

**AUTOMATED EMAIL SPAM TRAP**



ADIYA NAJMIN BINTI MOHD NOREZAN

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

**BORANG PENGESAHAN STATUS TESIS\***

JUDUL: AUTOMATED EMAIL SPAM TRAP SYSTEM

SESI PENGAJIAN: 2016/2017

Saya ADIYA NAJMIN BINTI MOHD NOREZAN

(HURUF BESAR)

mengaku membenarkan tesis (PSM/Sarjana/Doktor Falsafah) ini disimpan di Perpustakaan Fakulti Teknologi Maklumat dan Komunikasi dengan syarat-syarat kegunaan seperti berikut:

1. Tesis dan projek adalah hakmilik Universiti Teknikal Malaysia Melaka.
2. Perpustakaan Fakulti Teknologi Maklumat dan Komunikasi dibenarkan membuat salinan untuk tujuan pengajian sahaja.
3. Perpustakaan Fakulti Teknologi Maklumat dan Komunikasi dibenarkan membuat salinan tesis ini sebagai bahan pertukaran antara institusi pengajian tinggi.
4. \*\* Sila tandakan (/)  
\_\_\_\_\_ SULIT (Mengandungi maklumat yang berdarjah keselamatan atau kepentingan Malaysia seperti yang termaktub di dalam AKTA RAHSIA RASMI 1972)  
\_\_\_\_\_ TERHAD (Mengandungi maklumat TERHAD yang telah ditentukan oleh organisasi/badan di mana penyelidikan dijalankan)  
\_\_\_\_\_ / \_\_\_\_\_ TIDAK TERHAD



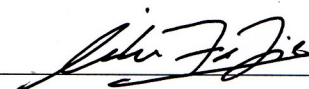
(TANDATANGAN PENULIS)

Alamat tetap: No 168, Jalan PJS 4/4,

Taman Medan, 46000 Petaling Jaya,

Selangor Darul Ehsan.

Tarikh: 24 Ogos 2016



(TANDATANGAN PENYELIA)

Raihana Syahirah binti Abdullah

Nama Penyelia

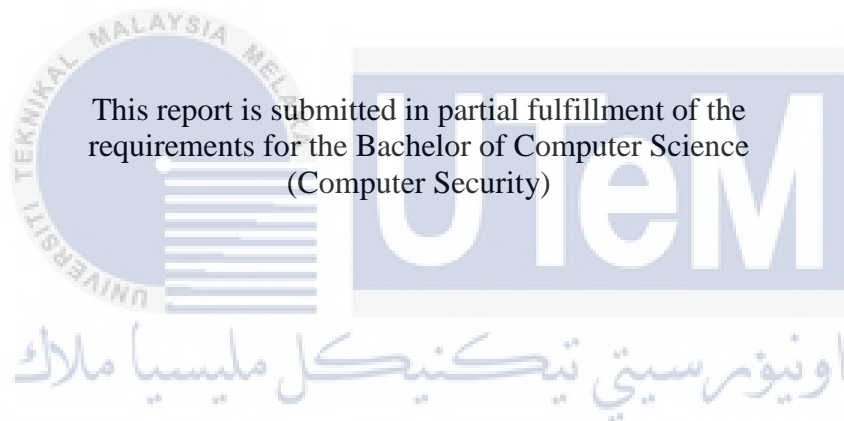
Tarikh: 24 Ogos 2016

CATATAN: \* Tesis dimaksudkan sebagai Laporan Akhir Projek Sarjana Muda (PSM)

\*\* Jika tesis ini SULIT atau TERHAD, sila lampirkan surat daripada pihak berkuasa.

# AUTOMATED EMAIL SPAM TRAP

ADIYA NAJMIN BINTI MOHD NOREZAN



This report is submitted in partial fulfillment of the requirements for the Bachelor of Computer Science (Computer Security)

FACULTY OF INFORMATION AND COMMUNICATION TECHNOLOGY

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

2016

## DECLARATION

I hereby declare that this project report entitled

### **AUTOMATED EMAIL SPAM TRAP SYSTEM**

is written by me and is my own effort and that no part has been plagiarized

without citations.

