Topic-based Story Archival and Search

ADNAN JAMEEL

01-134132-010

M. SAMEER AHMED

01-134132-208

Bachelor of Science in Computer Science

Supervisor: Dr. Arif Ur Rahman

Department of Computer Science
Bahria University, Islamabad

May 2017

Certificate

We accept the work contained in the report titled “Topic-based Story Archival and Search”, written by Mr. Adnan Jameel and Mr. M. Sameer Ahmed as a confirmation to the required standard for the partial fulfillment of the degree of Bachelor of Science in Computer Science.

Approved by . . . :

Supervisor: Dr. Arif Ur Rahman (Assistant Professor)

____

Internal Examiner: Dr. Asfand-e-yar (Assistant Professor)

____

External Examiner: Dr. Rashid Ahmad (Assistant Professor)

____

Project Coordinator: Dr. Arif Ur Rahman (Assistant Professor)

____

Head of the Department: Dr. Faisal Bashir (Associate Professor)

____

May 18^{th}, 2017

Abstract

Preserving News Stories from Web may helpful and important because of various reasons. They provide information about any event and incident which happened in the past. Keeping the stories may be important because of various reasons including cultural heritage, evidence of activities and scientific and historical reasons. However, the stories may be lost because of the constant change of the environment and the technologies used to present and publish. Certain individuals or institutes may be interested to collect information related to specific topic or event. The idea came from various research papers to overcome the need by capturing the stories which are relevant to the topic.

Information is provided to the system through a list of keywords. The tool automatically collects all the stories from the provided sources by calculating a score of similarity between the story and keywords.

The tool has the functionality of searching the already downloaded stories. Users may search stories by entering keywords which are automatically processed and all the relevant stories are retrieved from the archive. Users may then choose to open and read details about a story of their choice.

Acknowledgments

First of all, We are grateful to Almighty Allah for his blessings upon us. Who gave us the ability and courage to complete our project on time. He is the only one who we always looked at in the event of happiness and trouble and He always helped us in the time of need. With his blessing upon us we have completed our Work.

We would specially thank to our supervisor Dr. Arif Ur Rahman. Who remained the source of guidance till the end of this project. He gave us a continuous advice on the content of the report and project. He gave us too much time to guide each and every step of this project. We could not achieve the desired results without the guidance of our supervisor.

We would like to thank our parents who supported us in all our endeavors and saw successful persons in us.

Last but not the least we would like to thank all our friends whose silent support led us to complete this task and enjoy the four years stay at the university.

Adnan Jameel

Muhammad Sameer Ahmed
Islamabad, Pakistan

May 2017

"Coming together is a beginning, Keeping together is a progress,
Working together is a Success."

Henry Ford

Contents

List of Figures

List of Tables

Acronyms and Abbreviations

API	Application Programming Interface
CSV	Comma separated value
GUI	Graphical User Interface
JSON	JavaScript Object Notation
NLP	Natural Language Processing
POI	Poor Obfuscation Implementation
RAKE	Rapid Automatic Keyword Extraction
XML	Extensible Markup Language

Chapter 1

Introduction

Newspaper covers several types of stories and events like important cases in courts, political events, murder stories, sports and health related stories. In the past few years the trend of publishing newspapers on paper shifted to publishing online version. Technology change the version of Newspapers. The news generation in the digital organizations have the linear process with a fixed single output like printed newspaper. The News are rapidly generated and modified in continuous trend. So, because of frequently updating news and speed of generation of information, it has become vital need to preserve this digital news for the later use.

The lifespan of news stories published on different News website differ from one newspaper to another i.e. from one day to one week. Though there is backup of news stories and have the archived by the news publisher. In the future, it will be difficult for us to access those stories published in various papers about the same story. The issues are complicated if a story is navigating through the archive of many newspaper which requires different accessibility technologies. It may even more difficult to extract the meta data manually i.e. author name, publication date, location etc. which is not mention explicitly and the list of names used in an article. The focus of this report is to extract and normalize those information from different website which are publish online. Web resources can be use by using the feature like browsing, Web Crawling and by developing a tool. The technique includes the requirements what the need, what type of resources to be captured and extraction frequency.

So, the need of this type of software to extract the news story from various websites and analyze those stories. However, there is some issues may be face like every site have their own data structure in which they are define so our software will have the proposed solution to extract news stories from different news site.

In this report, we proposed a solution to this problem by introducing the application related to it. In Which it identifies the document from the News website. This application consists of many steps and procedure to accomplish that task. In which we distribute sentences into a chunk of words and compare them to determine the topic and check it is related to the topic or not. By applying score to it higher the score higher will be the possibility of the topic close to need otherwise ignore it. This application is generic can be used for any type of topic.

User can also search the topic through search engine which explicit the names of person, location and publisher name etc. and different feature will be available to it.

Chapter 2

Literature Review

2.1 Natural Language Processing

Natural language processing is a field of computer science used in Artificial Intelligence concerned with an interaction among human (natural language) and computers. for example, NLP is to the area of human computer interaction. Many challenges involve in this some of them involve like: natural language understanding, enabling computer to derive meaning from human or natural language input and other involve natural language generation.

This is important because it handle the task which have direct real-world application, while more commonly serve as sub-tasks that are used to aid in working and solving big problems.

Some of the popular uses of natural language processing like:

- **Machine Translation**
- **Search Engines**
- **Speech Recognition**
- **Stemming**
- **Text Simplification**
- **Text to Speech**
- **Query Expansion**
- **Natural Language Search**
- **Automated Essay Scoring**

2.1.1 Tokenization

In tokenization in which we give a character sequence and a defined document unit, tokenization is the task of chopping it up into pieces called tokens, perhaps throwing away certain characters, such as punctuation. A token is a sequence of characters in some particular document are grouped together as a useful semantic unit for processing.

2.1.2 Stemming

The process of removing prefixes and suffixes form words to reduce them to stems thus eliminating tag-of-speech and other verbal or plural inflection.

Stemming refers to mapping to words forms of stems or basic word forms. Word form may differ from stems due to morphological changes necessary for grammatical reasons. It is usually sufficient that related words map to the same stem, even if this stem is not itself valid root.

