AUTO TEXT SUMMARIZATION

Submitted in partial fulfillment of the requirements
of the degree of
Bachelor of Engineering in Information Technology

By

Vishal Akshali

Sayali Jadhav

Mandar Kambli

Prasad Kothavale

Supervisor:

Mrs. Asma Parveen I Sidhavattam

Department of Information Technology
Vivekanand Education Society's Institute of Technology
2013-14

Project Report Approval for B. E.

This project report entitled AUTO TEXT SUMMARIZATION by **Vishal Akshali, Sayali Jadhav, Mandar Kambli, Prasad Kothavale** is approved for the degree of **Bachelor of Engineering in Information Technology.**

Examiners

1.----------

2.----------

Supervisors

1.----------

2.----------

Date:

Place:

Declaration

I declare that this written submission represents my ideas in my own words and where others' ideas or words have been included, I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Vishal Akshali

Sayali Jadhav

Mandar Kambli

Prasad Kothavale

Date:

ACKNOWLEDGEMENT

This project would not have been successfully accomplished without encouragement from **Mrs. Asma Parveen I Sidhavattam**, our faculty advisor, Assistant Professor in Information Technology Dept. of Vivekanand Education Society's Institute of Technology. It was her faith in us that gave us the confidence to undertake this project.

We would also like to thank our **PRINCIPAL Dr. (Mrs.) J. M. Nair & Mrs. Vijayalakshmi Murlidhar**, **HEAD OF DEPARTMENT** and the college for extending its support and providing the entire necessary infrastructure our project demanded. Thanks and appreciation to the helpful people of Information Technology Dept. for their support. We would also like to thank our parents and friends for all the support and encouragement.

Abstract

Summarization is a complex task that requires understanding of the document content to determine the importance of the text. Lexical cohesion is a method to identify connected portions of the text based on the relations between the words in the text. Lexical cohesive relations can be represented using lexical chains. Lexical chains are sequences of semantically related words spread over the entire text. Lexical chains are used in variety of Natural Language Processing (NLP) and Information Retrieval (IR) applications. In current thesis, we propose a lexical chaining method that includes the glossary relations in the chaining process. These relations enable us to identify topically related concepts, for instance dormitory and student, and thereby enhances the identification of cohesive ties in the text. We then present methods that use the lexical chains to generate summaries by extracting sentences from the document(s). Headlines are generated by filtering the portions of the sentences extracted, which do not contribute towards the meaning of the sentence. Headlines generated can be used in real world application to skim through the document collections in a digital library. Multi-document summarization is gaining demand with the explosive growth of online news sources. It requires identification of the several themes present in the collection to attain good compression and avoid redundancy. In this thesis, we propose methods to group the portions of the texts of a document collection into meaningful clusters. Clustering enable us to extract the various themes of the document collection. Sentences from clusters can then be extracted to generate a summary for the multi-document collection. Clusters can also be used to generate summaries with respect to a given query. We designed a system to compute lexical chains for the given text and use them to extract the salient portions of the document

Table of Contents Page no.

LIST OF TABLES

LIST OF FIGURES