STUDENT :  \_\_\_\_\_ Date : 24 AUGUST 2016

(ADIYA NAJMIN BINTI MOHD NOREZAN)

  
UNIVERSITI TEKNIKAL MALAYSIA MELAKA

I hereby declare that I have read this project report and found

this project report is sufficient in term of the scope and quality for the award of

Bachelor of Computer Science (Computer Security) With Honours.

SUPERVISOR :  \_\_\_\_\_ Date : 24 AUGUST 2016

(RAIHANA SYAHIRAH BINTI ABDULLAH)

## DEDICATION

To my beloved parents who is always by my side in highs and lows, easy or hard times, supportive and encouraging for their time, money, and advice so that this thesis finishes successfully and satisfying although there are many challenges faced during the development of this project.



## ACKNOWLEDGEMENTS

First and foremost, I praise God, the almighty for giving me this opportunity and granting me the capability to proceed successfully. I offer my sincerest gratitude towards my supervisor, Madam Raihana Syahirah binti Abdullah who has supported me with the expertise, understanding, and patience added considerably to my graduate experience.

I would also like to thank my PA, Madam Khadijah binti Wan Mohd Ghazali for her initiative in helping her student's affairs especially in reviewing my proposal with Miss Zakiah binti Ayub. A special gratitude is given to Mr Suryanata and Madam Marliza Ramly for spending some time in evaluating my proposal for better documentation.

I would like to express my appreciation to the PSM Committee, Madam Robiah because has taken care of BITZ students throughout the process of making this project. Also a special thanks to all the PSM Committee for their commitment and responsibilities in keeping all of the matters run smoothly.

I warmly thank my roommate, who is also my housemate and my classmate for taking part in making my project finishes in time. I would also like to express my deepest gratitude towards my parents, Madam Norhuda binti Abdul Rahman and Mr Mohd Norezan bin Ibrahim for their prayers and motivation during my degree studies.

I would like to give a special remarks for all the lecturers that have teach me all the knowledge that are then use in this project which is Madam Maslita, Madam Emaliana, Madam Hafiezah, Madam Zaheera, Madam Rabiah, Madam Norashikin, Sir Gede, Sir Zaki, Sir Fairuz, Prof Dr Madya Nurazman, Prof Zaki, Sir Suhaimi, Sir Warusia and many more.

Last but not least, I express my gratitude to my sibling and close friends that have always been there for me whenever I needed some comfort, companion and strength to move on in finishing my project. Thank you so much to everyone, for everything.

In conclusion, I recognize that this project would not finish without all the equipments and facilities of Universiti Teknikal Malaysia Melaka (UTeM), Faculty of Information and Communication Technology (FTMK), MyCert, Cyber Security Malaysia (CSM) and other organizations that have contributed in this project.



## ABSTRACT

This project proposes solutions by developing an automated email spam trap to substitute the traditional way of managing email spam trap. The motivation is the spam email problem that keeps emerging from time to time thus prevention of this spam email is necessary. Therefore, the results from the analysis conducted will help to produce a system that helps to monitor spam email messages by giving a score towards the email messages received by user. The scope of this project involves identifying spam email messages that could harm and lower productivity level of an organization, examining email trap that help collect samples of unsolicited messages through ISPs and collect data about spam activities from resourceful Websites. The approach used in this project is by using simulation and prototype construction with Python programming language. The samples of email dataset are analyzed to study behavior of spam email messages from an email program. The important variables that are being controlled are the content type of spam email messages, scoring that are being measured and source that are ignored. The results are distinctive compared to other systems as it provides scoring for email messages using content based approach until an automated system are being developed. The results show that it is specific to a particular case not generalized as it concludes only content type view. In conclusion, this project helps to justify content type approach as one of the ways to build an automated email spam trap.



