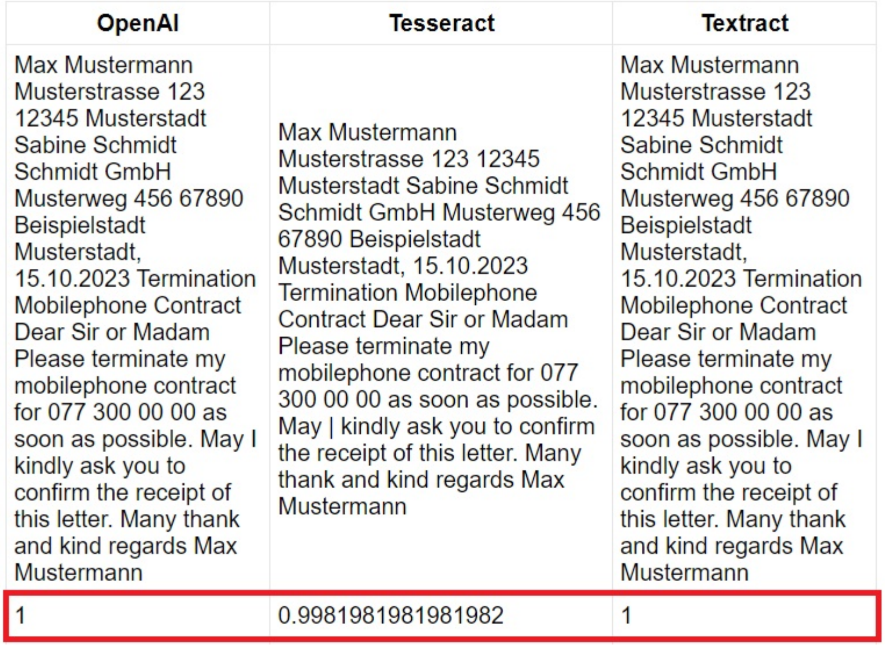
☑ Textract ☑ OpenAI ☑ Tesseract

File: Datei auswählen test_text.png

Text File: Datei auswählen test_text.txt

Figure 10.1: Upload section for ground truth file and image to test OCR accuracy

The string comparison is done with the Jellyfish library. The Jellyfish library offers several string comparison algorithms [2]. The Anonymsier app uses Jellyfish with the Jaro Simulartiy [3] to calculate the difference between the ground truth and the extracted image text.

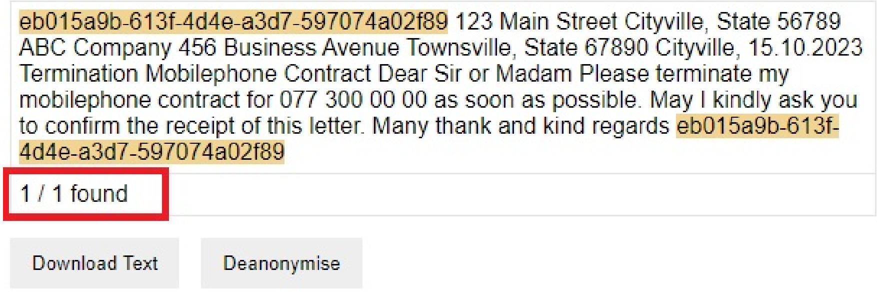


OpenAI	Tesseract	Textract
Max Mustermann Musterstrasse 123 12345 Musterstadt Sabine Schmidt Schmidt GmbH Musterweg 456 67890 Beispielstadt Musterstadt, 15.10.2023 Termination Mobilephone Contract Dear Sir or Madam Please terminate my mobilephone contract for 077 300 00 00 as soon as possible. May I kindly ask you to confirm the receipt of this letter. Many thank and kind regards Max Mustermann	Max Mustermann Musterstrasse 123 12345 Musterstadt Sabine Schmidt Schmidt GmbH Musterweg 456 67890 Beispielstadt Musterstadt, 15.10.2023 Termination Mobilephone Contract Dear Sir or Madam Please terminate my mobilephone contract for 077 300 00 00 as soon as possible. May | kindly ask you to confirm the receipt of this letter. Many thank and kind regards Max Mustermann	Max Mustermann Musterstrasse 123 12345 Musterstadt Sabine Schmidt Schmidt GmbH Musterweg 456 67890 Beispielstadt Musterstadt, 15.10.2023 Termination Mobilephone Contract Dear Sir or Madam Please terminate my mobilephone contract for 077 300 00 00 as soon as possible. May I kindly ask you to confirm the receipt of this letter. Many thank and kind regards Max Mustermann
1	0.9981981981981982	1

Figure 10.2: Result of the accuracy test

10.2 Identifying Named Entities Test

Analogous to the validation of the OCR accuracy, the user has also the opportunity to test the correctness of the NER output. The test should provide the user confidence that the NER model detects all relevant personal names in a text. For this purpose, the user can upload a text document, containing a list with name entities, which the model has to detect. During the test, it is checked whether the model detects all name entities provided in the text file. The check simply iterates over the list of entities from the text file and compares them with the list of extracted entities from the NER model.



eb015a9b-613f-4d4e-a3d7-597074a02f89 123 Main Street Cityville, State 56789 ABC Company 456 Business Avenue Townsville, State 67890 Cityville, 15.10.2023 Termination Mobilephone Contract Dear Sir or Madam Please terminate my mobilephone contract for 077 300 00 00 as soon as possible. May I kindly ask you to confirm the receipt of this letter. Many thank and kind regards eb015a9b-613f-4d4e-a3d7-597074a02f89

1 / 1 found

Figure 10.3: Result shows how many of the entities have been detected

CHAPTER 11

Findings During Implementation and Usage

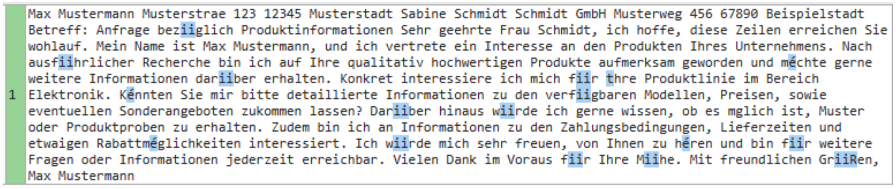
Following chapter outlines the findings done during the implementation and integration of the OCR, NER and anonymisation methods. To highlight the deviations from the output files of the OCR tools and the ground truth, a tool text compare was used. Furthermore, some of the test documents were generated by ChatGPT-V3.5.

11.1 Findings for Optical Character Recognition

This section will highlight the findings made during the implementation of OCR.

11.1.1 Tesseract Struggles With Mutated Vowels

During the implementation and also when using the Anoynmiser app, it became obvious that the initial setup of Tesseract is not able to handle mutated vowels like ä,ö,ü and the character ß, even though the German language is supported out of the box in the latest version.



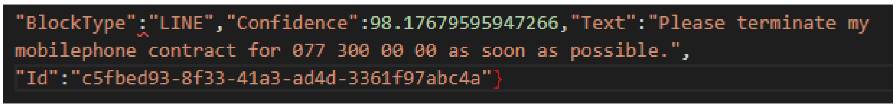
```
Max Mustermann Musterstrae 123 12345 Musterstadt Sabine Schmidt Schmidt GmbH Musterweg 456 67890 Beispielstadt
Betreff: Anfrage beziiglich Produktinformationen Sehr geehrte Frau Schmidt, ich hoffe, diese Zeilen erreichen Sie
wohlauf. Mein Name ist Max Mustermann, und ich vertrete ein Interesse an den Produkten Ihres Unternehmens. Nach
ausfiihrlicher Recherche bin ich auf Ihre qualitativ hochwertigen Produkte aufmerksam geworden und méchte gerne
weitere Informationen dariiber erhalten. Konkret interessiere ich mich fiir thre Produktlinie im Bereich
Elektronik. Kénnten Sie mir bitte detaillierte Informationen zu den verfiigbaren Modellen, Preisen, sowie
eventuellen Sonderangeboten zukommen lassen? Dariiber hinaus wiirde ich gerne wissen, ob es mglich ist, Muster
oder Produktproben zu erhalten. Zudem bin ich an Informationen zu den Zahlungsbedingungen, Lieferzeiten und
etwaigen Rabattméglichkeiten interessiert. Ich wiirde mich sehr freuen, von Ihnen zu héren und bin fiir weitere
Fragen oder Informationen jederzeit erreichbar. Vielen Dank im Voraus fiir Ihre Miihe. Mit freundlichen GriiRen,
Max Mustermann
```

Figure 11.1: Output from Tesseract where mutated vowels are not recognised

As shown in the above figure, Tesseract was not able to recognise the mutated vowels. To resolve this issue, the input parameters have to be changed i.e. German has to be added as language.

11.1.2 OCR Tools Handle Newline Characters Differently

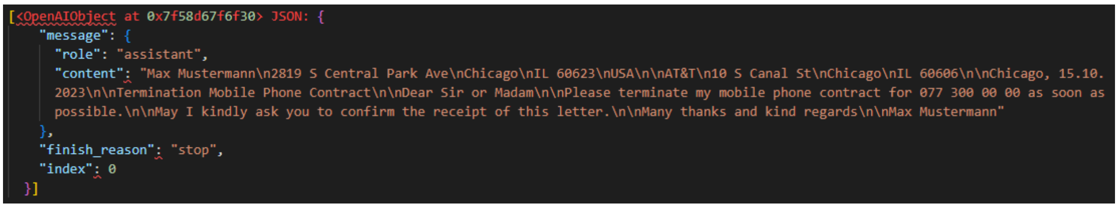
The output of the Optical Character Recognition tools have different formats and also the representation of the newline characters is done differently. Amazon Textract and GPT-4V provide a JSON response, whereas Tesseract returns a string. The JSON response from Textract provides each line of text stored as a key-value pair in combination with additional information like coordinates. Below JSON segment shows how Textract returns the extracted text.



```
"BlockType":"LINE","Confidence":98.17679595947266,"Text":"Please terminate my
mobilephone contract for 077 300 00 00 as soon as possible.",
"Id":"c5fbed93-8f33-41a3-ad4d-3361f97abc4a"}
```

Figure 11.2: Example how a text line is returned by Textract

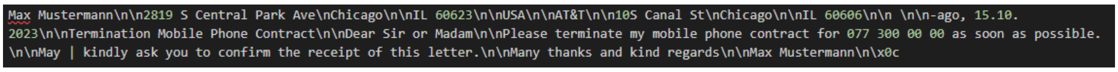
On the other hand, the JSON file from GPT-4V contains a string with the newline characters included.



```
[<OpenAIObject at 0x7f58d67f6f30> JSON: {
    "message": {
      "role": "assistant",
      "content": "Max Mustermann\n2819 S Central Park Ave\nChicago\nIL 60623\nUSA\n\nAT&T\n10 S Canal St\nChicago\nIL 60606\n\nChicago, 15.10.
      2023\n\nTermination Mobile Phone Contract\n\nDear Sir or Madam\n\nPlease terminate my mobile phone contract for 077 300 00 00 as soon as
      possible.\n\nMay I kindly ask you to confirm the receipt of this letter.\n\nMany thanks and kind regards\n\nMax Mustermann"
    },
    "finish_reason": "stop",
    "index": 0
  }]
```

Figure 11.3: Response provided by GPT-4V

Tesseract also includes the newline characters in the output string.



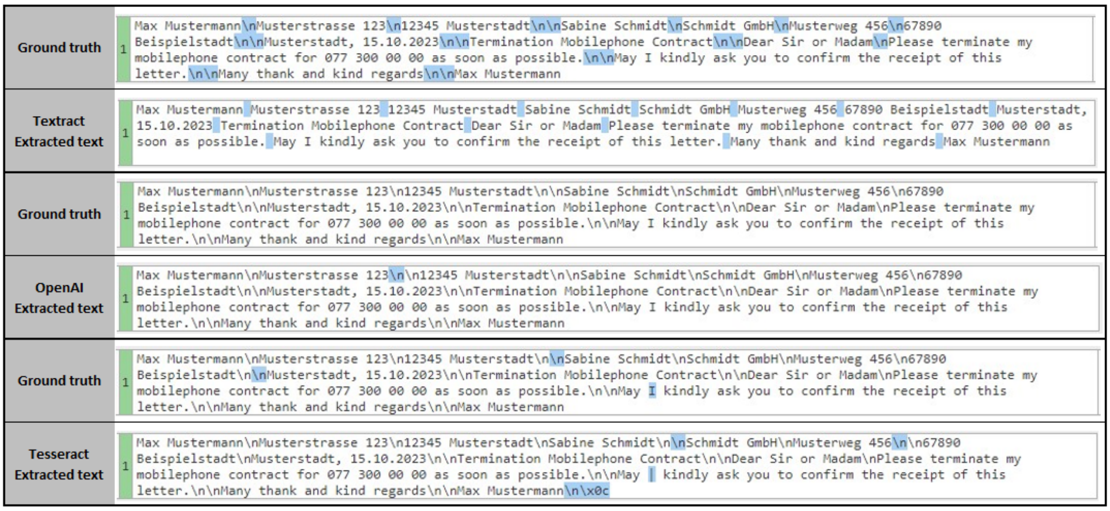
```
Max Mustermann\n\n2819 S Central Park Ave\nChicago\n\nIL 60623\n\nUSA\n\nAT&T\n\n10S Canal St\nChicago\n\nIL 60606\n\n \n\n-ago, 15.10.
2023\n\nTermination Mobile Phone Contract\n\nDear Sir or Madam\n\nPlease terminate my mobile phone contract for 077 300 00 00 as soon as possible.
\n\nMay | kindly ask you to confirm the receipt of this letter.\n\nMany thanks and kind regards\n\nMax Mustermann\n\x0c
```

Figure 11.4: String provided by Tesseract as result

Whereas GPT-4V and Tesseract include the newline characters in their response, in the response from Textract they are missing. These circumstances needs to be considered for subsequent processing.

11.1.3 Handling of Newline Characters

Previous section already introduced the findings which come along with the different return types. The handling of the newline characters is a challenging tasks. As Amazon Textract returns a JSON, the information about the number of newlines between text sections is missing. In addition, the number of newlines are also varying due to differences in the recognition, when comparing the output of GPT-4V and Tessarct. Below table compares the three OCR tools and how they recognise the newline characters.



Ground truth	Max Mustermann\nMusterstrasse 123\n12345 Musterstadt\n\nSabine Schmidt\nSchmidt GmbH\nMusterweg 456\n67890 Beispielstadt\n\nMusterstadt, 15.10.2023\n\nTermination Mobilephone Contract\n\nDear Sir or Madam\nPlease terminate my mobilephone contract for 077 300 00 00 as soon as possible.\n\nMay I kindly ask you to confirm the receipt of this letter.\n\nMany thank and kind regards\n\nMax Mustermann
Textract Extracted text	Max Mustermann Musterstrasse 123 12345 Musterstadt Sabine Schmidt Schmidt GmbH Musterweg 456 67890 Beispielstadt Musterstadt, 15.10.2023 Termination Mobilephone Contract Dear Sir or Madam Please terminate my mobilephone contract for 077 300 00 00 as soon as possible. May I kindly ask you to confirm the receipt of this letter. Many thank and kind regards Max Mustermann
Ground truth	Max Mustermann\nMusterstrasse 123\n12345 Musterstadt\n\nSabine Schmidt\nSchmidt GmbH\nMusterweg 456\n67890 Beispielstadt\n\nMusterstadt, 15.10.2023\n\nTermination Mobilephone Contract\n\nDear Sir or Madam\nPlease terminate my mobilephone contract for 077 300 00 00 as soon as possible.\n\nMay I kindly ask you to confirm the receipt of this letter.\n\nMany thank and kind regards\n\nMax Mustermann
OpenAI Extracted text	Max Mustermann\nMusterstrasse 123\n\n12345 Musterstadt\n\nSabine Schmidt\nSchmidt GmbH\nMusterweg 456\n67890 Beispielstadt\n\nMusterstadt, 15.10.2023\n\nTermination Mobilephone Contract\n\nDear Sir or Madam\nPlease terminate my mobilephone contract for 077 300 00 00 as soon as possible.\n\nMay I kindly ask you to confirm the receipt of this letter.\n\nMany thank and kind regards\n\nMax Mustermann
Ground truth	Max Mustermann\nMusterstrasse 123\n12345 Musterstadt\n\nSabine Schmidt\nSchmidt GmbH\nMusterweg 456\n67890 Beispielstadt\n\nMusterstadt, 15.10.2023\n\nTermination Mobilephone Contract\n\nDear Sir or Madam\nPlease terminate my mobilephone contract for 077 300 00 00 as soon as possible.\n\nMay I kindly ask you to confirm the receipt of this letter.\n\nMany thank and kind regards\n\nMax Mustermann
Tesseract Extracted text	Max Mustermann\nMusterstrasse 123\n12345 Musterstadt\nSabine Schmidt\n\nSchmidt GmbH\nMusterweg 456\n\n67890 Beispielstadt\nMusterstadt, 15.10.2023\n\nTermination Mobilephone Contract\n\nDear Sir or Madam\nPlease terminate my mobilephone contract for 077 300 00 00 as soon as possible.\n\nMay | kindly ask you to confirm the receipt of this letter.\n\nMany thank and kind regards\n\nMax Mustermann\n\x0c

Figure 11.5: Highlighting the differences in newline character recognition

Due to the fact, that the final output format of the Anonymiser app is a string, the issue with the newline characters can be resolved. For Tesseract and GPT-4V all newline characters in the string are replaced by whitespaces. As the number of newline characters varies, multiple whitespaces can follow after each other. To avoid having several whitespaces after each other, the string is converted in an array with the split function. The split function ignores multiple whitespaces and just takes the word between two whitespaces and puts it the array. Finally, the array is converted to a string again, by separating the words with only one whitespace. By doing so, it can be ensured, that multiple newlines recognised by the OCR do not impact the final output.

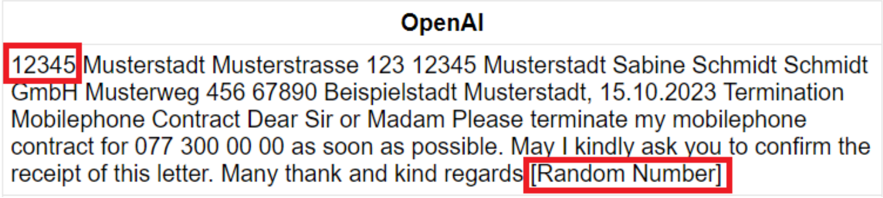
To retrieve the text from Textract, the text lines stored as key-value pairs in the JSON are extracted and converted to one string. There is no newline character handling needed.

11.1.4 Instructions for OpenAI GPT-4V

It is impressive how powerful GPT-4V is. To tell the tool what it has to do, a prompt is sent to the OpenAI API. This prompt is critical to get the needed information. In the scope of this work, OpenAI GPT-4V is used to extract the text within a image. However, the tool offers much more possibilities as all subsequent process steps could be done with it. The prompt has been adjusted as follows:

Extract the text within the image. Then identify the sender of the letter and anonymise the sender with a random number. Return only the anonymised text without any additional comments.

The response is not yet perfect, however GPT-4V was able to understand the instructions and to provide a result. The uploaded test document can be found in the appendix 18.3.



OpenAI

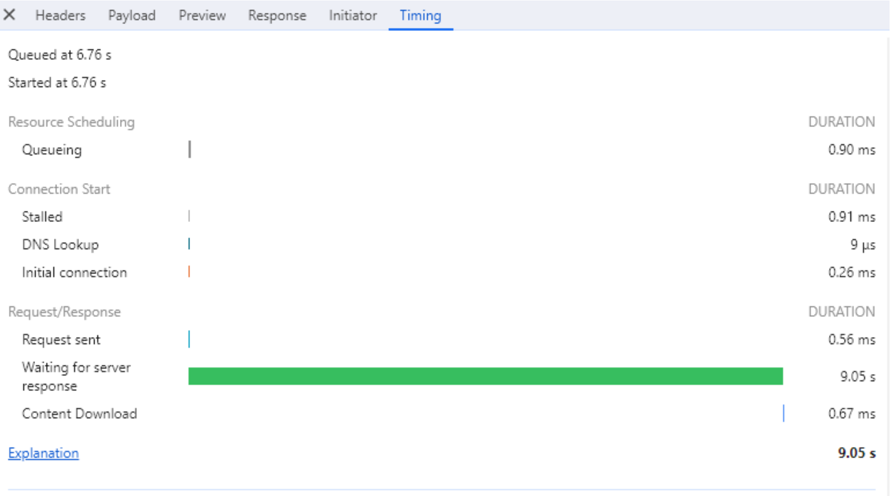
12345 Musterstadt Musterstrasse 123 12345 Musterstadt Sabine Schmidt Schmidt GmbH Musterweg 456 67890 Beispielstadt Musterstadt, 15.10.2023 Termination Mobilephone Contract Dear Sir or Madam Please terminate my mobilephone contract for 077 300 00 00 as soon as possible. May I kindly ask you to confirm the receipt of this letter. Many thank and kind regards [Random Number]

Figure 11.6: Result of the end-to-end anonymisation process done by GPT-4V

Notable is that GPT-4V was able to distinguish between the sender and the recipient of the letter.

11.1.5 Slow Response Times for OpenAI Requests

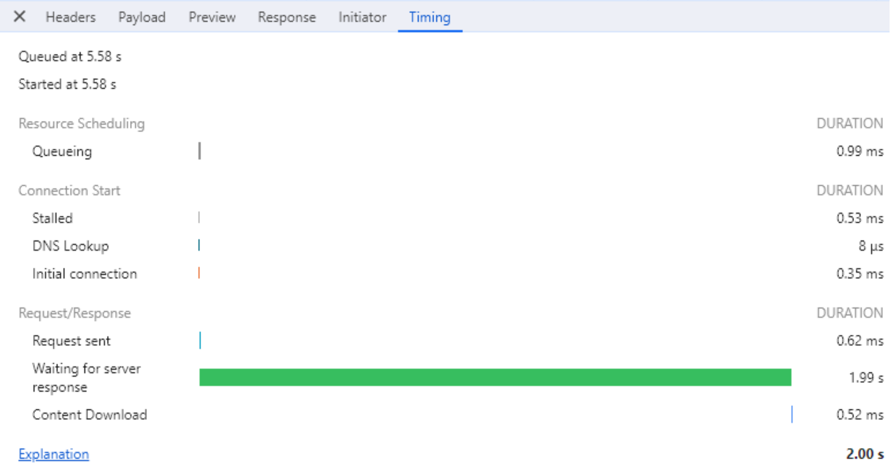
Drawback of OpenAI GPT-4V are the slow response times compared to the other tools.



Queued at 6.76 s
Started at 6.76 s

Resource Scheduling	DURATION
Queueing	0.90 ms
Connection Start	DURATION
Stalled	0.91 ms
DNS Lookup	9 µs
Initial connection	0.26 ms
Request/Response	DURATION
Request sent	0.56 ms
Waiting for server response	9.05 s
Content Download	0.67 ms
	9.05 s

Figure 11.7: Long response time for OpenAI service



Queued at 5.58 s
Started at 5.58 s

Resource Scheduling	DURATION
Queueing	0.99 ms
Connection Start	DURATION
Stalled	0.53 ms
DNS Lookup	8 µs
Initial connection	0.35 ms
Request/Response	DURATION
Request sent	0.62 ms
Waiting for server response	1.99 s
Content Download	0.52 ms
	2.00 s

Figure 11.8: Response time for Textract

In the scope of this POC, the response times are less relevant. For a productive environment, a more scalable and performant application is necessary. It might be re-evaluated, which tools are suitable as especially the slow response times are a known issue for GPT-4V [5].

11.1.6 OCR Quality of Tesseract

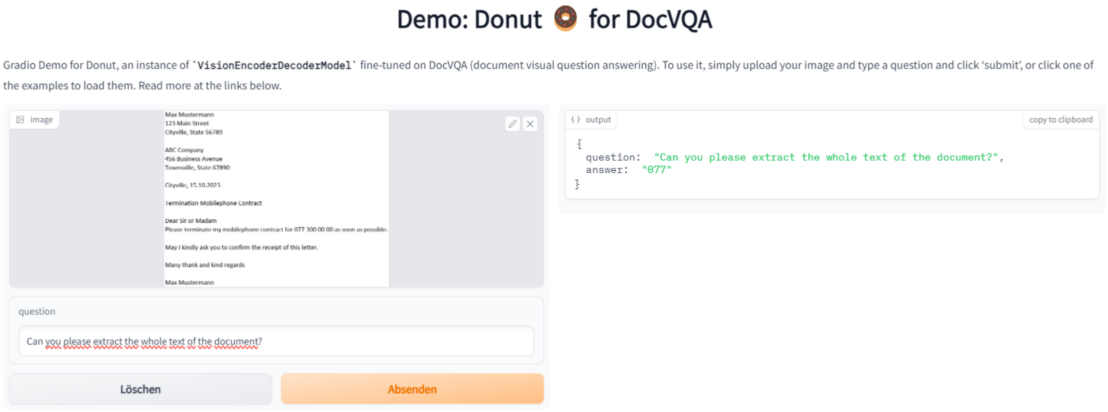
In all the test document uploaded to the Anonymiser app, Tesseract lacks behind GPT-4V and Textract in regard to OCR accuracy. Even though the text cases were simple letters without any special layout.

If the document structure is more complex, the input parameters for Tesseract have to be adjusted accordingly. For example could other segmentation methods be selected. An example test can be found in the appendix chapter 18.1

11.1.7 Donut Model Is Difficult to Integrate and Not Yet Suited

During the initial phase of this work, the Donut model was considered to be integrated. Even though there exists several tutorials, videos and blog posts, it is not trivial to get the application run in a adequate manner. There exist pre-trained Donut models, but they are not suited for the Anonymiser app. The models are often trained on the Consolidated Receipt Dataset or trained for

Visual Question Answering. Consequently, they are not able to extract a whole text from a document, which then can be anonymised. Furthermore, there were several issues with the dependencies to get the model run. However, as several OCR-free models have arisen, there is a lot of potential.



Demo: Donut 🍩 for DocVQA

Gradio Demo for Donut, an instance of `VisionEncoderDecoderModel` fine-tuned on DocVQA (document visual question answering). To use it, simply upload your image and type a question and click 'submit', or click one of the examples to load them. Read more at the links below.

Max Mustermann
123 Main Street
Cityville, State 56789

ABC Company
456 Business Avenue
Townsville, State 67890

Cityville, 15.10.2023

Termination Mobilephone Contract

Dear Sir or Madam
Please terminate my mobilephone contract for 077 300 00 00 as soon as possible.

May I kindly ask you to confirm the receipt of this letter.

Many thank and kind regards

Max Mustermann

question: Can you please extract the whole text of the document?

```
{
  question:  "Can you please extract the whole text of the document?",
  answer:  "077"
}
```

Figure 11.9: The Donut model trained for VQA is not able to extract the whole text of a document

11.2 Findings for Named Entity Recognition

This section discusses the findings made during the implementation of the bert-base-NER Transformer model.

11.2.1 Identifying the Most Suitable Model

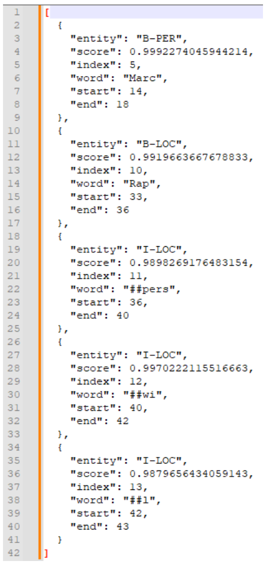
As of January 2024 there exist more than 14'000 models capable of token classification on Huggingface. Some models are pre-trained, others are already fine-tuned. Some of the models are trained to recognise entities in Chinese, others to understand the Danish language. There exist models, which are very well documented and a reference to the training set is provided. On the other hand, there exist models, which do not have a documentation at all. Also the tags for same entities or tokens are not standardised. Simply replacing a model, may require also code adjustments.

Due to the massive number of models, it is difficult to find the most suitable model. A good starting point is the number of downloads of the model. Afterwards, the documentation should be consulted and some models even offer a online demo.

11.2.2 Models Are Not Simply Off-the-Shelf

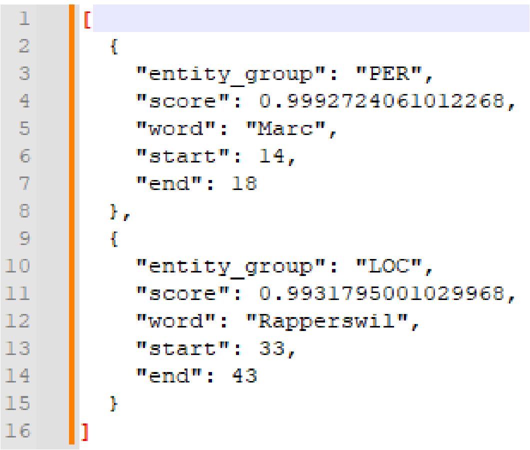
Integration of models into an application can be cumbersome. For the bert-base-NER Transformer model used in this application, numerous dependencies had to be downloaded and installed. This lead from time to time to dependency issues, so that the model did not work anymore. To overcome these issues, the application was ultimately set up in Docker.

Once a model is up and running, it has to be noticed that there are several options, which can be used to improve the output. For example, it can be decided if tokens of the same entity should be grouped. The entity *Rapperswil* will be tokenized in *"Rap", "##pers", "##wi" and "##l"*. In order to get the correct output, "grouped_entities=True" has been given as a parameter for the model pipeline.



```
[
  {
    "entity": "B-PER",
    "score": 0.9992274045944214,
    "index": 5,
    "word": "Marc",
    "start": 14,
    "end": 18
  },
  {
    "entity": "B-LOC",
    "score": 0.9919663667678833,
    "index": 10,
    "word": "Rap",
    "start": 33,
    "end": 36
  },
  {
    "entity": "I-LOC",
    "score": 0.9898269176483154,
    "index": 11,
    "word": "##pers",
    "start": 36,
    "end": 40
  },
  {
    "entity": "I-LOC",
    "score": 0.997022115516663,
    "index": 12,
    "word": "##wi",
    "start": 40,
    "end": 42
  },
  {
    "entity": "I-LOC",
    "score": 0.9879656434059143,
    "index": 13,
    "word": "##l",
    "start": 42,
    "end": 43
  }
]
```

Figure 11.10: Example response if grouped entity is not selected



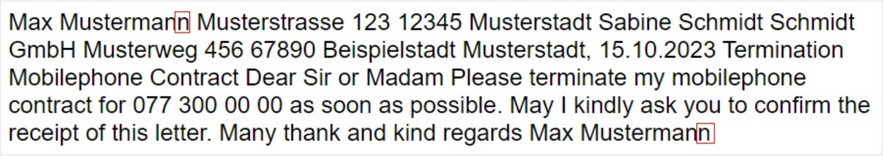
```
[
  {
    "entity_group": "PER",
    "score": 0.9992724061012268,
    "word": "Marc",
    "start": 14,
    "end": 18
  },
  {
    "entity_group": "LOC",
    "score": 0.993179500102996 8,
    "word": "Rapperswil",
    "start": 33,
    "end": 43
  }
]
```

Figure 11.11: Example response if grouped entity is selected

Several parameters exist, which have an impact on the output and on the performance of the model pipeline. Depending on what the result should look like, further adjustments have to be done.

11.2.3 Difficulties With Tokenization

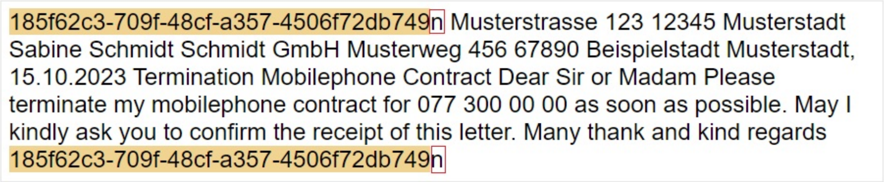
The entity recognition does not always work accurately, as can be seen in the following example:



Max Mustermann Musterstrasse 123 12345 Musterstadt Sabine Schmidt Schmidt GmbH Musterweg 456 67890 Beispielstadt Musterstadt, 15.10.2023 Termination Mobilephone Contract Dear Sir or Madam Please terminate my mobilephone contract for 077 300 00 00 as soon as possible. May I kindly ask you to confirm the receipt of this letter. Many thank and kind regards Max Mustermann

Figure 11.12: Text provided by OCR

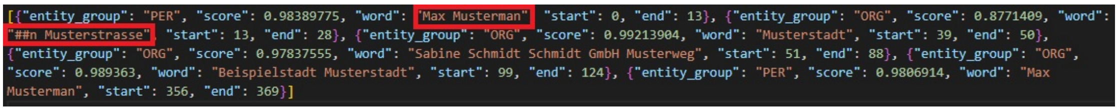
The anonymisation of "Mustermann" did not work properly as the last letter of the name was skipped.



185f62c3-709f-48cf-a357-4506f72db749n Musterstrasse 123 12345 Musterstadt Sabine Schmidt Schmidt GmbH Musterweg 456 67890 Beispielstadt Musterstadt, 15.10.2023 Termination Mobilephone Contract Dear Sir or Madam Please terminate my mobilephone contract for 077 300 00 00 as soon as possible. May I kindly ask you to confirm the receipt of this letter. Many thank and kind regards 185f62c3-709f-48cf-a357-4506f72db749n

Figure 11.13: Name entity was not properly recognised and hence the anonymisation is not accurate

To further investigate this case, the result of the entity recognition was checked. It can be seen that the last "n" of the name was skipped and added to the address.



```
[{"entity_group": "PER", "score": 0.98389775, "word": "Max Musterman"  "start": 0, "end": 13}, {"entity_group": "ORG", "score": 0.8771409, "word":
"##n Musterstrasse", "start": 13, "end": 28}, {"entity_group": "ORG", "score": 0.99213904, "word": "Musterstadt", "start": 39, "end": 50},
{"entity_group": "ORG", "score": 0.97837555, "word": "Sabine Schmidt Schmidt GmbH Musterweg", "start": 51, "end": 88}, {"entity_group": "ORG",
"score": 0.989363, "word": "Beispielstadt Musterstadt", "start": 99, "end": 124}, {"entity_group": "PER", "score": 0.9806914, "word": "Max
Musterman", "start": 356, "end": 369}]
```

Figure 11.14: JSON response provides an overview of the recognised name entities

One reason for this issue, is the token classification with the WordPiece algorithm. "Mustermann" is tokenized into the tokens "Must", "##erman" and "##n". This is highly possible due to the fact, that the bert-base-NER model is trained in English language. A model trained in German language might have done a different tokenization. However, why the token "##n" was added to the location has to be further analysed.

11.3 Findings for Anonymisation

In this section the experiences made during the implementation of the anonymisation process are outlined.

11.3.1 Using UUID

The UUID is a simple method to replace sensitive data with a randomly generated string. Only the native Python module UUID needs to be imported and UUID4 can be used. As the UUID4 cannot be reverted, the anonymisation is deemed to be secure. However, if the document is pseudonymised a database is required to store the original value and the corresponding UUID.

11.3.2 Varying Lenght of Name and UUID

The UUID consists of 128 bits. As the output of the Anonymiser app is always a string, the length of the UUID does not negatively impact the layout of the output. It does not matter that "Max Mustermann" (length < 128 bits) is replaced with a UUID. However, once the anonymisation takes place in the original document and the structure has to remain the same, the UUID might be not well suited as it would overwrite the subsequent text.

11.4 End-to-End Process Findings

Is the Anonymiser app able to perform the task end-to-end reliable? The answer to this question - it depends.

The Anonymiser app works quite well for documents and images with a simple layout and no noise. All three OCR tools are able to extract the text correctly. Also the entity recognition works and hence, the pseudonymisation as well. A simple example and the result can be found in the appendix chapter 18.3.

The end-to-end process for images of letters containing more content still works. However, for some documents minor errors occurred. Especially, in the address section of the letter. Even though, the OCR works perfectly, the NER model is not always able to recognise the person names correctly. This might be due to the fact, that bert-base-NER model is trained in English language and therefore not the most suitable model for the test documents (some test documents contain German names). The errors made during the entity recognition phase are then passed to the pseudonymisation phase, where not all entities were correctly pseudonymised. An example of a correct test can be found in the appendix chapter 18.4.

The last end-to-end test was done with a image of a real word invoice. The image is skewed, the lightning is suboptimal, there is noise and also the size is not ideal. Furthermore, the invoice is in German language. GPT-4V as well as Tesseract were not able to provide a valid output. The result provided by Textract was suitable and hence it was used to be anonymised. However, the NER model was not able to identify the persons correctly and hence the anonymisation process did also not work well. The test case can be found in the appendix chapter 18.5.

The Anonymiser app has shown, that it is able to fulfill the end-to-end process. However, based on the small number of test documents it is not possible to make a final statement on how well the app works. There are some indication that the bert-base-NER model is not the best suited NER model for the kind of data. To enhance the Anonymiser app and to overcome the issue of propagating wrong or suboptimal data to the next processing phase, it needs to be checked, if the bert-base-model is the most suitable model for this task.

11.4.1 Can the Anonymiser App Bring Digitalisation Further?

At the beginning of this work, two areas of application for the Anonymiser app were identified. One is the anonymisation of medical history data and the other area is data leakage prevention. The Anonymiser app should work as a POC for these areas. Even though, the app is far from being final, some of the main features work. The app is able to extract text from images or scanned documents, recognises name entities and anonymises these.

Nevertheless, further adjustment on the app have to be done in order to integrate it in the two areas. For the area of the anonymisation of medical history data, the documents have to be anonymised. This can simply be achieved by removing the data link component, so that anonymisation is irreversible. Additionally, other entities than names have to be identified and the optimal anonymisation techniques need to be evaluated, so that no relevant information is lost.

In regard to the data leakage prevention, additional interfaces need to be implemented, so that all client names are available for the Anonymiser app. In addition, the app has to be integrated in the email sending process, so that the attachments of the email can be scanned for client data.

CHAPTER 12

Conclusion and Outlook

12.1 Conclusion

It is impressive how powerful today's text extraction and NER models have become. The Anonymiser app was built using various tools to implement a process to anonymise text from an image. Three OCR engines have been integrated to extract the text from an image or PDF file. Each tool provides different features. However, different performance regarding OCR accuracy was noted. Amazon Textract performed the best for the small set of test documents. Not only was the accuracy assessed, but also the simplicity of integrating the tool. Tools like OpenAI or Textract are simple to integrate as only an API needs to be called. Hosting some code in an Amazon Lambda Function is also possible. Tesseract runs on the server side and is easy to install, but further adjustments are needed to get the optimal results.

The implemented bert-base-NER model provides overall good results. It can extract the name entity for a lot of the test documents. However, the integration of such models can become challenging as many dependencies have to be installed. Once the model runs, further parameters have to be passed to optimise the result. The bert-base-NER model was not the ideal model for the Anonymiser app, as the test documents mainly consisted of letters, and the model was trained on articles. Furthermore, the language also had an impact on the output quality.

Provided OCR and name entity recognition worked properly, pseudonymisation is the most straightforward part of the whole process. The pseudonymisation was implemented with the help of a UUID, which is stored with the entity in a PostgreSQL database. This also makes it possible to revert the pseudonymisation.

During the implementation of each process step, it became obvious that the steps depend highly on each other. If an error is made at the beginning of the process, the error is propagated to subsequent phases and will have a negative impact on the result. Challenges like newline character handling, improving OCR accuracy, finding the optimal NER model, or issues with tokenization need to be handled.

Nevertheless, the application demonstrated that it could automate the anonymisation of an image text or PDF file. The Anonymiser app is deemed a POC for anonymising medical history data and data leakage prevention.

12.2 Outlook

Using different tools and models that work independently is challenging. Ideally, the process of text extraction and entity recognition is executed by one model to avoid error propagation. This could be achieved by using an end-to-end OCR-free VDU model like Donut. Furthermore, the model should be fine-tuned in the respective language and for the specific types of documents. Research in this area is ongoing, and new models, e.g., DUBLIN, are coming up. Nevertheless, also the applicability needs to be further improved, so that the integration and training of models becomes easier.

Integrating such new tools in the Anonymiser app could improve the performance and capabilities. So far, the focus has been on detecting name entities. However, there are a lot more entities that have to be identified and anonymised in order to be really useful (i.e., medical history data and data leakage prevention). Once additional entities are detected, the anonymisation process will require much more attention as not all entities can simply be replaced by a UUID. Other techniques have to be included.

In addition to just anonymising a string, the anonymisation of sensitive data in the image or document could be of interest. Maintaining the structure of a document in combination with the anonymisation comes one step closer to the original vision of the Anonymiser app. A tool that effortless anonymises images and PDF documents.

Part II

Project Documentation

CHAPTER 13

Project Plan

The project plan was created according to the four phases of RUP. This means, the time available to finish the bachelor thesis was divided in the following phases:

Inception

During the inception phase, the repositories, templates, and Jira are set up. A project plan draft and some initial research has to be done.

Elaboration

In the elaboration phase, the focus is still on the literature research. In addition, it is planned to get the DONUT model to run, which will provide an impression how a transformer model works. Furthermore, the task description needs to be finalised.

Construction

Main focus during the construction phase is to build the Anonymiser app. There is still a permanent need for literature research, as the most suitable tools have to be identified. It also has to be started with writing the documentation.

Transition

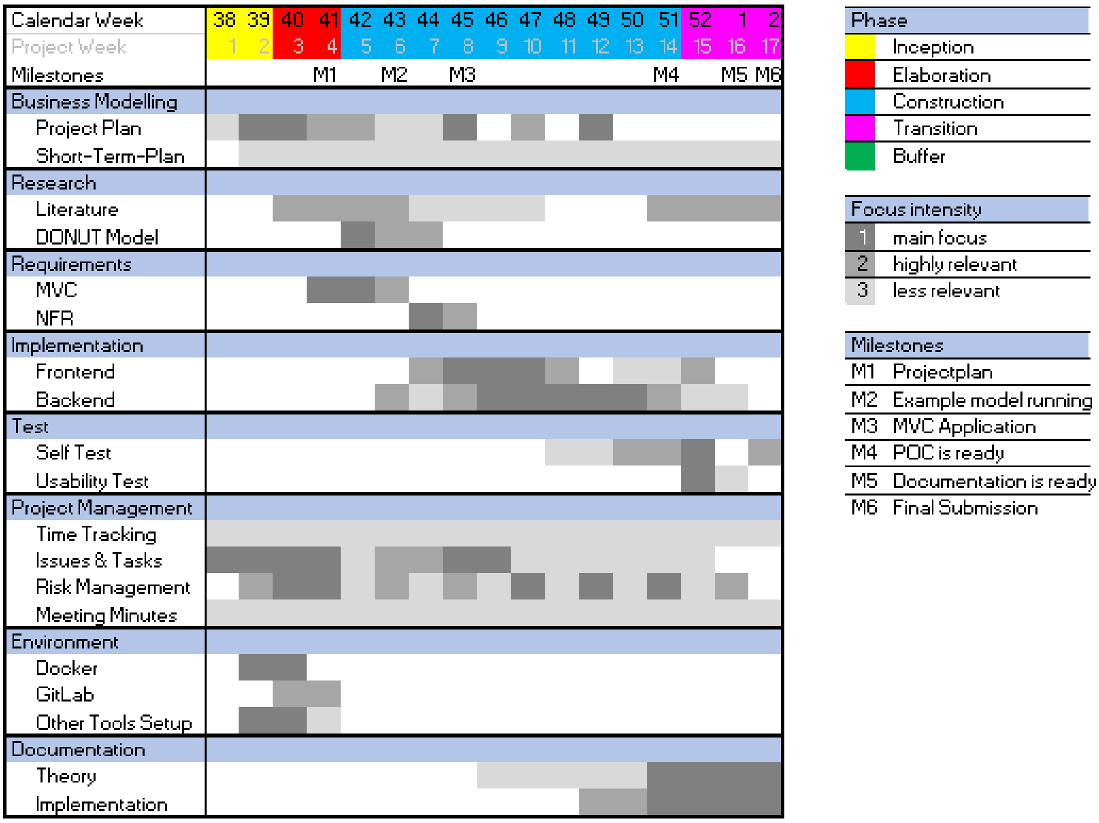
The transition phase is used to finalise the POC of the Anonymiser app. Also the writing on the documentation has to come to an end.

Milestones

To monitor the progress of the project and to ensure a timely hand in, six milestones have been defined:

- M1: Finalised project plan
- M2: Donut models runs on the local machine and valuable insights are gained
- M3: End-to-end pseudonisation process works
- M4: POC is ready incl. OCR/NER accuracy check and frontend
- M5: Documentation is ready
- M6: Final Submission

Based on the above defined milestones and project phases, a long term planning has been done.



Calendar Week	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	1	2
Project Week	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Milestones				M1		M2		M3						M4		M5	M6
Business Modelling																	
Project Plan	░	▓	▓	▒	▒	░	░	▓		▒		▓					
Short-Term-Plan		░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░
Research																	
Literature			▒	▒	▒	▒	░	░	░	░				▒	▒	▒	▒
DONUT Model					▓	▒	▒										
Requirements																	
MVC				▓	▓	▒											
NFR							▓	▒									
Implementation																	
Frontend							▒	▓	▓	▓	▒		░	░	▒		
Backend						▒	░	▒	▓	▓	▓	▓	▓	▒	░	░	
Test																	
Self Test											░	░	▒	▒	▓		▒
Usability Test															▓	░	
Project Management																	
Time Tracking	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░
Issues & Tasks	▓	▓	▓	▓	░	▒	▒	▓	▓	░	░	░	░	░	░		
Risk Management		▒	▓	▓	░	▒	░	▒	░	▓	░	▓	░	▓	░	▒	
Meeting Minutes	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░
Environment																	
Docker		▓	▓														
GitLab			▒	▒													
Other Tools Setup		▓	▓	░													
Documentation																	
Theory									░	░	░	░	░	▓	▓	▓	▓
Implementation												▒	▒	▓	▓	▓	▓

Phase	
🟨	Inception
🟥	Elaboration
🟦	Construction
🟪	Transition
🟩	Buffer

Focus intensity	
▓ 1	main focus
▒ 2	highly relevant
░ 3	less relevant

Milestones	
M1	Projectplan
M2	Example model running
M3	MVC Application
M4	POC is ready
M5	Documentation is ready
M6	Final Submission

Table 13.1: Longterm plan for the bachelor thesis

Part III

Appendix

List of Figures

List of Tables

Bibliography

[1] —. *Improving the quality of the output*. https://tesseract-ocr.github.io/tessdoc/ImproveQuality.html. accessed: 05.01.2024.

[2] —. *jamesturk/jellyfish*. URL: https://github.com/jamesturk/jellyfish. accessed: 09.01.2024.

[3] —. *Jaro–Winkler distance*. URL: https://en.wikipedia.org/wiki/Jaro%E2%80%93Winkler_distance. accessed: 09.01.2024.

[4] —. *NLP Course*. URL: https://huggingface.co/learn/nlp-course/chapter1/1. accessed: 09.01.2024.

[5] —. *OpenAI Why Are The API Calls So Slow? When will it be fixed?* URL: https://community.openai.com/t/openai-why-are-the-api-calls-so-slow-when-will-it-be-fixed/148339?page=2. accessed: 07.01.2024.

[6] —. *The C4 model for visualising software architecture*. URL: https://c4model.com/. accessed: 09.01.2024.

[7] Accenture. *What is digital transformation?* 2000. URL: https://www.accenture.com/us-en/insights/digital-transformation-index. accessed: 12.12.2023.

[8] Acrobat. *What is OCR and why is OCR software important?* URL: https://www.adobe.com/acrobat/guides/what-is-ocr.html. accessed: 02.01.2024.

[9] Sharp Cookie Advisors. *Anonymization and GDPR compliance; an overview*. 2020. URL: https://www.gdprsummary.com/anonymization-and-gdpr/. accessed: 16.12.2023.

[10] Ashish Vaswani et al. *Attention Is All You Need*. Tech. rep. Google Brain, 2017.

[11] Geewook Kim et al. *OCR-free Document Understanding Transformer*. Tech. rep. NAVER CLOVA, 2022.

[12] Kriti Aggarwal et al. *DUBLIN: Visual Document Understanding By Language-Image Network*. Tech. rep. Microsoft Corporation, 2023.

[13] Ze Liu et al. *Swin Transformer: Hierarchical Vision Transformer using Shifted Windows*. Tech. rep. Microsoft Research Asia, 2021.

[14] Zhengyuan Yang et al. *The Dawn of LMMs: Preliminary Explorations with GPT-4V(ision)*. Tech. rep. Microsoft Corperation, 2023.

[15] *Amazon Textract*. URL: https://aws.amazon.com/textract/. accessed: 05.01.2024.

[16] *Amazon Textract features*. URL: https://aws.amazon.com/textract/faqs/. accessed: 05.01.2024.

[17] Amnesia. *High accuracy Data Anonymization*. URL: https://amnesia.openaire.eu/index.html. accessed: 31.12.2023.

[18] Amnesia. *Learn the process*. URL: https://amnesia.openaire.eu/about-flow.html. accessed: 31.12.2023.

[19] PostgreSQL Anonymizer. *Anonymization & Data Masking for PostgreSQL*. URL: https://postgresql-anonymizer.readthedocs.io/en/stable/. accessed: 31.12.2023.

[20] ARX. *Anonymization tool.* URL: https://arx.deidentifier.org/anonymization-tool/. accessed: 31.12.2023.

[21] ARX. *Overview.* URL: https://arx.deidentifier.org/overview/. accessed: 31.12.2023.

[22] ARX. *Privacy models.* URL: https://arx.deidentifier.org/overview/privacy-criteria/. accessed: 31.12.2023.

[23] Greg Council. *Using OCR: How Accurate is Your Data?* URL: https://tdwi.org/articles/2018/03/05/diq-all-how-accurate-is-your-data.aspx. accessed: 11.01.2024.

[24] IBM Cloud Education. *What Is Optical Character Recognition (OCR)?* 2022. URL: https://www.ibm.com/blog/optical-character-recognition/. accessed: 02.01.2024.

[25] *Encoder models.* URL: https://huggingface.co/learn/nlp-course/chapter1/5?fw=pt. accessed: 04.01.2024.

[26] Leonie Ritscher Erich Richter. *Datenschutz: Eine Übersicht zum neuen Gesetz.* 2023. URL: https://www.economiesuisse.ch/de/artikel/datenschutz-eine-uebersicht-zum-neuen-gesetz-0. accessed: 23.12.2023.

[27] *Example: Thinning and Skeletonization.* URL: https://support.ptc.com/help/mathcad/r9.0/en/index.html#page/PTC_Mathcad_Help/example_thinning_and_skeletonization.html#. accessed: 02.01.2024.

[28] Ed Felten. *Does Hashing Make Data "Anonymous"?* 2012. URL: https://www.ftc.gov/policy/advocacy-research/tech-at-ftc/2012/04/does-hashing-make-data-anonymous. accessed: 01.01.2024.

[29] Python Software Foundation. *Secure hashes and message digests.* 2024. URL: https://docs.python.org/3/library/hashlib.html. accessed: 01.01.2024.

[30] Python Software Foundation. *UUID objects according to RFC 4122.* 2024. URL: https://docs.python.org/3/library/uuid.html. accessed: 01.01.2024.

[31] Michal Gasparik. *Understanding UUID: Purpose and Benefits of a Universal Unique Identifier.* 2023. URL: https://medium.com/@gaspm/understanding-uuid-purpose-and-benefits-of-a-universal-unique-identifier-59110154d897. accessed: 01.01.2024.

[32] Karez Abdulwahhab Hamad and Mehmet Kaya. "A Detailed Analysis of Optical Character Recognition Technology". In: *International Journal of Applied Mathematics, Electronics and Computers* Special Issue.4 (2016).

[33] *How do Transformers work?* URL: https://huggingface.co/learn/nlp-course/chapter1/4?fw=pt. accessed: 04.01.2024.

[34] MENTIS INC. *Anonymization vs. Pseudonymization.* 2019. URL: https://medium.com/@mentisinc/anonymization-vs-pseudonymization-faca97676eb2. accessed: 26.12.2023.

[35] *Input Documents.* URL: https://docs.aws.amazon.com/textract/latest/dg/how-it-works-documents.html. accessed: 05.01.2024.

[36] Joshua J. *How much does GPT-4 cost?* https://help.openai.com/en/articles/7127956-how-much-does-gpt-4-cost. 2023. accessed: 05.01.2024.

[37] Deolinda Rasteiro Joana Tomás and Jorge Bernardino. "Data Anonymization: An Experimental Evaluation Using Open-Source Tools". In: *Future Internet* 167.14 (2022).

[38] Forough Karandish. *The Comprehensive Guide to Optical Character Recognition (OCR).* URL: https://moov.ai/en/blog/optical-character-recognition-ocr. accessed: 02.01.2024.

[39] Yasin Küçükkaya. *7 zentrale Unterschiede zwischen dem nDSG und der DSGVO.* 2023. URL: https://www.adnovum.com/de/blog/datenschutzgesetz-2023-was-unterscheidet-das-neue-dsg-von-der-dsgvo. accessed: 26.12.2023.

[40] Yasin Küçükkaya. *Was ist das DSG? Das neue Schweizer Datenschutzgesetz 2023 im Überblick.* 2023. URL: https://www.adnovum.com/de/blog/das-datenschutzgesetz-2023. accessed: 23.12.2023.

[41] Konstantinos Limniotis (HDPA) Meiko Jensen (Kiel University) Cedric Lauradoux (INRIA). *Pseudonymisation techniques and best practices*. Tech. rep. European Union Agency for Cybersecurity, 2019.

[42] Dewi Nasien Muhammad Arif Mohamad Haswadi Hassan and Habibollah Haron. "A Review on Feature Extraction and Feature Selection for Handwritten Character Recognition". In: *International Journal of Advanced Computer Science and Applications* 6.2 (2015).

[43] Kevin Nisbet. *The False Allure of Hashing for Anonymization*. 2018. URL: `https://goteleport.com/blog/hashing-for-anonymization/`. accessed: 01.01.2024.

[44] M. Mealling P. Leach and R. Salz. *A Universally Unique IDentifier (UUID) URN Namespace*. 2005. URL: `https://datatracker.ietf.org/doc/html/rfc4122.html%5C#section-4.1`. accessed: 01.01.2024.

[45] Vijay Pandurangan. *On Taxis and Rainbow Tables: Lessons for researchers and governments from NYC's improperly anonymized taxi logs*. 2014. URL: `https://blogs.lse.ac.uk/impactofsocialsciences/2014/07/16/nyc-improperly-anonymized-taxi-logs-pandurangan/`. accessed: 01.01.2024.

[46] Susmith Reddy. *Segmentation in OCR !!* 2019. URL: `https://towardsdatascience.com/segmentation-in-ocr-10de176cf373`. accessed: 02.01.2024.

[47] Sabrina Frei Reto Stauffacher. *Neues Datenschutzgesetz in der Schweiz: Das müssen Unternehmen jetzt beachten*. 2023. URL: `https://www.gryps.ch/news/2023/6/13/webinar-datenschutzgesetz-it-treuhand`. accessed: 26.12.2023.

[48] Santiagogonzalezcuellar. *Hash Functions*. 2023. URL: `https://medium.com/@santiagogonzalezcuellar/hash-functions-33043182ddc2`. accessed: 01.01.2024.

[49] Olenka Van Schendel. *Data masking: Anonymisation or pseudonymisation?* 2000. URL: `https://www.grcworldforums.com/data-management/data-masking-anonymisation-or-pseudonymisation/12.article`. accessed: 26.12.2023.

[50] Gaurav Y. Tawde and Jayashree M. Kundargi. "An Overview of Feature Extraction Techniques in OCR for Indian Scripts Focused on Offline Handwriting". In: *International Journal of Engineering Research and Applications (IJERA)* 3.1 (2013).

[51] *Tesseract User Manual*. https://tesseract-ocr.github.io/tessdoc/#introduction. accessed: 05.01.2024.

[52] Amit Timalsina. *Analysis and Benchmarking of OCR Accuracy for Data Extraction Models*. 2023. URL: `https://www.docsumo.com/blog/ocr-accuracy#:`. accessed: 11.01.2024.

[53] *Transformer's Encoder-Decoder*. 2021. URL: `https://kikaben.com/transformers-encoder-decoder/`. accessed: 04.01.2024.

[54] *Transformer's Self-Attention*. 2021. URL: `https://kikaben.com/transformers-self-attention/`. accessed: 04.01.2024.

[55] European Union. *Opinion 05/2014 on Anonymisation Techniques*. 2014. URL: `https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216%5C_en.pdf`. accessed: 29.12.2023.

[56] University of Neuchâtel Università della Svizzera italiana. *Data anonymization*. URL: `https://www.epfl.ch/campus/services/data-protection/in-practice/privacy-in-research/data-anonymization/`. accessed: 28.12.2023.

[57] UUIDTools. *UUID Versions Explained*. URL: `https://www.uuidtools.com/uuid-versions-explained`. accessed: 01.01.2024.

[58] Huy Tran Uwe Zdun Rafael Capill and Olaf Zimmermann. *Sustainable Architectural Design Decisions*. 2014. URL: `https://www.infoq.com/articles/sustainable-architectural-design-decisions/`. accessed: 08.01.2024.

[59] Math Works. *Types of Morphological Operations*. URL: `https://www.mathworks.com/help/images/morphological-dilation-and-erosion.html`. accessed: 02.01.2024.

[60] Yu-Jin Zhang. *Image Acquisition Devices*. 2021. URL: `https://link.springer.com/chapter/10.1007/978-981-15-5873-3%5C_3`. accessed: 02.01.2024.

