Software Requirement Specification

Sri Lanka Institute of Information Technology.

Comprehensive Design and Analysis (CDAP)

15th May 2018

Project Id 18_007

Regular Intake

BSc (Hons) in Information Technology Specialized in Software Engineering

Gunarathna T.M.T.A

IT14145476

Automated Processing for Data in a Mass Emergency.

Sri Lanka Institute of Information Technology.

Comprehensive Design and Analysis (CDAP)

15th May 2018

Project Id 18_007

Regular Intake

BSc (Hons) in Information Technology Specialized in Software Engineering

Author

Gunarathna T.M.T.A

IT14145476

Supervisor

Mr. Nuwan Kuruwitaarachchi

External Supervisor

Dr Raj Prasanna

Declaration

I declare that this is my own work and this SRS does not incorporate without acknowledgement any material previously submitted for a degree or diploma in any other university or Institute of higher learning and to the best of our knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text.

Gunarathna T..M.T.A

Table to contents

List of tables

List of figures

1 Introduction

1.1 Purpose

The purpose of this Software Requirement Specification Document (SRS) is to provide a detailed description of the functionalities of automatic text summarization module of automated processing for social media data in a mass emergency project. This document will primarily cover the system's purpose, intended features while mainly focusing on how the automatic text summarization will be done. It will also explain the constraints, interfaces, hardware, software and other dependencies of the system. The document is primarily intended for the supervisor and subject coordinator to serve as a reference document while developing this system, but also be of interest to any parties invested in developing automatic text summarization. The document is written in a form that any person can read and understand the content while it will also be useful for former researchers who are interested in not only implementing similar projects but also interested in the study of automatic text summarization.

1.2 Scope

Main goal of this research work is to develop an open source application programming interface (API) for processing social media textual data at presence of a natural disaster to support individuals of natural disaster supporting teams. The end product would be composed with four modules or components which will be focussing on major or foremost and priorities aspects which can be expected from automated processing of social media data in a mass emergency. The four principal components are,

1. Developing an automatic text summarization component for processing social media posts in an emergency and generating related summaries.
2. Categorizing the information identified and prioritizing the information of social media posts in order to obtain filtered information.
3. Semantic analysis of information to measure how critical, the corresponding situation could be.

4. Validating the accuracy of each social media post by analyzing the follow up comments.

This document will be mainly focusing on developing an automatic text summarization component for processing social media posts which are emerging dramatically by the time and going parallel to an event of mass emergency where the current or ongoing status and critical informations are hidden inside of the respective event occured and generating respective summaries by identifying the core meaning of a particular post and extracting the summarized content over the bulk of social media posts while preserving or maintaining the core meaning of a particular post without damaging the actual core meaning of it.

1.3 Definitions, Acronyms, and Abbreviations

SRS	Software Requirement Specification Document
NLP	Natural Language Processing
NLG	Natural Language Generation
AWS	Amazon Web Services
API	Application Programming Interface

Table 1.0 Definitions, Acronyms, and Abbreviations

1.4 Overview

SRS document is intended to cover all functional and non-functional requirements of component for automatic text summarization of social media posts of the end product.In an emergency, the social media which are dedicated for posting the current or on-going status of natural disasters publish emerging number of huge datasets which are incapable to process them manually by individuals of natural disaster supporting teams. Because of that the research aspect of summarizing social media posts will be focused on identifying the core meaning of a particular post and extracting the candidate summarizing content over the bulk of social media posts. It will

be crucial to be responsible to maintain the core meaning of a particular post without damaging the actual meaning of it.

This document will have been divided mainly into three phases.The first phase of document will be explaining about the purpose of preparing this document. The scope describes clearly what the project team will do and not do. It describes the benefits, objectives or goals of the particular software. In overview it explains how the SRS is organized and describes what the rest of this document contains in a brief manner. The second phase of document describes the overall description in non-technical way which is understandable by the user. It includes product perspective, product functions, user characteristics, constraints, assumptions and dependencies and apportioning of requirements. Main target of product, perspective is to find whether the existing system is available in regard of developing application. Product functions are also described as a summary of all major functions of the application. In user characteristics, it describes the kind of people the typical user characteristics. In constraints sub section describes all conditions that may limit developers' options. Assumptions and dependencies sub section describe that any assumptions being made when developing the application. Under the third phase of the document, it describes developer point of view of the end product. External interface requirements, performance requirements, design constraints, application attributes and other requirements are also explained in advance.

