Temporal Document Clustering using News Dataset

Submitted By

Salman Khalid (BC-131071)
Omer Saddique (BC-131084)
Muhammad Umair Anjum (Expert Advisor)Consultant

Supervised By

Dr. Muhammad Azhar Iqbal

Department of Computer Science

Capital University of Science & Technology, Islamabad

8th semester/16th of January, 2017

Capital University of Science & Technology

Final Approval Certificate

It is to certify that this project meets the entire requirements and warrants its acceptance by CUST for the degree of BS (Major in Computer Science) for Salman khalid, BC131071, sammekh04@gmail.com and Omer Saddique, BC131084, muh.omer192@gmail.com.

Committee Signatures:

Consultant: _____

Muhammad Umair Anjum

Supervisor: _____

Dr. M. Azhar Iqbal

Project Coordinator: _____

Mr. Saleem Anwar

Head of Department: _____

Dr. NayyerMasood

Acknowledgement

We are greatly indebted to our supervisor Dr. Azhar Iqbal, for his determined guidance and dedication to help us complete this project. We are thankful to our supervisor for his patience and the time he had given us to see our project.

Declaration

We hereby declare that this undergraduate project, neither as a whole nor as a part there of has been copied out from any sources, wherever references have been provided. It is further declared that we have developed this undergraduate project and accompanied report entirely on the bases of our knowledge. If any part of this undergraduate project is proved to be copied out from any source and found to be reproduced of some other source without proper acknowledgement, we will stand by the consequences, and the university will take no responsibility.

Project Members:

Signature: _____

Salman Khalid (BC131071)

Signature: _____

Omer Saddique (BC131084)

Date: _____

DEDICATION

We would like to dedicate this project to our parents, our teachers especially to our supervisors Dr. M. Azhar Iqbal and Dr. M. Arshad Islam and to those of our friends who have helped us see through this project by being a great source of comfort and encouragement to us.

Table of Contents

Chapter 1

Introduction

In computer science research is growing rapidly day by day. Time clearly plays a central role in any information space and it has been studied in several areas like information extraction topic detection, query log analysis, and summarization. Time and temporal measurements can help recreating a particular historical period or a collection of documents. Content clustering is very useful for this purpose but there are several improvements that can be made.

1.1 Purpose of the Project

The purpose or aim is to develop a java application which tells us about the lifespan of news in different newspapers on given timeline and place them in different clusters. Clustering is the process of grouping similar things or people or documents on the bases of their any similarity. In a computer system, a cluster is a group of servers and other resources that act like a single system and enable high availability and, in some cases, load balancing and parallel processing. Given figure 1.1 is an example of how this project will work.