2.1.3 Stop Words

Table tab:stopwords (in Appendix A) presents the List of Stop words. Stop words are commonly use words in the document which do not have any importance and any worth to be used in search Problems. These types of words are filtered out because they do not give any importance to the content of the document. These words provide most of the time give us the unnecessary information. These words may be a (articles, prepositions) or these kinds of words.

2.2 Topic Identification

Topic identification is the method to identify or discover the topic require by the user for example if In this we use to analyse the topic and knowing that this topic is fulfil the user needs or it is equal to the topic the user want.

2.2.1 Single Document

Keywords are the special kind of words which have some importance in the document content, in which sequence of one or more characters are define in it, provide a compressed illustration of a document content. Basically, Keywords are used to summarize the content and to fetch the important words from the document. They are most commonly used to determine the queries with in the information retrieval (IR) system because they are easy to remember, learn, share and define them easily. In comparison to mathematical signatures, keywords are independent of any amount and can be useful across multiple quantities and IR systems [1].

2.2.2 Multiple Documents

Automatic keyword extraction is widely used in many applications, such as the selection of relational databases, the clustering of websites, the characterizing of news that uses several linguistic technologies [2]. We will give the measure of weight of a word in documents. Our discussion is made upon the assumption that all stop words have been removed from documents and word stemming has been done [2].

2.3 Keyword Extraction

The focus of the topic to extract the keyword from the document and that extracted data operate on a single document. Such document extract the same information with the help of comparison with keywords. This document-oriented method provides the enabling additional analytical methods, document feature, Context independent changes with a period. The methods of document-oriented suited to quantities that change, such as the publication of technical summaries collection that raise overtime or in the form of news articles. Furthermore, by applying on a single document, these techniques of massive collection and can be apply in many context to develop IR systems and analysis tools [1].

2.4 NLP Tools

The languages supported by the tools:
English, Chinese, Arabic, Swedish, German etc.

- **StanfordCore NLP:** Stanford CoreNLP provides a set of natural language analysis tools. It can give the base forms of words, their parts of speech, whether they are names of companies, people, etc., normalize dates, times, and numeric quantities, mark up the structure of sentences in terms of phrases and word dependencies, indicate which noun phrases refer to the same entities, indicate sentiment, extract or open-class relations between entity mentions, get quotes people said, etc.

- **Lucene:** Lucene is a java base library use to make it easy to add search functionality to an application or websites. It is a full-text search which is used to add content to a full-text content. Apply queries to the index, returning results ranking by either the relevance to the query or sorted by an arbitrary field such as last modified document date.

- **OpenNLP:** It is the basic NLP machine learning based toolkit for the processing of natural language. It supports most commonly use NLP task which is mostly use in our programs, such as tokenization, part-of-speech tagging, named entity, parsing and co reference resolution. These are usually use to build text processing services.

- **Maui:** Maui is used to automate your keyword extraction and subject indexing. It is an open source tool helps to analyse, categorize and access any kind of document, documentations, articles etc. It uses the combination of natural language processing(NLP) and machine learning to quickly and accurately assign meta data and extract the main topics of document.
- **RAKE:** RAKE stands for Rapid Automatic Keyword Extraction. It is an algorithm to automatically extract word/keywords from different documents. Keywords are sequences of one or more words that together, provide a compact representation of content. Rake is a widely use NLP mostly depends on factors like language in which the content is written, the domain but mostly aimed to be implemented in English.

2.5 Recall and Precision

Recall and precision are the machine learning methods are used mostly to measure the performance of the algorithm. Recall is the ratio of several events can correctly recall to several all correct events. Precision is the ratio of several events you can correctly recall to several all events you recall mixture of both correct and wrong recalls.

$$Recall = \frac{TruePositive}{(TruePositive + FalseNegative)}$$

$$Precision = \frac{TruePositive}{(TruePositive + FalsePositive)}$$

Chapter 3

Requirement Specifications

This chapter defines what functionality is to be perform by our system and what is to be develop and also tell the functional and non functional requirement of our project and also define who are the stakeholder what activity they have to perform.

3.1 Functional Requirements

- The application will provide links related to different News-stories websites.
- This tool will Compare a story or news with a Keywords which we extract manually from the main newspaper sites then filter those stories.
- Score is basically a comparison result how story words are close to our topic.
- Check if the score is greater than the limit we provided than story will be downloaded.
- It will generate an XML/text file which stores the content of that stories.
- User can also Navigate the News through search engine given in GUI.
- User can also search the multiple words at the same time.
- In the end the summarize version of a story will be displayed to the user which display the location, person name, nouns, publisher name and published date is mention in the summarized area.

3.2 Non-Functional Requirements

- Performance: In this it provides a user-friendly environment user can easily find and extract the News-stories related to sexual abuse provided on different websites. Web crawling method is use which can automatically provide the links for the user.

- Availability: In this it concerns with the Websites through which we get the stories is updated or modified with the new News-stories or not. Application perform to get that story and apply keyword extraction on it.

- Data integrity: It uses proper data structure method through which our data is store in a consistence way and the data which will be stored will be saved in sorted order by using its publish date.

- Accuracy: The system should be accurate enough that it could identify the right story.

- Portability: As there is a constraint that the internet facility should be there and this application could only be perform on desktops computers (PC, laptops etc.).

- Maintainability: Maintenance issues can only be handle by the developers.

3.3 Keywords Identification

We read many stories related to our project and those words which are frequently use in different stories we add them to our list of words or list of keywords because these words will help to identify the topic in different news-stories in different websites. Some of the list of words we use to identify the topic as shown in the table 3.1 and 3.2.

