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

[ACKNOWLEGMENTS 3](#_Toc168735954)

[TABLE OF CONTENTS 4](#_Toc168735955)

[LIST OF FIGURES 6](#_Toc168735956)

[LIST OF TABLES 7](#_Toc168735957)

[ABSTRACT 8](#_Toc168735958)

[Chapter 1 INTRODUCTION 9](#_Toc168735959)

[1.1 Background 9](#_Toc168735960)

[1.2 Problem Statement 9](#_Toc168735961)

[1.3 Scope and Objectives 9](#_Toc168735962)

[1.4 Structure of thesis 9](#_Toc168735963)

[Chapter 2 LITURATURE REVIEW 11](#_Toc168735964)

[2.1 Overview of Automatic Document Segmentation 11](#_Toc168735968)

[2.1.1 GROBID 12](#_Toc168735969)

[2.2 Overview of Natural Language Processing 13](#_Toc168735972)

[2.3 Overview of Text Summarization 15](#_Toc168735973)

[2.3.1 Extractive text summarization 16](#_Toc168735974)

[2.3.2 Abstractive text summarization 17](#_Toc168735975)

[2.3.2.1 Machine-learning-based Summary Generation 18](#_Toc168735976)

[Chapter 3 METHODOLOGY 19](#_Toc168735977)

[3.1 Overview 19](#_Toc168735978)

[3.2 System design 19](#_Toc168735979)

[3.2.1 Data format 20](#_Toc168735980)

[3.2.2 Document Segmentation Subsystem 22](#_Toc168735981)

[3.2.3 Text Summarization Subsystem 22](#_Toc168735982)

[Chapter 4 IMPLEMENT AND RESULTS 24](#_Toc168735983)

[4.1 Overview 24](#_Toc168735984)

[4.2 Implementation 24](#_Toc168735985)

[4.2.1 Document Segmentation 24](#_Toc168735986)

[4.2.2 Text summarization 26](#_Toc168735987)

[4.2.2.1 LSA 26](#_Toc168735988)

[4.2.2.2 TextRank 27](#_Toc168735989)

[4.3 Results 27](#_Toc168735990)

[Chapter 5 DISCUSSION AND EVALUATION 31](#_Toc168735991)

[5.1 Evaluation 31](#_Toc168735992)

[Chapter 6 CONCLUSION AND FUTURE WORK 32](#_Toc168735993)

[6.1 Conclusion 32](#_Toc168735994)

[6.2 Future work 32](#_Toc168735995)

[REFERENCES 33](#_Toc168735996)

[APPENDIX 34](#_Toc168735997)

# LIST OF FIGURES

[Figure 2‑1 - GROBID cascade of sequencing model [2] 12](#_Toc156380825)

[Figure 2‑2 - Dependency style parse tree [3] 15](#_Toc156380826)

[Figure 2‑3 - Extractive text summarization general framework [4] 17](#_Toc156380827)

[Figure 2‑4 - Abstractive text summarization general framework [4] 18](#_Toc156380828)

[Figure 3‑1 - General structure of the application 19](#_Toc156380829)

[Figure 3‑2 – The component diagram of the system 21](#_Toc156380830)

[Figure 3‑3 – The application’s activity diagram 23](#_Toc156380831)

[Figure 4‑1 – GROBID initialization code snippet 25](#_Toc156380832)

[Figure 4‑2 – PDF segmentation code snippet 25](#_Toc156380833)

[Figure 4‑3 – 5 first step of the process 26](#_Toc156380834)

[Figure 4‑4 – The last 2 step of the LSA summarization process 27](#_Toc156380835)

[Figure 4‑5 – TextRank summarization function 27](#_Toc156380836)

[Figure 4‑6 – Segmentation result 29](#_Toc156380837)

[Figure 4‑7 – TextRank summarization result 29](#_Toc156380838)

[Figure 4‑8 – LSA summarization result 30](#_Toc156380839)

[Figure 7‑1 – Segmentation full result 38](#_Toc156380840)

[Figure 7‑2 – TextRank summarization full result 38](#_Toc156380841)

[Figure 7‑3 – LSA summarization full result 39](#_Toc156380842)

# LIST OF TABLES

**No table of figures entries found.**

# 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.

# INTRODUCTION

## 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 said tools is the focus of the thesis.

## 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.

## 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.

## 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.

# LITURATURE REVIEW



## 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.

### GROBID

GROBID 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.

A diagram of a reference

Description automatically generated

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

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.



## 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 [3]. 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

A diagram of a tree

Description automatically generated

Figure 2‑2 - Dependency style parse tree [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)[4]. The framework’s visualization can be seen in Figure 2.3-1

A diagram of a process

Description automatically generated

Figure 2.3‑1 – Text summarization general framework [4]

### 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) [5], and Latent Semantic Analysis (LSA) (which made use of Single Value Decomposition technique to analyze the relationship between documents) [6]

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

Diagram of a process

Description automatically generated

Figure 2‑3 - Extractive text summarization general framework [4]

### 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

A diagram of a process

Description automatically generated

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

#### 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**) [7], or **T**ext-**t**o-**T**ext **T**ransfer **T**ransformer (**T5**) [8] 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 [9]. Inspired by how humans acquire new knowledge by applying and adapting old one, AI model are trained in a two-phase framework [10], 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 [11] (a T5-based model), the Facebook’s bart-large-cnn model [12] (a BART-based model).

# METHODOLOGY

## 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-preprocessing 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.

A diagram of a document segmentation

Description automatically generated

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.