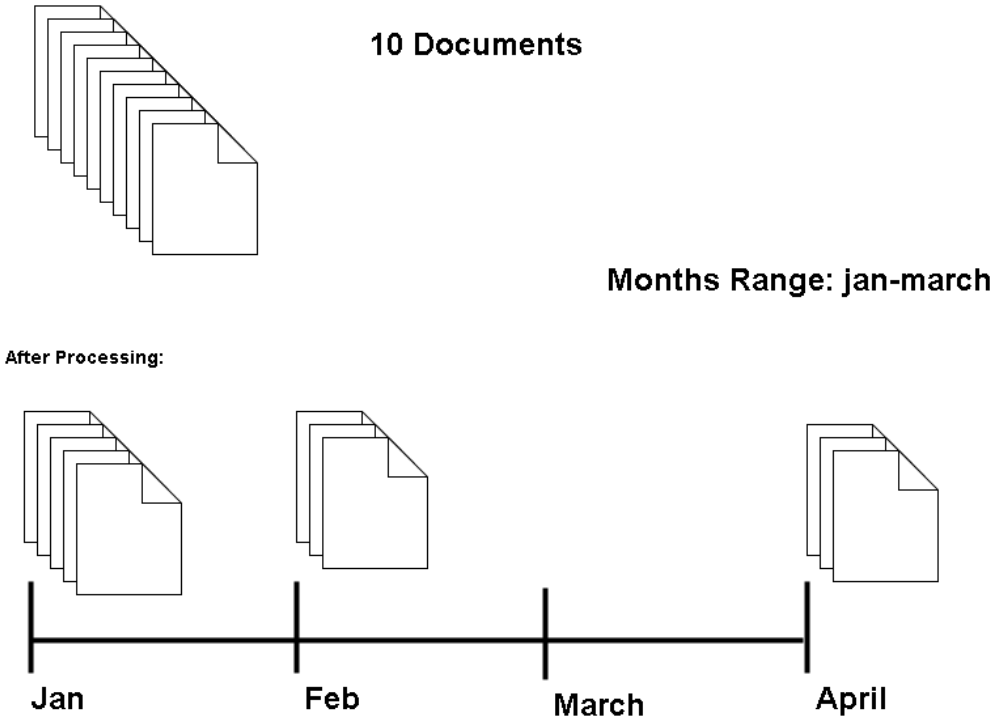

Figure 1.1 Process example

1.2 List of Deliverables

- Working Java application.
- Documented report of the project.
- Temporal visualization.

1.3 Business Scope

The scope of this project is to survey the lifespan of news in different newspapers within the given time, which is useful in journalism and in other public surveys.

1.4 Tools and Technologies Require

- Net beans IDE.
- Java APIs.
- Time library SDK.
- News archives.

1.5 Project Breakdown

1	Design Project-I Mid	● Vision Document. ● Understanding clustering and providing its examples ● Software Requirement Specification.
2	Design Project-I Final	● Extracting News time from the News archives and maintaining a list of all citation. ● Sorting the news according to given time
3	Design Project-II Mid	● Extracting the news span
4	Design Project-II Final	● Testing and Evaluation.

1.6 Architecture Diagram:

This architecture diagram show the flow of how the project will work, each box represent a result that was found in result of executing a function written on the arrow. In the end different cluster will form and documents will be placed in them.

