DECLARATION

I hereby declare that I carried out the work reported in this thesis in the Department of Information Science under the Faculty of Information and Communication Technology, University of Technology (Yatanarpon Cyber City), under the supervision of Dr. Nandar Win Min. I solemnly declare that to the best of my knowledge; no part of this thesis has been submitted here or elsewhere in a previous application for award of a degree. All sources of knowledge used have been duly acknowledged.

..…………………………

4th October, 2019

THAW DE ZIN

6IST-66

**APPROVAL**

This is to certify that the thesis titled **“WEB CRAWLING AND SCRAPING AGENT USING TF-IDF”** carried out by **THAW DE ZIN, 6IST-66** has been read and approved for meeting part of the requirements and regulations governing the award of the degree of Bachelor of Engineering (Information Science and Technology), Department of Information Science under the Faculty of Information and Communication Technology, University of Technology (Yatanarpon Cyber City), Myanmar.

1. Dr. Reenu

Pro-Rector (Academic)

University of Technology (Yatanarpon Cyber City) ………………….

(Chairman)

1. Dr. Hnin Aye Thant

Professor and Head

Department of Information Science

Faculty of Information and Communication Technology ………...…………

(Co-Chairman)

1. Dr. Htet Ne Oo

Lecturer and Course Coordinator

Department of Information Science

Faculty of Information and Communication Technology ....………………

(Course Coordinator)

1. Dr. Nandar Win Min

Lecturer

Department of Information Science

Faculty of Information and Communication Technology …………………

(Supervisor)

1. U Zaw Htet

Lecturer

Department of Information Science

Faculty of Information and Communication Technology …….……………

(Member)

**ACKNOWLEDGEMENTS**

First of all, I would like to express my special thanks to Dr. Aung Win, Rector, University of Technology (Yatanarpon Cyber City), for initiating the Bachelor programme and for his kind permission to complete this thesis at the University of Technology (Yatanarpon Cyber City).

I would like to express my respectful gratitude to Dr. Reenu, Pro-rector, University of Technology (Yatanarpon Cyber City), for giving enough consideration to her ideas and views.

I would like to express grateful thank to Dr. Hnin Aye Thant, Professor and Head of Department of Information Science, Faculty of Information and Communication Technology, University of Technology (Yatanarpon Cyber City), for her kind guidance, encouragement and supervision and patience in making my thesis to complete successfully.

I am very grateful to Dr. Htet Ne Oo, Lecturer and Course Coordinator of Final Year (Information Science and Technology), Department of Information Science, Faculty of Information and Communication Technology, University of Technology (Yatanarpon Cyber City), for her kind guidance and encouragement and supervision and patience in making my thesis to complete successfully.

I am very grateful to my supervisor, Dr. Nandar Win Min, Lecturer, Department of Information Science, Faculty of Information and Communication Technology, University of Technology (Yatanarpon Cyber City), for her kind guidance and encouragement. She has been very supportive in my thesis, and also guides a lot, particularly, at the level of quality of presentation.

Very Special thanks to Daw Khin Moe Moe Thu, Assistant Lecturer, English Department, University of Technology (Yatanarpon Cyber City) for her valuable supports from the language point of view in my thesis work.

I would like to thank a lot to all my teachers for their mentoring, encouragement, and recommending this dissertation.

Finally, I am grateful to my parents and friends who specially offered strong moral and physical support, care and kindness, during the year of my thesis study.

**ABSTRACT**

The need to scrape websites came with the popularity of the Internet, where a lot of data and content are shared. The first widely known scrapers were invented by search engine developers (like Google or Bing). These scrapers go through (almost) the entire Internet, scan every web page, extract information from it, and build an index that can be used for miscellaneous purposes. Web Crawling and Scraping is most interesting field in Computer Science. Web Scraping is an automated process for finding and extracting data from on-line sources. It is also referred to as automated web extracting, web content mining, or data collection. It is commonly used in price comparison, president election, brand promoting and miscellaneous purposes. The proposed system is to develop web agent for people to choose which products they should buy. The web crawler is used to get information about products across the web resources. In this system, during crawling the entire three website pages/links, TF-IDF is used to filter the pages/links about the product (laptop) and extract 7 features about processor, RAM, storage capacity, price, display, links and graphic information from each laptop. The most relevant laptops are generated according to the user input using EDIT DISTANCE algorithm.