## 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.

### 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.

A diagram of a computer system

Description automatically generated

Figure 3‑2 – The component diagram of the system

### 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, 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.

### Text Summarization Subsystem

Due to its ability to summarize text using different summarization techniques, a method to indicate the user’s need for a certain approach is required. As such, a flag of is provided when starting the program that enables user to do so. User can either choose to include “-a lsa” or “-a textrank” at the end of the command line to toggle between each approach. The provided flag is then processed in the report-gen.sh file to direct the flow of data. Data is first processed through the Document Segmentation Subsystem before being summarized. If LSA was chosen, then data is then flowed through the SegmentSummarization where LSA will be applied to each segment. If, on the other hand, TextRank was chosen, data will move through textrank.py, where it will be transformed using the TextRank implementation from the summa library. After which, in either case, the summary will be stored in a YAML file with a similar format to that of the segmentation output file. The general flow of the system through the eye of the user can be seen from the activity diagram in Figure 3-3

A diagram of a document

Description automatically generated

Figure 3‑3 – The application’s activity diagram

# IMPLEMENT AND RESULTS

## 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

## 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.

### Document Segmentation

As GROBID is a machine learning approach to the segmentation problem, its initialization requires the initialization of the GROBID container. To. 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

The next important step of the process is to partition the PDF files into multiple segments, this can be seen in Figure 4-2

public static **String** **pdfSegmenting**(**File** pdfFile) {

**String** result;

**String** tei = "";

    try {

        tei = engine.**fullTextToTEI**(pdfFile, analysisConfig);

    } catch (**Exception** e) {

        e.**printStackTrace**();

    }

    result = tei;

    return result;

}

Figure 4‑2 – PDF segmentation code snippet

### 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.

#### 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. Since each step has been implemented as a function, the summarization step only needs to invoke them one by one in the established order. This can be seen in the code snippet of Figure 4-3, with the exception of the word searching and summarization.

ngram.**generate**(list, 1);

ngram.**filterStopWords**();

ngram.**sort**();

ngram.**getSentenceUsingModel**(text);

Figure 4‑3 – 5 first step of the process

The summary is then generated by using the output from the steps above. This can be seen in Figure 4-4

for (int i = 0; i <= Ap.**getRowDimension**() - 1; i++) {

        for (int j = 0; j <= Ap.**getColumnDimension**() - 1; j++) {

            if (Ap.**getEntry**(i, j) >= **maxEntry**(Ap) / 2.5

&& !setSummarySentences.**contains**(ngram.**getSentences**()[j])) {

                setSummarySentences.**add**(ngram.**getSentences**()[j]);

            }

            if (setSummarySentences.**size**() == k) {

                break;

            }

        }

    }

    for (**String** string : setSummarySentences) {

        summary = summary.**concat**(" ").**concat**(string).**trim**();

    }

    return summary;

}

Figure 4‑4 – The last 2 step of the LSA summarization process

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

#### 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):

    return summarize(

        text=text,

        ratio=0.2,

        words=50,

        additional\_stopwords=**get\_stop\_word\_list**()

    )

Figure 4‑5 – TextRank summarization function

In said function, “text” refers to each PDF segment that the user wants to summarize. “ratio” refers to the proportion of the text when comparing the final summary to the original text. For instance, if the ratio is 0.2 and the total word count in a specific text is 500, then the summary will have a max limit of 100 words. “words” is the lower limit of the resulting summary and “additional\_stopwords” refers to the list of words that need to be filtered out before processing the summary.

## Results

The results from the process of the entire system can be condensed into 2 different YAML files with the first being the document segmentation result and the second being the summary. As an example, one research paper named “Using Text Mining Techniques for Extracting Information from Research Article” [13] has been chosen.

As the result of the document segmentation being the content of the entire paper, only the abstract and the first header will be showcased in Figure 4-6. The remaining result can be seen in the Appendix section.

abstractSeg: "Nowadays, research in text mining has become one of the widespread fields\

  \ in analyzing natural language documents. The present study demonstrates a comprehensive\

  \ overview about text mining and its current research status. As indicated in the\

  \ literature, there is a limitation in addressing Information Extraction from research\

  \ articles using Data Mining techniques. The synergy between them helps to discover\

  \ different interesting text patterns in the retrieved articles. In our study, we\

  \ collected, and textually analyzed through various text mining techniques, three\

  \ hundred refereed journal articles in the field of mobile learning from six scientific\

  \ databases, namely: Springer, Wiley, Science Direct, SAGE, IEEE, and Cambridge.\

  \ The selection of the collected articles was based on the criteria that all these\

  \ articles should incorporate mobile learning as the main component in the higher\

  \ educational context. Experimental results indicated that Springer database represents\

  \ the main source for research articles in the field of mobile education for the\

  \ medical domain. Moreover, results where the similarity among topics could not\

  \ be detected were due to either their interrelations or ambiguity in their meaning.\

  \ Furthermore, findings showed that there was a booming increase in the number of"

sectionList:

- header: "Introduction"

  content: "Nowadays, almost all of the existing information in different institutions\

    \ (e.g. government, business, industry, and others) is preserved in electronic\

    \ documents in which it contains semi-structured data. In these documents, the\

    \ \"abstract\" is an example of unstructured text component. Whereas, examples\

    \ of structured fields in a document are: author's name, publication date, title,\

    \ and category [1]. A study by [2] stated that text mining has become one of the\

    \ trendy fields that has been incorporated in several research fields such as\

    \ computational linguistics, Information Retrieval (IR) and data mining. Text\

    \ mining is different from data mining [3]. Data mining is focused on discovering\

    \ interesting patterns from large databases rather than textual information [4].\

    \ Information recovery methodologies like text indexing techniques have been developed\