2 Overall Descriptions

In an emergency, the social media which are dedicated for posting the current or on-going status of natural disasters publish emerging number of huge datasets which are incapable to process them manually. Because of that the research aspect of summarizing social media posts will be focused on identifying the core meaning of a particular post and extracting the summarized content over the bulk of social media posts. It will be crucial to be responsible to maintain the core meaning of a particular post without damaging the actual meaning of it. This volume of text is an invaluable source of information and knowledge which needs to be effectively summarized

to be useful [8] for the natural disaster supporting teams to take actions timely, effectively and efficiently. Summarization helps to gain required information in less time.

Automatic text summarization is very challenging, because when we as humans summarize a piece of text, we usually read it entirely to develop our understanding, and then write a summary highlighting its main points. Since computers lack human knowledge and language capability, it makes automatic text summarization a very difficult and non-trivial task[8].According to Radef et al. [6] a summary is defined as "a text that is produced from one or more texts, that conveys important information in the original text(s), and that is no longer than half of the original text(s) and usually, significantly less than that". Automatic text summarization is the task of producing a concise and fluent summary while preserving key information content and overall meaning. [8]

Text summarization approaches can be broadly divided into two groups: extractive summarization and abstractive summarization. Extractive summarizations extract important sentences or phrases from the original documents and group them to produce a summary without changing the original text. Abstractive summarization consists of understanding the source text by using linguistic method to interpret and examine the text. Abstractive methods need a deeper analysis of the text. These methods have the ability to generate new sentences, which improves the focus of a summary, reduce its redundancy and keeps a good compression rate.

Summaries produced by extractive summarization techniques are constructed by choosing a subset of sentences in the original text which is being the input for the text summarizer. The chosen sentences are supposed to be most important sentences of the input text corpus. According to the context of the research, the input text could be a social medial post with follow up comments. Extractive methods tend to be verbose and this is especially problematic as produced summaries should not be lengthy and be readable for natural disaster supporting teams. Thus, an informative and concise abstractive summary would be a better solution.

Existing work in abstractive summarization has been quite limited and can be categorized into two categories: (1) approaches using prior knowledge and (2) approaches using Natural

Language Generation (NLG) systems. The first category of work requires considerable amount of manual effort to define schema such as frames and templates that can be filled with the use of information extraction techniques. These systems were mainly used to summarize news articles. The second category of work uses deeper NLP analysis with special techniques for text regeneration. Both approaches either heavily rely on manual effort or are domain dependent. [7]

Because of the latter mention failures of using extractive and abstractive summarization for automatic text summarization of social media posts, a novel flexible summarization framework, Opinosis, can be proposed. That uses graphs to produce abstractive summaries of highly redundant opinions.

2.1 Product perspective

Tough social media is practically and widely used in financial business-oriented scenarios applications for other purposes are scarce increasing widespread use, popularity and large user base of social media had lead the way for researchers to identify various other uses of social media platforms. In fact, there is a lot of work to be done for the context of social media usage in an emergency.

Some organizations and government agencies have identified the use of social media as an important role in emergency response. For example, American Red Cross has deployed so called Digital Response Center in order to provide situational awareness information and help who are in need. Due to the lack of manpower, lack of funds to conduct proper research and criticality of a situation stakeholders believe that it is resource wasting unachievable task

The task of processing social media entries requires new means of information filtering, classifying and summarization. The lacking feature of most current systems available is the accuracy and the dependability of a given entry. Hybrid systems highly depend on crowdsourcing which requires volunteers so called digital volunteers. This affects the latency of the process. Existing systems are highly dependent on the Twitter. Extracting data from numerous sources other than Twitter streaming API is a challenging task to be completed. The unstructured data needs to be cleaned in order to be used in other stages. Finding appropriate

optimal number of categories to match the requirements of different parties (Organization, Government agencies etc.), identifying ways of calculating accuracy levels for entries, defining thresholds and finding the criticality of situation are major research areas which would be covered throughout the research project. Here is a comparison of existing systems which can be included under the domain of proposed system.

Table 1.3 Features of existing systems and proposed system comparison.

2.1.1 System interfaces

- Twitter streaming API for extracting social media posts of mass emergency for automatic text summarization.

2.1.2 User interfaces

Figure 1 Dashboard user interface.

2.1.3 Hardware interfaces

- Compatible smartphone / tablet pc.
- Personal computer / laptop with required specifications.

2.1.4 Software interfaces

- Virtual servers (Amazon EC2) in the cloud by Amazon Web Services (AWS)

- Docker
- Neo4J graph platforms
- MongoDB NoSQL database
- Firebase realtime database

2.1.5 Communication interfaces

- Modem (Built in GPRS/EDGE/3G/4G)
- Wi-Fi Router

2.1.6 Memory constraints

The system will be developed as microservices for each module or component, deployed on Amazon Web Services (AWS). Memory and storage usages will be allocated according to the consumption of resources of proposed system. Virtual server instances for microservices will be initiated with 8GB of RAM and 1TB of storage.

