

**TOPIC MODELING FOR
INDONESIAN TEXT
UNDERGRADUATE THESIS**

Submitted to the Faculty of Computer Science Universitas Amikom Yogyakarta as a partial fulfillment of the requirement for bachelor degree



**By
Rasyiid Indra Parmadi
15.61.0042**

**BACHELOR DEGREE
STUDY OF INFORMATICS
FACULTY OF COMPUTER SCIENCE
UNIVERSITAS AMIKOM YOGYAKARTA
YOGYAKARTA
2019**

APPROVAL

THESIS

**TOPIC MODELING FOR
INDONESIAN TEXT**

prepared and compiled by

Rasyiid Indra Parmadi

15.61.0042

has been approved by the Thesis Supervisor
on February 9, 2019

Supervisor,



Arief Setyanto, Dr., S. Si, M. T.

NIK. 190302036

ATTESTATION

THESIS

**TOPIC MODELING FOR
INDONESIAN TEXT**

prepared and compiled by

Rasyiid Indra Parmadi

15.61.0042

has been maintained in front of the Board of Examiners

on February 28, 2019

Board of Examiners

Examiners

Signature

Sudarmawan, S. T., M. T.

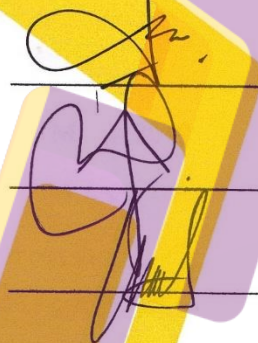
NIK. 190302035

Andi Sunyoto, M. Kom.

NIK. 190302052

Ike Verawati, M. Kom.

NIK. 190302237



This thesis has been accepted as one of the requirements to obtain a Bachelor of

Computer degree on March 19, 2019

DEAN OF THE FACULTY OF COMPUTER SCIENCE



Krisnayati, S. Si, M. T.

NIK. 190302038

STATEMENT

I, the undersigned, hereby declare that, this thesis is my own (ORIGINAL) work, and the contents of this thesis do not have works submitted by others to obtain an academic degree at any higher education institution, and as long as my knowledge is not works or opinions that have been written and / or published by others, except those written in this text and mentioned in the bibliography.

Everything related to the manuscript and the work that has been made is my personal responsibility.

Yogyakarta, 3 March 2019



Rasyiid Indra Parmadi

NIM. 15.61.0042

MOTTO

“Il est bon à savoir, it is good to know.”

(Den Dhimas)

“Do the best and pray. God will take care of the rest.”

(Den Dhimas)

“Use your youth as good as possible.”

(Den Dhimas)

“Everything will be okay in the end, if its not okay, its not the end.”

(Den Dhimas)

DEDICATION

I thank Allah God Almighty for giving His blessings, mercy and guidance so that I can finish this Thesis well. I also feel grateful to the people around me who have directly or indirectly helped me in working on this Thesis. . I present this thesis to :

1. My father, Suparmadi, My mother, Murni Indrawati, and one and only sister Zharifah Indra Parmadi who always prayed for, encouraged, and supported me.
2. Mr Arief Setyanto, Dr., S. Si, M. T. as a supervisor who always provides input and guidance in completing the Thesis.
3. To Raka Wichaksana and Ripto Sudyarno as a mentor who has given his knowledge to help conduct research.
4. My WTF friends, who always give encouragement and motivation to immediately complete the Thesis.
5. My friends in playing game who always help entertain and help in distress.
6. 15BCI classmates, AMIKOM assistants and organizations that have become my friends during college.

As well as all those who have helped and supported me that I cannot mention one by one.

ACKNOWLEDGEMENTS

Thank you we pray to Allah SWT for His blessings and gifts so that the writer can complete the thesis report in time with the title "TOPIC MODELING FOR INDONESIAN TEXT" This thesis was prepared to complete the final assignment of college and fulfill the graduation requirements of the Bachelor Informatics Education program at Amikom University Yogyakarta. During the education of Bachelor Informatics up to the Thesis completion process, various parties have provided facilities, assisted, fostered, and guided the writer for that especially to:

1. Mr. Prof. Dr. M. Suyanto, MM as Chancellor of the Amikom University in Yogyakarta who has provided many facilities in completing education..
2. Mr. Arief Setyanto, Dr., S. Si, M. T. as a supervisor who has spent a lot of time and energy guiding the author during the preparation of this thesis..
3. Mr / Mrs. Lecturer at Amikom University Yogyakarta who has provided writers with several useful disciplines..
4. Friends in arms of Bachelor Informatics Students in 2015, who have discussed and collaborate with writers during their education..

