

TABLE OF CONTENTS

APPROVAL PAGE	iii
STATEMENT	iv
FOREWORD	vi
TABLE OF CONTENTS	vii
LIST OF FIGURES	x
LIST OF TABLES.....	xiii
ABSTRACT.....	xv
CHAPTER I INTRODUCTION	1
1.1 Research Background	1
1.2 Research Problem	2
1.3 Research Scope	3
1.4 Research Objectives.....	3
1.5 Research Benefits.....	3
1.6 Research Methodology	4
1.7 Thesis Outline	5
CHAPTER II LITERATURE REVIEW	7
CHAPTER III THEORETICAL BASIS	13
3.1 Questionnaire Data.....	13
3.2 Word Representation.....	13
3.3 Text Clustering.....	18
3.4 DBSCAN	19
3.4.1 DBSCAN Process	21
3.4.2 Benefits of DBSCAN.....	22
3.4.3 Drawbacks of DBSCAN.....	23
3.4.4 Determining the parameters Eps and Minpts.....	23
3.4.5 Euclidean Distance	24
3.5 Accelerated HDBSCAN*	25
3.5.1 HDBSCAN*.....	25
3.5.2 Improvement to HDBSCAN*	26

3.6 Performance Evaluation.....	33
3.6.1 Silhouette Coefficient	33
3.6.2 Calinski-Harabaz Index.....	34
3.6.3 Run Time	35
CHAPTER IV ANALYSIS AND DESIGN	36
4.1 Problem Analysis.....	36
4.2 Analysis of Dataset.....	37
4.2.1 Dataset.....	38
4.2.2 Data Preparation Design	38
4.3 Word Representation Design.....	39
4.4 Parameter Initialisation Design	41
4.4.1 Parameter Initialisation for DBSCAN.....	41
4.4.2 Parameter Initialisation for Accelerated HDBSCAN*	42
4.5 Clustering Design	42
4.6 Evaluation Design.....	43
CHAPTER V IMPLEMENTATION	44
5.1 Hardware and Software Specification	44
5.3 Implementation of Word Vector Representation.....	45
5.4 Implementation of DBSCAN.....	48
5.5 Implementation of Accelerated HDBSCAN*	50
5.6 Implementation of Silhouette Coefficient.....	51
5.6 Implementation of Calinski-Harabaz Index.....	51
5.8 Implementation of Run Time.....	51
CHAPTER VI RESULT AND DISCUSSION	52
6.1 Evaluation on Sample of 5,000	52
6.1.1 Silhouette Coefficient for Sample of 5,000.....	52
6.1.2 Run Time for Sample of 5,000	55
6.2 Evaluation on Sample of 6,000	57
6.2.1 Silhouette Coefficient for Sample of 6,000.....	57
6.2.2 Run Time for Sample of 6,000	60
6.3 Evaluation on Sample of 7,000	62
6.3.1 Silhouette Coefficient for Sample of 7,000.....	62
6.3.2 Run Time for Sample of 7,000	66
6.4 Evaluation on Sample of 8,000	68
6.4.1 Silhouette Coefficient for Sample of 8,000.....	68
6.4.2 Run Time for Sample of 8,000	71
6.5 Evaluation on Sample of 9,000	74

6.5.1 Silhouette Coefficient for Sample of 9,000.....	74
6.5.2 Run Time for Sample of 9,000.....	77
6.6 Evaluation on Sample of 10,000.....	79
6.6.1 Silhouette Coefficient for Sample of 10,000.....	79
6.6.2 Run Time for Sample of 10,000.....	83
6.7 Average Silhouette Coefficient and Run Time.....	86
6.7.1 Average Silhouette Coefficient.....	86
6.7.2 Average Run Time.....	87
6.8 Calinski-Harabaz Index.....	88
6.8.1 Calinski-Harabaz Index for Sample of 6,000.....	89
6.8.2 Calinski-Harabaz Index for Sample of 9,000.....	91
6.9 Summary of Result and Discussion.....	93
CHAPTER VII CONCLUSION AND SUGGESTION.....	95
7.1 Conclusion.....	95
7.2 Suggestion.....	95
REFERENCES.....	96