2.1.7 Operations

- Upload text corpus to be processed through automatic text summarization.
- Generate textual graph by uploaded text corpus.
- Generate word cloud from a bulk of social media posts which are related to a particular context.
- Compare and contrast generated summarized text output over input text corpus.
- View generated summarized text output by means of graphs in a diagramatic way.

2.1.8 Site adaptation requirements

The user interfaces of proposed system are expected to be designed only in English language. Multilingual support will be available in future releases depending on end user requirements.

2.2 Product functions

2.2.1 Use Case Diagram

Use case diagram illustrates the requirements of the system from user's perspective. This will describe the behavior of the system from a user's standpoint, which provides functional description of the system and its major processes. This shows the graphical description of the users of the system and what kind of interactions to expect within the system and displays the details of the processes that occur within the application area. The following use case diagram describes the functionalities that are related to automatic text summarization component of proposed system.

Figure 2 Use case diagram

2.2.2 Use Case Scenarios

Use case name	Summarize text corpus automatically.
Goal	Generate summarized text from input text corpus.
Pre-condition	Text corpus to be summarized has been uploaded.
Actor	Social media post analyser.
Main success scenario	1. Use case starts when the end user click on upload button. 2. The end user uploads the text corpus to be summarized. 3. System will automatically generate the summarized text output.
Extension	

Table 2.0 Use case scenario 1

Use case name	Compare text input and output.
Goal	Comparison between the generated text output (summarized) and input text.
Pre-condition	Summarized text has been generated.
Actor	Social media post analyser.
Main success scenario	1. Use case starts when the end user click on compare button. 2. System will generate an output showing the generated summarized text output and uploaded text corpus.
Extension	

Table 2.1 Use case scenario 2

Use case name	Generate word cloud.
Goal	The word cloud that will show the principal words according the

	context will be generated.
Pre-condition	Bulk of social media posts have been uploaded.
Actor	Social media post analyser.
Main success scenario	1. Use case starts when the end user click on generate word cloud button. 2. System will analyze the uploaded bulk of social media posts and returns the word cloud according to the context.
Extension	

Table 2.2 Use case scenario 3

2.3 User characteristics

The user is likely to be an individual who is working for a natural disaster supporting team and expecting to generate automatic text summarization in order to perform their intended tasks in an effective and efficient manner and a service provider who is intended to process social media posts from a particular source of information will be able to make the use of proposed application programming interfaces (API) to get processed and produced required output upon the context of requirement.

2.4 Constraints

Since the application will be using modern and state of the art technologies and techniques, there has to be some constraints in order to provide end users with a satisfactory experience. High speed internet connectivity with a required bandwidth would be a must to have for the end users to use and operate the proposed application programming interface (API).

2.5 Assumptions and dependencies

- The end user must have an internet connectivity with a considerable higher speed.
- The users should have the basic knowledge using a computer with internet connectivity.

- Virtual server instances on AWS cloud should be up and running 24x7 without downtime.
- There should be sufficient memory and processing power to run the system.
- The end user should have a perfect knowledge at the context of the system is built on.

2.6 Apportioning of requirements

The primary requirements of proposed solution are described in sections 1 and 2 of this document. The section 1.4 of this document gives an overview of the system and it is further discussed under chapter 2. The way in which the system is implemented can be changed from the content mentioned in this document during the system design phase. However, the functional and nonfunctional requirements specified in this document will not vary at any point in time. The final implemented system will adhere to the specified requirements meanwhile achieving the objectives of the project.

3 Specific requirements

3.1 External interface requirements

3.1.1 User interfaces

Figure 1- Dashboard user interface.

This UI shows the state of the application of dashboard that the end user interacts with. The end user will be able to upload an input text corpus and view the generated summarized text output side by side and also this will give opportunity to the end user to access other user interfaces as well.

Note : As the final output of proposed system is an application programming interface (API), the user interfaces will depend on the application programming interface (API) user and on his or her preferences.

3.1.2 Hardware interfaces

Compatible smartphone / tablet pc and personal computer / laptop with required specifications - function as means of displaying frontend which is more specific to the service providers who are intended to consume the use of proposed application programming interface (API). The latter mentioned hardware interfaces should have the capability of connecting to the internet which is supposed to be high speed and bandwidth while required web browsers are installed.

3.1.3 Software interfaces

Virtual servers (Amazon EC2) in the cloud by Amazon Web Services (AWS) - Amazon Elastic Compute Cloud (Amazon EC2) provides scalable computing capacity in the Amazon Web Services (AWS) cloud. Using Amazon EC2 eliminates the user need to invest in hardware upfront, so the user can develop and deploy applications faster. The user can use Amazon EC2 to launch as many or as few virtual servers as the user need, configure security and networking, and manage storage. Amazon EC2 enables the user to scale up or down to handle changes in requirements or spikes in popularity, reducing the user need to forecast traffic.

