VIETNAM NATIONAL UNIVERSITY OF HOCHIMINH CITY
THE INTERNATIONAL UNIVERSITY
SCHOOL OF COMPUTER SCIENCE AND ENGINEERING

AUTOMATIC DOCUMENT SEGMENTATION AND SUMMARIZATION USING NLP-BASED METHODS

Student name: Lương Trí Vỹ (ITITIU20359)
Advisor: Dr. Nguyễn Thị Thanh Sang

A thesis submitted to the School of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
Bachelor of Computer Science

Ho Chi Minh City, Vietnam
January 2024

AUTOMATIC DOCUMENT SEGMENTATION AND SUMMARIZATION USING NLP-BASED METHOD

APPROVED BY:

______________________________ ,

Dr. Nguyễn Thị Thanh Sang

____

(*Typed Committee name here*)

____

(*Typed Committee name here*)

____

(*Typed Committee name here*)

____

(*Typed Committee name here*)

THESIS COMMITTEE

ACKNOWLEGMENTS

I would like to express my deepest gratitude towards Dr. Nguyễn Thị Thanh Sang for her support as the advisor throughout the thesis. During the making of the project, her guidance helped guide me toward the appropriate direction and not being lost throughout the researching process.

Additionally, it is with deep appreciation that I would like to acknowledge my family's support during this project. Without my family's presence, I would have lacked the necessary motivation to further develop and complete this project.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

In today's exceedingly information-rich world, the efficiency in which humans can absorb, consume, and integrate information into their own repertoire of knowledge is dependent upon one's reading speed and their own understanding of said information. As a result, one solution to said problem is for text to be summarized and information to be condensed into forms that can be quickly absorbed briefly. For the solution to be viable, the conversion of a PDF file of a research paper must be successfully converted into a document containing the summary of each segment of the paper. The main structure of the project, therefore, can be separated into 2 different sections: PDF Segmentation and text summarization of each segment.

Although information from a PDF can easily be extracted using tools such as PyPDF2, the structure of each paper remains an issue due to the wide variety in which research papers are formatted. As such, traditional methods cannot be relied upon but require the usage of machine learning models.

The process of text summarization is more varied as many approaches are available. In general, text summarization can be achieved either through extractive or abstractive methods. An extractive method can only generate a summary using the provided text. On the other hand, abstractive method's output is a summary that is made using text that is not present in the document itself.

This thesis's aim is to test multiple different approaches to text summary and determine the most suitable approach and a final product that can demonstrate said result.

Chapter 1 INTRODUCTION

1.1 Background

As the speed at which information is generated in the age of digitalization, the speed of information consumption and digestion must also be improved through the usage of tools and techniques. As such, the development of a tool that can help with that issue is the focus of the thesis.

1.2 Problem Statement

Due to the amount of information that each research paper holds are tremendous, the potential and possibilities to which a research paper can make an impact on the world can be lost. With topics ranging from Physics, Medicine, Psychology, etc., potential loss of valuable insights that could be gleamed from research papers or papers that are overlooked is immense.

Although methods to improve reading speed, such as Speed reading, Skimming, Active reading, etc., exist, all these methods aim to improve the person that is acquiring information. Even though this method is viable and can improve efficiency, much time is required to be spent for one person to acquire said skill. Human beings, innately, improve their existence and society through the development of tools. As such, developing a tool that can be used to assist said endeavor is this project's main purpose.

1.3 Scope and Objectives

For the thesis to be considered completed, the development of a tool that can be used to parse a PDF file, turn it into a document of different segments, then automatically summarize the information must be completed and presented as the result of the thesis.

1.4 Structure of thesis

The main structure of the paper can be divided into 6 different chapters, with each chapter having its own purpose in providing full coverage of the topic. Chapter 1 serves as the paper's introduction as the problem and aim of the thesis is introduced. Chapter 2 gives an overview of the information and knowledge used in the paper. Moving onto chapter 3, the methodologies and techniques used for the implementation are discussed in detail. The results of the project are shown and displayed in chapter 4. Said results are then evaluated and compared based on predefined measurements to determine the effectiveness of the model itself. Lastly, chapter 6 is the summary of the project, including the strengths, weaknesses and improvements that could be made to the thesis.

Chapter 2 LITURATURE REVIEW

2.1 Overview of Automatic Document Segmentation

The general goal of the Automatic Documentation Segmentation (ADS) is to scan through a document (in the form of a PDF file), identify each segment of the document, then output the result into a well-defined form for further use cases.

As most scientific texts in online databases, which are well-defined and documented, are in the form of a PDF file, this has become the priority in terms of input format. This choice is further enhanced due to an estimate of around 50 million research papers [1] that are stored as a PDF file in 2008. Since then, the number of research papers has only grown ceaselessly, further enhancing the decision of using PDFs as inputs.

Segments of a document referred to each part of a paper that is frequently required to be a part of any research paper, such as Abstract, Introduction, Methodologies, etc. However, the number of segments that are presented in the paper should not have any severe effects on the functionalities of the tool itself. As the aim of the application is not to adhere to a certain principle or guideline regarding the writing style, but to effectively summarize a paper, all segments of a paper are treated as equal, and can equally influence the reader.

The act of identifying each segment of the document referred to the process of acquiring the texts that are available in the paper, classifying it accordingly based on its segment, and formatting and storing it into a predefined form of storage. The process of text acquisition can be achieved through using multiple different methods. It can be as simple as extracting the text from the PDF as is without any regards for the structure of the paper, by making use of tools such as the PyPDF2 package in Python. The structure of a paper can be difficult to acquire due to the variety of methods and standards in which a paper is written. For example, whether the content is divided into two columns or one, whether the spacing on each line is 1.5 or Double, all aspects of a PDF can influence the way that the text is perceived by the machine. Currently, two methods in which the structure can be parsed are using Optical Character Recognition (OCR) and classification of content into predefined categories using Machine learning. While the first method is viable, its output is highly dependent on how the document is rendered. It is, therefore, the second method that will be the focus of the paper. To achieve this, **G**ene**R**ation **O**f **B**ibliographic **D**ata (or GROBID) is the tool of choice according to the criteria that has been set. Since GROBID integrates with a third-party file conversion tool called **pdftoxml** to convert file from PDF to XML, XML is chosen as the form of storage for the output of the extraction process.

2.1.1 GROBID

GROBID [2] makes use of a cascade of sequence labelling models to parse its documents. This means that a document must go through multiple different models, each specializing in a specific task, to achieve the result. In the case of extracting the segments of the entire document, the model must first go through each layer and get the result, then pass said result to the next layer. Each of these layers are then combined and formatted accordingly. The GROBID cascade of sequencing model can be seen from Figure 2.1.1-1.

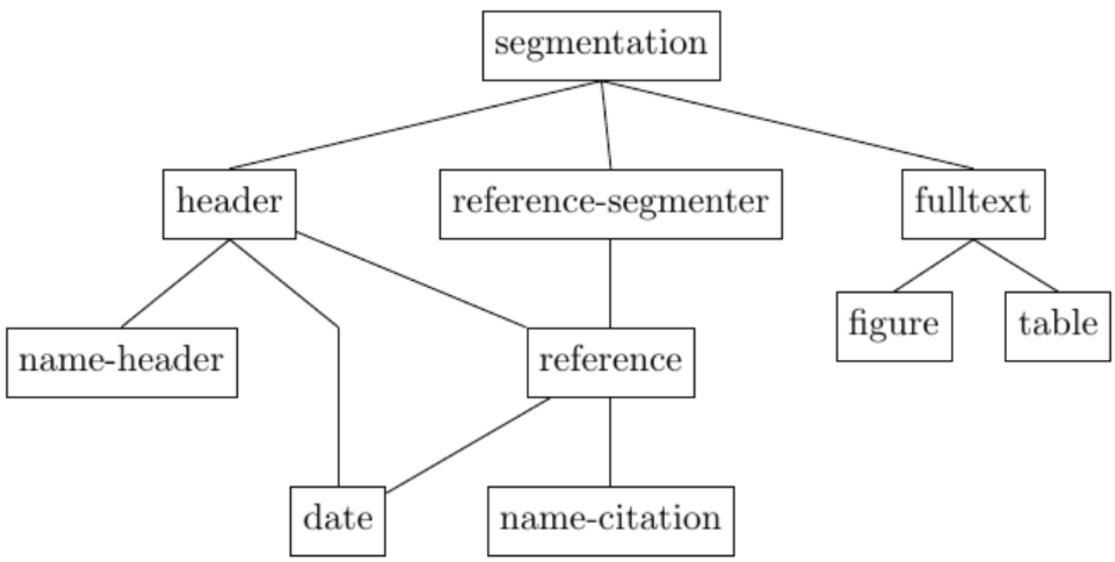

Figure 2-1 - GROBID cascade of sequencing model [3]

Although each model can create output from the original text, some models' purpose is only to provide the lower layer with input, creating a *cascade* of information. As a result, the accuracy of the model is highly dependent on the combination of multiple models. However, any errors that result from the higher layer will have a negative effect on the final output.

Although GROBID's implementation support the use of both a Deep Learning model and a *Conditional Random Field (CRF)* model, for the purpose of segmenting a scientific PDF paper, only the CRF model can be utilized as the number of parameters that are needed for processing the entire document exceeds the limit of what the deep learning model is currently capable of.

Although the CRF model is lacking in terms of accuracy in comparison to the Deep Learning model, it makes up for it through the speed with which it processes each PDF.

2.2 Overview of Natural Language Processing

Natural Language Processing (NLP) is an interdisciplinary field that focuses on the process of comprehending and manipulating human language. The general goal is to achieve better understanding of natural language using simple and durable techniques for fast processing of text.

Although differences exist in the function and usage of NLP between different use cases, he common Natural Language process typically include seven steps, which are Sentence segmentation, Word tokenization, Stemming, Lemmatization, Stop word analysis, Dependency parsing, and Part-of-speech tagging.

The first step of the process is to retrieve the list of sentences from the provided text through the process known as *sentence segmentation*. In most cases, a sentence ends with a dot, which can make processing this step simple as that is the most common separator of a sentence. Only sentences that end with an “e.g.”, “?”, and “etc.” should be taken note of as the that can happen occasionally.

After sentence segmentation, *word tokenization* is the next step. This step’s main purpose is to segment the sentences into separate words. These words are separated by white space, comma, dash, dot, etc. This process can be achieved simply through storing each word of a sentence into a data structure, then proceeding to filter out unwanted whitespace word.

Stemming is the process of stripping the words of its prefixes and suffixes, allowing the algorithm to obtain the basic form of a word. As an example, through stemming the word “Improvement”, “Improv” is the result of this step. Multiple different words can result in the same result through the stemming process as can be seen when applying this process to both “Improvement” and “Improvise”. An implementation of this process can be implemented through the Porter Stemmer algorithm, the Lancaster Stemmer algorithm, etc. The output of the process, however, does not bear resemblance to an English word.

