

WEB PAGE CLASSIFICATION

**Submitted in the partial fulfillment of requirements
for the award of degree of**

**BACHELOR OF TECHNOLOGY
IN
COMPUTER ENGG.**

**Under the supervision of
Mr. Faiyaz Ahmad**

**Submitted By
Akshay Kumar(10-CSS-06)
Niyas C(10-CSS-44)**



**Department of Computer Engg.
Faculty of Engineering and Technology
Jamia Millia Islamia
New Delhi-110025
2013-2014**

CERTIFICATE

This is to certify that project entitled "Web Page Classification" done by Akshay Kumar(10-CSS-06) and Niyas C(10-CSS-44) is an authentic work carried out under my guidance.

The matter embodied in this project has not submitted earlier for the award of any degree or diploma to the best of my knowledge and belief.

Date:

Mr. Faiyaz Ahmad
Assistant Professor
Department of Computer Engineering
Faculty of Engineering and Technology
Jamia Millia Islamia
New Delhi

ACKNOWLEDGEMENT

We are greatly indebted to our supervisor and guide Mr. Faiyaz Ahmad for his invaluable technical guidance, great innovative ideas and overwhelming moral support during the course of the project. We are grateful to our H.O.D. Professor M.N.Doja for his invaluable support throughout the project.

We are also thankful to Department of Computer Engineering and the entire faculty members especially Prof. Md. Sufiyan Beg, Mr. Sarfaraz Masood, Dr. Tanvir Ahmad, Dr. Bashir Alam, Dr. Amjad, Md. Zeeshan Ansari, Mr. Danish Raza Rizvi, Mr. Mumtaz Ahmad, Mr. Jawahar Lal and Mr. Shehzad and Mr Musheer Ahmad for their teachings, guidance and encouragement. We are also thankful to our classmates and friends for the valuable suggestions and active support.

We would like to extend a special thanks to our families for their constant motivation and encouragement throughout the tenure of this work.

Akshay Kumar

10-CSS-06

Department of Computer Engg.

Faculty of Engineering and Tech.

Jamia Millia Islamia

New Delhi-25

Niyas C

10-CSS-44

Department of Computer Engg.

Faculty of Engineering and Tech.

Jamia Millia Islamia

New Delhi-25

INDEX

1	Abstract	1
2	Introduction	2
3	Related Work	4
4	Classification of Web Pages	5
5	Bayes Algorithm	6
	a) Bayesian Classifiers	6
	b) Bayes Theorem	6
	c) Naive Bayes Classifier	7
	d) Modification on Naive Bayes for document classification	8
	e) Laplacian Correction	9
6	Experimental Setup	9
	a) Creation of dataset	9
	b) Cleaning HTML documents	9
	c) Vocabulary generation	13
	d) Testing and training Classifier	14
7	Experimental Results	17
8	Area of Application	20
9	Conclusion and Future Scope	21
10	References	22