

BUSITEMA UNIVERSITY
FACULTY OF ENGINEERING
DEPARTMENT OF COMPUTER ENGINEERING
FINAL YEAR PROJECT REPORT
An Online News Event Summarization System

BY
MUGANYIZI ALEX

Reg. No: BU/UG/2013/43

Email: alexmuganyizi8@gmail.com

Tel: +256700207096/0785226630

Supervisor: MR. ODONGTOO GODFREY

A project Submitted to the Department of Computer Engineering in Partial Fulfillment of the Requirements for the Award of a Bachelor's Degree in Computer Engineering of Busitema University

May, 2017

DECLARATION

I, Muganyizi Alex do hereby declare that this Project Report is original and has not been submitted for any other degree to any other University before.

Signature:Date:

Name: MUGANYIZI ALEX

Bachelor of Science in (BU)

Department of Computer Engineering

Busitema University

APPROVAL

This is to certify that the project report under the title “**An Online News event summarization system**” fully worked on and submitted to the department of computer engineering for examination under my supervision

Sign:

Date:

Mr. ODONGTOO GODFREY

Department of Computer Engineering

LIST OF ACRONYMS

NLP	Natural Language Processing
RST	Rhetorical Structure Theory
AKE	Automatic Key Phrase Extraction
TL; DR	Too Long; Didn't Read
MTT	Meaning-Text Theory
API	Application Programming Interface
PC	Personal Computer
RSS	Rich Site Summary
IDE	Integrated development Environment

TABLE OF FIGURES

Figure 3. 1 Proposed block diagram	20
--	----

ABSTRACT

The massive quantity of data available today in the Internet has reached such a huge volume that it has become humanly unfeasible to efficiently sieve useful information from it. With the advent of connected computing devices, we are being presented with a barrage of information every minute. This makes it increasingly important to consume as much information as possible in the least amount of time, while eliminating irrelevant and redundant data.

News is one major domain which falls prey to this information overload. With the emergence of lightning-fast news delivery through the various News Sources.

This project is aimed at developing a summarization system to generate abstracts of original news articles.

TABLE OF CONTENTS

DECLARATION	i
APPROVAL	ii
LIST OF ACRONYMS	iii
TABLE OF FIGURES	iv
ABSTRACT.....	v
CHAPTER ONE	10
INTRODUCTION	10
1.0 Introduction.....	10
1.1 Background	10
1.2 Problem statement:.....	11
1.3 Objectives	11
1.3.1 Main Objective	11
1.3.2 Specific objectives	11
1.4 Justification	11
1.5 Scope.....	12
1.5.1 Technical scope	12
1.5.2 Time scope	12
CHAPTER TWO	13
LITERATURE REVIEW	13
2.0 Introduction.....	13
2.1 Document Summarization	13
2.2 Types of summarization.....	13
2.2.1 Extraction based summarization	13
2.2.1.3 Deep parsing algorithms:	14
2.2.2 Abstractive summarization	14
2.3 Natural Language Processing.....	14
2.4 Text Summarization Techniques	15
2.4.1 Key Phrase Extraction	15
2.4.2 Lexical Chains	15
2.4.3 Co-reference Chains	15
2.4.4 Text Rank	16
2.5 Existing systems and related work.....	16

2.5.1 Cut and Paste Based Text Summarization	16
2.5.2 Tetranet Software (Extractor)	17
2.5.3 The Summly app	17
2.5.4 Text Analyst	17
2.5.5 Microsoft (MS Word Auto Summarization)	18
2.5.6 Other key phrase extraction systems	18
2.6 Proposed system.....	18
CHAPTER THREE	19
METHODOLOGY	19
3.0 Introduction.....	19
3.1 Requirements elicitation	19
3.2 Data collection	19
3.3 Data analysis	19
3.4 System design	20
3.4.1 Block diagram of the system	20
3.4.2 Tools to be used	20
3.5 System development.....	21
3.6 Testing and validation.....	22
3.6.1 Unit Testing	22
3.6.2 Integration testing	22
3.6.3 System testing	22
CHAPTER FOUR.....	23
SYSTEM ANALYSIS AND DESIGN.....	23
4.0 Introduction.....	23
4.1 functional Analysis	23
4.2 Requirements Analysis	23
4.2.1 Functional requirements	23
4.2.2 Non-functional requirements	24
4.3 Entity Relations.....	24
4.4 Use case Diagrams	25
CHAPTER FIVE	26
IMPLEMENTATION AND TESTING.....	26
5.0 Introduction.....	26

5.1 Development platforms.....	26
5.1.1 MySQL	26
5.1.2 Windows 10	26
5.2 Code Design.....	26
Code to rank sentences.....	27
Code to generate summaries	27
5.3 Verification	28
5.3.1 Testing	28
5.3.2 Computer specifications	28
5.4 Validation.....	28
5.5 Evaluation	28
CHAPTER SIX.....	29
RECOMMENDATIONS AND CONCLUSION.....	29
6.0 Introduction.....	29
6.1 Relevance of the findings.....	29
6.2 Critical analysis.....	29
6.3 Recommendation	29
6.3.1 Integration of the system into the local/native languages	29
6.3.2 Integrating the system with mobile phones	30
6.3.3 Working to solve the problem brought about by sites with cookies	30
REFERENCES	31
APPENDICES	33
APPENDIX A.....	33
APPENDIX B	34

CHAPTER ONE

INTRODUCTION

1.0 Introduction

This chapter comprises of background, problem statement, justification and objective of the study.

1.1 Background

The massive quantity of data available today in the Internet has reached such a huge volume that it has become humanly unfeasible to efficiently sieve useful information from it. With the advent of connected computing devices, we are being presented with a barrage of information every minute. This makes it increasingly important to consume as much information as possible in the least amount of time, while eliminating irrelevant and redundant data.

News is one major domain which falls prey to this information overload. With the emergence of lightning-fast news delivery through the various News Sources, It is a dire need of the day to save time and grasp just enough information that is required about current events[1].

Humans have also reported problems in understanding or making decisions when faced with excessive amounts of information mostly online, which is nowadays known as Information Overload problem. There is hardly any time to read everything and yet we have to make critical decisions based on whatever information is available volume that it has become humanly unfeasible to efficiently sieve useful information from it[2].

The abundance of information also makes the search for relevant information more complex like finding a needle in a haystack. At the same time, the abundance of information does not always cover relevant information to understand or make decisions, also known as, Information Scarcity problem. The scale and complexity of the Information Overload and Scarcity problems increased with the rise of modern computers in the 1960s. The modern computers were connected to create the Internet in the late 1970s. This network became global and brought access to a very large amount of information. Daily news articles and broadcast news are a good example of huge amounts of information daily published in the Internet. Again, people have more difficulty finding information to understand events in news documents, a reasonable solution is to generate a

REFERENCES

- [1] Zadbukey, "International Journal of Advanced Research in Computer Science and Software Engineering," pp. 124-127, March 2016.
- [2] R. M. Nabi, R. A. Mohammed, and R. M. Nabi, "International Journal of Advanced Research in Computer Science and Software Engineering," *International Journal*, vol. 3, no. 6, 2013.
- [3] A. T. Schutz., "Key-phrase Extraction from Single Documents in the Open Domain Exploiting Linguistic and Statistical Methods.," *Master's thesis, National University of Ireland*, 2008.
- [4] H. Zha, "Generic Summarization and Keyphrase Extraction using Mutual Reinforcement Principle and Sentence," in *25th Annual International ACM SIGIR Conference on Research and Development on Information Retrieval*, New York, NY, USA, 2002, pp. 113-120: ACM.
- [5] I. Mani and M. T. Maybury, *Advances in automatic text summarization*. MIT Press, 1999.
- [6] R. M. Aliguliyev, "Automatic document summarization by sentence extraction," *Вычислительные технологии*, vol. 12, no. 5, 2007.
- [7] E. D. Liddy, "Natural Language Processing. In Encyclopedia of Library and Information Science," vol. 2nd 2001.
- [8] K. S. Hasan and V. Ng, "Conundrums in unsupervised keyphrase extraction: making sense of the state-of-the-art," in *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, 2010, pp. 365-373: Association for Computational Linguistics.
- [9] H. T. Jakub Piskorski, Martin Atkinson, and Erik Van Der Goot. Clustercentric, "Clustercentric Approach to News Event Extraction," in *New Trends in Multimedia and Network Information Systems*, Amsterdam, Teh Nertherlands, 2008, pp. 276-290.
- [10] "Recognition of Named-Event Passages in News Articles," Mumbai,India, 2012, pp. 329-336: ACL.
- [11] "Keyphrase Cloud Generation of Broadcast News," in *12th Annual Conference of the International Speech Communicatioon Association*, 2011.
- [12] T. H. Byron Y-L Kuo, Benjamin M . Good, and Mark D. Wilkinson, "Tag Clouds for

- Summarizing Web Search Results," in *The 16th international conference on World Wide Web, WWW '07*, New York, 2007, pp. 1203-1204: ACM Press.
- [13] P. Lal, *Text Summarization*. 2002.
- [14] K. H. a. R. J. G. S. Azzam, "Using Coreference Chains in Text SUMmarization," in *In Proceedings of the Workshop on Coreference and it's Application*, 1999.
- [15] L. C. d. S. Marujo, *Event-based Multi-document Summarization*. Lisbon Portugal: Tecnico Lisboa, 2015.
- [16] H. J. a. K. R. McKeown, *Cut and Paste Based Text Summarization*. New York, NY: Columbia University.
- [17] P. Turney, *Learning to Extract Keyphrases from Text*. National Research Council of Canada, 1999.
- [18] A. Inc, "Summly Technology," ed: Apple Inc, 2012.
- [19] M. S. A. B. C. C. Mitray, "Automatic text summarization by paragraph extraction," vol. 22215, 1997.
- [20] B. Krulwich, and Burkey, C, *Learning User Information Interests through the Extraction*. Carlifornia: AAAI Press, 1996.