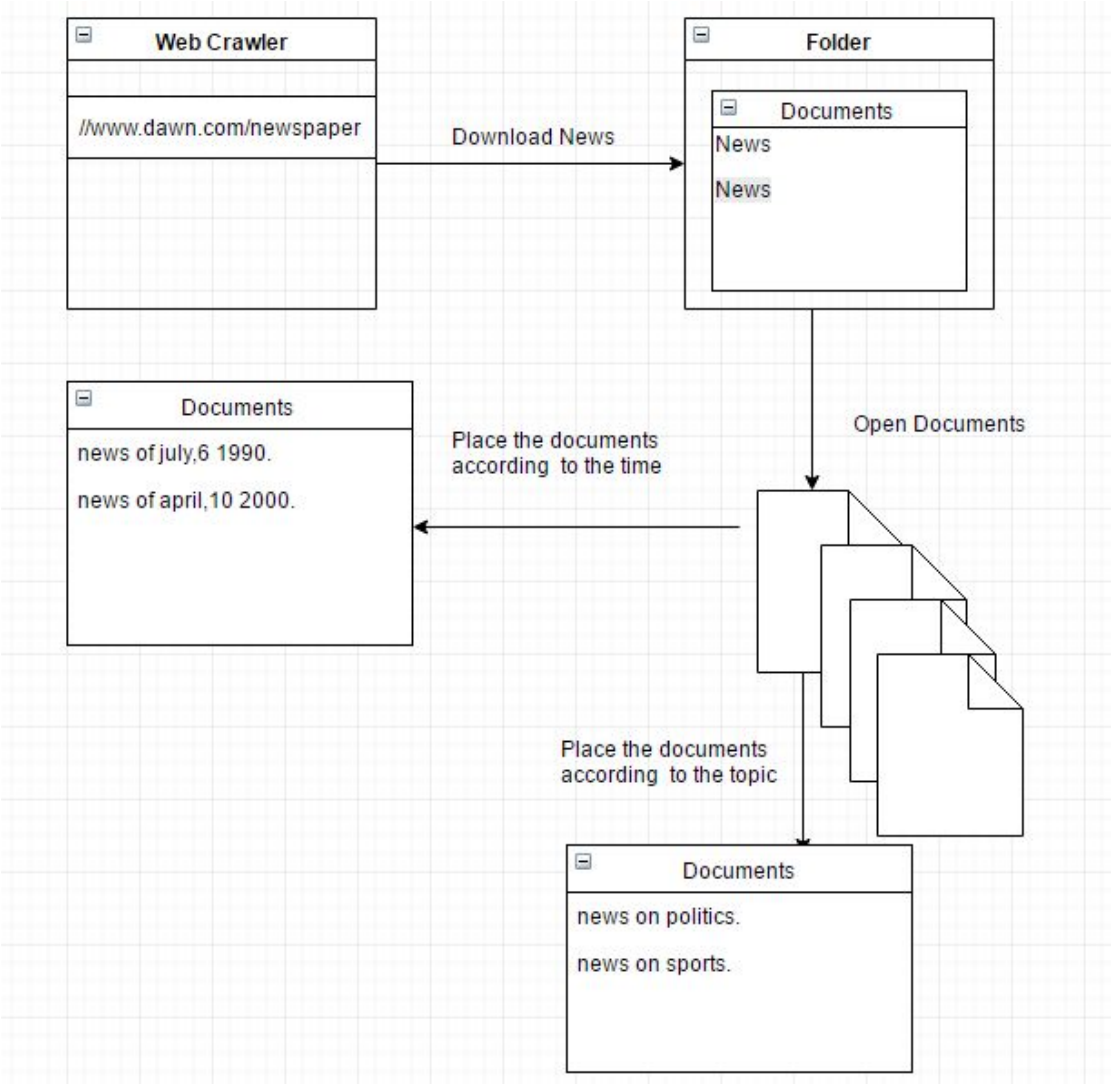

Figure 1.2 System Architecture Diagram

Chapter 2

Software Requirement Analysis

2.1 Purpose

Requirements analysis, also called requirements engineering, is the process of determining user expectations for a new or modified product. These features, called requirements, must be quantifiable, relevant and detailed. The requirements defined in this section will later be used to validate and verify the project. Software requirements for this project are characterized into only one type, on the basis of the devices used for the project that is Computers.

2.2 Scope

We are developing this project for the general public use to check the lifespan of news in newspapers and also compare it with different newspapers.

2.3 Overall Description

2.3.1 Product Perspective

The main purpose of this system is to provide the survey of lifespan of news in different newspapers. To achieve this we have to understand the clustering.

Clustering:

Clustering is the task of grouping a set of objects in such a way that objects in the same group are more similar to each other than to those in others groups. The algorithm which we used to implement the clustering is k-means algorithm.

K-means algorithm:

K-means algorithm is used to make the groups of data points, as shown in figure 2.1.

Figure 2.1 Data set for k-mean example

Suppose we want to divide these points into 3 clusters so in this case the Value of k=3 (clusters).

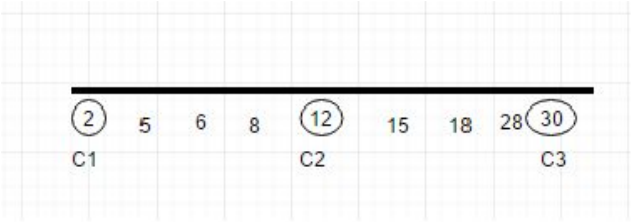

Figure 2.2: Selecting centroids from the data set

The first step is to find the clusters center from these data points as shown in figure 2.2. The criteria to select the clusters center are that centers are farthest apart from each other.

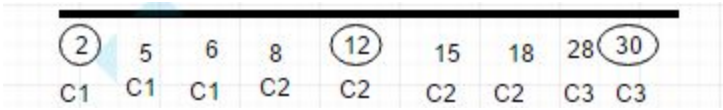

Figure2.3: assigning values to their closest centroid

Now we have to consider all the data points and see which point is closest to which cluster. For Example if we consider 5 from data points we can see that 5 are closest to c1 so we assigned 5 the level of c1 as shown in figure 2.3.