Input from the stemming function is then piped into the *Lemmatization* function. This process made improvements upon the stemming process by turning it into a *lemma*, which is the canonical form of a word. Instead of “Improv”, which is not a proper word, “Improve” is returned instead.

After the word list has been refined and reduced to contain a list of lemmas, words that are not considered important to the linguistic analysis process are removed from the list through the process of *stop-word filtering*. Words such as “a”, “the”, “an”, etc., appear frequently throughout the content of a document, and are considered as *stop words* since they do not contribute any additional meaning or context to the document. Filtering these values can improve the accuracy of the overall process.

Through filtering stop words, the word list can now be considered clean, as it contains only lemmas that influence the meaning of the document and can be used for the *dependency parsing process*. During the process, the structure of a sentence can be broken down into its basic form, from Sentences to Noun phrase and Verb phrase, to Verb and Determiner, etc. The sentence is parsed, creating a tree like structure with each lower layer being more specific than the one above it.

Part-of-speech tagging improves upon the dependency parsing functionality by assigning a part-of-speech label to each sequence of words [4]. Tags can be divided into *closed class* words, which are frequent and ambiguous, and *open-class* words, which are divided differently depending on the set of tags that is being used. Typically, a tag set can include tags such as Nouns, Verbs, adjectives, etc. The result from both dependency parsing and part-of-speech tagging can be seen from the Figure 2.2-1

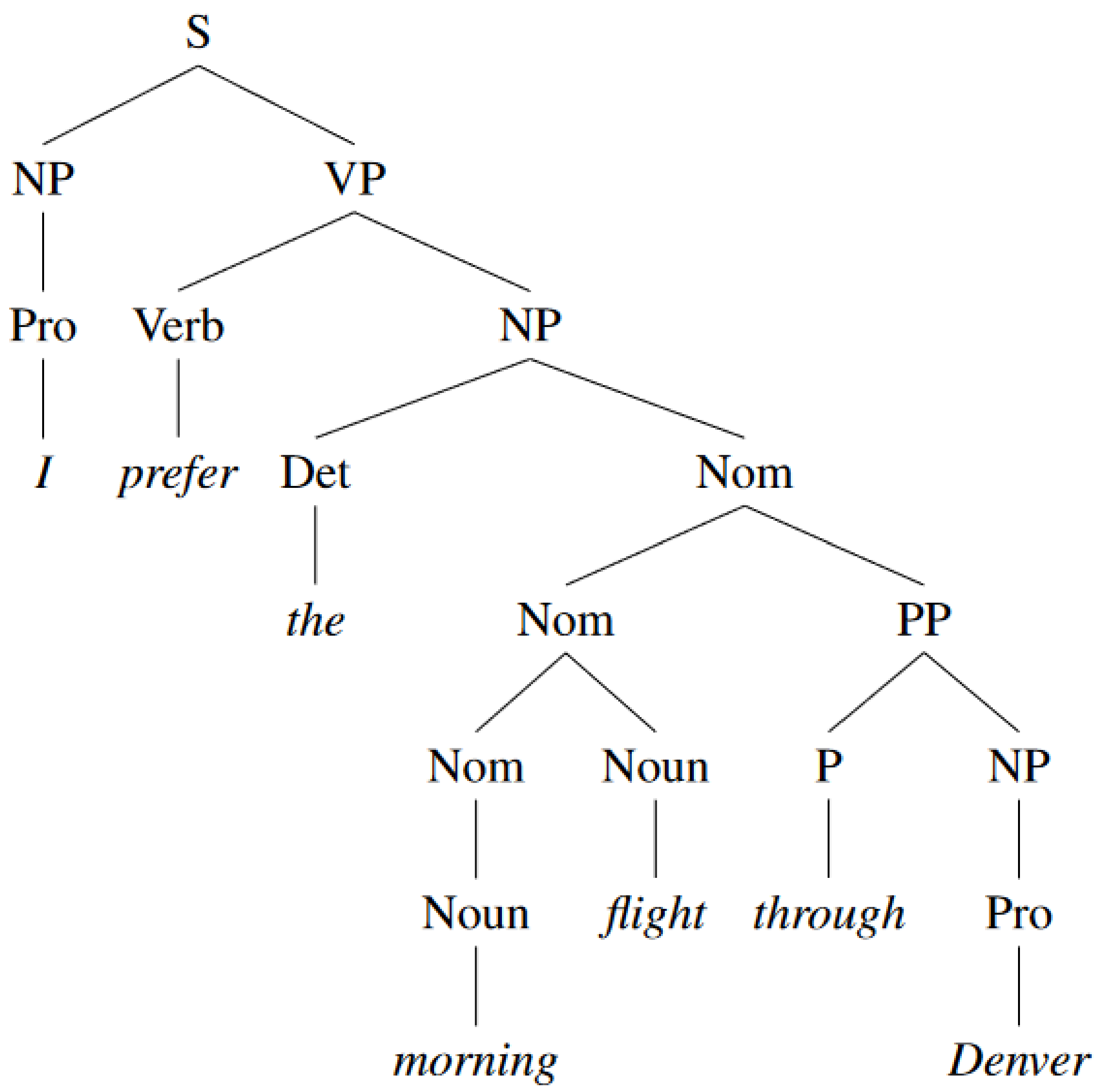

Figure 2-2 - Dependency style parse tree [4]

2.3 Overview of Text Summarization

Text summarization, as the name suggests, is the process of parsing a single or multiple documents and constructing a summary for each respective document. Applications of this system are widely available, from summarization of news, opinion, sentiment, story and novels to scientific paper summaries in all fields.

The general framework of the system involves a pre-processing step, a processing step and a post-processing step with the final step being optional in some cases. Text summarization system can be classified based on many of its aspects including input size (single-document or multi-document), approach (extractive, subtractive, or hybrid), nature of output summary (generic-based or query-based), language (monolingual, multilingual, or cross-lingual), type (headline, sentence-level, highlights, or full summary), domain (generic or domain specific)[5]. The framework's visualization can be seen in Figure 2.3-1

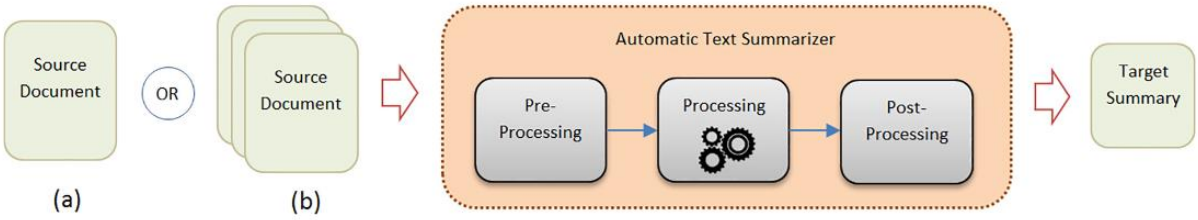

Figure 2.3-1 – Text summarization general framework [5]

2.3.1 Extractive text summarization

The *extractive text summarization* approach selects the most important sentence using one approach amongst many to create the output. Among the available methods, approaches that have proven to produce positive results include statistical-based methods, concept-based methods, graph-based methods, deep-learning methods, etc.

Due to only reusing words from the original text, some advantages that this approach may provide include having faster response time in comparison with abstractive text summarization techniques and having higher accuracy in its output. However, for the same reason, some disadvantages of the techniques were created. The resulting text can contain redundancy due to having two closely-related-in-meaning sentences included. Chosen sentences may also be too lengthy to serve as a summary or the chosen sentence may lack the required context that is necessary for the reader to fully comprehend the summary. Additionally, sentences with different temporal status may conflict with each other when included in the same paragraph. Some notable extractive methods include TextRank (which made use of a graph-based system to determine the importance of words) [6], and Latent Semantic Analysis (LSA) (which made use of Single Value Decomposition technique to analyze the relationship between documents) [7]

The general framework of a text summarization system in the previous section can then be updated to include general steps in an extractive system. These changes can be seen in Figure 2.3.1-1

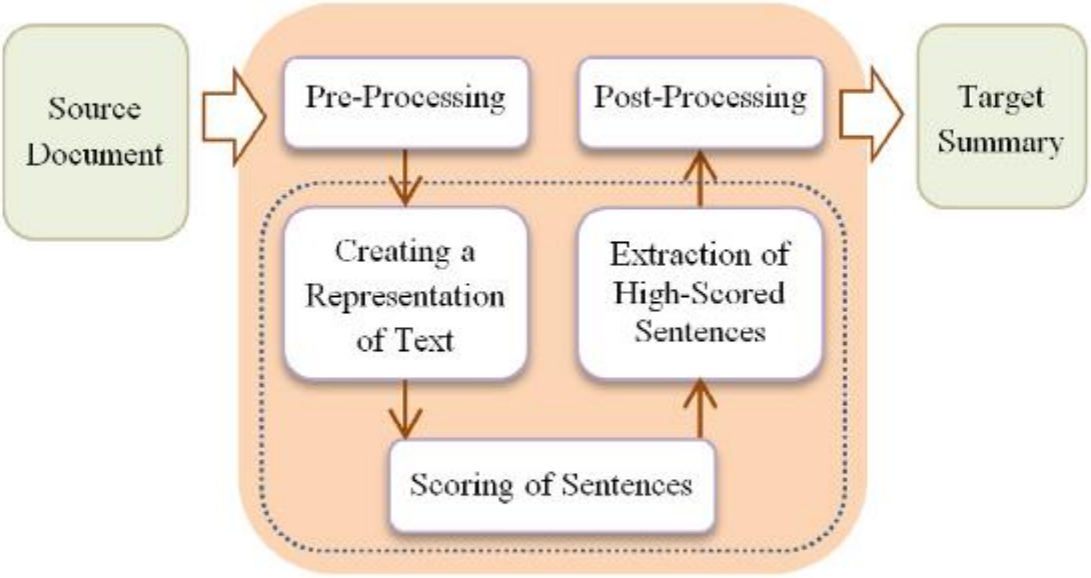

Figure 2-3 - Extractive text summarization general framework [5]

2.3.2 Abstractive text summarization

Abstractive text summarization differs from the extractive method in that before the input is processed, it is turned into one of the intermediate forms that will then be the input of the summarization process. Similar to the extractive text summarization, multiple different approaches can be used to the summary generation process, ranging from graph-based, tree-based, template-based, machine-learning-based, etc.

An abstractive text summarization system can generate greater summary with words that do not belong to the original text. For a summary generation system, the ability to create new sentences defines what an abstractive text summary is. The output of this type of system is generally more similar to summary created by humans in comparison to the extractive type. However, due to the highly advanced and complex nature of this system, creating one is typically difficult and can require the use of natural language generation or machine learning, which is still a growing field.