## ABSTRAK

Projek ini mencadangkan penyelesaian dengan membangunkan perangkap e-mel spam automatik untuk menggantikan cara tradisional menguruskan perangkap e-mel spam. Masalah e-mel spam yang berlaku dari semasa ke semasa memberi motivasi bahawa pencegahan adalah perlu. Oleh itu, keputusan daripada analisis yang dijalankan akan membantu untuk menghasilkan satu sistem yang membantu untuk memantau mesej e-mel spam dengan memberi skor terhadap mesej e-mel yang diterima oleh pengguna. Skop projek ini melibatkan mengenal pasti mesej spam e-mel yang boleh memudaratkan dan merendahkan tahap produktiviti sesebuah organisasi, memeriksa perangkap e-mel yang membantu sampel mengumpul mesej yang tidak diminta melalui ISP dan mengumpul data mengenai aktiviti spam dari laman web yang pintar. Pendekatan yang digunakan dalam projek ini adalah dengan menggunakan simulasi dan pembinaan prototaip dengan bahasa pengaturcaraan Python. Sampel set data e-mel dianalisis untuk mengkaji tingkah laku mesej e-mel spam daripada program e-mel. Pembolehubah penting yang dikawal adalah jenis kandungan mesej e-mel spam, pemarkahan yang sedang diukur dan sumber yang diabaikan. Sistem automatik yang dibangunkan ini menggunakan pendekatan bahawa keputusan menggunakan pendekatan jenis kandungan menghasilkan impak yang berbeza berbanding jaringan yang lain. Keputusan menunjukkan bahawa ia adalah khusus untuk kes tertentu tidak umum kerana ia menyimpulkan hanya kandungan jenis paparan. Kesimpulannya, projek ini membantu mewajarkan pendekatan jenis kandungan sebagai salah satu cara untuk membina perangkap e-mel spam automatik.

## Table of Contents

<b>DECLARATION</b> .....	<b>i</b>
<b>DEDICATION</b> .....	<b>ii</b>
<b>ACKNOWLEDGEMENTS</b> .....	<b>iii</b>
<b>ABSTRACT</b> .....	<b>v</b>
<b>ABSTRAK</b> .....	<b>vi</b>
<b>TABLE OF CONTENTS</b> .....	<b>vii</b>
<b>LIST OF TABLES</b> .....	<b>xi</b>
<b>LIST OF FIGURES</b> .....	<b>xiii</b>
<b>CHAPTER I</b> .....	<b>1</b>
1.0 INTRODUCTION .....	1
1.1 Introduction .....	2
1.2 Problem Statement (PS) .....	3
1.3 Project Question (PQ).....	3
1.4 Project Objective (PO) .....	4
1.5 Project Scope .....	4
1.6 Project Contribution .....	5
1.7 Thesis Organization.....	6
1.8 Conclusion.....	7
<b>CHAPTER II</b> .....	<b>8</b>
2.0 LITERATURE REVIEW .....	8
2.1 Introduction .....	8
2.2 Related Work/Previous Work.....	9
2.2.1 Website Resources .....	11
2.3 Critical Review of Current Problem and Justification.....	21

2.3.1 History of Spam .....	23
2.3.2 Methodologies.....	26
2.3.3 Techniques .....	26
2.3.4 Parameters/Attributes.....	29
2.3.5 Software/Hardware .....	30
2.4 Proposed Solution/Further Project.....	35
2.5 Conclusion.....	35
<b>CHAPTER III .....</b>	<b>36</b>
3.0 PROJECT METHODOLOGY.....	36
3.1 Introduction.....	36
3.2 Methodology.....	37
3.3 Project Flow.....	39
3.4 Project Milestones .....	40
3.5 Conclusion.....	44
<b>CHAPTER IV .....</b>	<b>45</b>
4.0 ANALYSIS AND DESIGN .....	45
4.1 Introduction .....	45
4.2 Problem Analysis.....	46
4.3 Requirement Analysis.....	47
4.3.1 Data Requirement .....	47
4.3.2 Functional Requirement.....	48
4.3.3 Non-functional Requirement .....	48
4.3.4 Others Requirement .....	48
4.4 High-Level Design .....	49