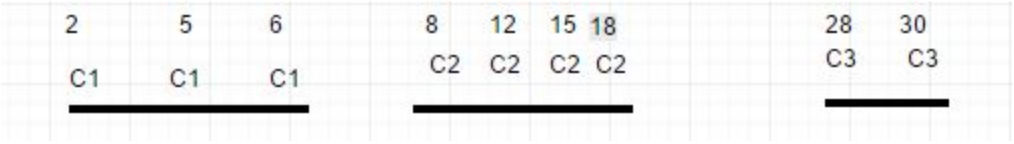

Figure2.4: creating different datasets base on centroids

Third step is to find the mean of clusters the mean of C1 is 4.3, C2 is 13.25 and C3 is 29. The mean which we have calculated are considered as the new cluster center for each of the cluster as shown in the figure 2.4.

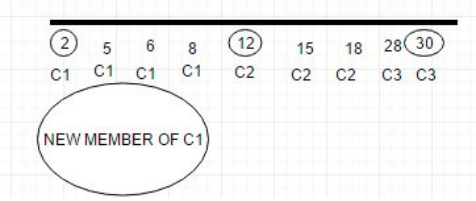

Figure2.5: change values positions according to new mean

The Fourth step is an iterative process, again we have to consider all the data points, calculate the distance of all data points from each of the cluster center and assign all data points to the level of cluster center which is closest to that data point. So in this case the assignment of some data points will be changed. For Example if we consider 8 from data points previously this is closest to C2 that why we assigned it level C2 but now these new values of clusters centers. We can see the distance of this point is least from C1 so the level of this point changed to C1.This will not be member of C2 because the distance of this point is 3.7 from C1 and 5.2 from C2.We will do this process for all data points after reassignment of cluster levels to each data point. We again have to calculate mean of cluster center of each cluster maybe shift towards right or left so this will again change the membership of different data points. We repeat the same process till we observe the shifting in cluster center has stopped as shown in the figure 2.5.

Example # 2:

Let us take another example of k-means algorithm with two dimensional data which is given below.

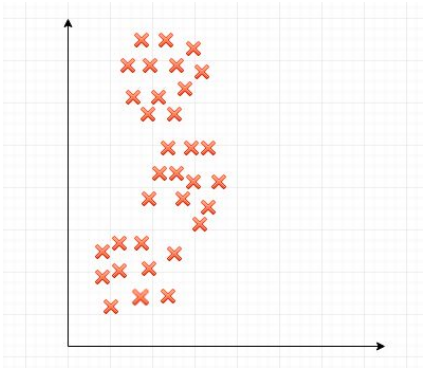

Figure 2.6: Data set for 2D k-mean example

Suppose we want to divide these points into 3 clusters so in this case the value of k=3(clusters).

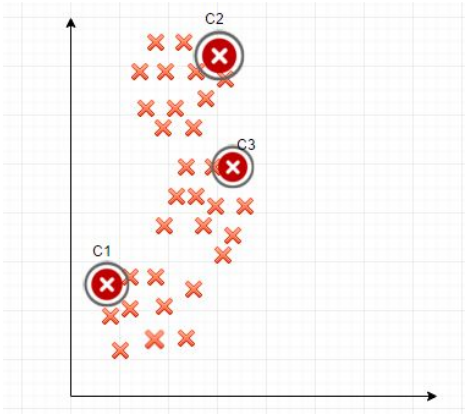

Figure2.7: Selecting Centroids from the data set

The first step is to find the clusters center from these data points. The criteria to select the clusters center are that centers are farthest apart from each other.