CHAPTER 17

Mockup of the Image to Text Anonymiser Frontend

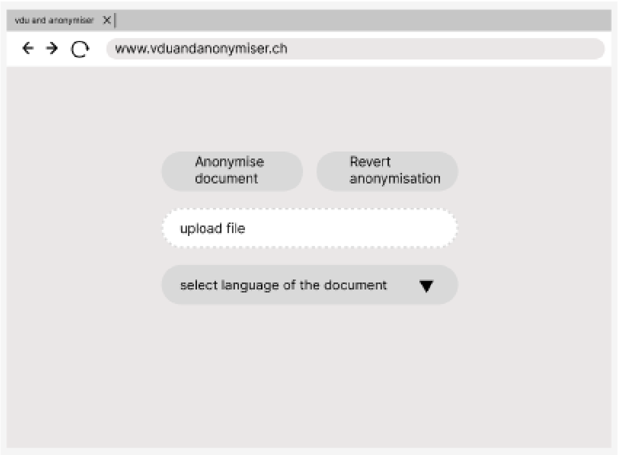
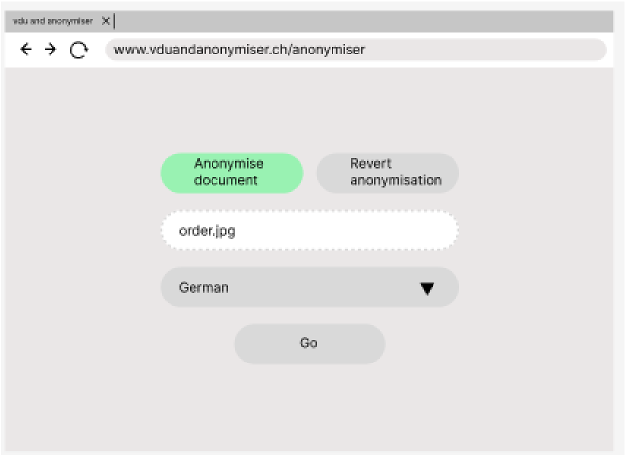


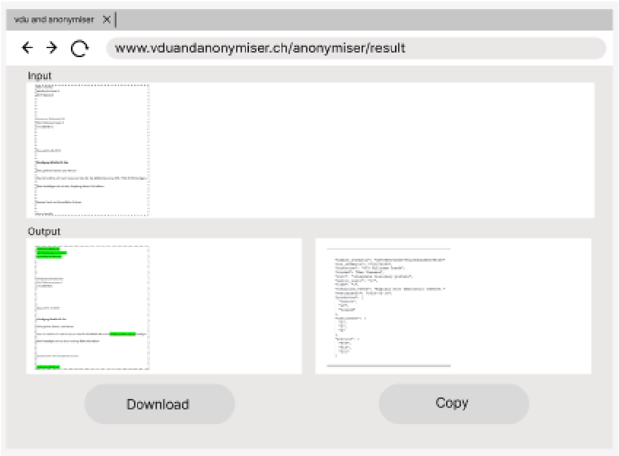
Figure 17.1: Mockup landing page



order.jpg

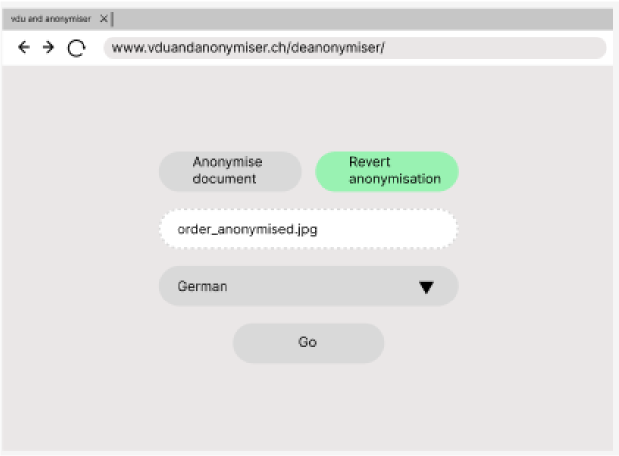
German

Figure 17.2: Mockup document upload



www.vduandanonymiser.ch/anonymiser/result

Figure 17.3: Mockup result page

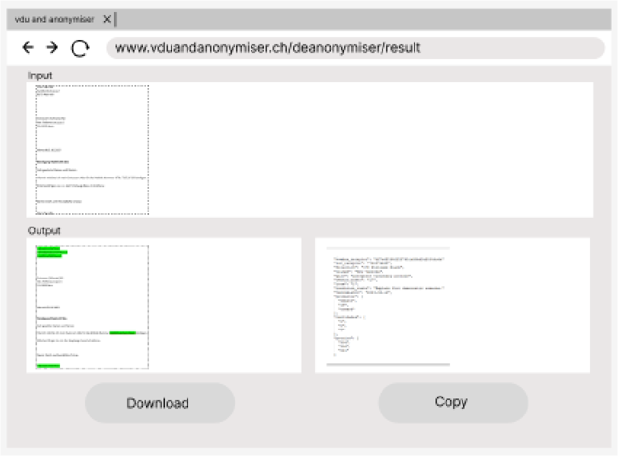


www.vduandanonymiser.ch/deanonymiser/

order_anonymised.jpg

German ▼

Figure 17.4: Mockup upload of pseudonymised document



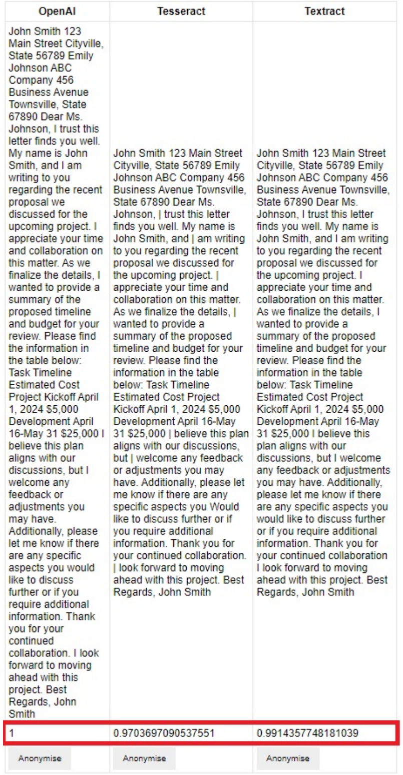
www.vduandanonymiser.ch/deanonymiser/result

Figure 17.5: Mockup result pseudonymisation reverted

CHAPTER 18

Test Documents

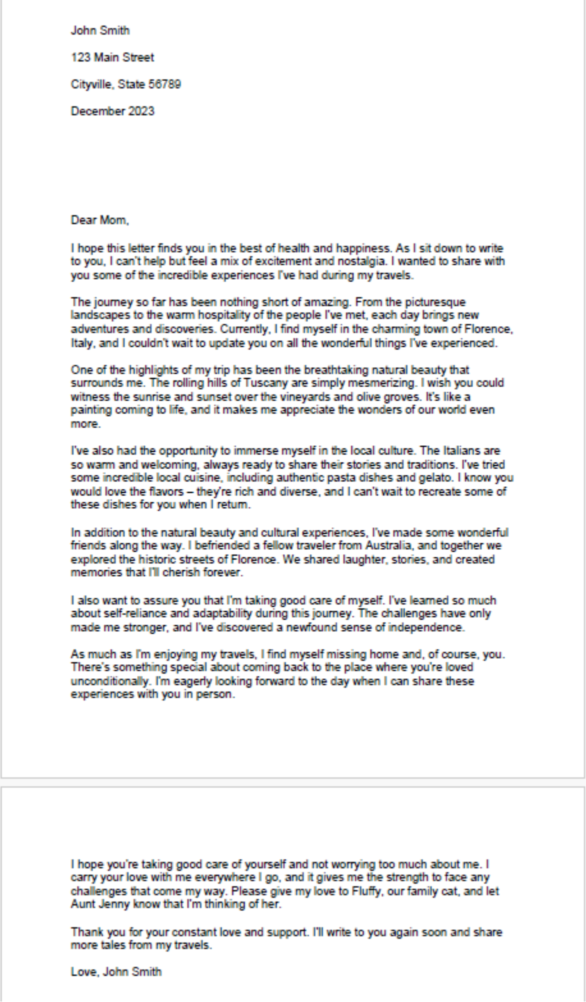
18.1 OCR Accuracy of the Three Engines



OpenAI	Tesseract	Textract
John Smith 123 Main Street Cityville, State 56789 Emily Johnson ABC Company 456 Business Avenue Townsville, State 67890 Dear Ms. Johnson, I trust this letter finds you well. My name is John Smith, and I am writing to you regarding the recent proposal we discussed for the upcoming project. I appreciate your time and collaboration on this matter. As we finalize the details, I wanted to provide a summary of the proposed timeline and budget for your review. Please find the information in the table below: Task Timeline Estimated Cost Project Kickoff April 1, 2024 $5,000 Development April 16-May 31 $25,000 I believe this plan aligns with our discussions, but I welcome any feedback or adjustments you may have. Additionally, please let me know if there are any specific aspects you would like to discuss further or if you require additional information. Thank you for your continued collaboration. I look forward to moving ahead with this project. Best Regards, John Smith	John Smith 123 Main Street Cityville, State 56789 Emily Johnson ABC Company 456 Business Avenue Townsville, State 67890 Dear Ms. Johnson, | trust this letter finds you well. My name is John Smith, and | am writing to you regarding the recent proposal we discussed for the upcoming project. | appreciate your time and collaboration on this matter. As we finalize the details, | wanted to provide a summary of the proposed timeline and budget for your review. Please find the information in the table below: Task Timeline Estimated Cost Project Kickoff April 1, 2024 $5,000 Development April 16-May 31 $25,000 | believe this plan aligns with our discussions, but | welcome any feedback or adjustments you may have. Additionally, please let me know if there are any specific aspects you Would like to discuss further or if you require additional information. Thank you for your continued collaboration. | look forward to moving ahead with this project. Best Regards, John Smith	John Smith 123 Main Street Cityville, State 56789 Emily Johnson ABC Company 456 Business Avenue Townsville, State 67890 Dear Ms. Johnson, I trust this letter finds you well. My name is John Smith, and I am writing to you regarding the recent proposal we discussed for the upcoming project. I appreciate your time and collaboration on this matter. As we finalize the details, I wanted to provide a summary of the proposed timeline and budget for your review. Please find the information in the table below: Task Timeline Estimated Cost Project Kickoff April 1, 2024 $5,000 Development April 16-May 31 $25,000 I believe this plan aligns with our discussions, but I welcome any feedback or adjustments you may have. Additionally, please let me know if there are any specific aspects you would like to discuss further or if you require additional information. Thank you for your continued collaboration I look forward to moving ahead with this project. Best Regards, John Smith
1	0.9703697090537551	0.9914357748181039
Anonymise	Anonymise	Anonymise

Figure 18.1: OCR accuracy provided by the three engines for a document with a table

18.2 Example Upload PDF File With Multiple Pages



John Smith

123 Main Street

Cityville, State 56789

December 2023

Dear Mom,

I hope this letter finds you in the best of health and happiness. As I sit down to write to you, I can't help but feel a mix of excitement and nostalgia. I wanted to share with you some of the incredible experiences I've had during my travels.

The journey so far has been nothing short of amazing. From the picturesque landscapes to the warm hospitality of the people I've met, each day brings new adventures and discoveries. Currently, I find myself in the charming town of Florence, Italy, and I couldn't wait to update you on all the wonderful things I've experienced.

One of the highlights of my trip has been the breathtaking natural beauty that surrounds me. The rolling hills of Tuscany are simply mesmerizing. I wish you could witness the sunrise and sunset over the vineyards and olive groves. It's like a painting coming to life, and it makes me appreciate the wonders of our world even more.

I've also had the opportunity to immerse myself in the local culture. The Italians are so warm and welcoming, always ready to share their stories and traditions. I've tried some incredible local cuisine, including authentic pasta dishes and gelato. I know you would love the flavors – they're rich and diverse, and I can't wait to recreate some of these dishes for you when I return.