Table 3.1: Keywords identify from different stories related to Sexual abuse

abduction	threatened	brutal	child	defense
abuse	tortured	depression	disturbing	emotional
addiction	trapped	feminist	fondled	frightened
adults	urban	gang	gay	group
alone	victim	guilt	HIV	molesting
ashamed	vulnerable	pornography	rampant	rape
refused	rejection	rural	secret	sex
shelter	silent	society	STD?s	strangers
threat				

Table 3.2: Keywords identify from different stories related to health

aid	attack	blood	cancer	cell
cholesterol	clotting	diabetes	diet	disease
expire	fatal	genetic	group	health
heart	inherited	insulin	invoking	medication
medicine	neurone	organ	HIV	risk
surgery	therapy	threshold	exercise	weak

3.4 Match Keywords with Story

The keywords which we obtain from different websites then these keywords will match with the News-stories which we get from different websites. We apply formula and obtain the score. Score is basically the obtaining number which we get after applying the formula. For example, we are using the following formula in our project to match the keyword with story.

$$Score = \frac{MatchingnumberofTokens}{TotalNumberoftokeninthestory}$$

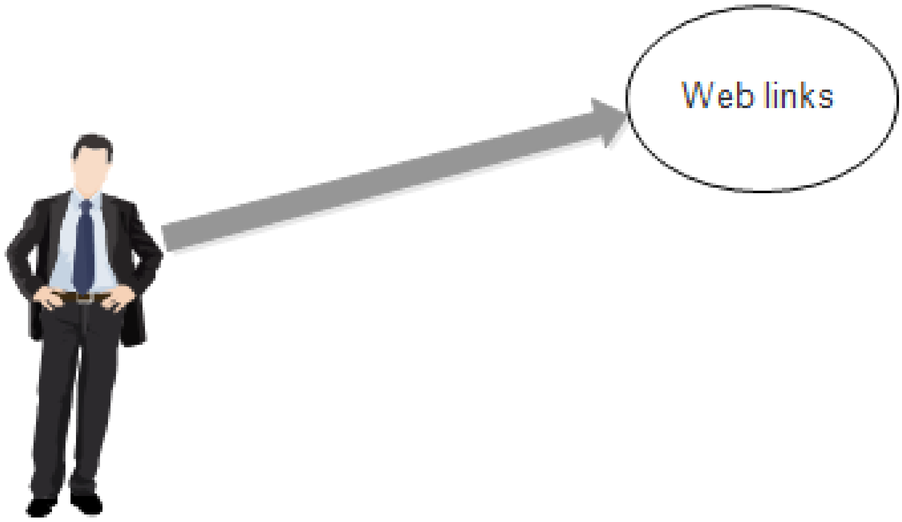
From the above formula, we obtain the score which means how much News-stories are close to our topic. The story which have the higher score will be closer or equal to our project topic and those which show very less comparison result and the comparison result is very low then we assume that this topic is not related to us so we ignore it.

3.5 Use Case

Table 3.3: Usecase Table

Case id	1
Actors	User
Precondition	Program should be running
Assumption	Internet should be available
Post Condition	output should display the comparison score
Exception	If the Internet is not there then application cannot work.

Figure 3.1: Use case Diagram.



Chapter 4

Design

In this chapter we define our system design how our system will look like what activity it would be perform. What are the data, modules, graphical user interface(GUI) and components for a system to meets the user needs. which are define as follow:

4.1 System Architecture

The Architecture of this system is interactive and simple.It provide user friendly interface which contains a different websites links user can select the different links by using web crawling. The high level diagram of this system as shown in the 4.3 and the Flow diagram of the given system as shown in 4.4. In which it display how the system performs its task step by step.

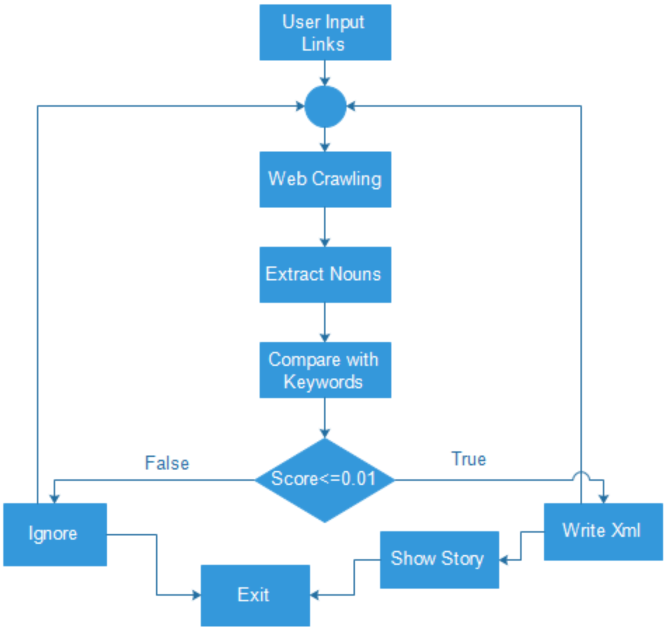
4.2 Design Constraints

1. Internet facility must be available.
2. Targeted users are educated enough to operate a computer and understand the system as well as the application to be perform.

4.3 Design Methodology

Using incremental model, The incremental process model is a process of software development where the product is designed, implemented and tested incrementally. At each increment a new and little more things or feedback is added until the product is to be develop. It includes both maintenance and development. When all the requirement is to be

Figure 4.1: System Flow Diagram



done then product is said to be finished. This model comprises the elements of both the waterfall model and Iterative idea of prototyping.

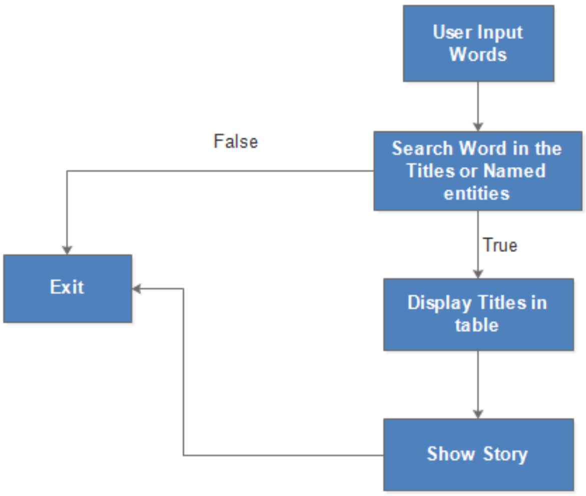
So, by Using Incremental model, we divide our project into four different phases through which project is completed in different increments until the product is finished.

1. Downloading story for analysis.
2. Extract information from the story to verify its relevance to the topic.
3. Uniquely Identify each story as the same story may be reported in multiple newspapers and multiple times on the same website (redundancy will be remove).
4. Extract name, location ,publication date and publisher name information of story will be displayed in the form of summarized version.
5. Search engine is given in which searching the nouns like (name of person, location, publisher and publication date etc) in story by entering in the search engine. The flow diagram of search engine is shown in 4.2.

4.4 High Level Design

High level diagram which shown in 4.3. In which user will input the link where he/she want to extract the story. Then after selection it will move to the processing part where it process the link accordingly. Then after being process it will display the output score result of comparison.

Figure 4.2: System Flow Diagram of searching



4.5 Low Level Design

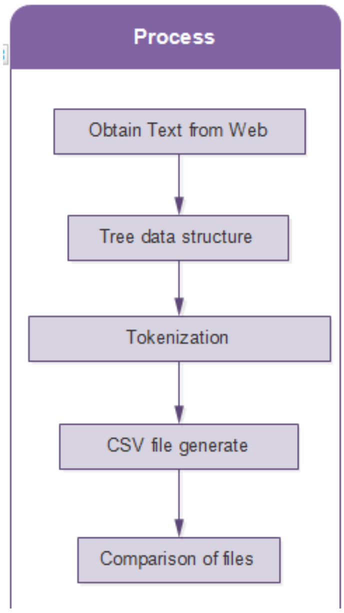
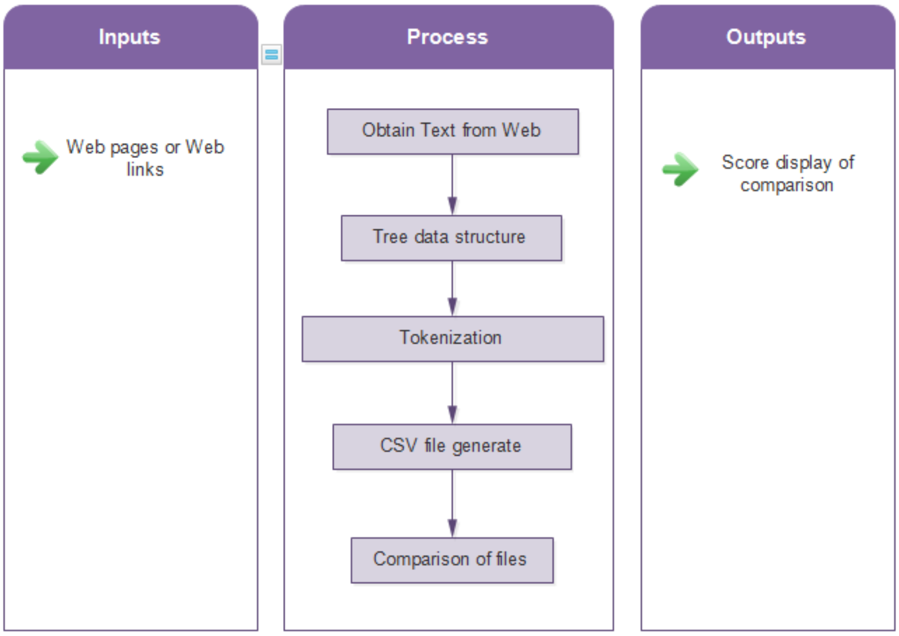


Figure 4.4: Low level design

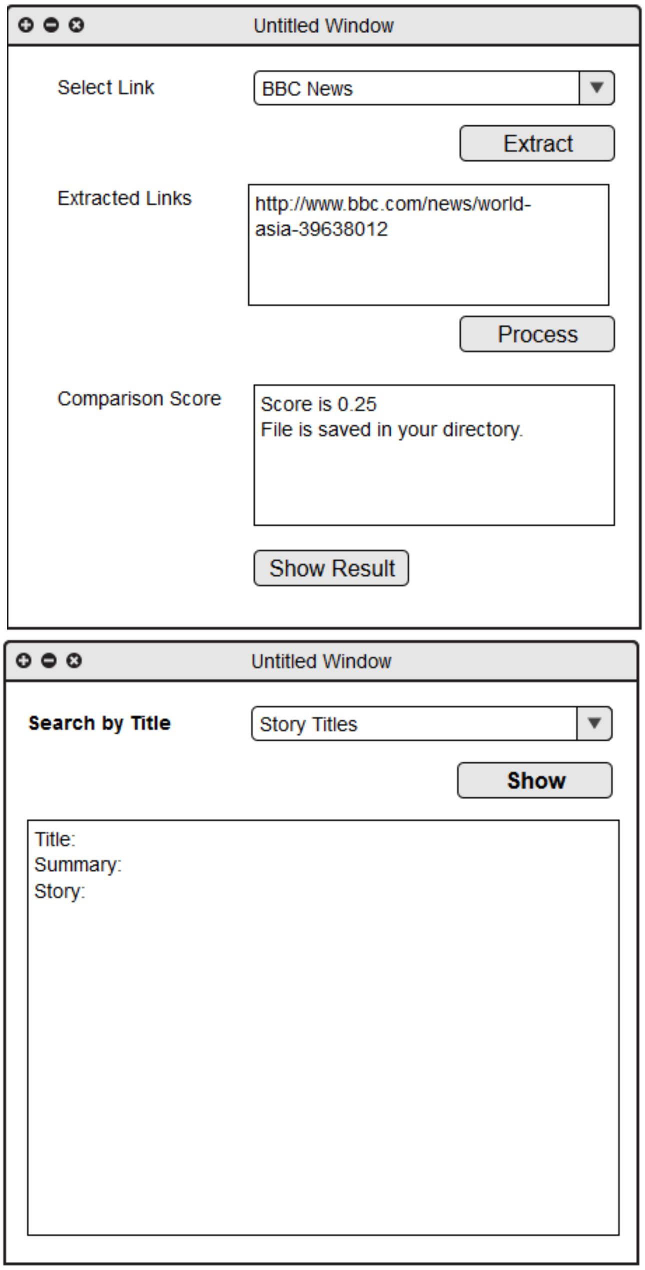
This diagram shows the processing part of the application. From the first part where we obtain the text from the website by using Jsoup API which can copy all the text from the website and paste it into the text file Which shown in the 4.4figure 4.3. Process will make a tree of words by using tokenization of words which was extracted from the site. CSV (comma separated value) file is created for both the words extracted from the web and the words which was manually extracted from the different stories and in the end comparison of both the CSV files and calculate their Score of comparison by using formula.