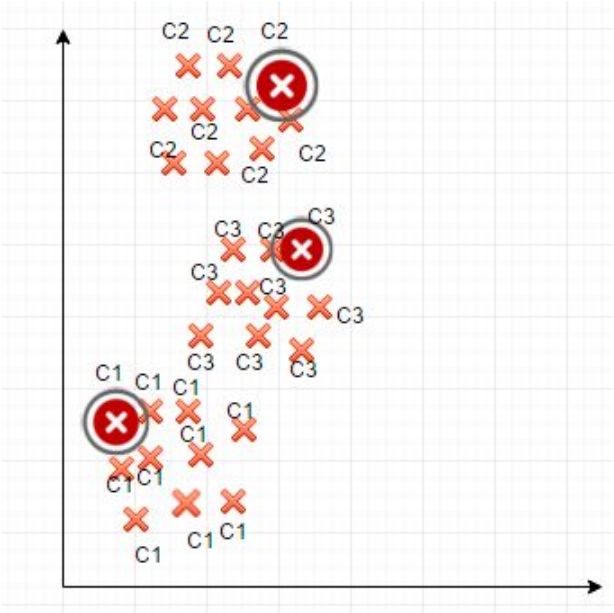

Figure2.8: assigning values to their closest centroid

Now we have to consider all the data points and see which point is closest to which cluster.

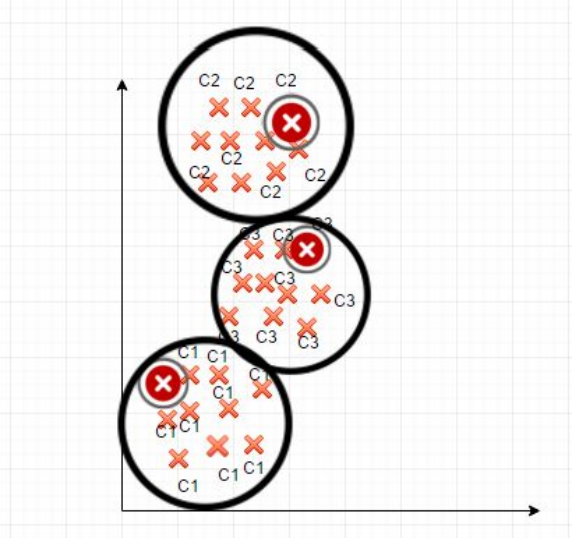

Figure2.9: assigning values to their closest centroid

Third step is to find the mean of clusters. The mean which we have calculated are considered as the new cluster center for each of the cluster.

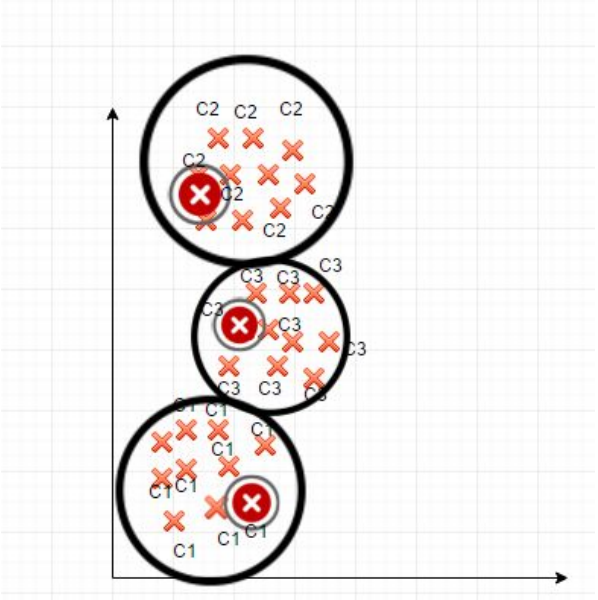

Figure2.10: change values positions according to new mean

The fourth step is an iterative process, again we have to consider all the data points, calculate the distance of all data points from each of the cluster center and assign all data points the level of cluster center which is closest to that data point. We will do this process for all data points after reassignment of cluster levels to each data point. We again have to calculate mean of cluster center of each cluster maybe shift towards right or left so this will again change the

membership of different data points. We repeat the same process till we observe the shifting in cluster center has stopped.