4.4.1 System Architecture.....	49
4.4.2 User Interface Design .....	50
4.5 Analysis Result .....	52
4.5.1 Keyword.....	53
4.6 Conclusion.....	57
<b>CHAPTER V .....</b>	<b>59</b>
5.0 IMPLEMENTATION.....	59
5.1 Introduction .....	59
5.2 Software Development Environment Setup .....	60
5.2.1 Software.....	60
5.2.2 Hardware.....	61
5.3 Software Configuration Management .....	61
5.3.1 Configuration Environment Setup.....	61
5.3.2 Version Control Procedure .....	62
5.4 Implementation Status .....	63
5.5 Discussion.....	63
5.6 Conclusion.....	64
<b>CHAPTER VI.....</b>	<b>65</b>
6.0 TESTING.....	65
6.1 Introduction .....	65
6.2 Test Plan .....	66
6.2.1 Test Organization.....	66
6.2.2 Test Environment.....	66
6.2.3 Test Schedule .....	68

6.3 Test Strategy .....	68
6.3.1 Classes of Tests .....	69
6.4 Test Design .....	70
6.4.1 Test Description .....	70
6.4.2 Test Data .....	70
6.5 Test Results and Analysis .....	70
6.6 Discussion .....	71
6.7 Conclusion .....	72
<b>CHAPTER VII .....</b>	<b>73</b>
7.0 PROJECT CONCLUSION .....	73
7.1 Introduction .....	73
7.2 Project Summarization .....	73
7.3 Project Contribution .....	74
7.4 Project Limitation .....	75
7.5 Future Works .....	75
7.6 Conclusion .....	75
<b>REFERENCES .....</b>	<b>77</b>
<b>APPENDICES .....</b>	<b>80</b>

## LIST OF TABLES

<b>TABLE</b>	<b>TITLE</b>	<b>PAGE</b>
1.1	Summary of Problem Statement	3
1.2	Summary of Project Question	3
1.3	Summary of Project Objective	4
1.4	Summary of Project Contribution	5
2.1	Direct and Indirect Economic Harm Affecting Participants	11
2.3	Reported Incidents based on General Incidents Classification Statistics 2014: All incidents	12
2.4	Reported Incidents based on General Incidents Classification Statistics 2014: Spam only	14
2.5	Reported Incidents based on General Incidents Classification Statistics 2015: All incidents	15
2.10	MyCERT Spam Email Statistics 2016 for business people and organization.	20
3.1	Project Milestones	40
4.1	The spam value based on previous analysis	52
4.2	Keyword occurrence and percentage of spam	

Keyword	53
5.1 List of software and its availability for this Project	60
5.2 Progress development status for each component/module	63
6.1 Classes of tests for this project	69
6.2 The expected result for each module for test	
Description	70
Summary of Gantt Chart PSM 1	80



## LIST OF FIGURES

<b>FIGURE</b>	<b>TITLE</b>	<b>PAGE</b>
1.1	Thesis Organization	6
2.1	Direct and Indirect Economic Harm Affecting Participants	11
2.3	Reported Incidents based on General Incidents Classification Statistics 2014: All incidents	12
2.4	Reported Incidents based on General Incidents Classification Statistics 2014: Spam only	14
2.7	Comparison Graph of incidents during Quarter 2 of 2015: April until June	17
2.8	Comparison Pie Chart of incidents during Quarter 2 of 2015: April until June	18
2.9	Automated Malware Collection of Honeypots etc from isecLAB	19
2.10	MyCERT Spam Email Statistics 2016 throughout year 2016	20
2.11	Spam filtering path	23
2.12	Spam evolution from time to time	23
2.14	Basic architecture of email spam	25