Figure 4.3: High level design



4.6 GUI Design

Figure 4.5: GUI Design



Select Link: BBC News

Extracted Links: http://www.bbc.com/news/world-asia-39638012

Comparison Score: Score is 0.25
File is saved in your directory.

Search by Title: Story Titles

Title:
Summary:
Story:

The Graphical user interface of this application is simple and consistent because it is used for general user easy to understand and easy to analysis the steps taken by the application. User will easily identify which button will perform which action easily. GUI design which is meant to be look like is shown in 4.5.

- In this we have two windows one is to add news stories and the other is use for the searching purposes. in first window user will select the given link and start extracting the story and then apply process to it then it will compare and fetch the relevant topic after getting it will save the relevant topic in XML format.

- In second window user have the facility that he/she can search the nouns like words in the story like publisher name, person name, location and publish date in the story. Search the story by title then desire result will be display in the text area.

Chapter 5

System Implementation

Implementation is the process in which we putting our decision and our plan into reality, effect or in execution. So in our case we tell you which techniques and software components we perform to develop this project into execution.

5.1 System Architecture

The system Architecture of this application is generic can be applicable to any News Story and can easily download those stories by using this tool. We just need to create a list of keywords which we want to compare with the News story. In this application, we extract News from BBC News website so every website has its own structure may be different to each other so for extracting News from other websites we follow data structure accordingly. Application uses the components, tools and techniques of java language. In which all the built-in Libraries and API's are used. Tools and techniques which are used in this Project for example Tokenization, stop words, Topic Identification, Keywords Extraction, single word or double words comparison file, Stanford Core NLP tools, API (Jsoup,CSV, POI,json).

5.2 Tokenization

We use this technique in which sentence are chopped into chunk of words and Now each word is deal with 1 token and throw away certain characters like punctuation etc. Token is a sequence of characters from a specific document join to make a semantic unit of processing.

5.3 stop Words

Table tab:stopwords (in Appendix A) presents the list of stop words. We eliminate Stop words from our story by using Core NLP tool because these words provide extra information or may be some may be sometime they provide unnecessary information which may cause story out of there boundary so by using the tool NLP we eliminate these types of words and makes our story to the point.

5.4 Stanford CoreNLP

By using this tool which deal with a human natural language use to analysis it. Their parts of speech (Noun, verbs, adjective etc.). By using it we can identify the story contains the company name, person, location etc. how many stop words are using all these things are captured by this NLP tool. We eliminate those words which provide any identity of anything. Program automatically extract the noun phrase and eliminate those words which have like person name, address, location etc. will be eliminated.

5.5 API

The following APIs are used in the project.

5.5.1 Json

Json is a (JavaScript object notation) java API use in our project. This API is an open-standard format use human understandable text use to spread data consisting of values and attributes. This is basically how the program is interacting with the human.

5.5.2 Jsoup

Jsoup is a java library use to work with the HTML. Jsoup is use in our project to extract all the data or story from the News story website and store it in a system or a file system. All the content like heading, hrefs, paragraph will be extracted and create a copy of that story in a system files.

5.5.3 CSV

CSV stands for comma separated value use to read each word as a separate this API is use in our project to read comma separated values which are store in Excel.

5.5.4 POI

We use POI API which is used to read the Microsoft office files. In our case, we save text on Ms Excel CSV API use to separate them and POI API is used to read that Ms Excel text and use to display it and Manipulation on it.

5.6 Keyword Extraction

We read different stories from the website in which same topic have been discussed like sexual abuse. We took all those words which are mostly used and quite similar in different News stories. Storing those words on Excel sheet they are single and double words like (abuse, child abuse).

5.7 Topic Identification

We identify our topic by the keywords we extracted manually and the story we extract from Jsoup library. From this we compare the keywords with a story by using the following Formula.

$$Score = \frac{MatchingnumberofTokens}{TotalNumberoftokeninthestory}$$

this formula calculate the score of comparison which story gave the highest scoring value will be the topic which the user need.

5.8 Search Engine

We have created a search bar in our application which uses the regular expression methodology in which user enter the specific word or it could be enter multiple words at the same time in the search engine. which he/she want to search the word. This entered word is matched in the match function in which it matches with different titles, nouns in the story and ignore all the words which are present before and after that word due to using the regular expression. When it successfully matched either with the title, names or location. Then this will be assigned to the list of string this list of strings will displayed on the table where title of the matched story will be displayed. When you click on a title then related story will be displayed.

5.9 Multi-Threading

We are providing the facility of multi-threading in which user perform multiple task at the same time. By applying multi-threading during the processing is perform to calculate the score user can perform other task like show list of keywords. User can perform multiple task on the same window.

5.10 View Keywords

User can view the list of keywords which are store in the excel sheet on which the comparison is apply and score is generated after comparison. User can view it as well as he/she can edit that words and modified it according to their needs.

Chapter 6

System Testing and Evaluation

In this Chapter various testing techniques are used for evaluation and validation of this application. Testing plays an important roll in the software development process. It helps to validate the system will meet its requirements and the working of the application. Every project have some limitations and these limitations will be explored during the test cases which are discuss in this chapter.

6.1 Usability Testing

Usability testing provides the information about how much time will it require performing a specific task of the system. Usability testing is evaluated by the target audience of the application. In our application, our audience is general Users. This application is use for security surveillance and for formulating the total number of (sexual abuse) cases now a day. Usability testing tells that the system is performing the tasks that it is intended to do or not. The application performs the task the user want to perform from the application or not.

6.1.1 Easy to use

This project is easy to use and project is self-explanatory that the user can easily Interpret what the system is intend to do.

6.1.2 Easy to learn

Our system is very simple and consistent, visible and clear. The system is very easy to learn for the new users.

6.2 Software Performance Testing

Software performance testing use to check how efficiently the system performs the task through this application. This will help us to determine the system capability, reliability and efficiency. Following steps were taken to increase the application performance.