The author realizes, this thesis still has many weaknesses and weaknesses. Therefore, constructive criticism and suggestions will be welcomed, hopefully, the existence of this thesis can be useful and increase our insight, especially about Web Security.

Yogyakarta, 3 March 2019

Author

TABLE OF CONTENTS

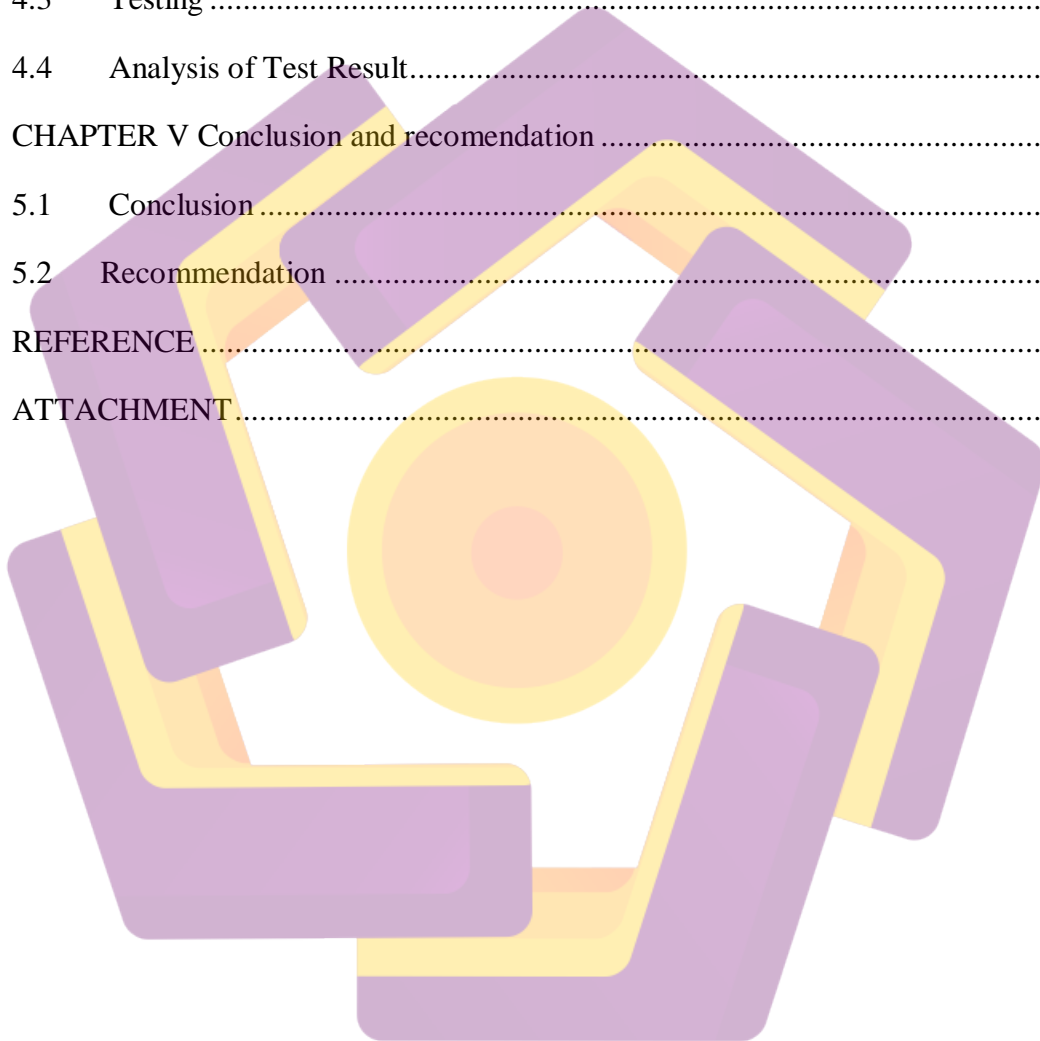
TITLE.....	ii
APPROVAL THESIS.....	iii
ATTESTATION THESIS.....	iv
STATEMENT.....	Error! Bookmark not defined.
MOTTO.....	v
DEDICATION.....	vii
ACKNOWLEDGEMENTS.....	viii
TABLE OF CONTENTS.....	ix
LIST OF TABLES.....	xiv
LIST OF FIGURES.....	xvi
ABSTRACT.....	xviii
CHAPTER I INTRODUCTION.....	1
1.1 Background.....	1
1.2 Statement of the Problem.....	2
1.3 Scope and Limitation.....	2
1.4 Objectives of the Research.....	3
1.5 Benefits of the Research.....	3
1.6 Research Methods.....	4
1.6.1 Literature Review.....	4
1.6.2 Data Collection.....	4
1.6.3 System Development.....	4