    \ for handling unstructured documents. In conventional researches, it is assumed\

    \ that a user mostly searches for known terms, which have been previously used\

    \ or written by someone else. The main problem is that the search results are\

    \ not relevant to the user's requirements. One solution is to use text mining\

    \ in order to find out relevant information, which is not indicated explicitly\

    \ nor written down so far. The procedure of text mining begins with gathering\

    \ documents through different resources. A particular document would be recovered\

    \ through text mining instrument and by checking its format and character sets;\

    \ it will be pre-processed by this instrument. The document would then pass through\

    \ a text analysis stage. Text analysis includes semantic analysis intended to\

    \ obtain high-quality information through text. Different text analysis methods\

    \ are available. Different methods can be used based on the organization's objective.\

    \ In some cases, text analysis methods are repeated until information is extracted.\

    \ The outcomes can be stored in a management information system that provides\

    \ a large amount of significant information for the user of that system.Text mining\

    \ intends to detect the information that was not recognized before through extracting\

    \ it automatically from various text-based sources. Structured data can be handled\

    \ through data mining tools while unstructured or semi-structured datasets like\

    \ full-text documents, emails, and HTML files can be handled through text mining.\

    \ Typically, the information will be kept in a natural form known as text. Text\

    \ mining is not similar to web mining. When something is explored on the web by\

    \ the user, it means that it is previously known and it was written by someone\

    \ else [5]. For example, in E-commerce, a major issue with web mining is buying\

    \ all the materials which are not relevant to the user's search and it will not\

    \ show unknown (hidden or implicit) information, while the major objective of\

    \ text mining is to find out the unknown information [6]; something that is not\

    \ recognized by anyone.Data is the basic kind of information, which is required\

    \ to be organized and mined for the knowledge generation. Discovering patterns\

    \ and trends from huge data is a significant challenge. Finding out the unknown\

    \ trends and patterns from databases properly is a major objective of data mining.\

    \ It is a method where data pre-processing is necessary before applying any other\

    \ method. Many approaches like clustering, classification, and decision trees\

    \ are involved in data mining. All the textual based information is stored by\

    \ electronic means, either on a client's personal computers or on a web server.\

    \ Due to the increasing growth in hardware storage devices, any computer or laptop\

    \ has the ability to store an enormous amount of data. Creating new information\

    \ can be simple while finding out relevant information from a huge amount of data\

    \ is challenging. In order to extract the relevant information, knowledge, or\

    \ patterns from various sources that are in unstructured form, text mining technique\

    \ can be employed. The common structure of text mining involves two consecutive\

    \ stages: text refining and knowledge distillation. In text refining, free-form\

    \ text documents are converted into an intermediate form, whereas in knowledge\

    \ distillation, patterns or knowledge are derived from intermediate form. Intermediate\

    \ form (IF) can be either semi-structured like the theoretical graph illustration\

    \ or structured like the relational data illustration. IF can be either a document-based\

    \ where every entity symbolizes document, it can be a concept-based where every\

    \ unit symbolizes an object or a concept of interest in a particular area.Various\

    \ research areas, techniques, and models are involved in different research domains.\

    \ The hottest topics of the research domains are the primary focus of many research\

    \ papers. The research results of a particular domain may influence other research\

    \ domains since some research domains may have similar topics. These research\

    \ topics always discuss such a promising research area that is worth studying.\

    \ Therefore, the trend of cross domain is determined in this research. The longitudinal\

    \ trends of academic articles in Mobile Leaning (ML) were explored in this research\

    \ with the help of text mining methods. We recovered and examined (300) refereed\

    \ journal articles and conference proceedings from various authentic databases.The\

    \ primary goals of this research are (1) Using text mining techniques for identifying\

    \ the topics of a scientific text related to ML research and developing a hierarchical\

    \ and evolutionary connection among these topics. (2) Using visualization tools\

    \ for presenting both the topics and the association among them as a convenient\

    \ way to help users to determine relevant topics.This paper is categorized as\

    \ follows: Sect. 2 provides an inclusive background concerning in the text mining\

    \ field. Other related studies are addressed by Sect. 3. Research methodology\

    \ is presented in Sect. 4. The results are demonstrated in Sect. 5. Conclusion\

    \ and future perspectives are presented in Sect. 6."

Figure 4‑6 – Segmentation result

The corresponding summary of the first 2 sections using TextRank can be found in Figure 4-7.

abstractSeg: 'In our study, we collected, and textually analyzed through various text

  mining techniques, three hundred refereed journal articles in the field of mobile

  learning from six scientific databases, namely: Springer, Wiley, Science Direct,

  SAGE, IEEE, and Cambridge.

  Experimental results indicated that Springer database represents the main source

  for research articles in the field of mobile education for the medical domain.'

sectionList:

- content: For example, in E-commerce, a major issue with web mining is buying all

    the materials which are not relevant to the user's search and it will not show

    unknown (hidden or implicit) information, while the major objective of text mining

    is to find out the unknown information [6]; something that is not recognized by

    anyone.Data is the basic kind of information, which is required to be organized

    and mined for the knowledge generation.

  header: Introduction

Figure 4‑7 – TextRank summarization result

A summarization made using the LSA technique can be seen in Figure 4-8.

abstractSeg: "The present study demonstrates a comprehensive overview about text mining\

  \ and its current research status. In our study, we collected, and textually analyzed\

  \ through various text mining techniques, three hundred refereed journal articles\

  \ in the field of mobile learning from six scientific databases, namely: Springer,\

  \ Wiley, Science Direct, SAGE, IEEE, and Cambridge."

sectionList:

- header: "Introduction"

  content: "A study by [2] stated that text mining has become one of the trendy fields\

    \ that has been incorporated in several research fields such as computational\

    \ linguistics, Information Retrieval (IR) and data mining. The outcomes can be\

    \ stored in a management information system that provides a large amount of significant\

    \ information for the user of that system.Text mining intends to detect the information\

    \ that was not recognized before through extracting it automatically from various\

    \ text-based sources."