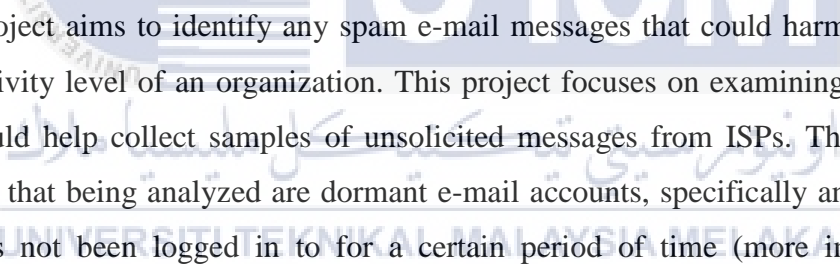


2.16	SpamTitan Spam Trap	30
2.17	UWO Spam Trap	31
2.18	InfoWest Spam Trap	32
2.19	ODU Spam Trap	33
2.20	SonicWall Email Security Spam Trap	34
3.1	Methodology	37
3.2	Project Process Flow	39
4.1	Synthesis for Chapter 4	46
4.2	Email dataset being analyzed	47
4.3	High-Level Design	49
4.4	System Architecture	50
4.5	Simple Graphical User Interface	50
4.6	Second interface	51
4.7	Distribution of spam value based on previous analysis	52
4.8	Keyword of Occurrence	56
4.9	Percentage of keyword occurrence	57
5.1	Outline diagram for Chapter 5	60
5.2	Code snippets	62
6.1	Outline chapter for Chapter 6	66

## CHAPTER I

### 1. INTRODUCTION

#### 1.1 Introduction



This project aims to identify any spam e-mail messages that could harm and lower the productivity level of an organization. This project focuses on examining an e-mail trap that could help collect samples of unsolicited messages from ISPs. The main type of account that being analyzed are dormant e-mail accounts, specifically an email account that has not been logged in to for a certain period of time (more info on specific timeframes below), then that account will be deemed dormant and will be a candidate for automatic deactivation. This project consists of programming language Python that is used to create an automated email spam trap system. This project focuses on Yahoo accounts as there are number of users using this account since ages.

Spam or known as unsolicited messages according to IBM, is an output message not associated with an input message. An unsolicited message results from a host application sending an unsolicited message. However, an unsolicited email which is known as spam is an unwanted email messages that force recipients to incur the cost of receiving, storing and removing it from one's email box. There are two terms related to bulk email which is legitimate bulk email and unsolicited (bulk) email. There is a main difference

between legitimate bulk email and unsolicited email that is very clear whereby legitimate bulk email requires sender to have verifiable permission from recipients before sending any messages to their email boxes.

The only way to fight unsolicited email (spam) is by spam trap. Email trap or spam trap is an email address traditionally used to expose illegitimate senders who add email addresses to their lists without permission which is not created for means of communication. It is also known that spam trap is email addresses that are activated by mailbox providers and others to find senders with poor data hygiene and collection practices. There are two types of spam traps which is recycled and pristine. The ones mentioned in this project is the recycled spam traps which is an abandoned email addresses mailbox providers recycled into spam traps that is dormant e-mail accounts.

Dormant e-mail accounts are due to an email account that has not been logged into for a certain period of time. Webmail accounts are free to register but they come with storage and other costs for their providers. Thus, when the email account is deemed dormant then it will become the candidate for automatic deactivation of the account. That is why most providers periodically deactivate accounts that are no longer used to maximize their resources. Yahoo mail for instance only allowed until minimum 6 months of inactivity before an email account are considered dormant.

Furthermore, the suggested programming language for this project is Python programming language because it is understandable and required simple coding. Python is widely used as high-level, general purpose, interpreted, dynamic programming language. Its design philosophy emphasizes code readability whereas its syntax allowed programmers to express concepts in fewer lines of code than would be possible in languages such as C++ or Java.

Moreover, this project creates an automated system which is a system that eliminates the need for human interference in order to complete a task. Several industries have used automated system to increase their production and reduce costs. Automated system helps speed up a process and handles a wide range of tasks. That is why this project could be an advantage towards industries and companies especially those involved in business.

## 1.2 Problem Statement (PS)

Problem statement states the issue occurred and problems that need to be solved from commencing on this project. It shows what current scenario that is happening until this project needs to be conducted.

Table 1.1: Overview of Problem Statement for this project that consists of two problems

PS	Problem Statement
PS1	Spam emails consume high usage of network bandwidth, slowing the data usage of user's email accounts and plummet user deliverability rate.
PS2	Current system has limitation on saving the spam email data for user thus it is not comprehensive enough.

## 1.3 Project Question (PQ)

Project questions are arises from questions that are being asked during project development. Each question should clarify the problem statement stated from the above. Project question helps to centralized main key point to be researched.