- Comparison between keywords and the web content will be perform on run time if story relevant than download otherwise ignore it this will increase the speed and consume less time.

- If there is no content on the link it simply ignore it and move forward.

- System will not take or waste time to extract or checking the videos and images because they are not relevant for system.

- System will take some time 2 to 3 minutes for processing the web content.

6.3 Compatibility Testing

Compatibility Testing is a type of non-functional testing. Compatibility means on what circumstances the system will perform well without any issues. This testing technique will help us to know the compatibility with which hardware and software resources need to use this application. Following are the compatibility feature need to be have:

- Processor must be fast cause it perform to many processes to be taken so processing must be fast for example core i3 and above and RAM 4gb.

- This application is developed on the Java. So system must have the JDK platforms to run this type of application.

6.4 Exception Handling

In this system there are many exceptions are to be handle.

- HTTP exception display if there is any issue regarding the internet connection exception will be displayed.

- Exception will be displayed if the text file is not there or the CSV format Keywords file is not present it will display the exception.

- Time out Socket exception is there as well.

6.5 Load Testing

Load testing is that to test the system under the abnormal situation applying stress to a software and determine the behaviour of the system under this type of situation. In this system stress may apply when Internet connection is disconnected again and again it will apply too much load on a system. In other case when a News story is too large it will too long to compare the process and calculate the score of comparison. And the last thing is when the processing speed is slow then it applies all the load on the processor it will slow down the speed and consume lots of the time.

6.6 Stress Testing

We apply stress test to our application by applying comparison on about 30,000 XML files. The application successfully perform the comparison on them it took about 14 seconds to completes its comparison on each file.

6.7 Security Testing

As our system Extracted the News from one of the biggest and authorize News website where security is there priority. Our Application not display those types of data which will break the security terms and condition. No one can misuse this type of data.

6.8 Installation Testing

Installation testing refers to the testing of installation of the application. To run the project, there must be the Microsoft operating system installed in the system. This project is developing on Java. so, system must have JDK platform and java platforms. These things must have to run this type of application.

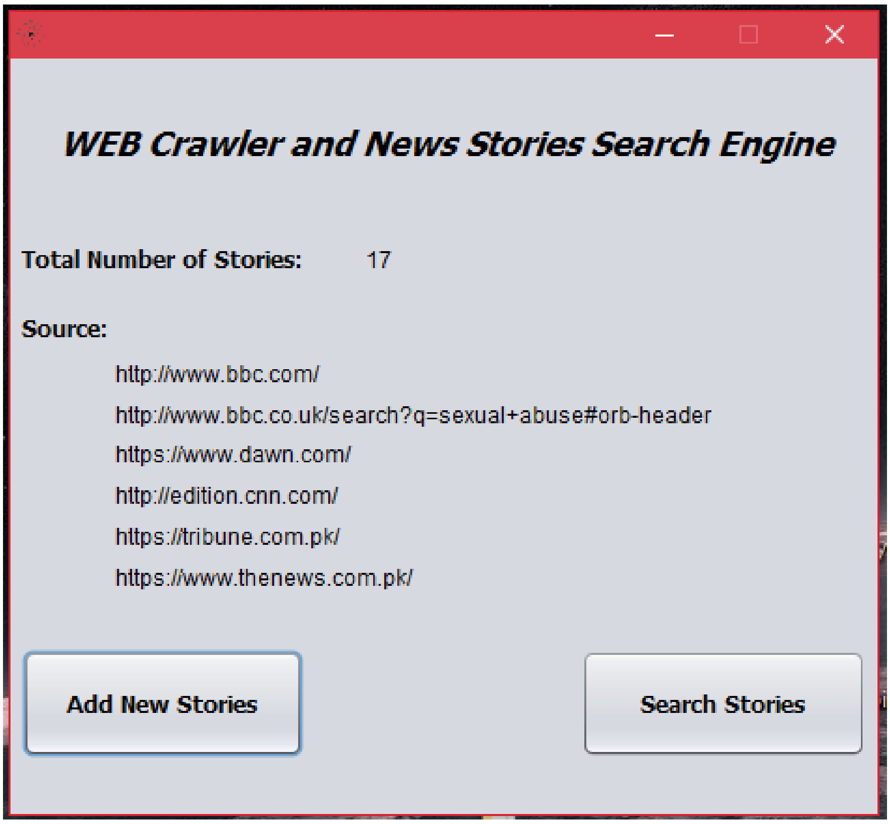
6.9 Formulas

A story consists of tokens. There are tokens which match with the identified vocabulary for Example the Noun phrase and verb phrase words are collectively make a matching number of tokens. This matching number of token will divide with the total number of the tokens which are available on the story. Then after applying calculation we calculate the score. Which tells how close the story is related to topic. Higher the score higher will be the possibility of the related topic. The score calculation done as follow:

$$score = \frac{MatchingNumberofTokens}{TotalNumberofTokensinaStory}$$

6.10 Graphical User Interface Testing

Figure 6.1: GUI Main Window



WEB Crawler and News Stories Search Engine

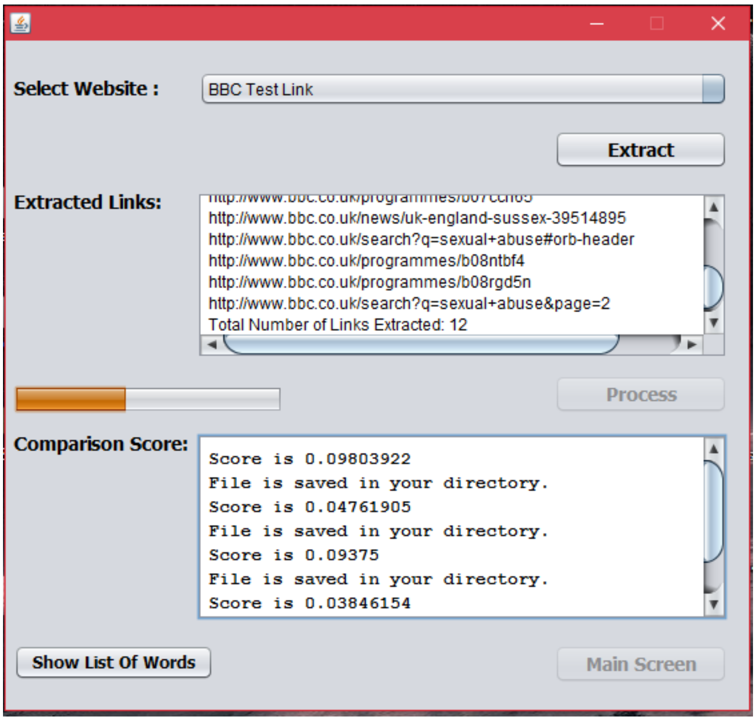
Total Number of Stories: 17

Source:

http://www.bbc.com/
http://www.bbc.co.uk/search?q=sexual+abuse#orb-header
https://www.dawn.com/
http://edition.cnn.com/
https://tribune.com.pk/
https://www.thenews.com.pk/

As shown in the Figure 6.1, 6.2, 6.3 where the Graphical user interface is displayed. GUI is developed in such a way that it is remain consistent in all the processes made by the application. The user interface of our system is self-explanatory. Our system contains tabs, drop-down, buttons, Text area etc. All things perform correctly. One Window will show the story which was extracted by system and second one is used for the searching purpose. The test on Search User interface is successful user can enter multiple words at the same time.

Figure 6.2: GUI Add New Link Window



Select Website : BBC Test Link

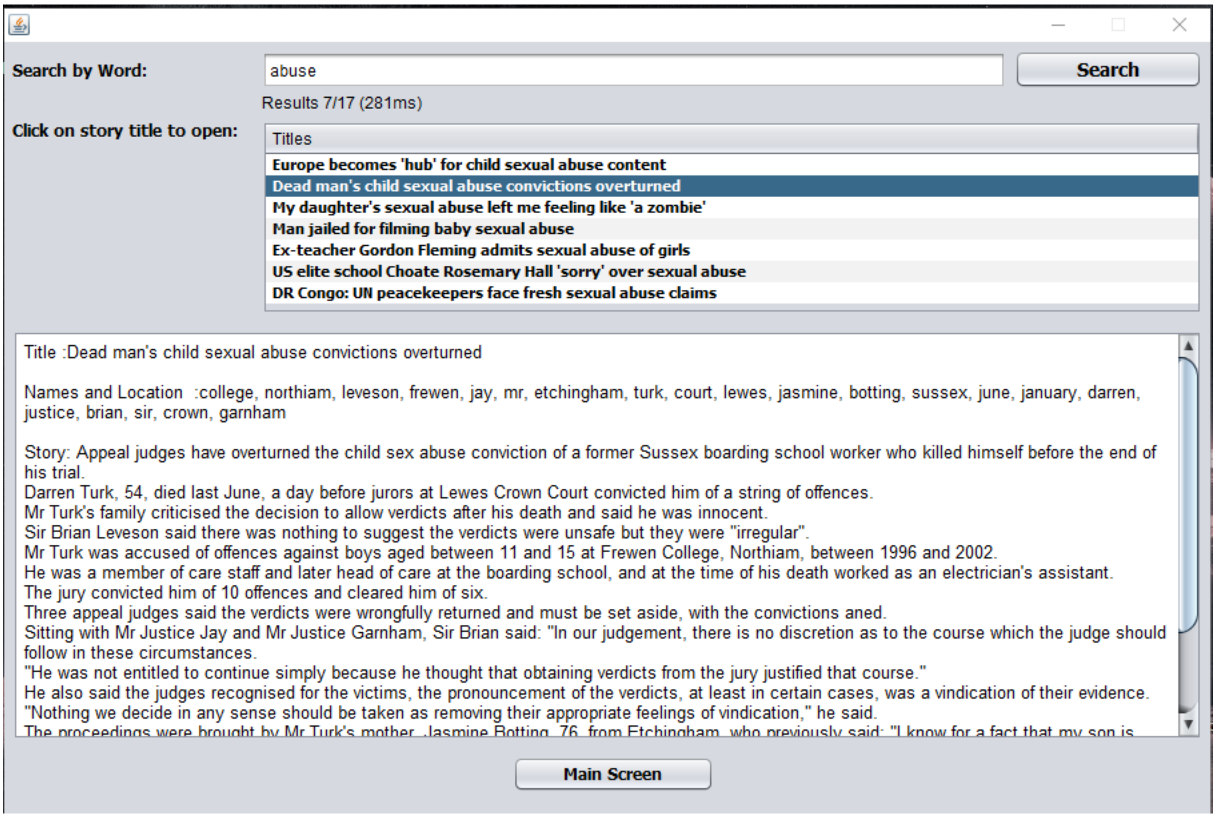
Extracted Links:

```
http://www.bbc.co.uk/programmes/b07ccn65
http://www.bbc.co.uk/news/uk-england-sussex-39514895
http://www.bbc.co.uk/search?q=sexual+abuse#orb-header
http://www.bbc.co.uk/programmes/b08ntbf4
http://www.bbc.co.uk/programmes/b08rgd5n
http://www.bbc.co.uk/search?q=sexual+abuse&page=2
Total Number of Links Extracted: 12
```

Comparison Score:

```
Score is 0.09803922
File is saved in your directory.
Score is 0.04761905
File is saved in your directory.
Score is 0.09375
File is saved in your directory.
Score is 0.03846154
```

Figure 6.3: GUI Searching Window



Search by Word: abuse

Results 7/17 (281ms)

Click on story title to open:

Titles
Europe becomes 'hub' for child sexual abuse content
Dead man's child sexual abuse convictions overturned
My daughter's sexual abuse left me feeling like 'a zombie'
Man jailed for filming baby sexual abuse
Ex-teacher Gordon Fleming admits sexual abuse of girls
US elite school Choate Rosemary Hall 'sorry' over sexual abuse
DR Congo: UN peacekeepers face fresh sexual abuse claims

Title :Dead man's child sexual abuse convictions overturned

Names and Location :college, northiam, leveson, frewen, jay, mr, etchingham, turk, court, lewes, jasmine, botting, sussex, june, january, darren, justice, brian, sir, crown, garnham