A new version of the general framework containing the architecture of an abstractive text summarization can then be visualized, resulting in Figure 2.3.2-1

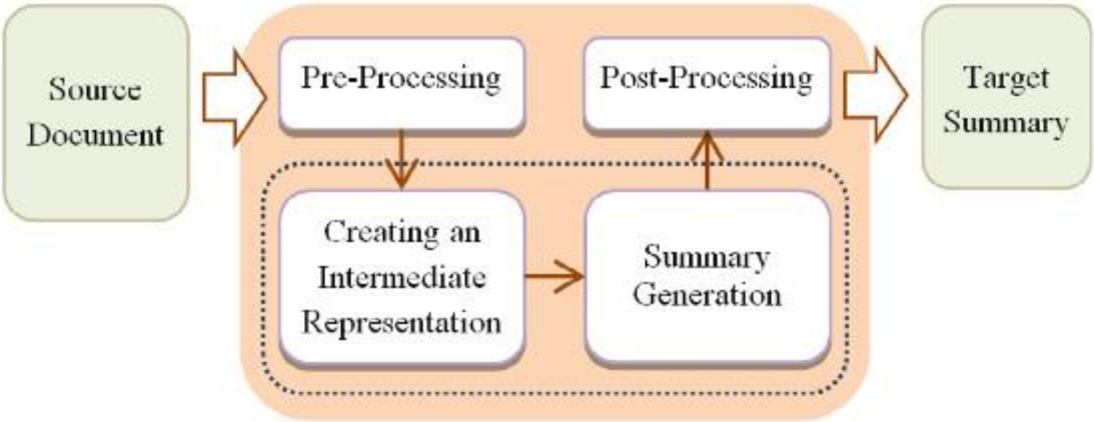

Figure 2-4 - Abstractive text summarization general framework [5]

2.3.2.1 Machine-learning-based Summary Generation

Due to the recent advancements in the progress of improving Artificial Intelligence (AI), techniques that make use of AI models have increasingly become both more effective and efficient at a rapid rate. Models such as **B**idirectional and **A**uto-**R**egressive **T**ransformers (**BART**) [8], or **T**ext-**t**o-**T**ext **T**ransfer **T**ransformer (**T5**) [9] have continuously improve the process of text summarizations.

Following the development of pretrained models, utilizing trained and optimized models has become trivial due to the reduction in training time required for these models. This is especially relevant in cases where resources are limited such as laptops or lightweight servers. The primary method in which this is achieved is through a method known as transferred learning [10]. Inspired by how humans acquire new knowledge by applying and adapting old one, AI model are trained in a two-phase framework [11], with the first being pre-training, allowing for the model to obtain knowledge from large labeled datasets, and the second being finetuning which allow a model to adjust to a specific task with less data required.

Amongst the models, three, which are possible to be run in a local environment efficiently, were chosen for this purpose. This includes the Falcon AI's text-summarization model [12] (a T5-based model), the Facebook's bart-large-cnn model [13] (a BART-based model). Even though another model which is not suitable for this use case, a relatively large model was chosen to for the benchmark, which was the Stability AI's stablelm-2-1_6b model.

Chapter 3 METHODOLOGY

3.1 Overview

Although this project does have exposed APIs and a minimal web user interface, the application only lacks refinement that can effectively serve its purpose. Currently, most core sections of the project are well defined, resulting in a fully featured tool that can be utilized, albeit, not very user-friendly. The general structure of the desired application can be surmised to be the combination of the document segmentation system and the text summarization system, with the output of the first system being the input of the next, with pre-processing being applied to the segmented text between the two system. The original input of the system should take into consideration both the single document and the multiple documents situations. As such, the general structure of the application can be described through Figure 3.1-1.

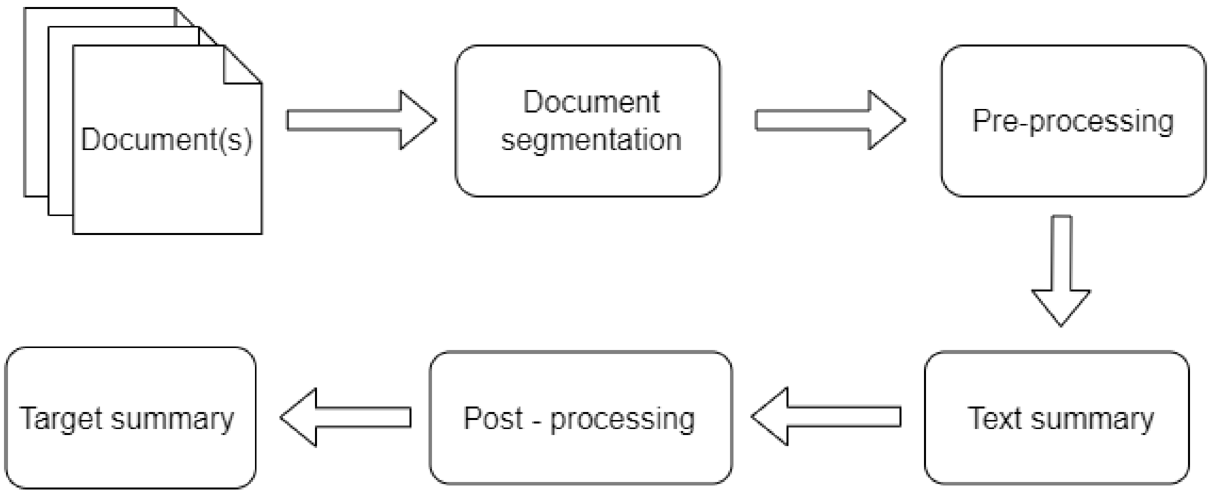

Figure 3-1 - General structure of the application

Since many summarizations techniques can be utilized in this situation, five were chosen to be used as the summarization methods. These methods would then be compared using specific benchmarks to determine the most suitable for this specific use case. Amongst them, aside from LSA and TextRank which is an extractive method, two previously mentioned pretrained machine learning models are utilized to achieve abstractive results, including text-summarization and bart-large-cnn.

3.2 Utilizing Extractive summarization methods for Pre-processing

During the normal pre-processing procedure for text, methods such as text normalization, whitespace trimming, etc. are frequently used. Although these methods can improve the input to the summarization system and clean up the input text, the remaining paragraphs can be exceedingly lengthy, spanning more than 1000 words each, which might include much unnecessary information. As such, filtering important sentences from the others can potentially improve the output of the whole system.

As such, extractive summarization methods such as LSA and TextRank have been decided to be used as methods for extracting the specified number of sentences to be passed into another summarization model. For the purpose of this approach, five sentences from each paragraph are chosen for testing purposes. Since both LSA and TextRank use different approaches to rank sentences, the final result can change drastically.

3.3 System design

Although Figure 3-1 depicts a multitude of steps involved in the process, the whole system can be condensed into 2 different parts, which are the segmentation subsystem, and the text summarization subsystem which includes all steps after segmenting the documents. The first subsystem contains the data path and logic behind the segmenting of a document using GROBID. Following, the output of the first subsystem is then preprocessed, summarized and postprocessed in the second subsystem. The subsystem would then deliver the response to the corresponding request, completing all tasks required. The overall design of the system can be expressed through the component diagram in Figure 3-2.

3.3.1 Data format

The usage of Docker as the main method for cross-platform compatibility is inevitable as the application needs to be flexible and adaptable to new system requirements. In addition, as each subsystem is designated as a different container, communication between each container is limited to HTTP request. Due to which, the application must, therefore, be data-driven, controlling its logic through changes made to the data. As a result, JSON file format was chosen as the common interface between each of the subsystem and processes.

Although the original output of the GROBID subsystem is an XML file, which can act as the interface for the overall system, a large amount of unnecessary data is included due to its highly generalized nature. Amongst the included data, tags containing header information and reference information, for instance, are of no use for the current application. Therefore, the truncation of said file is necessary to declutter and isolate the important information for further use. The resulting truncation is then reformatted into a JSON file due to the necessity of moving data between containers.

For the purposes of this application, a research paper can be divided into many sections, with each section being the header (i.e. Header 1. Introduction) of the paper. Each sub-header's content (i.e. Header 1.1's content) will be consolidated and appended into the parent header, creating a section. A section's name, however, is defined only by the parent header.

Since each paper can be divided into its sections and the section's content, necessary information of each paper includes the name of the header and the content of each header. The name of the file is also included in the final summary JSON file due to the need for differentiation between different files when multiple PDFs are uploaded. However, the abstract section is included in the JSON file due to its usefulness for testing purposes. As such, the JSON file contains the abstract section, the section list which includes multiple header and content.

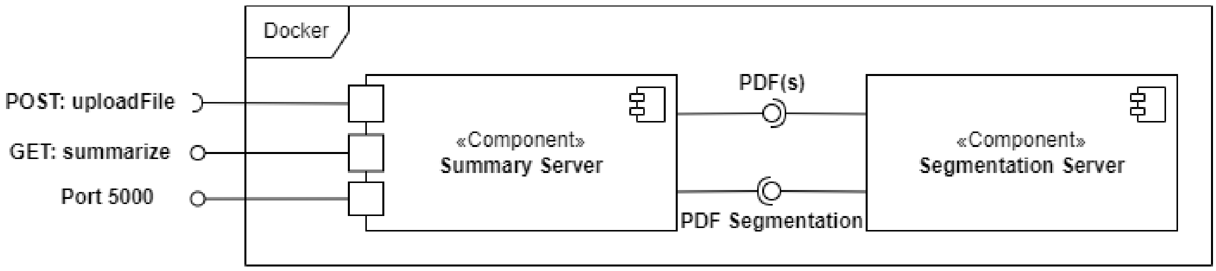

Figure 3-2 – The component diagram of the system

3.3.2 Document Segmentation Subsystem

As explained in the Literature Review chapter, GROBID has been chosen due to its high compatibility with the needs of the project. As a result, most classes and functionalities of this subsystem are for integrating the GROBID's functionalities with the main application. One of its main purposes is to initialize and configurate the basic GROBID settings such as the GROBID home (the GROBID source location), changing the input PDF file into an appropriate format, and getting the result after parsing. Said result, which is in the form of an XML file, then became the input for the XMLParser class and YAMLParser class, creating the final output of the subsystem, which is a YAML file containing the segmentation of the provided PDF.

3.3.3 Text Summarization Subsystem

As the text summarization subsystem is not only responsible for the text summarizer, but also both the pre-processing and postprocessing of the input and output, this subsystem is further divided into three main modules, with each being pre-processing, summary, and postprocessing correspondingly. Additionally, a wrapper module is used to encapsulate the whole subsystem, wiring everything together, and exposing the appropriate APIs and web interface.

In the pre-processing module, each section's content from the segmented paper is first trimmed and normalized. During the trimming process, excess white space from both between, before, and after words are removed, cleaning the text syntactically. After said process, all capitalized characters are lowercased in the word normalization process, allowing for the model to read the text more effectively. Each section and its name are then stored in a custom data object called Paper. Each paper contains the name of the currently processing PDFs and a list containing the multiple data objects called Section. A Section is responsible for containing the section's name and content. Utilizing the custom data object can help move data more effectively between each module.