.

**CONTENTS**

**Page**

DECLARATION i

APPROVAL ii

ACKNOWLEDGEMENTS iii

ABSTRACT iv

CONTENTS v

LIST OF FIGURES vii

LIST OF TABLES viii

LIST OF EQUATIONS ix

**CHAPTER 1 INTRODUCTION**

1.1 Objectives 2 1.2 Web Crawling and Web Scraping 2

1.3 Overview of the System 3

1.4 Organization of the Thesis 4

1.5 Summary 4

**CHAPTER 2 THEORETICAL BACKGROUND**

2.2 Web Crawling 5

2.1.1 Incremental Crawling 6

2.1.2 Depth First Tree Transversal 6

2.1.3 Visual and Programmatic Crawlers 9

2.1.4 Crawler Architecture and Identification 10

2.2 Web Scraping 11

2.2.1 Web Scraping Techniques 14

2.2.2 HTML Parsing 15

2.3 Preprocessing 19

2.3.1 Lowercase Conversion 19

2.3.2 Stopwords and Tags Removal 19

2.3.3 Ngrams 20

2.4 Information Retrieval 20

2.4.1 TF-IDF 21

2.5 Cosine Similarity 23

2.6 Edit Distance Algorithm 26

2.6.1 Levenshtein Distance 26

2.7 Summary 27

**CHAPTER 3 SYSTEM DESIGN AND IMPLEMENTAION**

3.1 System Design 28

3.2 System Flow 28

3.2.1 System Flow for Crawling 28

3.2.2 System Flow for Searching 29

3.3 System Implementation 30

3.3.1 System Implementation for Crawling 31

3.3.2 System Implementation for Searching 34

3.4 Output Result 39

3.5 Summary 39

**CHAPTER 4 CONCLUSION**

4.1 Benefits 40

4.2 Limitations 41

4.3 Further Extension 41

**REFERENCES** 42

# LIST OF FIGURES

**Figure**  **Page**

2.1 HTML Tags 8

2.2 HTML Tags to Binary Tree 9

2.3 The Difference between Web Crawler and Web Scraper 11

2.4 Cosine Similarity 24

2.5 Equivalent Degrees 25

2.6 Similarity Scores 25

3.1 System Flow for Crawling 29

3.2 System Flow for Searching 30

3.3 Home Page of the System 30

3.4 Crawling Page of the System 31

3.5 Detail Link Page of System 32

3.6 Response Body of the System 32

3.7 Response Body without HTML Tags of the System 33

3.8 Important Data with Tags 33

3.9 Simplified Form of Data 34

3.10 System Interface for Database View 34

3.11 System Interface for Searching with Two Options 35

3.12 System Interface for Price and Brand 35

3.13 System Interface for Price Range 36

3.14 System Interface for Price and Specifications 36

3.15 System Interface for Processor Generation and Brand Identifier 37

3.16 System Interface for Brand and Specifications 37

3.17 System Interface for Final Result 38

3.18 E-commerce Website Page 38

3.19 Product Specification from Website 39

## LISTS OF TABLES

**Table Page**

2.1 TF-IDF Demonstration 23

**LISTS OF EQUATIONS**

**Equations Page**

2.1 Term Frequency 22

2.2 Inverse Data Frequency 22

2.3 Term Frequency - Inverse Data Frequency 22

2.4 Dot Product of Two Vectors 23

2.5 Dot Product of 2 Dimension Vectors 24

2.6 Dot Product for 24

2.7 Rearranging Dot Product for 24