1.6.4	System Testing and Analysis	5
1.7	Outline	5
CHAPTER II Theoretical Review		7
2.1	Document.....	7
2.2	Automatic Text Summarization	7
2.2.1	Previous Research	8
2.3	Text Preprocessing	9
2.3.1	Case Folding	9
2.3.2	Filtering	10
2.3.3	Tokenizing	10
2.3.4	Removal Stopword.....	11
2.3.5	Stemming.....	11
2.3.6	TF-IDF (Term Frequency - Inverse Document Frequency) Methods.....	13
2.3.7	Cosine Similarity.....	14
2.4	Basic Concepts of Maximum Marginal Relevance (MMR).....	14
2.5	Evaluation of Text Summarization	16
2.6	Programing Language	18
2.6.1.	PHP.....	18
2.6.2	HTML.....	19
2.6.3	Javascript	19
2.6.4	CSS.....	20
2.6.5	Python.....	20
2.7	Supporting Software.....	21

2.7.1	XAMPP	21
2.7.2	Sublime Text 3.....	21
2.8	Software Model.....	22
2.8.1	DFD (Data Flow Diagram).....	22
CHAPTER III Analysis and design		25
3.1	Problem Analysis	25
3.2	System Analysis.....	25
3.2.1	Analysis of Input Data.....	26
3.2.2	Analysis of Preprocessing	27
3.2.2.1	Case Folding	27
3.2.2.2	Filtering	28
3.2.2.3	Tokenizing	30
3.2.2.4	Removal Stopword.....	34
3.2.2.5	Stemming.....	37
3.2.2.6	Metode TF-IDF (Term Frequency - Inverse Document Frequency)	40
3.2.2.7	Cosine Similarity.....	45
3.2.3	Maximum Marginal Relevance Methods (MMR)	50
3.2.3.1	Cosine Similarity Query to Sentence	50
3.2.3.2	Iteration Calculation.....	56
3.3	Summary Result	59
3.4	Analysis of Non-functional Requirements	60
3.4.1	Analysis of Hardware Requirements.....	60
3.4.2	Analysis of Software Requirements.....	60

3.4.3	Analysis of User Requirements	61
3.5	Analysis of Functional Requirements	61
3.6	System Planning.....	62
3.6.1	Context Diagram.....	62
3.6.2	Data Flow Diagram (DFD) Level 1	62
3.6.3	Data Flow Diagram (DFD) Level 2	63
3.6.4	Process Description.....	64
3.6.5	Interface Design	67
3.6.5.1	Interface of Summary Page	67
3.6.5.2	Interface of About Page	68
3.6.6	Procedural Design.....	68
3.6.6.1	Design of the Main Algorithm.....	68
3.6.6.2	Design of Case Folding	69
3.6.6.3	Design of Filtering	70
3.6.6.4	Design of Tokenizing.....	71
3.6.6.5	Design of Removal Stopword	72
3.6.6.6	Design of Stemming	72
3.7	Dataset	73
CHAPTER IV Implementation and testing.....		78
4.1	Interface Implementation.....	78
4.1.1	Interface of Summary Page	79
4.1.2	Interface of About Page.....	82
4.2	Program Implementation	84