Moving to the text summarization module, although many different pre-processing and summarization options are available, only a few options are included in the final product due to the speed at which these methods process the provided data. Especially of note are the text summarization methods involving the use of pretrained machine learning models, as these methods not only consume a significant amount of resource while running, but also during its start up. Amongst the aforementioned models, only Facebook's bart-large-cnn and Falcon AI's text-summarization are suitable for running in production, while Stability AI's stablelm-2-1_6b can only be used for testing purposes. Therefore, a list of available options for summarizing text includes: TextRank, LSA, text-summarization model, and bart-large-cnn. These methods are then connected to the overall system to produce the corresponding summaries.

Since the resulting summaries are formatted in different ways depending on the method in which the summary was produced. For example, stablelm-2-1_6b's output will start with "<|im_end|>" and end with "<|endoftext|>", which needs to be removed. As such it is the postprocessing module's responsibility to reformat the resulting text according to the corresponding summarization method. Furthermore, since each section of a paper is passed through the summarization method individually, the module must recompile all these sections and add them under the correct section header for the result, which is in the form of a JSON file.

As for the final module which is the application wrapper, APIs are constructed as the main method for communication with the system. Of note are two APIs, including a method for uploading a file and a method for summarizing the provided files. Port 5000 is also available by default as a method of choice for the system to be more accessible through a simple Web UI. Overall, an overview of the system can be observed through Figure 3-3.

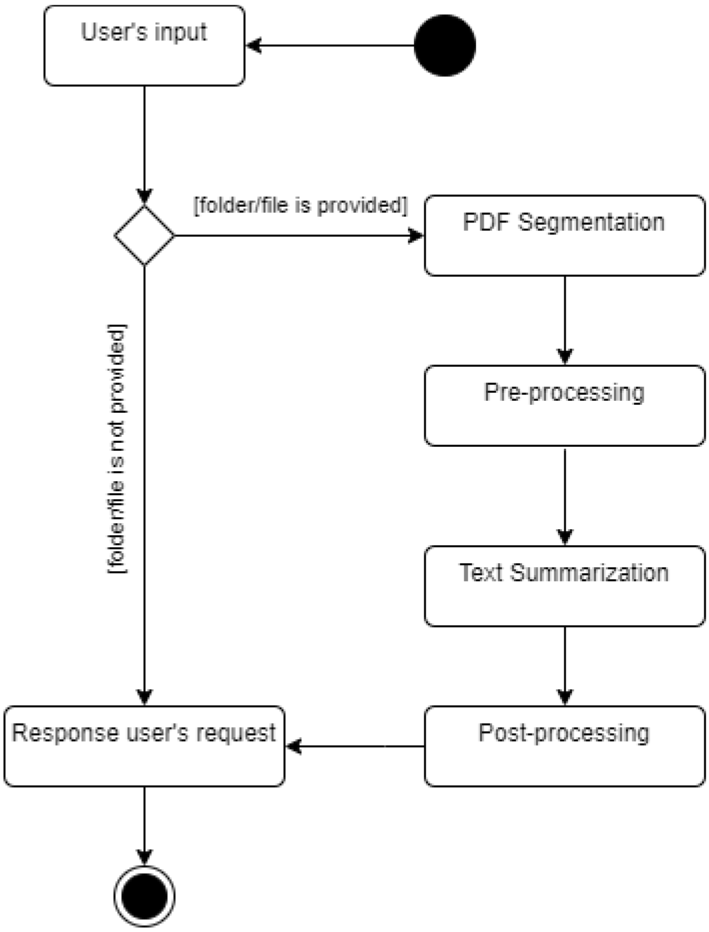

Figure 3-3 – The application's activity diagram

Chapter 4 IMPLEMENT AND RESULTS

4.1 Overview

This chapter provides explanations and descriptions of the work that has been done. Amongst them, most of the work can be divided into 2 distinct parts, which are implementing the document segmentation using GROBID and the development of the text summarization capability along with the multitude of summarization methods. Since the whole system is encapsulated in a cluster of Docker nodes, configured through the use of Docker Compose, the whole system can be reproduced

4.2 Implementation

Some code snippets explaining what has been discussed in Chapter 3 from each subsystem can be found in the following part of this chapter. By providing these code sections, the method in which the application is implemented can be better understood.

4.2.1 Document Segmentation

As GROBID is a machine learning approach to the segmentation problem, its initialization requires the initialization of the GROBID container. The code snippet in Figure 4-1 is the function through which the initialization is made.

```
async def parse_pdf() -> bool:
    """
    Connect to the GROBID and summarize the PDF in the folder.
    """
    with app.app.app_context():
        with concurrent.futures.ThreadPoolExecutor(max_workers=1):
            try:
                client = GrobidClient(config_path=GROBID_CONFIG_PATH)
                client.process(
                    "processFulltextDocument",
                    UPLOAD_FOLDER,
                    output=SEGMENT_FOLDER,
```

```
                    consolidate_citations=False,
                    tei_coordinates=False,
                    verbose=True
                )
            except:
                return False
    return True
```

Figure 4-1 – GROBID initialization code snippet

Since this method would find all PDF files available in the specified UPLOAD FOLDER, no additional set up is required except for having to upload the corresponding files to the correct folder. Toward that end, user can directly upload the PDF file by submitting it from the Web UI, or by utilizing the folder-uploading request, an example of which can be found in Figure 4-2.

```
@app.route("/upload", methods=['POST'])
def upload():
    if (request.method != 'POST'):
        return "No GET method", 404
    if 'files' not in request.files:
        return "No 'files' header in payload.", 404
    pdf_file = request.files['files']
    if pdf_file.filename == '':
        return "File has no name.", 404
    if not pdf_file:
        return "File does not exist.", 404
    if not allowed_file(pdf_file.filename):
        return "File does not end with .pdf", 404
    filename = secure_filename(pdf_file.filename)
    pdf_file.save(os.path.join(app.config['UPLOAD_FOLDER'], filename))
    return 'Successfully send files', 200
```

Figure 4-2 – PDF-uploading method

After the file content as well as its name is checked sufficiently, including the file extension, the content of the file, the filename. Additionally, the *secure_filename* function also provides a method to adjust the file name so that it does not violate any rules regarding the naming scheme of files from any operating system. The file is then saved to the corresponding folder, awaiting further usage.

4.2.2 Text summarization

As both LSA and TextRank are used during the development process, implementation of both systems is available. Even though TextRank was implemented through the usage of a library, its implementation is worth mentioning as some parameters and its choice have effects on the result of the summary. LSA, however, does not have a library that supports it from the start, so it needs to be implemented step by step. As for the pretrained models, HuggingFace [14] was utilized as the library of choice for implementation due to the large number of tools, support, and models available.

4.2.2.1 LSA

The LSA summarization approach goes through 7 steps in total, including Tokenizing, NGram generation, stop words filtering, word frequency sorting, sentence detection, word searching, and finally summarization. While the first 3 steps are a part of the *_create_dictionary* function, the remaining steps are distributed among the remaining function calls due to their complexity. The summary is then generated by using the output from the steps above. This can be seen in Figure 4-3.

```
def lsa (text:str, sent_num:int=SENT_NUM) -> str:
    if(len(text) == 0):
        return ""

    dictionary = _create_dictionary(text)
    sentences = sent_tokenize(text)

    matrix = _create_matrix(text, dictionary)
    matrix = _compute_term_frequency(matrix)

    u, sigma, v = singular_value_decomposition(matrix, full_matrices=False)
```

```
    ranks = iter(_compute_ranks(sigma, v))
    result = _get_best_sentences(
        sentences,
        sent_num,
        lambda s: next(ranks)
    )
    result_str = ''
    for sentence in result:
        result_str += sentence
    return result_str
```

Figure 4-3 – LSA Implementation

The summarization process above works by looping through each summary sentence, conducting word searching on it, and storing the results in a list. The summary is then concatenated and constructed from said list.

4.2.2.2 TextRank

As mentioned, the TextRank implementation is made using a library called *summa*, this can be seen through Figure 4-5, where the summarization function is defined.

```
def summarize_text(text:str, sent_num:int=SENT_NUM) -> str:
    return summarize(
        text = text,
        words = sent_num * (_get_max_length_sentence(text) + 5),
        additional_stopwords= summary.get_stop_word_list()
    )
```

Figure 4-4 – TextRank summarization function

In said function, "text" refers to each PDF segment that the user wants to summarize. "words" is the upper word limit of the resulting summary and "additional_stopwords" refers to the list of words that need to be filtered out before processing the summary. In order to control the number of sentences the resulting summary will have, the length of the longest sentence in the provided text must be calculated, which is the responsibility of the _get_max_length_sentence function. Utilizing which, only the text that needs to be summarized and the number of sentences needs to be passed in as parameter.

4.2.2.3 Pretrained models

Even though small differences exist when implementing different models with regards to how each one is initialized and utilized, a familiar structure can be distilled. Due to loading a model requiring an extended amount of time, it is preloaded and cached, allowing for subsequent usage requiring less memory and time to complete a task. An example of preloading the model can be found in Figure 4-4.

```
async def preload_stable_lm_chat_1_6b():
    global tokenizer
    global model
    tokenizer = AutoTokenizer.from_pretrained(
        'stabilityai/stablelm-2-1_6b-chat'
   )
    model = AutoModelForCausalLM.from_pretrained(
        'stabilityai/stablelm-2-1_6b-chat',
        device_map="auto",
    )
```

Figure 4-5 – Model preloading

The text is then passed into the chosen model and further processed to be suitable for the required text generation. The input is first passed into a prompt constructor, where it will instruct the model for the required text generation. The prompt would then be tokenized in some models such as stablelm-2-1_6b-chat as per the requirement of each model. Finally, the text is then passed into the model where it would be processed, and the result can be obtained. The process can be seen in Figure 4-5 in the case of stablelm-2-1_6b-chat.

```python
async def stable_lm_chat_1_6b(content:str) -> str:
    prompt = [{'role': 'user', 'content': prompt_constructor(content)}]
    inputs = tokenizer.apply_chat_template(
        prompt,
        add_generation_prompt=True,
        return_tensors='pt'
    )
    tokens = model.generate(
        inputs.to(model.device),
        max_new_tokens=get_max_length(content),
        temperature=0.01,
        do_sample=True
    )
    output = tokenizer.decode(
        tokens[:, inputs.shape[-1]:][0], skip_special_tokens=False)
    return output.rstrip("<|im_end|>\n<|endoftext|>")
```