In addition to the natural beauty and cultural experiences, I've made some wonderful friends along the way. I befriended a fellow traveler from Australia, and together we explored the historic streets of Florence. We shared laughter, stories, and created memories that I'll cherish forever.

I also want to assure you that I'm taking good care of myself. I've learned so much about self-reliance and adaptability during this journey. The challenges have only made me stronger, and I've discovered a newfound sense of independence.

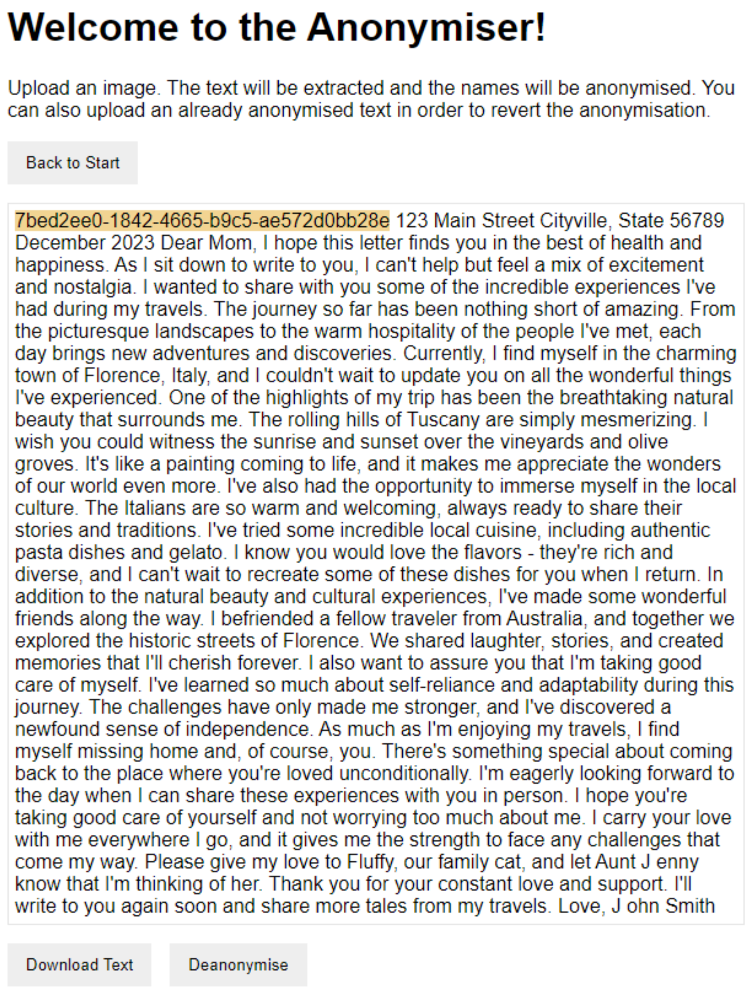
As much as I'm enjoying my travels, I find myself missing home and, of course, you. There's something special about coming back to the place where you're loved unconditionally. I'm eagerly looking forward to the day when I can share these experiences with you in person.

I hope you're taking good care of yourself and not worrying too much about me. I carry your love with me everywhere I go, and it gives me the strength to face any challenges that come my way. Please give my love to Fluffy, our family cat, and let Aunt Jenny know that I'm thinking of her.

Thank you for your constant love and support. I'll write to you again soon and share more tales from my travels.

Love, John Smith

Figure 18.2: PDF file with two pages



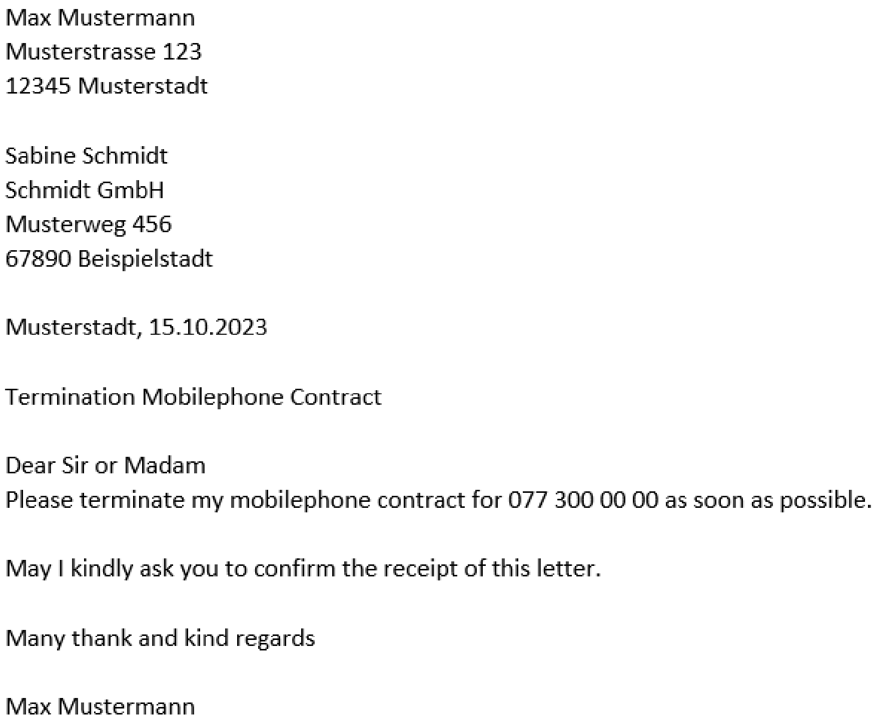
Welcome to the Anonymiser!

Upload an image. The text will be extracted and the names will be anonymised. You can also upload an already anonymised text in order to revert the anonymisation.

7bed2ee0-1842-4665-b9c5-ae572d0bb28e 123 Main Street Cityville, State 56789 December 2023 Dear Mom, I hope this letter finds you in the best of health and happiness. As I sit down to write to you, I can't help but feel a mix of excitement and nostalgia. I wanted to share with you some of the incredible experiences I've had during my travels. The journey so far has been nothing short of amazing. From the picturesque landscapes to the warm hospitality of the people I've met, each day brings new adventures and discoveries. Currently, I find myself in the charming town of Florence, Italy, and I couldn't wait to update you on all the wonderful things I've experienced. One of the highlights of my trip has been the breathtaking natural beauty that surrounds me. The rolling hills of Tuscany are simply mesmerizing. I wish you could witness the sunrise and sunset over the vineyards and olive groves. It's like a painting coming to life, and it makes me appreciate the wonders of our world even more. I've also had the opportunity to immerse myself in the local culture. The Italians are so warm and welcoming, always ready to share their stories and traditions. I've tried some incredible local cuisine, including authentic pasta dishes and gelato. I know you would love the flavors - they're rich and diverse, and I can't wait to recreate some of these dishes for you when I return. In addition to the natural beauty and cultural experiences, I've made some wonderful friends along the way. I befriended a fellow traveler from Australia, and together we explored the historic streets of Florence. We shared laughter, stories, and created memories that I'll cherish forever. I also want to assure you that I'm taking good care of myself. I've learned so much about self-reliance and adaptability during this journey. The challenges have only made me stronger, and I've discovered a newfound sense of independence. As much as I'm enjoying my travels, I find myself missing home and, of course, you. There's something special about coming back to the place where you're loved unconditionally. I'm eagerly looking forward to the day when I can share these experiences with you in person. I hope you're taking good care of yourself and not worrying too much about me. I carry your love with me everywhere I go, and it gives me the strength to face any challenges that come my way. Please give my love to Fluffy, our family cat, and let Aunt J enny know that I'm thinking of her. Thank you for your constant love and support. I'll write to you again soon and share more tales from my travels. Love, J ohn Smith

Figure 18.3: Anonymiser app provides a result if multiple pages are uploaded

18.3 Simple Testcase for End-to-End Findings



Max Mustermann
Musterstrasse 123
12345 Musterstadt

Sabine Schmidt
Schmidt GmbH
Musterweg 456
67890 Beispielstadt

Musterstadt, 15.10.2023

Termination Mobilephone Contract

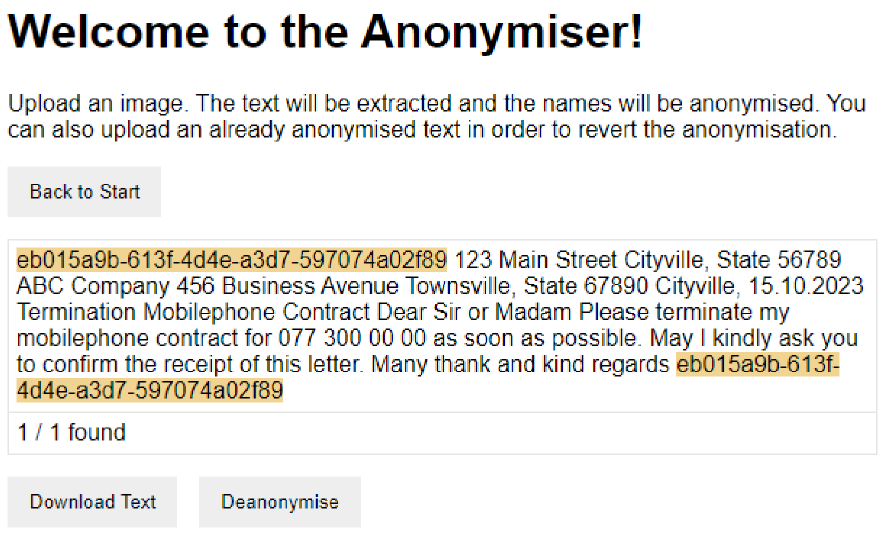
Dear Sir or Madam
Please terminate my mobilephone contract for 077 300 00 00 as soon as possible.

May I kindly ask you to confirm the receipt of this letter.

Many thank and kind regards

Max Mustermann

Figure 18.4: Simple test document



Welcome to the Anonymiser!

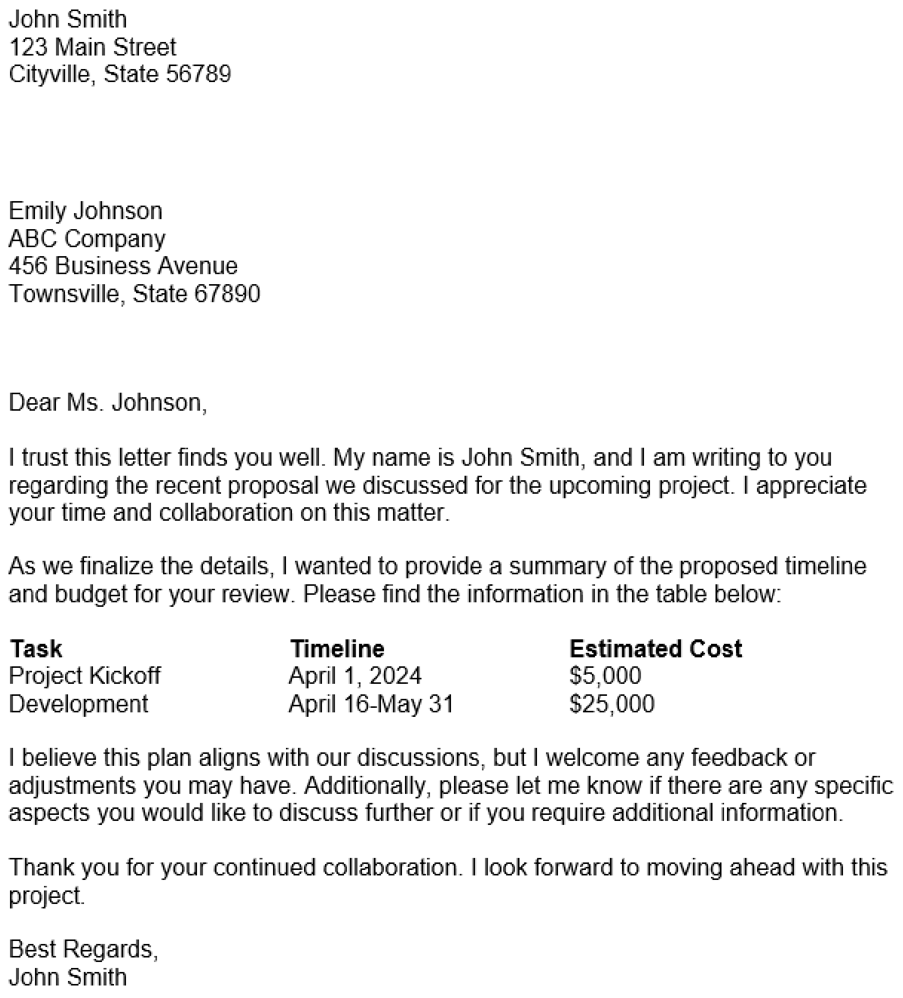
Upload an image. The text will be extracted and the names will be anonymised. You can also upload an already anonymised text in order to revert the anonymisation.