2.3.2 Product Functions

The system has following functional and non-functional requirements

Table 1: Functional & Non-Functional Requirements

Ref. #	Functional Requirements	Type	Non-Functional Requirements	Category
1.0	System should verify the dataset query and time range provided.	**Primary**	System should notify the user about failure or success in less than 5 seconds.	• Reliability • Performance
2.0	System should place documents in respective cluster folder and add their time in their name.	**Primary**	The system should provide a user interface to access cluster	• Reliability • Data • Integrity
3.0	System should open the list of cluster when user moves his mouse over it	**Primary**	The system should show all the documents names in that cluster in less than 2 seconds	• Usability
4.0	System shall open a document only when user clicks on it	**Primary**	System should verify, if that document is available to open else notify the user	• Reliability • Availability • Usability

2.4 Use Case Model:

A use case defines a set of use-case instances, where each instance is a sequence of actions a system performs that yields an observable result of value to a particular actor.

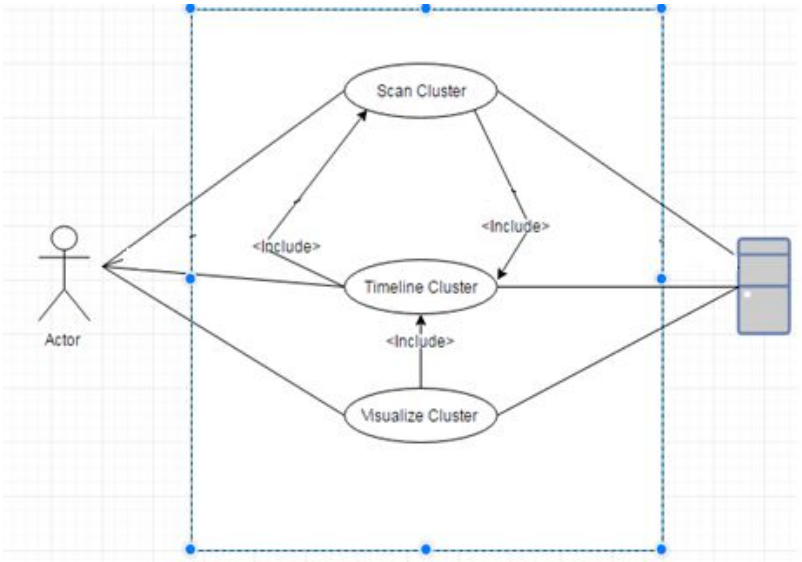

Figure2.11 Use Case Diagram for this project

2.5 Use Case Description:

2.5.1 Brief description of Use Case 'Scan Cluster':

Use Case ID:	1		
Use Case Name:	Scan cluster		
Created by:	Omer Saddique Salman Khalid	**Last Updated by:**	27 March 2016
Date Created:	7 April 2016	**Last Revision Date:**	1 April 2016
Actors:	User and System.		
Description:	User provides a data set which system divides into clusters and shows them a timeline (interface).		
Trigger:	User need to Click the start button.		
Pre-conditions:	User must have a data set location address.		
Post-conditions:	Folders are created for each cluster.		
Normal Flow:	Actor. 1. Actor types the address of data set in the given field.	System 4. System validates the data set location.	

	2. Actor type the time range in months in the given field. 3. Actor press the " Start Clustering" button	**5**. System validates the provided time range. **6**. System shows the writing screen. **7**. System displays the time line.
Alternative Flows:	**4.1.**If the time range validation fails. **4.1.1.**Actor is asked to enter time range again. **5.1.**If provided location is wrong. **5.1.1.**Actor is asked to end data set address again.	
Exceptions:		

2.5.2 Brief description of Use Case Timeline Cluster

Use Case ID:	2		
UseCase Name:	Timeline Cluster.		
Created by:	Omer Saddique Salman Khalid	**Last Updated by:**	14 April 2016
Date Created:	7 April 2016	**Last Revision Date:**	14 April 2016
Actors:	User and System.		
Description:	Actor clicks on the document name to open it.		
Trigger:			
Pre conditions:	Mouse pointer is on the cluster and a list is shown.		
Post conditions:	Document is opened.		
Normal Flow:	Actor. **1.**Actor clicks on the document name	System	