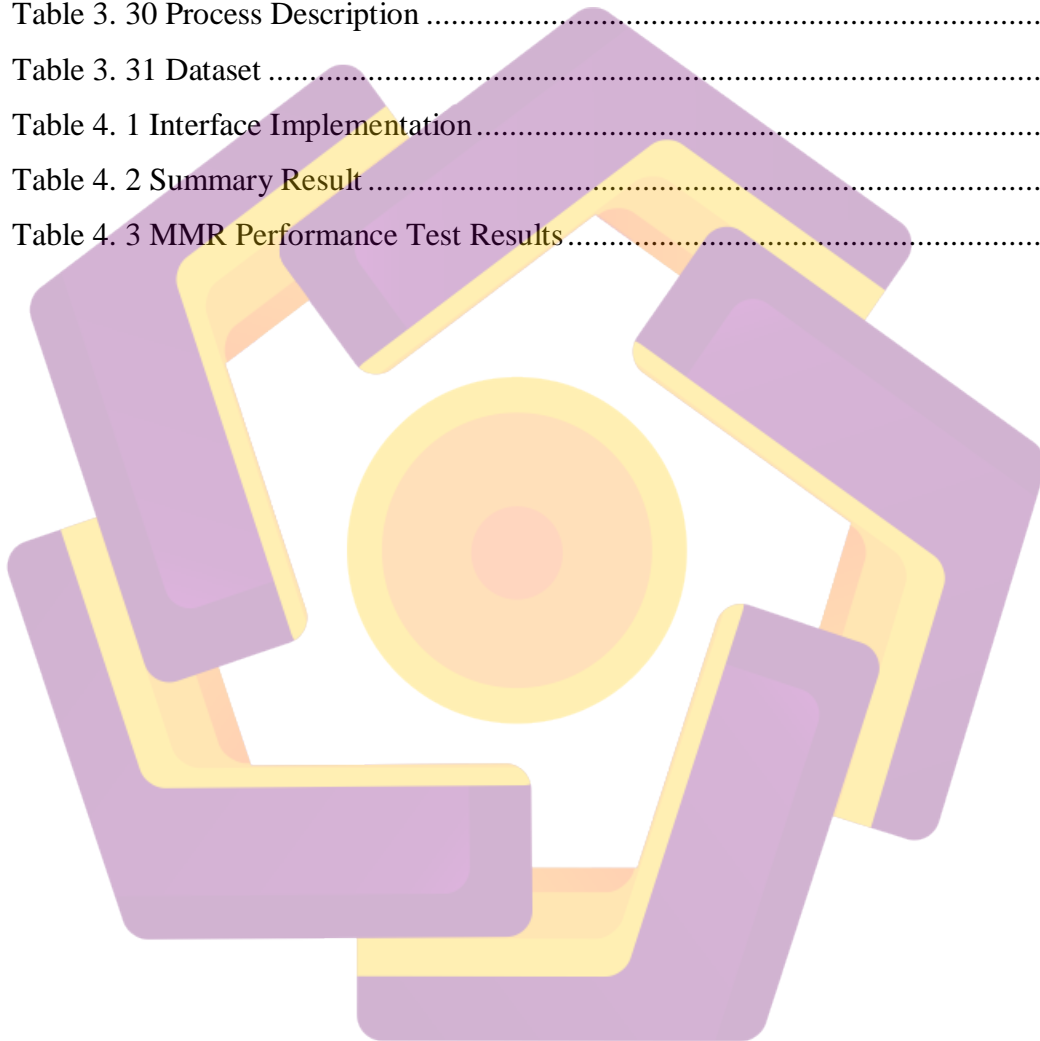
4.2.1	Implementation of Upload File	84
4.2.2	Implementation of Text Summarization.....	87
4.2.3	Implementation of Script	91
4.3	Testing	92
4.4	Analysis of Test Result.....	96
CHAPTER V Conclusion and recommendation		97
5.1	Conclusion	97
5.2	Recommendation	97
REFERENCE		98
ATTACHMENT		101