eb015a9b-613f-4d4e-a3d7-597074a02f89 123 Main Street Cityville, State 56789 ABC Company 456 Business Avenue Townsville, State 67890 Cityville, 15.10.2023 Termination Mobilephone Contract Dear Sir or Madam Please terminate my mobilephone contract for 077 300 00 00 as soon as possible. May I kindly ask you to confirm the receipt of this letter. Many thank and kind regards eb015a9b-613f-4d4e-a3d7-597074a02f89

1 / 1 found

Figure 18.5: Simple test result

18.4 Medium Difficult Testcase for End-to-End Findings



John Smith
123 Main Street
Cityville, State 56789

Emily Johnson
ABC Company
456 Business Avenue
Townsville, State 67890

Dear Ms. Johnson,

I trust this letter finds you well. My name is John Smith, and I am writing to you regarding the recent proposal we discussed for the upcoming project. I appreciate your time and collaboration on this matter.

As we finalize the details, I wanted to provide a summary of the proposed timeline and budget for your review. Please find the information in the table below:

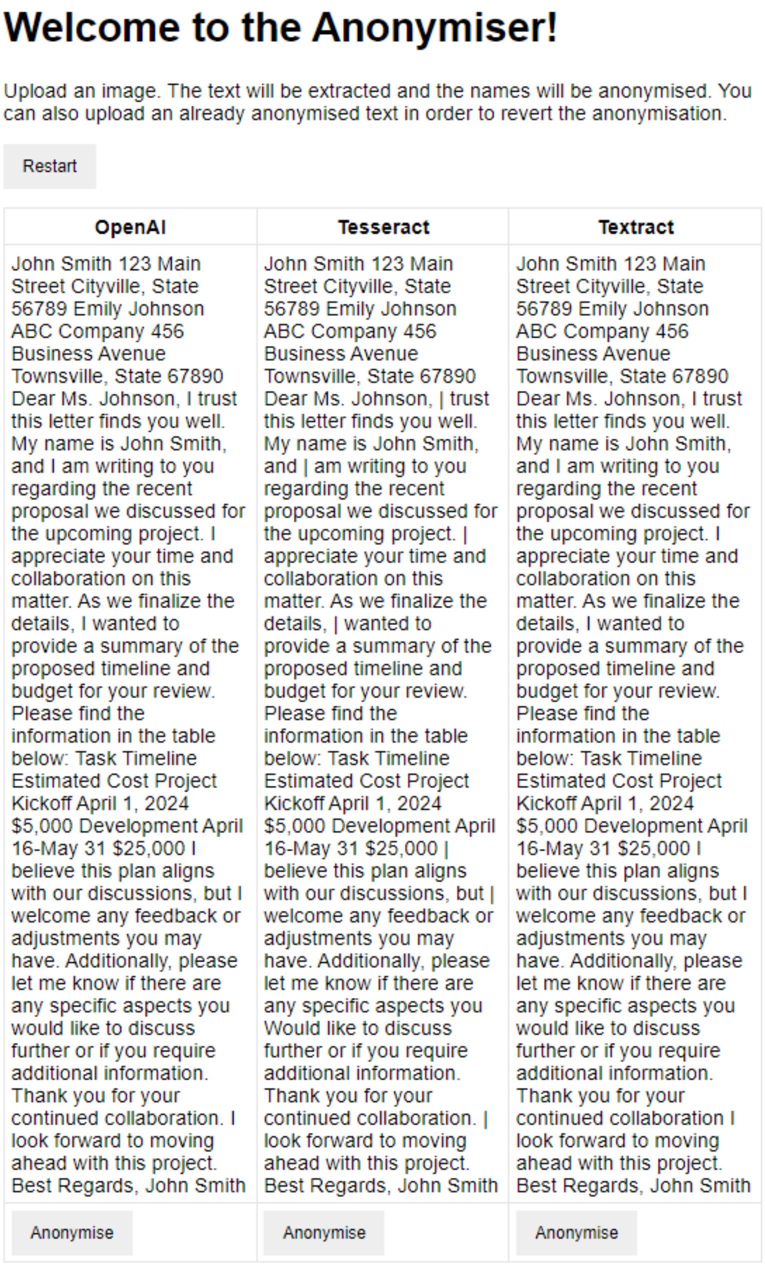
Task	**Timeline**	**Estimated Cost**
Project Kickoff	April 1, 2024	$5,000
Development	April 16-May 31	$25,000

I believe this plan aligns with our discussions, but I welcome any feedback or adjustments you may have. Additionally, please let me know if there are any specific aspects you would like to discuss further or if you require additional information.

Thank you for your continued collaboration. I look forward to moving ahead with this project.

Best Regards,
John Smith

Figure 18.6: Medium difficult test document

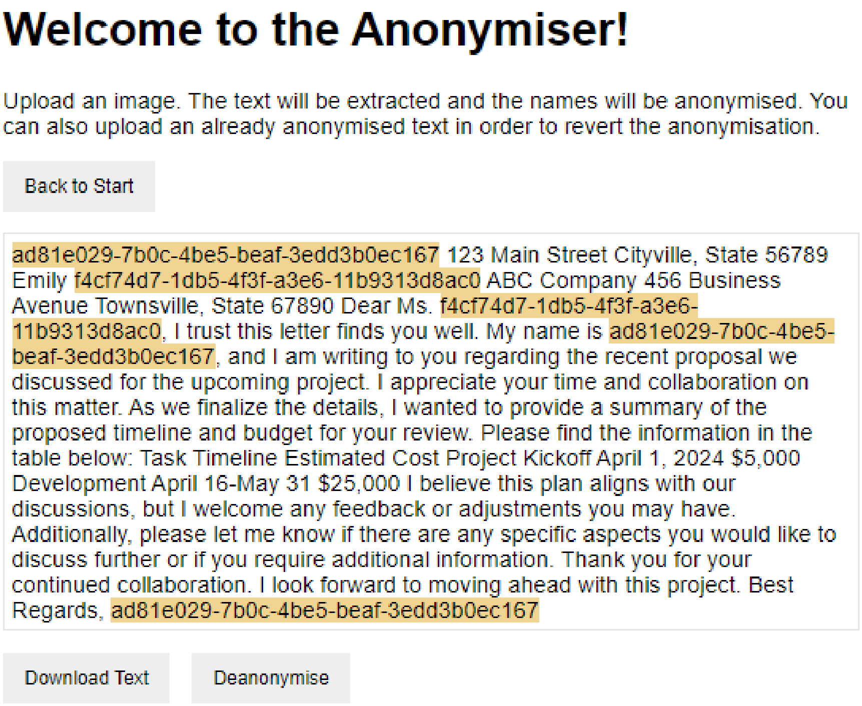


Welcome to the Anonymiser!

Upload an image. The text will be extracted and the names will be anonymised. You can also upload an already anonymised text in order to revert the anonymisation.

<table>
<tr><th>OpenAI</th><th>Tesseract</th><th>Textract</th></tr>
<tr><td>John Smith 123 Main Street Cityville, State 56789 Emily Johnson ABC Company 456 Business Avenue Townsville, State 67890 Dear Ms. Johnson, I trust this letter finds you well. My name is John Smith, and I am writing to you regarding the recent proposal we discussed for the upcoming project. I appreciate your time and collaboration on this matter. As we finalize the details, I wanted to provide a summary of the proposed timeline and budget for your review. Please find the information in the table below: Task Timeline Estimated Cost Project Kickoff April 1, 2024 $5,000 Development April 16-May 31 $25,000 I believe this plan aligns with our discussions, but I welcome any feedback or adjustments you may have. Additionally, please let me know if there are any specific aspects you would like to discuss further or if you require additional information. Thank you for your continued collaboration. I look forward to moving ahead with this project. Best Regards, John Smith</td><td>John Smith 123 Main Street Cityville, State 56789 Emily Johnson ABC Company 456 Business Avenue Townsville, State 67890 Dear Ms. Johnson, | trust this letter finds you well. My name is John Smith, and | am writing to you regarding the recent proposal we discussed for the upcoming project. | appreciate your time and collaboration on this matter. As we finalize the details, | wanted to provide a summary of the proposed timeline and budget for your review. Please find the information in the table below: Task Timeline Estimated Cost Project Kickoff April 1, 2024 $5,000 Development April 16-May 31 $25,000 | believe this plan aligns with our discussions, but | welcome any feedback or adjustments you may have. Additionally, please let me know if there are any specific aspects you Would like to discuss further or if you require additional information. Thank you for your continued collaboration. | look forward to moving ahead with this project. Best Regards, John Smith</td><td>John Smith 123 Main Street Cityville, State 56789 Emily Johnson ABC Company 456 Business Avenue Townsville, State 67890 Dear Ms. Johnson, I trust this letter finds you well. My name is John Smith, and I am writing to you regarding the recent proposal we discussed for the upcoming project. I appreciate your time and collaboration on this matter. As we finalize the details, I wanted to provide a summary of the proposed timeline and budget for your review. Please find the information in the table below: Task Timeline Estimated Cost Project Kickoff April 1, 2024 $5,000 Development April 16-May 31 $25,000 I believe this plan aligns with our discussions, but I welcome any feedback or adjustments you may have. Additionally, please let me know if there are any specific aspects you would like to discuss further or if you require additional information. Thank you for your continued collaboration I look forward to moving ahead with this project. Best Regards, John Smith</td></tr>
<tr><td>Anonymise</td><td>Anonymise</td><td>Anonymise</td></tr>
</table>

Figure 18.7: Result medium difficult test



Welcome to the Anonymiser!
Upload an image. The text will be extracted and the names will be anonymised. You can also upload an already anonymised text in order to revert the anonymisation.
Back to Start
Download Text
Deanonymise

ad81e029-7b0c-4be5-beaf-3edd3b0ec167 123 Main Street Cityville, State 56789 Emily f4cf74d7-1db5-4f3f-a3e6-11b9313d8ac0 ABC Company 456 Business Avenue Townsville, State 67890 Dear Ms. f4cf74d7-1db5-4f3f-a3e6-11b9313d8ac0, I trust this letter finds you well. My name is ad81e029-7b0c-4be5-beaf-3edd3b0ec167, and I am writing to you regarding the recent proposal we discussed for the upcoming project. I appreciate your time and collaboration on this matter. As we finalize the details, I wanted to provide a summary of the proposed timeline and budget for your review. Please find the information in the table below: Task Timeline Estimated Cost Project Kickoff April 1, 2024 $5,000 Development April 16-May 31 $25,000 I believe this plan aligns with our discussions, but I welcome any feedback or adjustments you may have. Additionally, please let me know if there are any specific aspects you would like to discuss further or if you require additional information. Thank you for your continued collaboration. I look forward to moving ahead with this project. Best Regards, ad81e029-7b0c-4be5-beaf-3edd3b0ec167

Figure 18.8: Result anonymisation medium diffiult test

18.5 Difficult Testcase for End-to-End findings