Figure 4‑8 – LSA summarization result

# DISCUSSION AND EVALUATION

## Evaluation

As this project is currently being a prototype, no evaluation method has been made when comparing the accuracy of the text summarization approaches that has been made. This, however, is part of what driven the development of both methods since the evaluation is an important step in determining which methods is more suitable for the task required of the application itself. As such, the evaluation and comparison are part of the future development plan.

# CONCLUSION AND FUTURE WORK

## Conclusion

Although the project has not been fully implemented, it is currently capable of producing a readable and comprehensible summary, albeit, in a form factor that is not user-friendly. Throughout the process of compiling this paper, we have successfully implemented a pipeline that is capable of receiving PDF files from users, segmenting the content in each file of the folder through the use of GROBID and summarizing it depending on the needs of the user. Each of the output from these approaches are then condensed and stored in a human-readable format that is YAML.

## Future work

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

1. Creating a user interface that can be used in tandem with an application that is used for reference management. One example that can be thought of is the creation of a plugin that can be used to automate the generation of a summary in an application that manages references. Zotero is a potential candidate for this approach since it is open sourced and highly support plugins.
2. Evaluations methods for comparing each approach is essential to the finalization of the project. This is due to the need to choose the most appropriate summarization techniques for this project.
3. Since both techniques in the summarization system are extractive approaches, a subtractive approach could be implemented in the future as a means to improving the quality of the final summarization result. Obviously, the result should also be evaluated with other methods.

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

[3] 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.

[4] 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.

[5] 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

[6] 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.

[7] 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.

[8] 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

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

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

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

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

[13] “Using Text Mining Techniques for Extracting Information from Research Articles | SpringerLink.” Accessed: Jan. 17, 2024. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-319-67056-0\_18

# APPENDIX

The full result from the Segmentation process

---

abstractSeg: "Nowadays, research in text mining has become one of the widespread fields\

  \ in analyzing natural language documents. The present study demonstrates a comprehensive\

  \ overview about text mining and its current research status. As indicated in the\

  \ literature, there is a limitation in addressing Information Extraction from research\

  \ articles using Data Mining techniques. The synergy between them helps to discover\

  \ different interesting text patterns in the retrieved articles. In our study, we\

  \ collected, and textually analyzed through various text mining techniques, three\

  \ hundred refereed journal articles in the field of mobile learning from six scientific\

  \ databases, namely: Springer, Wiley, Science Direct, SAGE, IEEE, and Cambridge.\

  \ The selection of the collected articles was based on the criteria that all these\

  \ articles should incorporate mobile learning as the main component in the higher\

  \ educational context. Experimental results indicated that Springer database represents\

  \ the main source for research articles in the field of mobile education for the\

  \ medical domain. Moreover, results where the similarity among topics could not\

  \ be detected were due to either their interrelations or ambiguity in their meaning.\

  \ Furthermore, findings showed that there was a booming increase in the number of"

sectionList:

- header: "Introduction"

  content: "Nowadays, almost all of the existing information in different institutions\

    \ (e.g. government, business, industry, and others) is preserved in electronic\

    \ documents in which it contains semi-structured data. In these documents, the\

    \ \"abstract\" is an example of unstructured text component. Whereas, examples\

    \ of structured fields in a document are: author's name, publication date, title,\

    \ and category [1]. A study by [2] stated that text mining has become one of the\

    \ trendy fields that has been incorporated in several research fields such as\

    \ computational linguistics, Information Retrieval (IR) and data mining. Text\

    \ mining is different from data mining [3]. Data mining is focused on discovering\

    \ interesting patterns from large databases rather than textual information [4].\

    \ Information recovery methodologies like text indexing techniques have been developed\

    \ for handling unstructured documents. In conventional researches, it is assumed\

    \ that a user mostly searches for known terms, which have been previously used\

    \ or written by someone else. The main problem is that the search results are\

    \ not relevant to the user's requirements. One solution is to use text mining\

    \ in order to find out relevant information, which is not indicated explicitly\

    \ nor written down so far. The procedure of text mining begins with gathering\

    \ documents through different resources. A particular document would be recovered\

    \ through text mining instrument and by checking its format and character sets;\

    \ it will be pre-processed by this instrument. The document would then pass through\

    \ a text analysis stage. Text analysis includes semantic analysis intended to\

    \ obtain high-quality information through text. Different text analysis methods\

    \ are available. Different methods can be used based on the organization's objective.\

    \ In some cases, text analysis methods are repeated until information is extracted.\

    \ The outcomes can be stored in a management information system that provides\

    \ a large amount of significant information for the user of that system.Text mining\

    \ intends to detect the information that was not recognized before through extracting\

    \ it automatically from various text-based sources. Structured data can be handled\

    \ through data mining tools while unstructured or semi-structured datasets like\

    \ full-text documents, emails, and HTML files can be handled through text mining.\

