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