

## TABLE OF CONTENTS

|  |      |
|--|------|
| APPOVAL PAGE .....                     | iii  |
| STATEMENT .....                        | iv   |
| FOREWORD .....                         | vi   |
| TABLE OF CONTENTS .....                | viii |
| LIST OF TABLES .....                   | xi   |
| LIST OF FIGURES .....                  | xii  |
| ABSTRACT .....                         | xiii |
| INTISARI .....                         | xiv  |
| CHAPTER I INTRODUCTION .....           | 1    |
| 1.1 Research Background .....          | 1    |
| 1.2 Research Problem .....             | 4    |
| 1.3 Research Scope .....               | 4    |
| 1.4 Research Objective .....           | 4    |
| 1.5 Research Advantage .....           | 4    |
| 1.6 Research Methodology .....         | 5    |
| 1.7 Thesis Organization .....          | 6    |
| CHAPTER II LITERATURE REVIEW .....     | 8    |
| CHAPTER III THEORETICAL BASES .....    | 13   |
| 3.1 IT Auditing .....                  | 13   |
| 3.2 Audit Evidence .....               | 14   |
| 3.3 Data Mining .....                  | 14   |
| 3.4 Information Retrieval .....        | 16   |
| 3.5 Text Mining .....                  | 16   |
| 3.6 Term Weighting .....               | 18   |
| 3.6.1 Term Frequency .....             | 18   |
| 3.6.2 Document Frequency .....         | 18   |
| 3.6.3 Inverse Document Frequency ..... | 18   |
| 3.6.4 TF-IDF .....                     | 18   |
| 3.7 Clustering .....                   | 19   |
| 3.8 Document Clustering .....          | 20   |

|   |    |
|---|----|
| 3.9 Latent Semantic Analysis .....                    | 20 |
| 3.9.1 Singular Value Decomposition .....              | 21 |
| 3.10 K-Means Clustering .....                         | 22 |
| 3.10.1 Euclidean distance .....                       | 22 |
| 3.11 Evaluation Method .....                          | 23 |
| 3.11.1 Sum of Squared Error .....                     | 23 |
| 3.11.2 Silhouette Coefficient .....                   | 24 |
| CHAPTER IV ANALYSIS AND DESIGN .....                  | 26 |
| 4.1 General Overview .....                            | 26 |
| 4.2 General Research Design .....                     | 27 |
| 4.3 Data Preparation .....                            | 29 |
| 4.3.1 Case Folding .....                              | 30 |
| 4.3.2 Stop-word removal .....                         | 30 |
| 4.3.3 Stemming .....                                  | 31 |
| 4.3.4 Tokenization .....                              | 32 |
| 4.4 TF-IDF .....                                      | 33 |
| 4.5 Singular Value Decomposition .....                | 36 |
| 4.6 K-Means Clustering .....                          | 39 |
| 4.7 Evaluation .....                                  | 42 |
| CHAPTER V IMPLEMENTATION .....                        | 44 |
| 5.1 Specification .....                               | 44 |
| 5.2 Preprocessing Implementation .....                | 44 |
| 5.3 TF-IDF Implementation .....                       | 45 |
| 5.4 Matrix Decomposition .....                        | 47 |
| 5.5 K-Means Clustering Implementation .....           | 47 |
| 5.6 Sum squared error code implementation .....       | 48 |
| 5.7 Silhouette coefficient evaluation .....           | 49 |
| CHAPTER VI RESULT AND DISCUSSION .....                | 50 |
| 6.1 Data Preparation Result .....                     | 50 |
| 6.2 Term-document Matrix Result .....                 | 51 |
| 6.3 Evaluation results .....                          | 53 |
| 6.3.1 First experiments with 80% reduced matrix ..... | 53 |

|  |   |    |
|--|---|----|
| 6.3.2                                  | Second Experiment with 85% reduced matrix ..... | 54 |
| 6.3.3                                  | Third experiment with 90% reduced matrix .....  | 55 |
| 6.3.4                                  | Fourth experiment with 80% reduced matrix ..... | 56 |
| 6.3.5                                  | Fifth experiment with 85% reduced matrix .....  | 56 |
| 6.3.6                                  | Sixth experiment with 90% reduced matrix .....  | 57 |
| 6.4                                    | Clustering results.....                         | 58 |
| 6.4.1                                  | SSE Evaluation .....                            | 58 |
| 6.4.2                                  | Silhouette Coefficient Evaluation .....         | 59 |
| 6.5                                    | Clustering results without SVD.....             | 61 |
| 6.5.1                                  | Sum of Squared Error score .....                | 62 |
| 6.5.2                                  | Silhouette score .....                          | 63 |
| 6.6                                    | Summary of Experiments .....                    | 64 |
| CHAPTER VI RESULT AND DISCUSSION ..... |   | 66 |
| 7.1                                    | Conclusions .....                               | 66 |
| 7.2                                    | Future Works .....                              | 67 |
| REFERENCES.....                        |   | 68 |