Story: Appeal judges have overturned the child sex abuse conviction of a former Sussex boarding school worker who killed himself before the end of his trial.
Darren Turk, 54, died last June, a day before jurors at Lewes Crown Court convicted him of a string of offences.
Mr Turk's family criticised the decision to allow verdicts after his death and said he was innocent.
Sir Brian Leveson said there was nothing to suggest the verdicts were unsafe but they were "irregular".
Mr Turk was accused of offences against boys aged between 11 and 15 at Frewen College, Northiam, between 1996 and 2002.
He was a member of care staff and later head of care at the boarding school, and at the time of his death worked as an electrician's assistant.
The jury convicted him of 10 offences and cleared him of six.
Three appeal judges said the verdicts were wrongfully returned and must be set aside, with the convictions aned.
Sitting with Mr Justice Jay and Mr Justice Garnham, Sir Brian said: "In our judgement, there is no discretion as to the course which the judge should follow in these circumstances.
"He was not entitled to continue simply because he thought that obtaining verdicts from the jury justified that course."
He also said the judges recognised for the victims, the pronouncement of the verdicts, at least in certain cases, was a vindication of their evidence.
"Nothing we decide in any sense should be taken as removing their appropriate feelings of vindication," he said.
The proceedings were brought by Mr Turk's mother, Jasmine Botting, 76, from Etchingham, who previously said: "I know for a fact that my son is

Chapter 7

Conclusions and Future Work

7.1 Conclusion

The project is all about to extract news from different website which are related to the topic and the user gets desired topics by using this application and a summarize version is given to it. User can also search the Keywords in already downloaded stories. User may choose to open and read a detail about a story about their choice.

In this we are Recognizing the topic which can reduce the time read, complexity and simplify the searching method for those who want a relevant topic in a specific domain. The proposed automated system will help you to find out the News story document in a textual format. The main method is used in this algorithm to divide the problem in to small chunks and overcome those problems to address the main issue. So that's why, we divide the text into sentences or in the form of words and compare the topic from each chunk of word to identify the desire result by obtaining its suitable syntactic parts. So, by each comparison we calculate the score of comparison and consider the topic which scored greater than the criteria. We reorganized its steps and modified its selection policy. We select Noun Phrase and Verb Phrase instead of noun-noun and noun-verb pairs from each sentence. By this modification, we achieved 80% and above of matching for both total and partial matching among 30 and above random documents from the different News websites.

7.2 Future Work

This project will further be implemented for further News links as we consider the BBC News website. It is easily applicable to be implemented for several News Websites and for various News topics like we have currently focusing on sexual abuse and health related news. We can further Expand it for other types of News and analyse easily which type of news a user need and program will except there needs and process it. Extract the latest News for the user according to their selected topic.

- In further extend we create a graph for it. In which for each reporting it will display the levels by using graphs. User will easily identify ups and downs in related topic.
- We can also use this as a story summarizer. By neglecting all the stop Words and the stuff which are not an important part or not creating immense difference on it will be neglected and the story will be to the point. The reader will easily understand what the story is all about.
- Applying some more modification on this project can also be used to Identify the total number of cases in different cities and many other things can be applying on the Next version of this project for the investigation purposes.
- For further operation and functionality may apply to know the number of Victims and abuser in the News Story and in case of health news total number of patients of a particular disease could be identify.

Appendix A

User Manual

A.1 Code Description

The following class and methods in the classes are written.

- 🟥 Web-Crawler: This class is use to extract the Web content
  - 🟦 Constructor:This function provide the connection between the web and the application
  - 🟦 Get-BBC-News-links: This method will extract all links from web page
  - 🟦 Get-News-from BBC URL: It opens link one by one and extract information from it
    - 🟩 String-URL: Use to store the URL
    - 🟩 List Links:It store the List of Links
- 🟥 Extract Noun: Use to Extract Nouns and Adjective from the News Story and compare it with the manually extracted keywords.
  - 🟦 Noun1: This method will extract the Noun and adjective.
  - 🟦 Comparison: This method will compare Verbs and adjective with the extracted keyword
    - 🟩 String News: The body of the News Story
    - 🟩 Double List-String L Phrase: It will store the Nouns and Verbs List
- 🟥 WriteXMLFILE: Use to store a story in XML format.
  - 🟦 Main Method: It will store the Title and paragraph of the News story.
    - 🟩 String news: It store the story from the web
    - 🟩 Title: Store the Title of the story

Table A.1: Common English Stopwords

a
about
above
after
again
against
all
am
an
and
any
are
aren’t
as
at
be
because
been
before
being
below
between
both
but
by
cannot
can’t
could
couldn’t
did
didn’t
do
does
doesn’t
doing
into
she’s
we’d
each
it
for
we’ve
further
me
hadn’t
more
hasn’t
mustn’t
haven’t
myself
he
nor
he’ll
of
here
on
hers
only
he’s
other
himself
our
how
ourselves
i
over
if
same
i’m
she
don’t
down
is
few
its
shouldn’t
itself
i’ve
let’s
had
has
the
have
theirs
having
no
he’d
not
her
off
here’s
once
herself
or
him
ought
his
ours
how’s
out
i’d
own
i’ll
very
in
she’d
wasn’t
isn’t
should
it’s
we’re
what’s
such
than
that
when
most
where
my
which
themselves
who
there
who’s
these
why’s
they’d
won’t
they’re
wouldn’t
you
those
you’ll
to
you’re
under
yourself
up
you’ve
was
she’ll
we
during
we’ll
were
so
weren’t
from
what
some
that’s
when’s
their
where’s
them
while
then
whom
there’s
why
they
with
they’ll
would
they’ve
this
you’d
through
your
too
yours
until
yourselves
shan’t

References

[1] Stuart Rose, Dave Engel, Nick Cramer, and Wendy Cowley. *Automatic Keyword Extraction from Individual Documents*, pages 1–20. John Wiley & Sons, Ltd, 2010. Cited on pp. 4 and 5.

[2] Kun Yue, Wei-Yi Liu, and Li-Ping Zhou. Automatic keyword extraction from documents based on multiple content-based measures. *Computer Systems Science and Engineering*, 26(2):133, 2011. Cited on p. 5.