Figure 4-6 – Model text generation

4.3 Results

The results from the process of the entire system can be condensed into 2 different JSON files with the first being the document segmentation result and the second being the summary. As an example, one research paper named has been chosen. As the result of the document segmentation being the content of the entire paper, only the first header will be showcased in Figure 4-7. Since many different methods of summarizations were used, only LSA and TextRank are shown as an example of the data format, which can be seen in Figure 4-8 and Figure 4-9.


```json
{
  "name": "Hansen_et_al._-_2024_-_Productivity_and_quality-adjusted_life_years_QALY.grobid.tei.xml",
  "segments": [
    {
      "header": "Introduction",
      "content": "Few aspects concern human beings more than health. But resources are scarce and, as we face demographic changes, with increased demand from retirees and a constrained labour market due to shrinking working age share of the population, there is a pressing need to protect the health and productivity of the economically active population. Therefore, critical decisions on health care interventions, as well as occupational health and safety policies, have to be made constantly. The evidence from clinical trials and observational studies, in addition to assessments about the productivity consequences, are crucial to make
```

```
those decisions. The purpose of this paper is to develop a unified framework for the
measurement and valuation of outcomes of such programmes and policies.It is widely accepted
that the health benefit a patient derives from a particular health care intervention can be
defined according to two natural dimensions: quality of life and quantity of life. An
alternative to QALYs is the so-called productivity-adjusted life years (in short, PALYs),
which are calculated by multiplying a productivity index by years lived. The productivity
index ranges from 0 (completely unproductive) to 1 (completely productive), and may take into
consideration factors such as absence from work due to ill health (absenteeism), reduced
productivity while at work (presenteeism) and premature exit from the workforce (e.g.,
Economic evaluation of policies to improve occupational health and safety is one field of
research where productivity outcome measures following the broader PALY idea are applied
extensively (e.g., Our approach builds upon the framework introduced in Hougaard et al. We
conclude this introduction stressing that our model treats health and productivity as
different individual attributes. In doing so, we obviously depart from the literature that
considers only one of them, but also from the simplistic assumption that both concepts are
perfectly correlated (which would allow to use a reduced model). The precise relationship
between health and productivity is complex and the anticipated correlation might actually be
positive or negative, depending on the viewpoint. For instance, The rest of the paper is
organized as follows. In Section 2, we introduce the framework and the basic common axioms
that all our evaluation functions will satisfy. In Section 3, we characterize the focal (and
somewhat polar) evaluation functions QALYs and PALYs. In Section 4, we characterize classes of
evaluation functions which compromise among the previous two. In Section 5, we characterize
more general functional forms, that evolve around the notion of healthy productive years
equivalent. In Section 6, we discuss our contribution with a special emphasis on the choice
among the different evaluation functions we characterize. Finally, in Section 7, we provide
some concluding remarks providing further connections to related literature and pointing out
possible extensions of our work. For a smooth passage, we defer all proofs to the Appendix."
    },…
}
```

Figure 4-7 – Segmentation result

```
{
  "name": "Hansen_et_al._-_2024_-_Productivity_and_quality-
adjusted_life_years_QALY.grobid.tei.xml",
  "segments": [
    {
      "header": "Introduction",
      "content": "The productivity index ranges from 0 (completely unproductive) to 1
(completely productive), and may take into consideration factors such as absence from work due
to ill health (absenteeism), reduced productivity while at work (presenteeism) and premature
exit from the workforce (e.g., Economic evaluation of policies to improve occupational health
and safety is one field of research where productivity outcome measures following the broader
PALY idea are applied extensively (e.g., Our approach builds upon the framework introduced in
Hougaard et al."
    },…
}
```

Figure 4-8 – TextRank summarization result

```
{
  "id": "Hansen_e",
  "name": "Hansen_et_al._-_2024_-_Productivity_and_quality-
adjusted_life_years_QALY.grobid.tei.xml",
  "segments": [
    {
      "header": "Introduction",
      "content": "The precise relationship between health and productivity is complex and the
anticipated correlation might actually be positive or negative, depending on the viewpoint."
    },…
}
```

Figure 4-9 – LSA summarization result

Chapter 5 DISCUSSION AND EVALUATION

5.1 Evaluation

As text segmentation utilized GROBID, which is a tool that has been used and benchmark through multiple sources ([15], [16], [17], [18]). As such the main focus of the evaluation section would be on the summarization method.

Evaluating and comparing the chosen summarization methods are an important task so as to choose the appropriate sentences for the product to be adaptable for a range of topics that can be covered for by a research paper. Additionally, the provided summarization methods must be efficient enough to be able to be executed by an average user. As such, additional information such as the evaluation method, the dataset, the specification of the test machine, and the evaluation results will be included in the following sections.

5.1.1 Evaluation methods and Dataset

Due text summarization being highly dependent upon the perspective and knowledge of the reader. As such, having testers with the appropriate knowledge to evaluate the result is the most desirable outcome as this can realistically reflect the opinion of normal readers. Otherwise, a dataset containing a reference summary result can also be used to compare to the generated text of the system being tested.

However, since the current summarization use case is highly specific, no dataset containing the required referenced summary were found. Since most open-sourced datasets focus on summarizing the entire paper's content to a small paragraph, this could not be applied for this current use case since the PDF is segmented and each segment is further summarized. Additionally, many datasets were rejected due to them being based on sources that are not research papers, including newspapers and books, this can be seen amongst the list of frequently used dataset for this purpose [5] as most of them are from the news and books domain.

As such, the only available option is to utilize a statistical calculation in order to determine the value of each summary. The calculation of choice was ROUGE-1, ROUGE-2, and ROUGE-L [19]. Although some traditional measures such as Recall, Precision, and F1 can be used to calculate quantitatively how much of the summary has recapture the context of the original text, and how much of that context is relevant, ROUGE measure provides more granularity as it has the benefits of the simpler measures, but also applies it to n-gram and longest matching sequence. Using these measures, the process of evaluating text summarization can be done automatically, allowing for faster feedback and improvement to the system.

In order to evaluate the system, a random set including 10 research papers PDF relating multiple topics (medicine, astrophysics, machine learning, etc.). These papers were all randomly chosen by browsing through Arvix, finding different topics, and choosing a paper at random. By having a more diversified list of PDFs, the dataset can be less biased overall. Furthermore, since the system needs to be automated, TextRank summaries are used as the method to obtain the referenced summaries to be compared to.

5.1.2 Evaluation results and Machine specifications

The machine being used to evaluate the system is currently an AMD RYZEN 5 5600 with 16 of RAM. After running ROUGE-1, ROUGE-2, and ROUGE-L measures, the results are tabulated in the table 5-1.

Pre-processing method	Summarization method	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**
None	**LSA**	0.576021	0.245746	0.576021
None	**Falcon AI's text-summarization**	0.437884	0.263082	0.437884
None	**Facebook's bart-large-cnn**	0.598212	0.367898	0.598212
None	**Stability AI's stablelm-2-1_6b**	0.539534	0.224551	0.539534
TextRank	**Falcon AI's text-summarization**	**0.661115**	0.425214	**0.661115**
TextRank	**Facebook's bart-large-cnn**	0.648827	**0.455075**	0.648827
TextRank	**Stability AI's stablelm-2-1_6b**	0.579882	0.260122	0.579882
LSA	**Falcon AI's text-summarization**	0.566133	0.315426	0.566133
LSA	**Facebook's bart-large-cnn**	0.604226	0.371978	0.604226
LSA	**Stability AI's stablelm-2-1_6b**	0.547256	0.217316	0.547256

Table 5-1 - ROUGE-1, ROUGE-2, and ROUGE-L evaluation results.

From the evaluation result, the combination of having TextRank as the pre-processing method and using the resulting sentences to be summarized by the Falcon AI's text-summarization model obtained the most positive result, being the top of both the ROUGE-1 and ROUGE-L measure. The combination between TextRank and Facebook's bart-large-cnn model also yields positive results since it claims the top of the ROUGE-2 measure.

Chapter 6 CONCLUSION AND FUTURE WORK

6.1 Conclusion

Although the project can be considered as functional, much is left to be desired as the project does not consider much of the user experience. Additionally, even though pre-trained models were used, finetuning the model to fit with the current use case was not possible due to limited time and resources. Overall, the project can be considered as the steppingstone for future improvements.

6.2 Future work

Ideas for extending the current work that can be done in the future:

1. Finetuning the model may allow for better summarization results, which can further improve the evaluation metrics
2. More PDFs can be added to the pool of dataset, allowing for more accurate results.
3. Human-evaluated reference PDF can be considered moving forward to improve the quality of the current dataset.
4. A self-developed model can be considered as having a simple model that is tailored to the current task can improve both the application's efficiency and output if done correctly.

REFERENCES

[1] A. Jinha, “Article 50 million: An estimate of the number of scholarly articles in existence,” *Learn. Publ.*, vol. 23, pp. 258–263, Jul. 2010, doi: 10.1087/20100308.

[2] P. Lopez, “GROBID: Combining Automatic Bibliographic Data Recognition and Term Extraction for Scholarship Publications,” in *Research and Advanced Technology for Digital Libraries*, M. Agosti, J. Borbinha, S. Kapidakis, C. Papatheodorou, and G. Tsakonas, Eds., Berlin, Heidelberg: Springer, 2009, pp. 473–474. doi: 10.1007/978-3-642-04346-8_62.

[3] “How GROBID works - GROBID Documentation.” Accessed: Jan. 13, 2024. [Online]. Available: https://grobid.readthedocs.io/en/latest/Principles/

[4] D. Jurafsky and J. H. Martin, *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition*. in Prentice Hall series in artificial intelligence. Upper Saddle River, N.J: Prentice Hall, 2000.

[5] W. S. El-Kassas, C. R. Salama, A. A. Rafea, and H. K. Mohamed, “Automatic text summarization: A comprehensive survey,” *Expert Syst. Appl.*, vol. 165, p. 113679, Mar. 2021, doi: 10.1016/j.eswa.2020.113679.

[6] R. Mihalcea and P. Tarau, “TextRank: Bringing Order into Text,” in *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, D. Lin and D. Wu, Eds., Barcelona, Spain: Association for Computational Linguistics, Jul. 2004, pp. 404–411. Accessed: Nov. 05, 2023. [Online]. Available: https://aclanthology.org/W04-3252

[7] P. W. Foltz, “Latent semantic analysis for text-based research,” *Behav. Res. Methods Instrum. Comput.*, vol. 28, no. 2, pp. 197–202, Jun. 1996, doi: 10.3758/BF03204765.

[8] M. Lewis *et al.*, “BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension.” arXiv, Oct. 29, 2019. doi: 10.48550/arXiv.1910.13461.

[9] C. Raffel *et al.*, “Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer.” arXiv, Sep. 19, 2023. Accessed: Jun. 08, 2024. [Online]. Available: http://arxiv.org/abs/1910.10683

[10] F. Zhuang *et al.*, “A Comprehensive Survey on Transfer Learning.” arXiv, Jun. 23, 2020. doi: 10.48550/arXiv.1911.02685.