LIST OF TABLES

Table 3. 1 Input Data	26
Table 3. 2 Case Folding	28
Table 3. 3 Filtering.....	29
Table 3. 4 Sentence Tokenizing	30
Table 3. 5 Word Tokenizing.....	32
Table 3. 6 Removal Stopword.....	34
Table 3. 7 Stemming	37
Table 3. 8 Stemming Result	39
Table 3. 9 TF-IDF Calculation	42
Table 3. 10 $\sum TF - IDF$ S1 Calculation	45
Table 3. 11 Vector Calculation.....	46
Table 3. 12 Result of the Multiplication Between S1 and Other Sentences (S1*Si) ..	47
Table 3. 13 Cosine Similarity Calculation of Sentence 1 (S1).....	48
Table 3. 14 Cosine Similarity Result Between Sentences	49
Table 3. 15 Result of total cosine similarity each sentence	50
Table 3. 16 Filtering Query	51
Table 3. 17 Case Folding	52
Table 3. 18 Tokenizing	52
Table 3. 19 Removal Stopword.....	53
Table 3. 20 Stemming	53
Table 3. 21 $\sum TF - IDF$ 2 Query Calculation.....	54
Table 3. 22 Multiplication Results Between Queries and Other Sentences.....	54
Table 3. 23 Cosine Similarity Calculation Between Queries and Other Sentences ...	55
Table 3. 24 Cosine Similarity Sentences That Will be Processed.....	56
Table 3. 25 1 st MMR Iteration Calculation	57

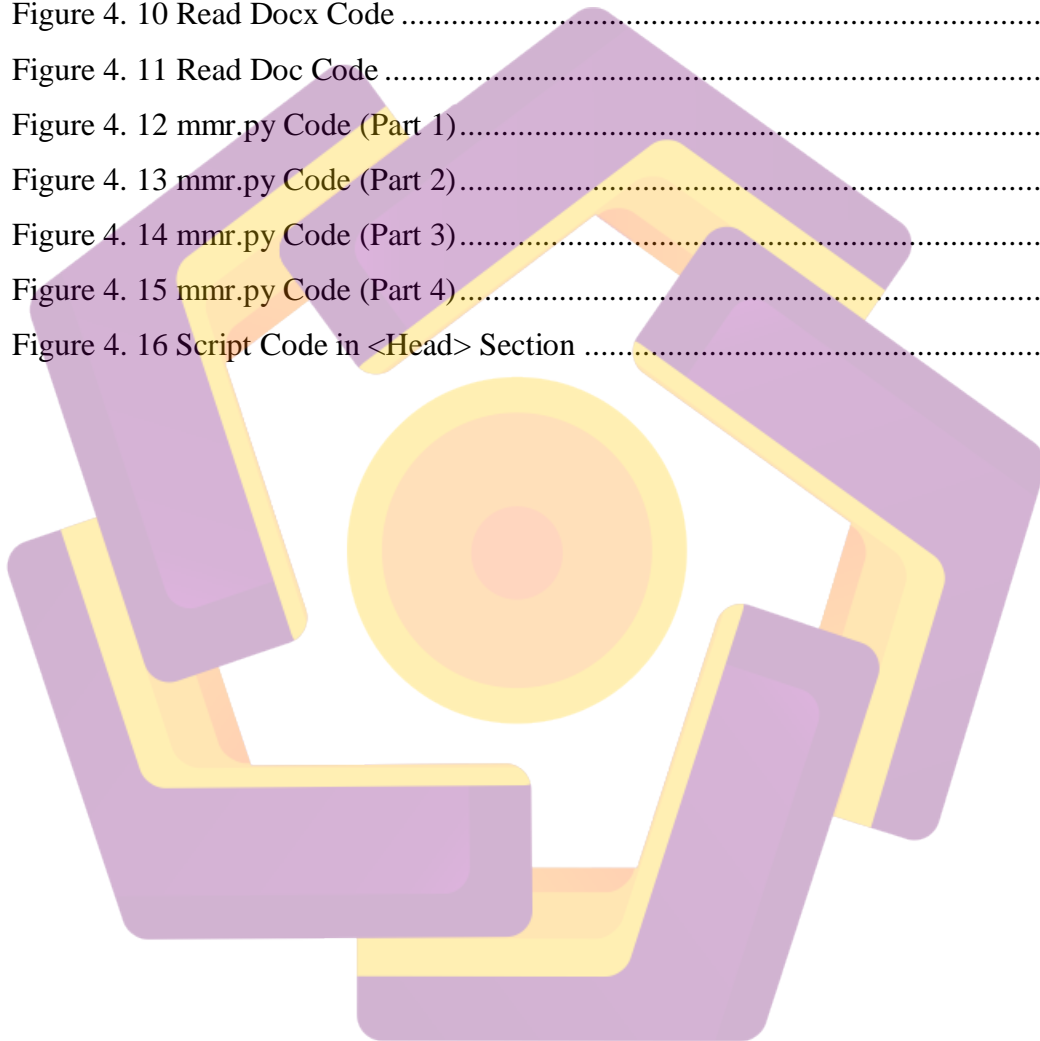
Table 3. 26 Cosine Similarity Sentence 5 (S5).....	58
Table 3. 27 2 nd MMR Iteration Calculation	58
Table 3. 28 Summary	59
Table 3. 29 Summary Result	59
Table 3. 30 Process Description	64
Table 3. 31 Dataset	73
Table 4. 1 Interface Implementation.....	78
Table 4. 2 Summary Result	92
Table 4. 3 MMR Performance Test Results.....	94



LIST OF FIGURES

Figure 2. 1 Preprocessing Process	9
Figure 3. 1 System Overview	26
Figure 3. 2 Preprocessing Process	27
Figure 3. 3 Case Folding Process	27
Figure 3. 4 Filtering Process.....	29
Figure 3. 5 Sentence Tokenizing Process	30
Figure 3. 6 Word Tokenizing Process.....	31
Figure 3. 7 Removal Stopword Process	34
Figure 3. 8 Stemming Process.....	37
Figure 3. 9 ContextDiagram.....	62
Figure 3. 10 DFD Level 1	63
Figure 3. 11 DFD Level 2 Text Summarization Process	63
Figure 3. 12 Summary Page Design	67
Figure 3. 13 About Page Design.....	68
Figure 3. 14 Flowchart System of MMR Text Summarization.....	69
Figure 3. 15 Flowchart Case Folding.....	70
Figure 3. 16 Flowchart Filtering.....	70
Figure 3. 17 Flowchart Tokenizing.....	71
Figure 3. 18 Flowchart Removal Stopword	72
Figure 3. 19 Flowchart Stemming	73
Figure 4. 1 Summary Page Code (Part 1)	79
Figure 4. 2 Summary Page Code (Part 2)	80
Figure 4. 3 Summary Page Code (Part 3)	81
Figure 4. 4 Summary Page Preview.....	81
Figure 4. 5 About Page Code (Part 1).....	82

Figure 4. 6 About Page Code (Part 2).....	83
Figure 4. 7 About Page Code (Part 3).....	83
Figure 4. 8 About Page Preview	84
Figure 4. 9 Upload File Code	85
Figure 4. 10 Read Docx Code	86
Figure 4. 11 Read Doc Code	86
Figure 4. 12 mmr.py Code (Part 1).....	88
Figure 4. 13 mmr.py Code (Part 2).....	89
Figure 4. 14 mmr.py Code (Part 3).....	90
Figure 4. 15 mmr.py Code (Part 4).....	90
Figure 4. 16 Script Code in <Head> Section	91



ABSTRACT

Reading is the activity of perceptual, analyzing, and interpreting what is done by the reader to get the message to be conveyed by the author in the writing media. Along with current technological developments, technology has changed the way a person reads a text, especially text that is in electronic media. Generally, a reader must read the text that is in a particular file thoroughly to find out what topics can be obtained from the text. This becomes a problem in terms of the time and magnitude of the effort used by the reader.

This research will make a document summarization program, especially in Indonesian text using the Maximum Marginal Relevance or MMR algorithm. The MMR algorithm is a simple and efficient text summarization algorithm. With text summarizing programs, change someone to find and find out what topics are available in the text without having to read the text thoroughly.

This program will produce topics or conclusions from a text and help someone to find out what topics are in the text without reading it. This program will help someone who doesn't like the process of reading a lot of text. After getting the topic, he will look for other explanations about the topic from the internet to learn manually.

Keywords : Summarization, Indonesian Text, Maximum Marginal Relevance (MMR).