Document Clustering and Ranking

A Project report submitted
in partial fulfillment of requirements
for the award of degree of

Bachelor of Technology
In
Information Technology

By

K.Sai Shilpa	**(Reg No:14131A1252)**
A.Supriya	**(Reg No:14131A1202)**
A.Yedukondalu	**(Reg No:14131A1204)**
G.Sai Krishna	**(Reg No:14131A1226)**

Under the esteemed guidance of

Mr. S.Y.PAVAN KUMAR
(Assistant Professor)
Department of Information Technology

Department of Information Technology
GAYATRI VIDYA PARISHAD COLLEGE OF ENGINEERING(AUTONOMOUS)
(Affiliated to JNTU-K, Kakinada)
VISAKHAPATNAM

2017 - 2018

Gayatri Vidya Parishad College of Engineering (Autonomous)

Visakhapatnam

CERTIFICATE

This report on ***"Document Clustering and Ranking"*** is a bonafide record of the project report submitted

By

K.Sai Shilpa	**(Reg No:14131A1252)**
A.Supriya	**(Reg No:14131A1202)**
A.Yedukondalu	(Reg No:14131A1204)
G.Sai Krishna	**(Reg No:14131A1226)**

in their VIII semester in partial fulfillment of the requirements for the Award of Degree of

Bachelor of Technology

In

Information Technology

During the academic year 2017-2018

Mr. S.Y.Pavan Kumar	Dr.K.B.Madhuri
Assistant Professor	Head of the Department
Project Guide	Department of Information Technology

External Examiner

DECLARATION

We hereby declare that this project entitled "**Document Clustering And Ranking**" is a bonafide work done by us and submitted to Department of Information Technology G.V.P College of Engineering (Autonomous) Visakhapatnam, in partial fulfillment for the award of the degree of B. Tech is of our own and it is not submitted to any other university or has been published any time before.

PLACE: Visakhapatnam K Sai
Shilpa(14131A1252)

A Supriya

(14131A1202)

A Yedukondalu

(14131A1204)

G Sai

Krishna(14131A1226)

ACKNOWLEDGEMENT

We would like to take this opportunity to extent our hearty gratitude to our esteemed institute “Gayatri Vidya Parishad College of Engineering (Autonomous)“ where we got the platform to fulfill our cherished desire.

We express our sincere thanks to Prof. A.B.KOTESWARA RAO, Principal of Gayatri Vidya Parishad College of Engineering (Autonomous), for his support and encouragement during the course of this project.

We express our deep sense of gratitude to Prof. DR.K.B.MADHURI, Head of Department, Department of Information Technology, for her constant encouragement.

We also thank Asst.Prof.D.NAGATEJ, project coordinator, Department of Information Technology, for guiding us throughout the project and helping us in completing the project efficiently.

We are obliged to **MR.S.Y.Pavan Kumar**, Department of Information Technology, who has been our guide, whose valuable suggestions, guidance and comprehensive assistance helped us a lot in realizing the project.

We would like to thank all the members of teaching and non-teaching staff of Department of Information Technology, for all their support.

Lastly, we are grateful to all my friends, for their relentless support in augmenting the value of work, our family, for being considerate and appreciative throughout.

Project Members-

K Sai Shilpa(14131A1252)
A Supriya (14131A1202)
A Yedukondalu (14131A1204)
G Sai Krishna(14131A1226)

ABSTRACT

An application to collect the documents, form groups out of the given documents based upon the users request, later retrieve the exact and the most relevant document(s) among the clusters, depending upon the search query from the user. We are expecting to provide the user with the most relevant document for the given query through this application and pdf to text convertor as this application is purely text based and generally the ebooks are of pdf type.

The existing system has a facility that is when given a query; it searches only the title of the book and author of the book for the particular word given by the user. This process may not give the accurate result of what the user is expecting and we know that every user expects to retrieve the search result in less time.

So, we propose this work of document clustering and ranking, which is an automatic grouping of documents (.pdf) convert them into text files later on divide them into clusters in order to reduce the time of searching of the documents. Retrieving the information and documents as per the user requested queries which are relevant to what the user is expecting based on the ranking is provided to the users.

CONTENTS