[11] X. Han *et al.*, “Pre-Trained Models: Past, Present and Future.” arXiv, Aug. 11, 2021. doi: 10.48550/arXiv.2106.07139.

[12] “Falconsai/text_summarization · Hugging Face.” Accessed: Jun. 12, 2024. [Online]. Available: https://huggingface.co/Falconsai/text_summarization

[13] “facebook/bart-large-cnn · Hugging Face.” Accessed: Jun. 12, 2024. [Online]. Available: https://huggingface.co/facebook/bart-large-cnn

[14] T. Wolf *et al.*, “Transformers: State-of-the-Art Natural Language Processing,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Q. Liu and D. Schlangen, Eds., Online: Association for Computational Linguistics, Oct. 2020, pp. 38–45. doi: 10.18653/v1/2020.emnlp-demos.6.

[15] M. Lipinski, K. Yao, C. Breitinger, J. Beel, and B. Gipp, “Evaluation of header metadata extraction approaches and tools for scientific PDF documents,” in *Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries*, Indianapolis Indiana USA: ACM, Jul. 2013, pp. 385–386. doi: 10.1145/2467696.2467753.

[16] P. Lopez, C. Du, J. Cohoon, K. Ram, and J. Howison, “Mining Software Entities in Scientific Literature: Document-level NER for an Extremely Imbalance and Large-scale Task,” in *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, in CIKM ’21. New York, NY, USA: Association for Computing Machinery, Oct. 2021, pp. 3986–3995. doi: 10.1145/3459637.3481936.

[17] J. M. Nicholson *et al.*, "scite: a smart citation index that displays the context of citations and classifies their intent using deep learning." bioRxiv, p. 2021.03.15.435418, Mar. 16, 2021. doi: 10.1101/2021.03.15.435418.
[18] M. Grennan and J. Beel, "Synthetic vs. Real Reference Strings for Citation Parsing, and the Importance of Re-training and Out-Of-Sample Data for Meaningful Evaluations: Experiments with GROBID, GIANT and Cora." arXiv, Apr. 25, 2020. Accessed: Nov. 03, 2023. [Online]. Available: http://arxiv.org/abs/2004.10410
[19] C.-Y. Lin, "ROUGE: A Package for Automatic Evaluation of Summaries," in *Text Summarization Branches Out*, Barcelona, Spain: Association for Computational Linguistics, Jul. 2004, pp. 74–81. Accessed: May 06, 2024. [Online]. Available: https://aclanthology.org/W04-1013

APPENDIX

The full result from the Segmentation process


```
{
  "name": "Hansen_et_al._-_2024_-_Productivity_and_quality-
adjusted_life_years_QALY.grobid.tei.xml",
  "segments": [
    {
      "header": "Introduction",
      "content": "Few aspects concern human beings more than health. But resources are scarce
and, as we face demographic changes, with increased demand from retirees and a constrained
labour market due to shrinking working age share of the population, there is a pressing need
to protect the health and productivity of the economically active population. Therefore,
critical decisions on health care interventions, as well as occupational health and safety
policies, have to be made constantly. The evidence from clinical trials and observational
studies, in addition to assessments about the productivity consequences, are crucial to make
those decisions. The purpose of this paper is to develop a unified framework for the
measurement and valuation of outcomes of such programmes and policies.It is widely accepted
that the health benefit a patient derives from a particular health care intervention can be
defined according to two natural dimensions: quality of life and quantity of life. An
alternative to QALYs is the so-called productivity-adjusted life years (in short, PALYs),
which are calculated by multiplying a productivity index by years lived. The productivity
index ranges from 0 (completely unproductive) to 1 (completely productive), and may take into
consideration factors such as absence from work due to ill health (absenteeism), reduced
productivity while at work (presenteeism) and premature exit from the workforce (e.g.,
Economic evaluation of policies to improve occupational health and safety is one field of
research where productivity outcome measures following the broader PALY idea are applied
extensively (e.g., Our approach builds upon the framework introduced in Hougaard et al. We
conclude this introduction stressing that our model treats health and productivity as
different individual attributes. In doing so, we obviously depart from the literature that
considers only one of them, but also from the simplistic assumption that both concepts are
perfectly correlated (which would allow to use a reduced model). The precise relationship
between health and productivity is complex and the anticipated correlation might actually be
positive or negative, depending on the viewpoint. For instance, The rest of the paper is
organized as follows. In Section 2, we introduce the framework and the basic common axioms
that all our evaluation functions will satisfy. In Section 3, we characterize the focal (and
somewhat polar) evaluation functions QALYs and PALYs. In Section 4, we characterize classes of
evaluation functions which compromise among the previous two. In Section 5, we characterize
more general functional forms, that evolve around the notion of healthy productive years
equivalent. In Section 6, we discuss our contribution with a special emphasis on the choice
among the different evaluation functions we characterize. Finally, in Section 7, we provide
some concluding remarks providing further connections to related literature and pointing out
possible extensions of our work. For a smooth passage, we defer all proofs to the Appendix."
    },
    {
      "header": "Preliminaries",
      "content": "Let a population consisting of n individuals be identified with the set N =
{1, ..., n}. Each individual i \u2208 N is described by a profile, formalized by a triple d i
= (a i , p i , t i ), where a i \u2208 A is a health state, p i \u2208 [0, 1] is the
productivity level, andWe can think of the health state a i as a chronic or representative
health state over time. The productivity p i is measured by any chosen indicator. For
instance, it can be an indicator of absence from work (e.g., number of sick days per year for
a person). Note that such an indicator may reflect productivity and contributions to society
in a broad sense. For example, work may include both labour market activities and domestic
work. Alternatively, a measure could be chosen that reflects the value of the work contributed
by the individuals, as would (very roughly) be the case if measured by e.g., (relative)
earnings (e.g., Finally, there are two plausible interpretations of time in our model. On the
one hand, it could be identifying the individual total lifetime. On the other hand, it could
be identifying incremental individual lifetime from a given status quo up to the end of life
or retirement.Let d = (d 1 , ..., d n ) denote a distribution of individual profiles, as
described above, and let D denote the set of possible distributions. We now give an example in
which two hypothetical distributions are presented. We shall return to this example later in
the text several times to illustrate how these two hypothetical distributions can be
(relatively) evaluated, by means of various evaluation functions we consider.Example 1
Consider the following two distributions, involving five individuals each (that could be
interpreted as representative agents of five different groups).In the first distribution (d
\u2206 ), all individuals are experiencing full health. The first one is also experiencing
```

maximum productivity as well as forty years of lifetime (until retirement, or the end of life), i.e., d \u2206 1 = (a * , 1, 40). The second individual is experiencing 50% of maximum productivity and forty years of lifetime, i.e., d \u2206 2 = (a * , 0.5, 40). The third individual is experiencing zero productivity and forty years of lifetime, i.e., d \u2206 3 = (a * , 0, 40). The fourth individual is experiencing 50% of maximum productivity and ten years of lifetime, i.e., d \u2206 4 = (a * , 0.5, 10). The last individual is experiencing zero productivity and lifetime, i.e., d \u2206 5 = (a * , 0, 0). In the second distribution (d \u039b), the first individual is also experiencing full health and maximum productivity as well as forty years of lifetime, i.e., Note that if E represents then any strictly increasing transformation of E would also do so.An evaluation function E may be interpreted as an effect measure if it is used for the economic evaluation of health care or working environment interventions.In what follows, we present some basic axioms for social preferences in the current context, that will be common to all the evaluation functions we consider in this paper. In this section, we present a set of seven axioms that forms the necessary conditions for the theorems presented in the remaining sections of this paper. These are termed the COMMON axioms. The axioms reflect basic principles adapted to our framework that are widely accepted in economics. In the following sections, additional axioms are presented, which together with the COMMON axioms close the characterizations of the evaluation functions we highlight.The first three COMMON axioms apply to all three attributes in the same way; whereas the latter four COMMON axioms introduce some conditions on time which distinguish it from the other two attributes.The first axiom, anonymity, reflects the principle of impartiality, with a long tradition in the theory of justice (e.g., The third axiom, continuity, is the adaptation of a standard operational condition to our context. It says that small changes in productivity or life years should only produce small changes in the evaluation of the distribution. i , tWe then move to the second group of COMMON axioms. First, the social zero condition, which is reminiscent of a well-known condition for individual utility functions on health (e.g., LMFHP: For each d \u2208 D and each i \u2208 N, such that (a i , p i) = (a * , 1) and each where q : A \u2192 [0, 1] is a function satisfying 0 \u2264 q(a i) \u2264 q(a *) = 1, for each a i \u2208 A.The unweighted aggregation of individual QALYs, as specified in This evaluation function ignores productivity. More precisely, it satisfies the following axiom, productivity independence, which states that for a fixed health state and lifetime, the productivity is irrelevant for the evaluation of the distribution.It also satisfies the time invariance at common health and full productivity axiom, which states that for two individuals at common health and maximum productivity, extra life years are interchangeable. That is, it does not matter to the social planner which individual (among those with common health and maximum productivity) gets extra life years.TICHFP: For each d \u2208 D, each pair i, j \u2208 N with a i = a j = a and p i = p j = 1, and eachOur first result states the QALY evaluation function is characterized by the combination of the previous two axioms and the COMMON axioms. 10 Theorem 1 The following statements are equivalent:1. is represented by a QALY evaluation function (1).A possible interpretation of Theorem 1 could be a situation where a social planner decides to evaluate and prioritise among a set of interventions using the QALY evaluation function A counterpart axiom of productivity independence is health independence, which states that, for fixed productivity and lifetime, the health state of an individual is irrelevant for the evaluation.And, likewise, a counterpart of time invariance at common health and full productivity is time invariance at full health and common productivity, which states that for two individuals at full health and common productivity, it does not matter to the social planner who receives the extra life years. 11 As all theorems in this paper require the COMMON axioms, these will not be mentioned again in the interpretations of the remaining theorems.TIFHCP: For each d \u2208 D, each pair i, j \u2208 N with a i = a j = a * and p i = p j = p, and eachAs the next result states, if the previous two axioms replace their counterparts at Theorem 1, we characterize the following generalized PALY evaluation function, which evaluates distributions by means of the aggregation of individual PALYs in society, when submitted first to a continuous function (v). Formally,whereThe following statements are equivalent:1. is represented by a generalized PALY evaluation function (2).The generalized PALY evaluation function is a counterpart of the QALY evaluation function.As such, neither the v function nor the q function (in their respective functional forms) has a monotonic structure. This makes sense in the latter case because the domain of health statesA does not have a mathematical structure. But the domain of productivity levels is naturally ordered and, therefore, it would make sense to impose v a non-decreasing structure. The following axiom will guarantee such a feature as a byproduct.The axiom productivity invariance at full health and common time states that, for any two individuals with common lifetime and full health, it makes no difference who gains in productivity.PIFHCT: For each d \u2208 D, each pair i, j \u2208 N with a i = a j = a * and t i = t j = t, and eachAs the next result states, adding this axiom to those in Theorem 2 we characterize the affine PALY evaluation function, which evaluates distributions by means of the aggregation of individual PALYs in society, when submitted first to an affine and non-decreasing function.Formally,where \u03b1 \u2208 [0, 1].The following statements are equivalent:1. is represented by an affine PALY evaluation function (3).satisfies COMMON, HI, TIFHCP, and PIFHCT.The previous families are obvious