Docker - Docker containers will be used for implementing microservices architecture

Neo4J graph platforms - Neo4J graph platforms will be used for automatic summarization module.

MongoDB NoSQL and Firebase realtime databases - NoSQL databases will be used for store training dataset of social media posts.

3.1.4 Communication interfaces

Modem (Built in GPRS/EDGE/3G/4G) and Wi-Fi Router - Used to connect to the internet to access web services.

3.2 Performance requirements

The system will be developed as microservices for each module or component, deployed on Amazon Web Services (AWS). Memory and storage usages will be allocated according to the consumption of resources of proposed system. Virtual server instances for microservices will be initiated with 8GB of RAM and 1TB of storage. API calls should not take more than 5 seconds as immediate responses of proposed system will be important to perform in a mass emergency. Longer response times will make bad impressions about the application on the end user's minds. Moreover, the proposed system is supposed to save time of the end user.

3.3 Design constraints

During the design stage the major constraint will be facing is the limitation of available time. System design will be prepared and it helps to specify the requirements and helps to define the overall system architecture. The major obstacle that we have to meet as a development group was total completion of the project according to the given schedule and defined deadlines of different milestones. Almost all the major features in the proposed system being totally or partially separate research areas pushed these limitations further.

3.4.1 Reliability

System should be reliable for information it provides to each user and also should be consistent when changes occur. The system must be capable of recovering from any crashes without making any harm to the system or data losses. Proper backup systems must be implemented to ensure the data is safe and it can be used in case of any failure of the system.

3.4.2 Availability

System is designed with high availability architecture therefore system may not have a downtime. To get high availability, system will maintain a backup. The system must be up and running for 24x7. The users will have the access to data of the system and design using the objects and tools available. The data available to the users must be up-to- date.

3.4.3 Security

There will be authentication and authorization techniques used by the system to prevent unauthorized access. Authentication will be done using Login Password and there will be access privilege system for the authenticated users. There will be strict security mechanisms used by this system to prevent unauthorized access.

3.4.4 Maintainability

Code will be kept with proper standards to increase the maintainability and everything will be documented and monitored. There will be some features given to the system admins to make the system more maintainable like adding users and system parameters. Proper programming practices and coding standards are followed to make the application maintainable. So the application welcomes any modification after deployment with ease.

4 Supporting information

4.1 Appendices

REFERENCES

[1] A.T.M Shahjahan and Kutub Uddin Chisty "Social Media Research and Its Effect on Our Society " World Academy of Science, Engineering and Technology International Journal of Information and Communication Engineering Vol:8, No:6, 2014
[2] Muhammad Imran, Carlos Castillo, Fernando Diaz, and Sarah Vieweg. 2015. Processing social media messages in mass emergency: A survey. ACM Comput. Surv. 47, 4, Article 67 (June 2015), 38 pages.DOI: http://dx.doi.org/10.1145/2771588

[3] Bing Liu. Sentiment Analysis and Opinion Mining. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, 2012.
[4] Yelena Mejova, Ingmar Weber, and Michael W Macy. Twitter: A Digital Socioscope. Cambridge University Press, 2015.
[5] Ahmed Nagy and Jeannie Stamberger. Crowd sentiment detection during disasters and crises. In Proceedings of the 9th International ISCRAM Conference, pages 1–9, 2012
[6] Dragomir R Radev, Eduard Hovy, and Kathleen McKeown. 2002. Introduction to the special issue on summarization. Computational linguistics 28, 4 (2002), 399–408

[7] Ganesan, K., C. Zhai and J. Han, 2010. Opinosis: A graph-based approach to abstractive summarization of highly redundant opinions. Proceedings of the 23rd International Conference on Computational Linguistics

[8] M. Allahyari, S. Pouriyeh, M. Assefi, S. Safaei, E. D., J. B., and K. Kochut, "Text Summarization Techniques: A Brief Survey," International Journal of Advanced Computer Science and Applications, vol. 8, no. 10, 2017.

[9] Y. J. Kumar, O. S. Goh, H. Basiron, N. H. Choon, and P. C. Suppiah, "A Review on Automatic Text Summarization Approaches," Journal of Computer Science, vol. 12, no. 4, pp. 178–190, Jan. 2016.
[10]https://www.google.lk/search?q=semantic+analysis+social+media&rlz=1C1CHBD_enLK771LK771&source=lnms&tbm=isch&sa=X&ved=0ahUKEwiU98SQ4JPaAhXFtI8KHUUoDD8Q_AUICigB&biw=1366&bih=613#imgrc=sT1wUMlUtiK_GM: