# An Approach for Extracting and Replicating Table Data from PDF Sources

MASTER THESIS

Submitted by

**Ankita Sarkar**
Reg. No. 22MDT1067

in partial fulfillment for the award of the degree of

M.Sc. Data Science

**Department of Mathematics**
**School of Advanced Sciences**
Vellore Institute of Technology, Chennai
Vandalur - Kelambakkam Road, Chennai - 600 127

April - 2024

# DECLARATION

I hereby declare that the project entitled **An Approach for Extracting and Replicating Table Data from PDF Sources** submitted by me to the Department of Mathematics, School of Advanced Sciences, Vellore Institute of Technology, Chennai, 600 127 in partial fulfillment of the requirements of the award of the degree of **Master of Science in Data Science** is a bonafide record of the work carried out by me under the supervision of **Dr. David Raj Micheal**. I further declare that the work reported in this project, has not been submitted and will not be submitted, either in part or in full, for the award of any other degree or diploma of this institute or of any other institute or University.

**Place** : Chennai

**Date** :

Ankita Sarkar

Reg. No: 22MDT1067

# CERTIFICATE

This is to certify that the thesis entitled **An Approach for Extracting and Replicating Table Data from PDF Sources** is prepared and submitted by **Ankita Sarkar (Reg. No. 22MDT1067)** to Vellore Institute of Technology, Chennai, in partial fulfillment of the requirement for the award of the degree of **Master of Science in Data Science** is a bona-fide record carried out under my guidance. The project fulfills the requirements as per the regulations of this University and in my opinion meets the necessary standards for submission. The contents of this report have not been submitted and will not be submitted either in part or in full, for the award of any other degree or diploma and the same is certified.

**Place** : Chennai

**Date** :

Signature of Guide

**Dr. David Raj Micheal**

Signature of HOD

**Dr. K Muthunagai**

# Acknowledgement

With immense pleasure and a deep sense of gratitude, I wish to express my sincere thanks to my supervisor *Dr. David Raj Micheal*, Assistant Professor, School of Advanced Sciences, Vellore Institute of Technology (VIT), Chennai for his motivation and continuous encouragement, this project would not have been successfully completed.

I am grateful to the Chancellor of VIT, *Dr. G.Viswanathan*, the Vice Presidents *Mr. Sankar Viswanathan*, *Dr. Sekar Viswanathan*, *Dr. G V Selvam*, the Executive Director, *Dr. Sandhya Pentareddy* , the Assistant Vice President *Ms. Kadhambari S. Viswanathan* the Vice Chancellor *Dr. V. S. Kanchana Bhaaskaran*, the Pro Vice Chancellor of VIT Chennai *Dr. T. Thyagarajan* and the Additional Registrar *Dr. P. K. Manoharan* for motivating me to carry out the project at Vellore Institute of Technology, Chennai.

I express my sincere thanks to *Dr. S. Mahalakshmi*, Dean, School of Advanced Sciences, VIT, Chennai and *Dr. K. Muthunagai*, HOD, Mathematics and Computing, School of Advanced Science, VIT, Chennai for their support and encouragement.

I wish to extend my profound sense of gratitude to *my parents* for providing me the moral support and encouragement whenever required.

|  |  |  |  |
|---|---|---|---|
| **Place** | : | Chennai | Ankita Sarkar |
| **Date** | : |  | Reg. No: 22MDT1067 |

# Abstract

Abstract to your work. Smith (2022)

**Keywords:** Machine Learning, Deep Learning...

# Contents

# List of Tables

# List of Figures

# Chapter 1

# Introduction

The introduction is a shorter version of the rest of the report, and in many cases the rest of the report can also have the same flow that summarizes the major contributions of the project. The chapter should provide a critical and concise outline of the subject to be covered by the dissertation and indicate how this study will contribute to the subject. This chapter should include the descriptions such as: (not necessarily in that order, but what is given below is a logical order). Example of a citation: Smith (2022)

- Background [The setting of the scene of the problem].
- Statement [Exact problem you are trying to solve].
- Motivation [Importance of the problem].
- Post/Related work [Existing methods including pros and cons of the methods should be cited wherever possible].
- Challenges [Difficulty in the problem solving].
- Essence of your approach [Your method of problem solving].
- Statement of assumptions [The conditions under which your solution is applicable].
- Organization of the report.

- Aim(s) and Objective(s)
- Avoid ‘routine’ background e.g. the C programming language.
- Don’t cite endless sources that are irrelevant or that you haven’t read.

# Chapter 2

# Overview / Literature Review

This chapter should include the brief description of the whole-proposed software system that is to be developed, system preliminary design, system planning and the details of the hardware & software used. System analysis & design vis-à-vis user requirements (Preliminary design) should also be represented as a block diagram. System planning is represented as either as PERT chart or as Gantt chart. A thorough review of the literature with respect to the chosen field should be projected. Should include earlier and current reports along with author citation and year. In other words it should be a collection and a record of past land recent work. Summarize major contributions of significant studies and articles related to your field under review, maintaining the focus established in the introduction. Evaluate current "state of art". Point out major gaps, inconsistencies in theory and findings. Conclude by providing some insight into the relationship between the central topic of the literature review and the areas / issues pertinent to future study.

## 2.1 Some section

some dummy data is written here... Please change it according to your need...

# Chapter 3

# System Design

This chapter should describe the engineering specifications and targets critically evaluating the existing benchmarks and specifically identifying the gaps which the project is intended to fill; It should show how the concepts evolved and were evaluated also should describe and justify the formation of the final product which may include possibly a number of subsections such as:

- Details of the development. System architecture indicating various modules / components and their interaction.
- Feasibility assessment report.
- Entity relationship diagram / analysis / DFD / State Transition Diagram.

## 3.1 Some section

some dummy data is written here... Please change it according to your need...

If you adopt an object-oriented method, you will include the following in this chapter:

- Sequence diagrams for each module and entire system.
- Class diagrams or any other UML diagram for each module and entire system.

# Chapter 4

# Implementation of System/ Methodology

This chapter should reflect development of the project such as: implementation, experimentation, optimization, evaluation etc. and unit integration testing should be discussed in detail. The unit test cases and system test cases should describe the input, expected output and output obtained. It can also include the details of the tools used for implementation, justification for the selected tool and the detailed description of implemented modules. Screen shots, Pseudocode etc. In case of simulation, modeling, programming techniques, programming steps, flow-charts, simulation results, verification of the approach followed and the like depending on the nature of the project.

## 4.1 Some section

some dummy data is written here... Please change it according to your need...

The materials required, techniques followed, sample preparations, research design and methods should be clearly mentioned. The experimental procedure should be clearly defined.

# Chapter 5

# Results and Discussions

This is part of the set of technical sections, and is usually a separate section for experimental/design papers. This chapter should include:

- Performance metrics.
- Parameters under study
- Comparison of cases/ studies with respect to existing and proposed work / algorithm/ design–comparison/ with the published data and deviations / improvements if any as expected in the aims and objectives
- Expected and obtained results- Analysis of the results- statistical analysis, plots, simulated results, synthesis of process, interpretation of the results
- Detailed results for each logical component of the project with an accompanying discussion section [Can include screen shots, graphs etc.].
- The results can be tabulated, graphically presented and photographs to be displayed if any.
- Discuss the results which should include an interpretation of the results and their relationship to the aims and objectives.

# Chapter 6

# Summary and Future Work

This chapter should summarize the key aspects of your project (failures as well as successes) and should state the conclusions you have been able to draw. Outline what you would do if given more time (future work). Try to pinpoint any insights your project uncovered that might not have been obvious at the outset. Discuss the success of the approach you adopted and the academic objectives you achieved. Avoid meaningless conclusions, [e.g. NOT " I learnt a lot about C++ programming "]. Be realistic about potential future work. Avoid the dreaded: "All the objectives have been met and the project has been a complete success". You have to crisply state the main take-away points from your work. Describe how your project is performed against planned outputs and performance targets. Identify the benefits from the project. Be careful to distinguish what you have done from what was there already. It is also a good idea to point out how much more is waiting to be done in relation to a specific problem, or give suggestions for improvement or extensions to what you have done. Future scope of the work for improvement may also be included

# Appendix A

# Appendix Chapter

Here is the appendix chapter. Usually, the code and the other related items are given here.

# References

Smith, J. (2022), 'The effects of climate change', *Scientific American* **327**(1), 44–49. This is a sample entry for an article in a magazine.

VIT®
CHENNAI
Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

School of Advanced Sciences
Vellore Institute of Technology, Chennai
Vandalur – Kelambakkam Road, Chennai - 600 127
(www.chennai.vit.ac.in)