    \ Typically, the information will be kept in a natural form known as text. Text\

    \ mining is not similar to web mining. When something is explored on the web by\

    \ the user, it means that it is previously known and it was written by someone\

    \ else [5]. For example, in E-commerce, a major issue with web mining is buying\

    \ all the materials which are not relevant to the user's search and it will not\

    \ show unknown (hidden or implicit) information, while the major objective of\

    \ text mining is to find out the unknown information [6]; something that is not\

    \ recognized by anyone.Data is the basic kind of information, which is required\

    \ to be organized and mined for the knowledge generation. Discovering patterns\

    \ and trends from huge data is a significant challenge. Finding out the unknown\

    \ trends and patterns from databases properly is a major objective of data mining.\

    \ It is a method where data pre-processing is necessary before applying any other\

    \ method. Many approaches like clustering, classification, and decision trees\

    \ are involved in data mining. All the textual based information is stored by\

    \ electronic means, either on a client's personal computers or on a web server.\

    \ Due to the increasing growth in hardware storage devices, any computer or laptop\

    \ has the ability to store an enormous amount of data. Creating new information\

    \ can be simple while finding out relevant information from a huge amount of data\

    \ is challenging. In order to extract the relevant information, knowledge, or\

    \ patterns from various sources that are in unstructured form, text mining technique\

    \ can be employed. The common structure of text mining involves two consecutive\

    \ stages: text refining and knowledge distillation. In text refining, free-form\

    \ text documents are converted into an intermediate form, whereas in knowledge\

    \ distillation, patterns or knowledge are derived from intermediate form. Intermediate\

    \ form (IF) can be either semi-structured like the theoretical graph illustration\

    \ or structured like the relational data illustration. IF can be either a document-based\

    \ where every entity symbolizes document, it can be a concept-based where every\

    \ unit symbolizes an object or a concept of interest in a particular area.Various\

    \ research areas, techniques, and models are involved in different research domains.\

    \ The hottest topics of the research domains are the primary focus of many research\

    \ papers. The research results of a particular domain may influence other research\

    \ domains since some research domains may have similar topics. These research\

    \ topics always discuss such a promising research area that is worth studying.\

    \ Therefore, the trend of cross domain is determined in this research. The longitudinal\

    \ trends of academic articles in Mobile Leaning (ML) were explored in this research\

    \ with the help of text mining methods. We recovered and examined (300) refereed\

    \ journal articles and conference proceedings from various authentic databases.The\

    \ primary goals of this research are (1) Using text mining techniques for identifying\

    \ the topics of a scientific text related to ML research and developing a hierarchical\

    \ and evolutionary connection among these topics. (2) Using visualization tools\

    \ for presenting both the topics and the association among them as a convenient\

    \ way to help users to determine relevant topics.This paper is categorized as\

    \ follows: Sect. 2 provides an inclusive background concerning in the text mining\

    \ field. Other related studies are addressed by Sect. 3. Research methodology\

    \ is presented in Sect. 4. The results are demonstrated in Sect. 5. Conclusion\

    \ and future perspectives are presented in Sect. 6."

- header: "Related Work"

  content: "Many research works contributed to the field of IE through the use of\

    \ various techniques. The primary focus of these researches was to determine how\

    \ different text mining procedures can be utilized as the structured data sets\

    \ exist in the text document format. This part begins with defining the topic\

    \ of the research, evaluating previous researches, and then major techniques are\

    \ applied using information extraction and text mining. In order to determine\

    \ the topic of each research area and to develop an evolutionary and hierarchical\

    \ connection between these topics, [35] used the method of text mining. Topics\

    \ are presented through visualization tools.Moreover, these tools are used in\

    \ order to show the connection between these topics and to offer interactive functions\

    \ so that users can effectively find the cross-domain topics and know the trends\

    \ of cross-domain research. Moloshnikov et al. [36] developed an algorithm for\

    \ finding documents on a particular topic depending on a selected reference collection\

    \ of documents. In addition, the context-semantic graph for visualization themes\

    \ in search results was also developed. The algorithm depends on the incorporation\

    \ of a group of entropic, probabilistic and semantic developers for mining of\

    \ weighted keywords and set of words that explain the specified topic. Results\

    \ indicated that the average precision is 99% and the recall is 84%. A unique\

    \ technique was also created for making graphs on the basis of the algorithm,\

    \ can remove key phrases with weights. It offers the opportunity to show an arrangement\

    \ of sub-topics in huge sets of documents in compact graph form.In order to offer\

    \ a reference for additional researches of other researchers, [37] discussed the\

    \ research status of text mining technology when it was used in the biomedical\

    \ field that covers 10 years. Biomedical text mining literature incorporated in\

    \ SCI from 2004 to 2013 were recovered, filtered and then examined from the viewpoint\

    \ of research institutions, yearly changes, research areas, local distribution,\

    \ journals sources, and keywords. A prominent increase in the amount of worldwide\

    \ biomedical text mining literature is observed. Among this global literature,\

    \ a huge percentage is taken up by literature related to named entity recognition,\

    \ entity relation extraction, text categorization, text clustering, abbreviations\

    \ extraction, and co-occurrence analysis. Studies carried out in USA and UK are\

    \ considered to be present in the primary position.In order to extract inter-language\

    \ clusters through multilingual documents depending on Closed Concepts Mining\

    \ and vector model, a new statistical approach was suggested by [38]. Formal Concept\

    \ Analysis methods are used for mining Closed Concepts from similar corpora and\

    \ later these Closed Concepts and vector models are utilized in the clustering\

    \ and arrangement of multilingual documents. An experimental assessment is carried\

    \ out over a set of French-English bilingual documents of CLEF's 2003. With a\

    \ notable comparability score and in order to remove the bilingual classes of\

    \ documents, results revealed that the interaction between vector model and Formal\

    \ Concept Analysis is very useful.Santosh [39] suggested the graph mining-based\

    \ document content (i.e. text fields) exploitation. That is, the query generated\

