



FINAL YEAR PROJECT REPORT

Web Mining & Text Classification

By

Abdul Haseeb Khan (19364)

Supervised by

Mr. Furqan Hussain Essani

Bahria University (Karachi Campus)

2013

Acknowledgment

First and foremost, I would like to thank my supervisor, Mr. Furqan Hussain Essani for his consistent help and advice in the creation of this report and for his continuing support throughout the project. I would also like to thank my friends and family for the various ways in which they have supported me throughout my final year.

Finally, I would like to thank you, the reader, for taking the time to glance through this work. I sincerely hope that it is both comprehensive and informative, and that the conclusions drawn are worthwhile.

Abstract

The process of categorizing text data can be quite tedious. With the growth of text data on the internet, computer scientists have come up with algorithms related to machine learning that help in text categorization. In this project I have automated the process of document classification and indexing by the use of machine learning algorithms. The main aim of the project is to help users, who have large amounts of text data, in classification and indexing of documents.

Key words:

Text Mining, Machine Learning, Naïve Bays, Categorization, Indexing, Lucene, MALLET

Contents

1	INTRODUCTION.....	6
1.1	PURPOSE	6
1.2	PROBLEM	7
1.3	SOLUTION.....	7
1.4	SCOPE OF THE PROJECT.....	7
1.5	ORGANIZATION OF REPORT.....	8
2	BACKGROUND AND LITERATURE REVIEW	9
2.1	BACKGROUND	9
2.2	TEXT EXTRACTION FROM THE WEB VIA TEXT-TO-TAG RATIO.....	9
2.3	PERFORMANCE ANALYSIS AND OPTIMIZATION ON LUCENE	10
2.4	PARTIALLY SUPERVISED CLASSIFICATION OF TEXT DOCUMENTS.....	10
2.5	AUTOMATIC TEXT CATEGORIZATION BY UNSUPERVISED LEARNING	10
2.6	AN IMPROVED TF-IDF APPROACH FOR TEXT CLASSIFICATION	11
3	REQUIREMENTS.....	12
3.1	SYSTEM ENVIRONMENT	13
3.2	FUNCTIONAL REQUIREMENTS SPECIFICATION.....	13
3.2.1	<i>Content Extractor</i>	13
3.2.2	<i>Indexer</i>	14
3.2.3	<i>Supervised Classification</i>	16
3.2.4	<i>Content Extractor</i>	18
3.2.5	<i>Indexer</i>	21
3.2.6	<i>Supervised Classification</i>	23
3.3	USER CHARACTERISTICS	30
3.4	NON-FUNCTIONAL REQUIREMENTS.....	30
3.5	STAKEHOLDERS & POSSIBLE CUSTOMERS	30
3.6	SCHEDULING	31
3.6.1	<i>Spreadsheet</i>	32
3.7	FEASIBILITY REPORT.....	33
3.7.1	<i>Operational Feasibility</i>	33
3.7.2	<i>Economical Feasibility</i>	34
➤	<i>Development Cost</i>	34
3.8	PROBLEMS FACED	35
3.8.1	<i>Problems Faced in Content Extraction</i>	35
3.8.2	<i>Problems Faced in Indexing</i>	35
3.8.3	<i>Problems Faced in Classification</i>	36
4	METHODOLOGY AND DESIGN	37
4.1	OVERALL SYSTEM STRUCTURE AND USED DATA STRUCTURES	37
4.2	PROCESS MODEL	42
4.2.1	<i>Planning</i>	42

4.2.2	<i>Requirements</i>	43
4.2.3	<i>Analysis & Design</i>	43
4.2.4	<i>Implementation</i>	43
4.2.5	<i>Testing</i>	44
4.2.6	<i>Deploy / Integrate</i>	44
4.2.7	<i>Evaluate</i>	44
4.3	GRAPHICAL USER INTERFACE (GUI).....	45
4.3.1	<i>GUI of Content Extractor</i>	45
4.3.2	<i>GUI of Create SVM Files</i>	46
4.3.3	<i>GUI of Create Classifier</i>	47
4.3.4	<i>GUI of Classify Files</i>	48
4.3.5	<i>GUI of View Results</i>	49
4.3.6	<i>GUI of Search Index</i>	51
5	IMPLEMENTATION	53
5.1	IMPLEMENTATION OF CONTENT EXTRACTOR MODULE.....	53
5.1.1	<i>Implementation of Link Extractor Class</i>	53
5.1.2	<i>Implementation of Content Extractor Class</i>	56
5.2	IMPLEMENTATION OF INDEXER MODULE	59
5.2.1	<i>Implementation of Lucene Indexer Class</i>	59
5.3	IMPLEMENTATION OF SUPERVISED CLASSIFICATION MODULE.....	66
5.3.1	<i>Implementation of Converting Files to Vectors</i>	66
5.3.2	<i>Implementation of TF-IDF Vectors</i>	68
5.3.3	<i>Implementation of Classification Class</i>	72
5.3.4	<i>Implementation of Results</i>	74
5.4	IMPLEMENTATION OF COSINE SIMILARITY MODULE.....	75
5.5	CONCLUSION OF IMPLEMENTATION.....	78
6	TESTING AND EVALUATION	79
6.1	WHITE BOX TESTING	80
6.1.1	<i>Testing on Content Extractor</i>	81
6.1.2	<i>Testing on Indexer</i>	81
6.1.3	<i>Testing on Classification</i>	82
6.2	BLACK BOX TESTING	84
6.3	EVALUATION.....	85
7	CONCLUSIONS AND FUTURE WORK	87
7.1	FUTURE WORK.....	87
REFERENCES		89
RESEARCH PAPERS.....		89
WEBSITES		89
APPENDICES		91
APPENDIX A.....		91
APPENDIX B.....		92

APPENDIX C.....	93
APPENDIX D.....	94
GLOSSARY.....	95