Table 1.2: Overview of Project Question that arises for this project

PS	PQ	Project Question
PS1	PQ1	How can spam emails consume network bandwidth?
		What does spam emails did to slow the data usage of user?
		Why does user deliverability rate plummet by spam emails?
PS2	PQ2	When should a spam email data be saved by user?

## 1.4 Project Objective (PO)

Project objectives are the main goals that are going to be focused upon developing this project. A project objective gives a clearer view about what is this whole project about. The process of completing this project depends on the project objective outlined.

Table 1.3: Overview of Project Objective that are focused on this project

PS	PQ	PO	Project Objective
PS1	PQ1	PO1	To investigate samples of unsolicited messages to lure spam away using mechanism of detecting and capturing spam e-mails.
			To measure and analyze the spam e-mails that threatens user usage or requirements and network effectiveness.
PS2	PQ2	PO2	To develop automated system that provides scoring for an email messages to considered it as spam.

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

## 1.5 Project Scope

- i) This project aims to identify any spam e-mail messages that could harm and lower the productivity level of an organization.
- ii) This project focuses on examining an e-mail trap that could help collect samples of unsolicited messages from ISPs.
- iii) This project collect data about spam activities from resourceful Websites such as isecLAB, MyCERT and many more.
- iv) This project uses Python programming language to produce an automated email spam trap.

- v) This project comprises of UTeM facilities and equipments as well as Yahoo accounts user.

## 1.6 Project Contribution

This project will investigate samples of unsolicited messages to lure spam away from trusted websites such as MyCERT, ZDNet, TechRepublic, CNet and many more to get the exact data of spam email. Other than that, this project will measure and analyse the spam e-mails to categorize whether it is harmful or not.

Furthermore, an automated trap e-mail spamming is developed using Python after making analysis on its characteristics and ability to expand more. Thus, this project will contribute to a better automated e-mail spamming that helps user to use e-mail safely and effectively especially towards business organization, institutes or even agencies.

Table 1.4: Overview of Project Contribution after considering its implementation

PS	PQ	PO	PC	Project Contribution
PS1	PQ1	PO1	PC1	Proposed an analysis to categorize spam email
		PO2	PC2	Proposed Python as new programming language in automated email spam trap

## 1.7 Thesis Organization

### CHAPTER 1: INTRODUCTION

- This chapter focuses on a general description about this project. This chapter also focuses on the problem statement, project question, project objectives, project scope, project contribution and thesis organization.

### CHAPTER 2: LITERATURE REVIEW (LR)

- This chapter highlight the previous work contributes towards the research of this project which is the Literature Review (LR). This project also discusses several methodologies and proposed solution based on discovery from this project.

### CHAPTER 3: PROJECT METHODOLOGIES

- This chapter discusses the methodologies used in development of this project. This chapter also provides the project milestones and additional appendices about this project flow chart.

### CHAPTER 4: ANALYSIS AND DESIGN

- This chapter goes deeper into analysis and design which state all related analysis of this project such as problem analysis and requirement analysis, high-level design and detailed design.

### CHAPTER 5: IMPLEMENTATION

- This chapter comprises of implementation part whereby software development environment setup, software configuration management and implementation status is elaborated.

### CHAPTER 6: TEST PLAN

- This chapter outlines the testing for this project. This project also includes test plan, test strategy, test design, and test results and analysis.

### CHAPTER 7: PROJECT CONCLUSION

- This chapter summarizes about this project. This chapter also describes more on project contribution, project limitation and future works for this project.

Figure 1.1: Thesis Organization for Project

## 1.8 Conclusion

In this chapter, the overview about this project is specified which is an automated email spam trap. There are a few specifications and criteria that need to be considered which is the problem arise with spam emails, data about spam email from trusted website, the programming language that are going to be used, the expected output, project contributions and also the scope. Basically, spam email occurs in an organization that is gigantic and deal with business matters. This is because spammers would like to attack the organization operations to lower their productivity and overwork the network traffic so that it causes monetary loss value towards the specified organization (Dhinakaran, Lee, and Nagamalai 2007). That is why this project is proposed to create a better email spam trap to avoid spam emails. In the next chapter which is Chapter 2, the system, techniques, requirements, methodology and milestones for this project are discussed deeply.