    \ the graph depending on the users' requirements. This is an easy and effective\

    \ graph mining method to extract similar patterns through the documents and changed\

    \ the query graph into model graphs which are utilized when the users are not\

    \ present. An intelligent solution for document information exploitation has been\

    \ created. This is characterized by simplicity, ease of use, accuracy, ease of\

    \ development, and flexibility. In order to understand graph models, it does not\

    \ need a huge collection of document images. Moreover, since model learning consumes\

    \ less than 10 s for an input pattern per class on average, changes, amendments,\

    \ and replacements can be done in the input patterns. Information exploitation\

    \ average performance is shown to have 86.64% as Precision, and 90.80% as a Recall.\

    \ However, the suggested technique failed to offer inclusive and accurate solutions\

    \ for the patterns that have a huge collection of fields in a zigzags arrangement\

    \ due to the query graph intricacy.Sirsat et al. [23] proposed two techniques\

    \ for mining text through online sources. The first technique dealt with the knowledge\

    \ that is required to be shown directly in the documents that need to be mined.\

    \ Text mining and IE are considered as the only effective tools for performing\

    \ that technique. The second one concerned with the documents that hold an actual\

    \ data in unstructured format instead of nonfigurative knowledge. IE can help\

    \ to change the unstructured data presented in the document corpus into structured\

    \ one. In order to discover nonfigurative patterns in the extracted data, data\

    \ mining algorithms and techniques can be used.Song and Kim [40] presented the\

    \ first attempt to apply text mining approaches to a huge collection of full-text\

    \ articles for discovering the knowledge structure of the area. Instead of depending\

    \ on the citation data presented in Web of Science, PubMed Central full-text articles\

    \ have been used for bibliometric examination. Above all, this assisted the creation\

    \ of text mining routines in order to develop a custom-made citation database\

    \ following the full-text mining. Findings showed that most of the documents that\

    \ were published in bioinformatics area were not cited by others. Additionally,\

    \ a constant and linear rise has been observed in the amount of publications across\

    \ publication years. Results revealed that the majority of the retrieved studies\

    \ were inspired by USA-based institutes followed by European institutes. Results\

    \ reported that the major primary focus of the important topics was on biological\

    \ factors. However, according to PageRank, the top 10 articles were highly concerned\

    \ with the computational factors.In order to facilitate the accurate extraction\

    \ of text from PDF files of research articles that can be utilized in text mining\

    \ applications, a \"Layout-Aware PDF Text Extraction\" (LA-PDF Text) system was\

    \ presented by [41]. Text blocks are mined from PDF-formatted fill-text research\

    \ articles under this system and then the system categorizes them into logical\

    \ units depending on rules that typify particular sections. Only the textual content\

    \ of the research articles is focused in the LA-PDF Text system. This system serves\

    \ as a basis for new experiments into more developed extraction methods dealing\

    \ with multi-modal content like images and graphs. The system goes through three\

    \ phases: (1) Identifying contiguous text blocks with the help of spatial layout\

    \ processing in order to discover blocks of contiguous text, (2) Categorization\

    \ of text blocks into metaphorical categories with the help of a rule-based method,\

    \ and (3) Joining categorized text blocks together by arranging them accurately\

    \ which results in the extraction of text from section-wise grouped blocks. An\

    \ evaluation of the accuracy of the block discovery algorithm used in step 2 was\

    \ performed. It was also shown that the system can identify and classify them\

    \ into metaphorical categories with Recall = 0.89%, Precision = 0.96%, and F =\

    \ 0.91%. Moreover, the accuracy of the text mined with the help of LA-PDF Text\

    \ is compared to the text from an Open Access subset of PubMed Central. This accuracy\

    \ is then compared with the text that was mined using the PDF2Text system. These\

    \ are the two frequently used techniques to extract text from PDF.Mooney and Bunescu\

    \ [42] described two techniques for using the natural language information extraction\

    \ for text mining. First, general knowledge can be mined directly from the text.\

    \ A project where a knowledge base of 6580 human protein interactions was extracted\

    \ by mining around 750,000 Medline abstracts in which reconsidered as an example\

    \ of this technique. Second, structured data can be mined through text documents\

    \ or web pages. In order to find out patterns in the mined data, traditional KDD\

    \ methods can be applied. The performed work on the DiscoTEX system and its application\

    \ to Amazon book descriptions, computer science job postings, and resumes were\

    \ considered as an example of this technique. In order to discover units and relations\

    \ in text, research in IE keeps on creating more efficient algorithms. Valuable\

    \ and significant knowledge can be mined effectively from the constantly developing\

    \ body of electronic documents and web pages by using modern approaches in human\

    \ language technology and computational linguistics, and linking them with the\

    \ modern techniques used in machine learning and conventional data mining techniques.\

    \ IE deals with determining a particular set of relevant items through natural\

    \ language documents. In order to discover topics that recur in articles of text\

    \ corpus, another method TopCat (Topic Categories) was proposed by [22]. IE was\

    \ used by this technique in order to discover named entities in individual articles\

    \ and to characterize them as a collection of items of an article. Therefore,\

    \ through recognition of frequent item sets which commonly occurred with named\

    \ entities, the issues in data mining or database context were studied. Association\

    \ rule data mining technique is used by TopCat to discover these frequent item\

    \ sets. By using a hypergraph splitting technique, TopCat further clusters the\

    \ named entities which discovers a collection of frequent item sets with significant\

    \ overlie. In order to discover documents regarding the topic, IR technique was\

    \ used. Different technologies like IE for named entity extraction, association\

    \ rule data mining, clustering of association rules, IR techniques were used in\

    \ this method. TopCat discovers topics that have a logical accuracy with reasonable\