		2. System opens that document in it respective application.
Alternative Flows:	**2.1.** If that document has been moved from the directory **2.1.1.** System shows “Document is removed” error.	
Exceptions:		

2.5.3 Brief description of Use Case Visualize Cluster:

Use Case ID:	3		
UseCase Name:	Visualize Cluster.		
Created by:	Omer Saddique Salman Khalid	**Last Updated by:**	14 April 2016
Date Created:	7 April 2016	**Last Revision Date:**	14 April 2016
Actors:	User and System.		
Description:	System creates clusters on the basis of news discussions.		
Trigger:			
Pre-conditions:	Given data set address is correct.		
Post-conditions:	System creates new folders which will be clusters and move documents.		
Normal Flow:	Actor.	System	

		1. System open documents and analyze them. 2. System open documents in their respective clusters on the basis of news. 3. System add news document in the document name. 4. System return an address Where it has created and stored clusters.
Alternative Flows:		
Exceptions:		

2.6 System Sequence Diagram:

System sequence diagrams are created to show the sequence of events among different objects in the system.

2.6.1 Scan Cluster

In the sequence diagram as shown in Figure 2.12, user input the query string and system will scan the cluster.

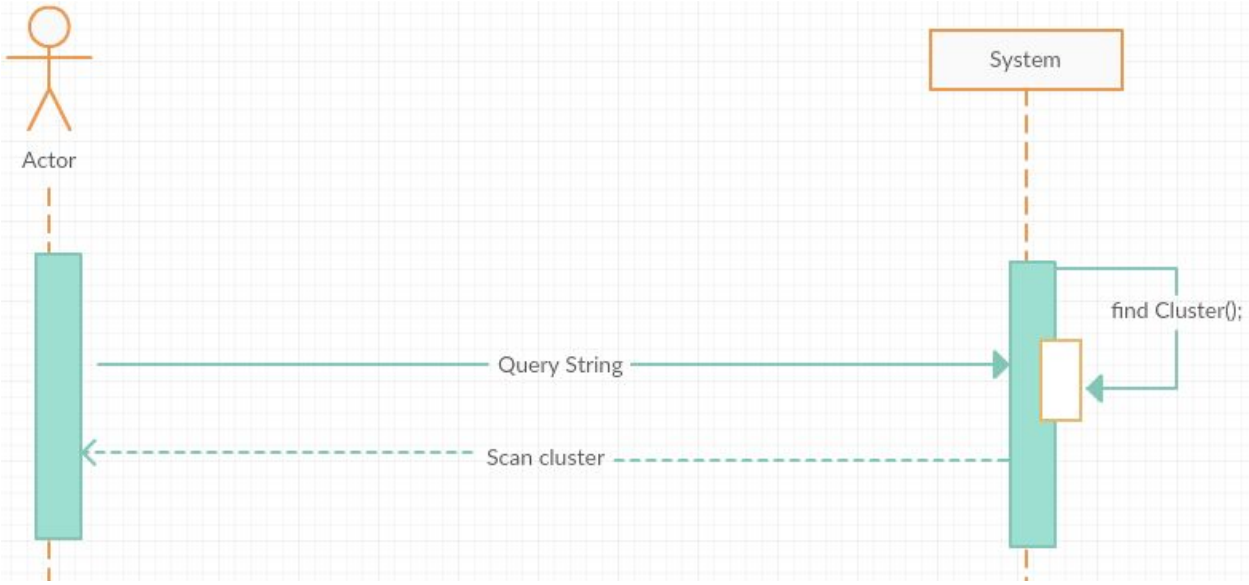

Figure 2.12 Scan Cluster

2.6.2 Timeline Cluster

In the sequence diagram as shown in Figure 2.14, an event will occur and system will timeline the cluster.