generalizations of the focal linear PALY evaluation function, which evaluates distributions by means of the unweighted aggregation of individual PALYs in society, or, in other words, by the weighted (through productivity levels) aggregate time span the distribution yields. Formally,The unweighted aggregation of individual PALYs as specified in (TIUP: For each d \u2208 D and each i \u2208 N such that p i = 0, and eachThe following statements are equivalent:1. is represented by a PALY evaluation function (4).satisfies COMMON, HI, TIFHCP, PIFHCT, and TIUP.According to Theorem 4, a social planner wishing to conduct an economic evaluation of working environment interventions using the PALY evaluation function (4) will also hold a number of specific values in the process of priority setting. These value choices include, for example, that changes in health among individuals following an intervention (axiom HI) and that increases in lifetime among unproductive individuals (axiom TIUP) both have no influence on the choice of intervention.We conclude this section applying the evaluation functions characterized in this section to the distributions from Example 1. More precisely, we summarize the computations in the next table. We infer from there that the first distribution is preferred from the viewpoint of QALYs, whereas the second distribution is preferred from the viewpoint of the PALYs-based evaluation functions."
    },
    {
      "header": "Compromising between QALYs and PALYs",
      "content": "As mentioned above, the evaluation functions characterized in the previous section ignore one dimension of our model. In other words, they all rely on a very demanding axiom of (productivity or health) independence. The purpose of this section is to dismiss those axioms, while obtaining characterizations of natural compromises between those focal (albeit polar) evaluation functions.For instance, the productivity-and-quality-adjusted life years (PQALY) evaluation function evaluates distributions by means of the weighted (through productivity and health levels) aggregate time span the distribution yields, so that health, productivity and lifespan of individuals enter the evaluation function multiplicatively. Formally,where q : A \u2192 [0, 1] is a health state quality weight satisfying 0 \u2264 q(a i) \u2264 q(a *) = 1 for eachAs the next result states, this evaluation function is characterized when we dismiss health independence in the previous result (characterizing PALYs), and strengthen the other two independence axioms to consider the following ones.First, time invariance at common health and productivity, which states that for two individuals at common health and productivity, it does not matter to the social planner who receives the extra life years.TICHP: For each d \u2208 D, each pair i, j \u2208 N with a i = a j = a and p i = p j = p, and each c > 0,Second, productivity invariance at common health and time, which says that for two individuals at common health and time, it does not matter who gains in productivity.PICHT: For each d \u2208 D, each pair i, j \u2208 N with a i = a j = a and t i = t j = t, and each c > 0Theorem 5 The following statements are equivalent:1. is represented by a PQALY evaluation function (5)."
    },
    {
      "header": "satisfies COMMON, TICHP, PICHT, and TIUP.",
      "content": "A social planner may view both health effects and productivity effects as important outcomes of interventions and may therefore choose an evaluation function like the PQALY where health status and productivity of individuals enter the evaluation function multiplicatively. This implies according to Theorem 5 that one of the values applied by the social planner in this situation is that increases in lifetime among unproductive individuals following an intervention (axiom TIUP) will not increase the desirability of that intervention. The social planner subscribes to two further values regarding invariance between different effect components where the first states that if an intervention leads to extra life years, it does not matter to the social planner which particular individual (among individuals with the same level of health and productivity) receives these extra life years (axiom TICHP). The second value specifies that if an intervention leads to improved productivity, it does not matter which particular individual (among individuals with the same level of health and lifespan) is able to perform better in the workplace (axiom PICHT).Our next result states that dismissing the TIUP axiom in the previous statement we obtain the following alternative intriguing compromise (dubbed QALY-PQALY) which evaluates distributions by means of a convex combination of the QALYs and PQALYs the distribution yields. Formally,where q, r : A \u2192 [0, 1] are health state quality weight functions satisfying 0 \u2264 q(a i) \u2264 q(a *) = 1 and 0 \u2264 r(a i) \u2264 r(a *) = 1 for each a i \u2208 A, and \u03b4 \u2208 [0, 1].The QALY-PQALY evaluation function (The parameter \u03b4 measures the relative importance that the social planner puts on pure health effects and productivity-and-quality adjusted life years resulting from an intervention.The following statements are equivalent:1. is represented by a QALY-PQALY evaluation function (6).Similar to the previous evaluation function, a social planner choosing the QALY-PQALY evaluation function (6) considers both health effects and productivity effects as important outcomes of interventions. In contrast to the previous theorem, Theorem 6 implies a rejection of axiom TIUP according to which improvements in lifetime among unproductive individuals following an intervention do not increase the

desirability of that intervention. Apart from that, the social planner subscribes to the same two axioms regarding invariance between different effect components as above (TICHP and PICHT).The previous family of evaluation functions (6) includes a natural sub-family of evaluation functions (QALY-PALY) that evaluate distributions by means of the convex combinations of the QALYs and PALYs that the distribution yields. Formally,where q : A \u2192 [0, 1] is a health state quality weight function satisfying 0 \u2264 q(a i) \u2264 q(a *) = 1, for each a i \u2208 A, and \u03c3 \u2208 [0, 1].Health effects and productivity effects enter the QALY-PALY evaluation function This family of evaluation functions (PICT: For each d \u2208 D, each pair i, j \u2208 N with t i = t j = t, and each c > 0 such thatThe following statements are equivalent:1. is represented by a QALY-PALY evaluation function all members of the family if and only if \u03b4 < 1 3 . As for the family (where But the family of evaluation functions (8) can also accommodate other evaluation functions that have not been introduced above. For instance, suppose that w is a semimultiplicative function in which productivity enters via a power function, whereas health enters via QALYs, i.e., w(a i , p i) = q(a i)p \u03b3 i , for each (a i , p i) \u2208 A \u00d7 [0, 1], where \u03b3 \u2208 (0, 1).where q : A \u2192 [0, 1] is a health state quality weight function satisfying 0 \u2264 q(a i) \u2264 q(a *) = 1, for each a i \u2208 A, and \u03b3 \u2208 (0, 1).Note that the previous evaluation function is formalizing a concern for the dispersion of productivity levels (as it is a concave function of those levels).The above is somewhat reminiscent of the focal welfare function within the literature on life-cycle preferences over consumption and health status. Therein, multiplicative separability from consumption and health is typically assumed. It turns out that the general family of evaluation functions (The following statements are equivalent:1. is represented by an evaluation function (8).Theorem 8 implies that if the social planner endorses the view that if an intervention leads to extra life years, it does not matter which particular individual (among individuals with the same level of health and productivity) receives these extra life years (axiom TICHP), then the evaluation will be via a weighted aggregation of the lifetimes the intervention yields. And the weight for each individual lifetime will be obtained via a general function of the health and productivity levels they face.A weaker axiom than time invariance at common health and productivity, is time invariance at full health and productivity, which states that extra years can be interchangeable among individuals with full health and maximum productivity.TIFHP: For each d \u2208 D, each pair i, j \u2208 N with a i = a j = a * and p i = p j = 1, and each c > 0,If we replace time invariance at common health and productivity by time invariance at full health and productivity we characterize a more general family of evaluation functions, which extend to this context the notion of healthy years equivalent (e.g., More precisely, the healthy productive years equivalent (HPYE) evaluation function evaluates distributions by the unweighted aggregation of HPYEs the distribution yields. Formally,where f : A \u00d7 [0, 1] \u00d7 R + \u2192 R + is continuous with respect to its second and third variables andwhere, for each (a i , p i , tThe evaluation function (Note that this family (9) includes the previous one As the next result states, the general family of evaluation functions (The following statements are equivalent:1. is represented by a HPYE evaluation function (9).Theorem 9 implies that if the social planner endorses the view that if an intervention leads to extra life years for individuals with full health and maximal productivity, it does not matter which particular individual receives these extra life years (axiom TIFHP), then the evaluation will be via the unweighted aggregation of the HPYEs the intervention yields (which are well defined due to COMMON).Finally, we can define the so-called generalized HPYE evaluation function by the unweighted aggregation of the image of HPYEs the distribution yields to a certain function. Formally,where g : R + \u2192 R is a strictly increasing and continuous function, and f :is continuous with respect to its second and third variables and for each dwhere, for each (a i , p i , tOur last result states that the generalized HPYE evaluation function is precisely characterized by the set of COMMON axioms.The following statements are equivalent:1. is represented by a generalized HPYE evaluation function Theorem 10 implies that if the social planner dismisses the TIFHP axiom, while still endorsing COMMON, then the evaluation will be via a general (but strictly increasing and continuous) function of the HPYEs the intervention yields (which are well defined due to COMMON). Thus, the evaluation will not necessarily be via a simple unweighted aggregation of HPYEs, which was the consequence of the TIFHP axiom, as formalized by the evaluation function "
    },
    {
      "header": "Discussion",
      "content": "The results presented above in the form of evaluation functions and their required axioms may be utilised in empirical applications, for example as part of an economic evaluation of a health care or working environment intervention. The data collection of the economic evaluation will typically involve capturing information on all individuals in the intervention and control group regarding their costs, health status and productivity level during the follow-up period of the study.As illustrated throughout the text with the two distributions from Example 1, the choice of evaluation function matters to a large extent when it comes to rank different distributions. Instead of making that choice directly based on