    \ identifiers. Callan and Mitamura [43] presented a new technique for named entity\

    \ detection, called KENE. In order to understand the extraction rules, the knowledge-based\

    \ technique is used in it. Generate-and-test approach is used for named entity\

    \ extraction from structured documents.We can observe from the surveyed literature\

    \ that there is a limitation in addressing the issue of IE from research articles\

    \ using data mining techniques. The synergy between these approaches (i.e. IE\

    \ with data mining techniques) helps to discover different interesting text patterns\

    \ in the retrieved articles. This approach could be applied to a variety of research\

    \ topics, where in each topic can generate a wide range of knowledge patterns.\

    \ Mobile learning (M-learning) has become one of the trendy fields in the higher\

    \ education [44][45][46][47][48][49][50]. In accordance to the existing literature,\

    \ we can perceive that IE and data mining techniques were never applied to the\

    \ M-learning field. This creates a need for collecting several research articles\

    \ in the field of M-learning from different scientific databases and applies the\

    \ synergic approach on them. Additionally, we are trying to respond to the following\

    \ research questions: RQ1: What are the most frequent keywords in the collected\

    \ articles? RQ2: What are the most frequent terms among the collected articles?\

    \ RQ3: What are the most common topics among the collected articles? RQ4: How\

    \ are the articles interrelated to each other? RQ5: How are the articles distributed\

    \ in terms of publication year?"

- header: "Experimental Results"

  content: "We have developed our customized framework which is inspired by the designed\

    \ framework proposed by [51], see Fig. The research articles were collected from\

    \ six scientific databases, namely: Springer, Wiley, Science Direct, SAGE, IEEE,\

    \ and Cambridge. The search term used for data collection is simply \"Mobile Learning\

    \ in higher education\". Based on that, 300 research articles in the field of\

    \ mobile learning were collected. These articles are categorized into six folders,\

    \ where each folder represents the database where these articles were retrieved.The\

    \ presence of the linguistic noise is a common problem in the content of the extracted\

    \ articles and we have dealt with. Then, the cleaned data are uploaded into RapidMiner\

    \ tool while the misplaced and unnecessary data have been removed from the dataset.\

    \ In order to improve the performance and data quality, all the irrelevant characteristics\

    \ are debarred while the data is being uploaded into RapidMiner tool. The major\

    \ steps involve the separation of the document into tokens; this task is called\

    \ Tokenization [52]. The next step is concerned with the transformation process\

    \ of all the characters where each document title is created in a lower case.\

    \ Stop words filtering is involved in the third step, where English is filtered\

    \ through this operator. A single English word is required to be signified by\

    \ each token. All tokens that were similar to stop words were eradicated from\

    \ the provided document by an operator. The document must have only one stop word\

    \ per line. The last step is concerned with the text processing phase that involves\

    \ filtering the tokens according to the length. The minimum number of the characters\

    \ that the token should have is 4, while the maximum number is 25 characters.The\

    \ application of various text mining techniques on the collected articles presents\

    \ different results and suggestions. In the present study, we are trying to apply\

    \ almost all of the text mining techniques that were mentioned in the literature\

    \ on the collected articles. Nevertheless, these techniques have not been applied\

    \ to the research articles concerning mobile learning in higher education; the\

    \ reason that makes this study is unique and adds a value to the research community."

- header: "Conclusion"

  content: "The present study demonstrates a comprehensive overview about text mining\

    \ and its current research status. According to the surveyed literature, there\

    \ is a limitation in discussing the issue of information extraction from research\

    \ articles using data mining techniques. The synergy between information extraction\

    \ and data mining techniques helps to discover different interesting text patterns\

    \ in the retrieved articles. This approach could be applied to a variety of research\

    \ topics, where in each topic it can generate a wide range of knowledge patterns.\

    \ Mobile learning has become one of the trendy fields in the higher education.\

    \ Accordingly, we can perceive that information extraction and data mining techniques\

    \ were never applied to the mobile learning field. This creates a need for collecting\

    \ a dataset that consists of several research articles in the field of mobile\

    \ learning from different scientific databases, and applying the proposed approach\

    \ on them.Three hundred refereed journal articles from six scientific databases\

    \ were collected, and textually analyzed through text mining techniques. The six\

    \ databases are Science Direct, IEEE, Wiley, Cambridge, SAGE, and Springer. The\

    \ selection of the collected articles was based on the criteria that all these\

    \ articles should incorporate mobile learning as the main component in the higher\

    \ educational context. In the present study, text clustering, association rule,\

    \ word cloud, and word frequency are the main tasks used for text analysis.By\

    \ applying the word cloud and the word frequency techniques, results indicated\

    \ that \"Learning\" is the most frequent keyword across all the collected articles;\

    \ followed by \"Patients\" and \"Students\", respectively. The increasing number\

    \ of the words: \"learning\" and \"students\" could be attributed to the fact\

    \ that learning and students form the core of the higher educational processes.\

    \ In addition, results revealed that the words: \"patients\", \"care\", \"medical\"\

    , and \"clinical\", were frequently mentioned in Springer database. These results\

    \ indicate that the most frequent linked words are those focused on studies targeting\

    \ mobile learning in medical education. Springer database represents the richest\

    \ source that contains these words followed by Wiley and Science Direct, respectively.\

    \ That is, researchers who are specialized in mobile medical education should\

    \ benefit from these results as it shows them that Springer database is the topmost\

    \ among other databases for finding research articles in this field.By applying\

    \ the association rule technique, findings showed that the term \"Education\"\

    \ is shown as being central to the tree structure having all the relevant words\

    \ connected to it. This could be referred to the fact that the text acquired from\

    \ the collected research articles is mainly concentrated on the learning field.\

    \ In addition, we performed the similarity measure on the collected articles in\

    \ order to identify the topics that are highly similar to each other. Results\

    \ revealed that the similarity operator could not detect a clear similarity among\

    \ some topics the reason is that these topics are interrelated and similar in\

    \ meaning to each other (i.e. all the articles are discussing the topic of mobile\

    \ learning in higher education).By applying the clustering technique, we used\

    \ the k-means algorithm through the use of different k values. Results indicated\

    \ that there were six clusters. Almost all of the articles (N = 285) were accumulated\

    \ in one cluster; this indicates that these articles are discussing the main studied\

    \ topic (i.e. mobile learning in higher education). On the other side, by further\

    \ investigating the remaining articles (N = 15) that are accumulated in the other\

    \ clusters; it has been found that these articles are discussing other topics\

    \ in learning and education rather than the studied topic. By distributing the\

    \ collected research papers across their years of publication, findings showed\

    \ that there was a booming increase in the number of published articles during\

    \ the years 2015 through 2016. This could be referred to the reason that mobile\

    \ learning has witnessed in these years an enormous attraction from a lot of scholars\

    \ who published many articles that contribute to the evolvement of mobile learning.As\

    \ a future work, we are interested in collecting articles from various research\

    \ topics, i.e. not to focus on one area. This will help us to find more interesting\

    \ patterns in these articles and how such articles are distributed among the targeted\

    \ databases. In addition, this will allow the similarity operator to work properly\

    \ and to draw a clear relationship among the articles."

Figure 0‑1 – Segmentation full result

The full summary using TextRank:

abstractSeg: 'In our study, we collected, and textually analyzed through various text

  mining techniques, three hundred refereed journal articles in the field of mobile

  learning from six scientific databases, namely: Springer, Wiley, Science Direct,

  SAGE, IEEE, and Cambridge.

  Experimental results indicated that Springer database represents the main source

  for research articles in the field of mobile education for the medical domain.'

sectionList:

- content: For example, in E-commerce, a major issue with web mining is buying all

    the materials which are not relevant to the user's search and it will not show

    unknown (hidden or implicit) information, while the major objective of text mining

    is to find out the unknown information [6]; something that is not recognized by

    anyone.Data is the basic kind of information, which is required to be organized

    and mined for the knowledge generation.

  header: Introduction

- content: In order to discover nonfigurative patterns in the extracted data, data

    mining algorithms and techniques can be used.Song and Kim [40] presented the first

    attempt to apply text mining approaches to a huge collection of full-text articles

    for discovering the knowledge structure of the area.

  header: Related Work

- content: 'The minimum number of the characters that the token should have is 4,

    while the maximum number is 25 characters.The application of various text mining

    techniques on the collected articles presents different results and suggestions.

    In the present study, we are trying to apply almost all of the text mining techniques

    that were mentioned in the literature on the collected articles.'

  header: Experimental Results

- content: Results revealed that the similarity operator could not detect a clear

    similarity among some topics the reason is that these topics are interrelated

    and similar in meaning to each other (i.e. all the articles are discussing the

    topic of mobile learning in higher education).By applying the clustering technique,

    we used the k-means algorithm through the use of different k values.

  header: Conclusion

Figure 0‑2 – TextRank summarization full result

The full summary using LSA:

---

abstractSeg: "The present study demonstrates a comprehensive overview about text mining\

  \ and its current research status. In our study, we collected, and textually analyzed\

  \ through various text mining techniques, three hundred refereed journal articles\

  \ in the field of mobile learning from six scientific databases, namely: Springer,\

  \ Wiley, Science Direct, SAGE, IEEE, and Cambridge."

sectionList:

- header: "Introduction"

  content: "A study by [2] stated that text mining has become one of the trendy fields\

    \ that has been incorporated in several research fields such as computational\

    \ linguistics, Information Retrieval (IR) and data mining. The outcomes can be\

    \ stored in a management information system that provides a large amount of significant\

    \ information for the user of that system.Text mining intends to detect the information\

    \ that was not recognized before through extracting it automatically from various\

    \ text-based sources."

- header: "Related Work"

  content: "The primary focus of these researches was to determine how different text\

    \ mining procedures can be utilized as the structured data sets exist in the text\

    \ document format. It offers the opportunity to show an arrangement of sub-topics\

    \ in huge sets of documents in compact graph form.In order to offer a reference\

    \ for additional researches of other researchers, [37] discussed the research\

    \ status of text mining technology when it was used in the biomedical field that\

    \ covers 10 years."

- header: "Experimental Results"

  content: "These articles are categorized into six folders, where each folder represents\

    \ the database where these articles were retrieved.The presence of the linguistic\

    \ noise is a common problem in the content of the extracted articles and we have\

    \ dealt with. The minimum number of the characters that the token should have\

    \ is 4, while the maximum number is 25 characters.The application of various text\

    \ mining techniques on the collected articles presents different results and suggestions."

- header: "Conclusion"

  content: "This creates a need for collecting a dataset that consists of several\

    \ research articles in the field of mobile learning from different scientific\

    \ databases, and applying the proposed approach on them.Three hundred refereed\

    \ journal articles from six scientific databases were collected, and textually\

    \ analyzed through text mining techniques. The selection of the collected articles\

    \ was based on the criteria that all these articles should incorporate mobile\

    \ learning as the main component in the higher educational context."

Figure 0‑3 – LSA summarization full result