their functional forms, we rather believe the choice should be guided by the axioms they satisfy. Hence the interest of our axiomatic approach.If the analyst (working on behalf of a social planner) is of the view that health effects and productivity effects are both important outcomes when assessing the benefit of the working environment intervention, evaluation functions (Intervention A: 1000 individuals obtaining 1 year in full health and having zero productivity.Intervention B: x individuals obtaining 1 year in health state a and having zero productivity.In particular, each respondent would be asked to identify the number of individuals x in Intervention B to be indifferent between both interventions. The quality weight can then be derived as q(a) = 1000x . The parameter \u03c3 may be elicited using a different version of the person trade-off technique.To wit, respondents would now be asked to state which of the following interventions are most desirable for society: Intervention C: one individual obtaining 1 year in full health and having zero productivity.Intervention D: one individual obtaining y years in health state a and having maximum productivity.Each respondent would then be asked to identify the duration y in Intervention D to be indifferent between the two interventions. Once y is identified, it follows from evaluation function the need for using a more flexible evaluation function (for example the QALY-PQALY). If people largely agree with this axiom (i.e., a majority is largely indifferent between adding productivity to one type or another, or roughly participants are split into those that favor persons of one type and those that favor persons of the other type) it provides support for using the QALY-PALY. 15 The reader is referred to "
    },
    {
      "header": "Concluding remarks",
      "content": "Some of our models, particularly the more general functional forms (They can also be seen as prioritarian evaluation functions (also known as prioritarian social welfare functionals), which rank well-being vectors according to the sum of a strictly increasing (dubbed equivalent health-adjusted lifespan) that is sensitive to inequality in both age-specific health and health-adjusted lifespan. It is a life years metric that nests health-adjusted life expectancy.In the present framework considering health states, productivity and life years, we could also impose a power transformation of the individual components in each of the structured evaluation functions (1)-(have no effects on (labour market) productivity. This might render QALYs more appropriate for the evaluation of these treatments. But using the same evaluation function for younger patients misses productivity effects. This is a possible motivation for age weights (another motivation is the fair innings argument mentioned above). Our hybrid evaluation functions, such as PQALYs, QALYs-PALYs, QALYs-PQALYs or the more general functional form We conclude mentioning that our framework allows for alternative plausible interpretations, as well as for further generalizations.Regarding the former, we focused on chronic health states, for ease of exposition, but our analysis also allows to consider time-varying health. To wit, we made no assumptions regarding the domain of health states A (except for the existence of a maximal element). In particular, this allows for time trajectories rather than fixed levels of health (with the trajectory determined by t i). That is, a i = a i (\u2022), where a i (s) denotes the health status of individual i at time s \u2264 t i .This would require a reinterpretation of some of the axioms we considered above. 18   Likewise, instead of assuming that p i \u2208 [0, 1] captures the productivity of individual i, we could assume that it captures the probability to succeed in life that individual i has. Mariotti and Veneziani (2018) refer to this as \"chances of success\" and characterize a multiplicative form to evaluate social profiles of \"chances of success\" (also known as \"boxes of life\"). 19 We could also derive their characterization results in our model upon endorsing first the axiom of health independence and a counterpart axiom of time independence (not considered in this paper). We, 17 A natural way to start would be by modifying the TICHP axiom, so as to prefer to give the extra life years to the shorter-lived individual. 18 Bossert and D'Ambrosio (2013) is a nice example of an axiomatic analysis of time trajectories (streams) of wealth (rather than health). 19 See also nevertheless, acknowledge that their main axiom is one formalizing a non-interference principle that we do not consider in this paper. 20   As for further generalizations, we stress that the scope of our theory can be enlarged to account for more general evaluations. To wit, our theory deals with the evaluation of population distributions where individuals can be characterized by two instantaneous attributes (one qualitative and one quantitative) and a duration. These attributes can indeed be interpreted as health (qualitative) and productivity (quantitative), as we do in this paper. But there are other potential interpretations (such as happiness or well-being, to name a few). Our results could therefore provide interesting lessons for those settings too.where 0 First, we prove that for each d \u2208 D and each i \u2208 N, there existsIf t i = 0, then it follows from ZERO that t * i = t i = 0. Therefore, assume t i > 0. We prove that t * i exists by contradiction. Therefore, assume that t * i does not exist. Then, T = A \u222a B, whereBy FHPS, [(a * , 1, t i), d N \\{i}] d, implying that either t * i = t i (a contradiction), or t i \u2208 B. Assume the latter. Thus, B is a non-empty set.By PLD and ZERO, it follows that either t * i = 0 (a contradiction), or 0 \u2208 A. Again, assume the latter. Thus, A is a non-empty set.By

```
CONT, A and B are open sets relative to T . Altogether, it follows that T is not a connected
set, which is a contradiction. Thus, t * i exists, and due to LMFHP, it is uniquely
determined. Finally, by SEP, we can determine each t * i separately. Therefore, let f i :
A\u00d7[0, 1]\u00d7R + \u2192 R be such that f i (a i , p i , t i ) = t * i for each i \u2208
N. By ANON, f i () = f j () = f () for each i, j \u2208 N. By CONT, f is continuous with
respect to its second and third variable and, by the above, we know that 0 \u2264 f (a i , p i
, t i ) \u2264 t i , so the range of f is a connected subset of R. Also, by FHPS, f (a i , p i
, t i ) \u2264 f (a i , 1, t i ), andwhich implies that social preferences only depend on the
profile of HPYEs, and, by CONT, they do so continuously. As in the models of where g : R +
\u2192 R is strictly increasing.Suppose first that is represented by a PHEF satisfying
andThus,Conversely, assume now that preferences satisfy all the axioms in COMMON as well as
TIFHP. Then, by Theorem 10, for each pair d, d \u2032 \u2208 D,where g : R + \u2192 R is
strictly increasing. Now, for each pair t i , t j \u2208 R + , and for each c > 0, it follows
by TIFHP that g(f (a * , 1, t i + c)) + g(f (a * , 1, t j )) = g(f (a * , 1, t i )) + g(f (a *
, 1, t j + c)).Or, equivalently,In particular,for each x, y \u2265 0. As g is continuous and
strictly increasing, it follows from Theorem 1 in Suppose first that is represented by a PHEF
satisfying Then, by Theorem 10, for each pair d, d \u2032 \u2208 D,where = \u03d5(a, p,=
\u03d5(a, p,In particular, \u03d5(a, p, t \u2192 R be such that w(a, p) = \u0175(a,p) \u0175(a
* ,1) , for each (a, p) \u2208 A \u00d7 [0, 1]. Thus, it follows that 1 = w(a * , 1) \u2265
w(a, p) \u2265 0, for each (a, p) \u2208 A \u00d7 [0, 1]. Then, we may write: \u03d5(a, p, t i
) = w(a, p)t i , where 0 \u2264 w(a, p) \u2264 w(a * , p) \u2264 w(a * , 1) = 1, and 0 \u2264
w(a, p) \u2264 w(a, 1) \u2264 w(a * , 1) = 1 for each (a, p) \u2208 A \u00d7 [0, 1], as
desired.Suppose first that is represented by a PHEF satisfying Then, for each c >
0,Conversely, assume now that preferences satisfy all the axioms in the statement of Theorem
where g : R + \u2192 R is strictly increasing. PI and TICHFP together imply TICHP. Thus,
preferences satisfy all the axioms in the statement of Theorem 8. Thus, for each d \u2208
D,where 0 \u2264 w(a, p) \u2264 w(a * , p) \u2264 w(a * , 1) = 1, and 0 \u2264 w(a, p) \u2264
w(a, 1) \u2264 w(a * , 1) = 1 for each (a, p) \u2208 A \u00d7 [0, 1]. Now, by PI, w(a, p) =
w(a, 1), for each a \u2208 A. Let q : A \u2192 R be such that q(a i ) = w(a i , 1), for each a
i \u2208 A. Then, it follows that 1 = q(a * ) \u2265 q(a) \u2265 0, for each a \u2208 A.And we
may write:where 0 \u2264 q(a) \u2264 q(a * ) = 1, for each a \u2208 A, as desired.Suppose
first that is represented by a PHEF satisfying Then, for each c > 0,Thus, as v( and each c > 0
such that p i + c, p j + c \u2208 [0, 1]. In particular, Suppose first that is represented by
a PHEF satisfying Then, for each c > 0 such that p i + c, p j + c \u2264 1, E[(a, p i + c, t),
(a, p j , t), d N \\{i,j} ] = q(a)(p i + c)t + q(a)p j t + k\u2208N \\{i,j} q(a k )p k t k ,
and E[(a, p i , t), (a, p j + c, t), d N \\{i,j} ] = q(a)p i t + q(a)p j (t + c) + k\u2208N
\\{i,j} q(a k )p k t k .Thus, [(a, p i + c, t), (a, p j , t), d N \\{i,j} ] \u223c [(a, p i ,
t), (a, p j + c, t), d N \\{i,j} ].Conversely, assume now that preferences satisfy all the
axioms in the statement of Theorem 5. Then, they also satisfy the axioms of Theorem 8. Thus,
for each pair d, d \u2032 \u2208 D, and, therefore, is indeed represented by an evaluation
function satisfying (5), as desired.Suppose first that is represented by a PHEF satisfying
Conversely, assume now that preferences satisfy all the axioms in the statement of Theorem for
each a \u2208 A, it follows that is represented by an evaluation function satisfying Case 3.
\u03b1(a * ) = 0 = \u03b2(a * ).Let q, r : A \u2192 R be such that r(a) = \u03b1(a) \u03b1(a *
) and q(a) = \u03b2(a) \u03b2(a * ) , for each a \u2208 A. Then, by FHPS, 1 = q(a * ) \u2265
q(a) \u2265 0, and 1 = r(a * ) \u2265 r(a) \u2265 0. Furthermore, 0 < \u03b1(a * ), \u03b2(a *
) < 1. Now, if we let \u03b4 = \u03b2(a * ) \u2208 (0, 1), we have w(a, p) = \u03b4q(a) +
(1\u03b4)r(a)p, for each p \u2208 [0, 1], and each a \u2208 A. Thus, for each pair d, d \u2032
\u2208 D,where q, r : A \u2192 [0, 1] are such that 1 = q(a * ) \u2265 q(a) \u2265 0, and 1 =
r(a * ) \u2265 r(a) \u2265 0 and \u03b4 \u2208 (0, 1). Consequently, is indeed represented by
an evaluation function satisfying Suppose first that is represented by a PHEF satisfying where
q, r : A \u2192 [0, 1] are such that 1 = q(a * ) \u2265 q(a) \u2265 0, and 1 = r(a * ) \u2265
r(a) \u2265 0 andBy PICT, [(a i , p i + c, t), (a j , p j , t), d N \\{i,j} ] \u223c [(a i , p
i , t), (a j , p j + c, t), d N \\{i,j} ],for each d \u2208 D, and i, j \u2208 N with t i = t
j = t. Now, as E[(a i , p i + c, t), (a j , p j , t), d N \\{i,j} ] = \u03b4(q(a i )t + q(a j
)t) + (1\u03b4)(r(a i )(p i + c)t + r(a j )p j t) + k\u2208N \\{i,j} \u03b4q(a k )t k +
(1\u03c3)r(a k )p k t k , and E[(a i , p i , t), (a j , p j + c, t), d N \\{i,j} ] =
\u03b4(q(a i )t + q(a j )t) + (1\u03b4)(r(a i )p i t + r(a j )(p j + c)t) + k\u2208N \\{i,j}
\u03b4q(a k )t k + (1\u03c3)r(a k )p k t k . Thus, r(a i ) = r(a j ), for each pair a i , a j
\u2208 A. As, 1 = r(a * ), it follows that r(a) = 1, for each a \u2208 A. Thus, letting \u03c3
= \u03b4, we obtain that is indeed represented by an evaluation function satisfying "
    }
  ]
}
```


Figure 0-1 – Segmentation full result