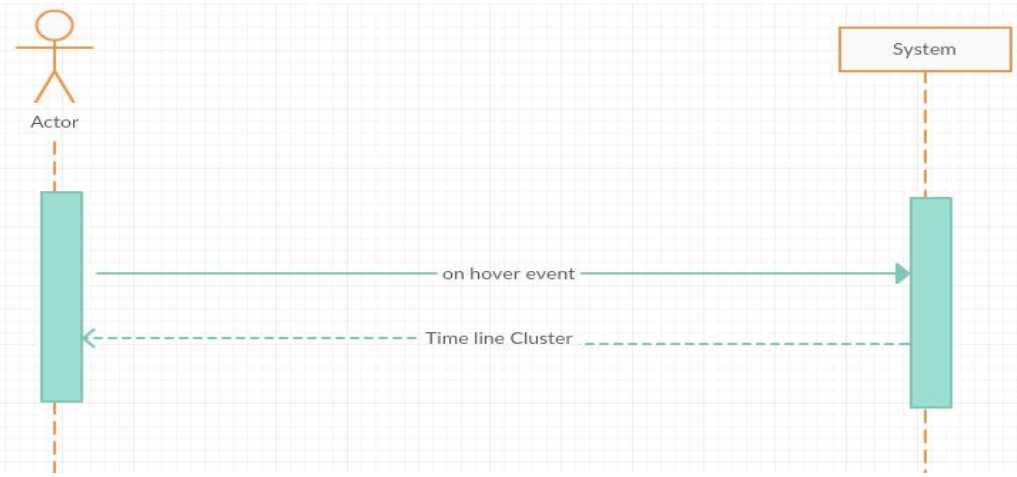

Figure 2.13 Time Line Cluster

2.6.3 Visualize Cluster:

In this sequence diagram as shown in Figure 2.12 below user input the query string and system will scan the cluster.

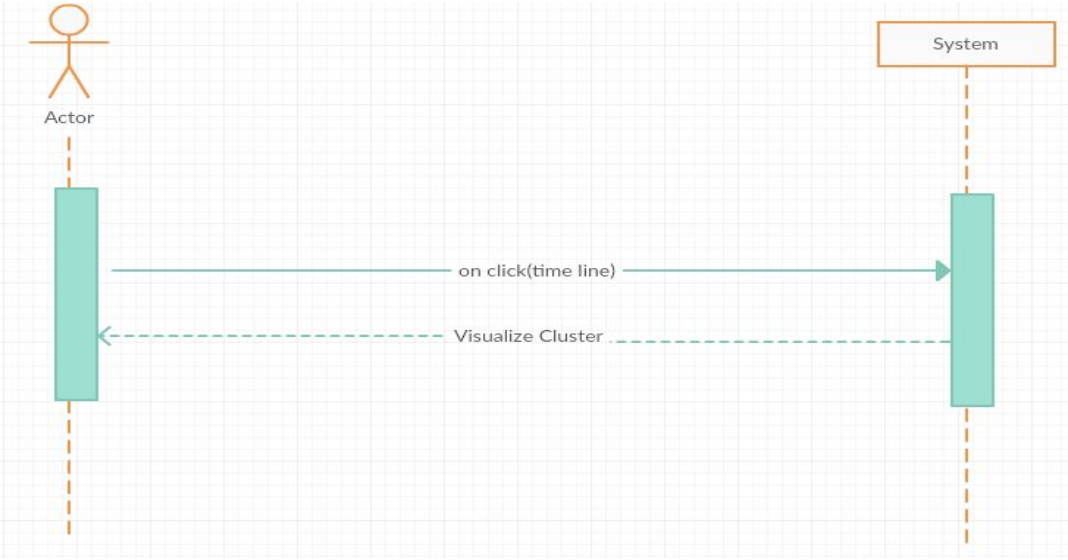

Figure 2.14 Visualize Cluster

Chapter 3

System Design

3.1 Architecture diagram

The Architecture diagram explains the general architecture of the whole system as shown in figure 3.1.Time and temporal measurements can help recreating a particular historical period or a collection of documents. Content clustering is very useful for this purpose but there are several improvements that can be made.

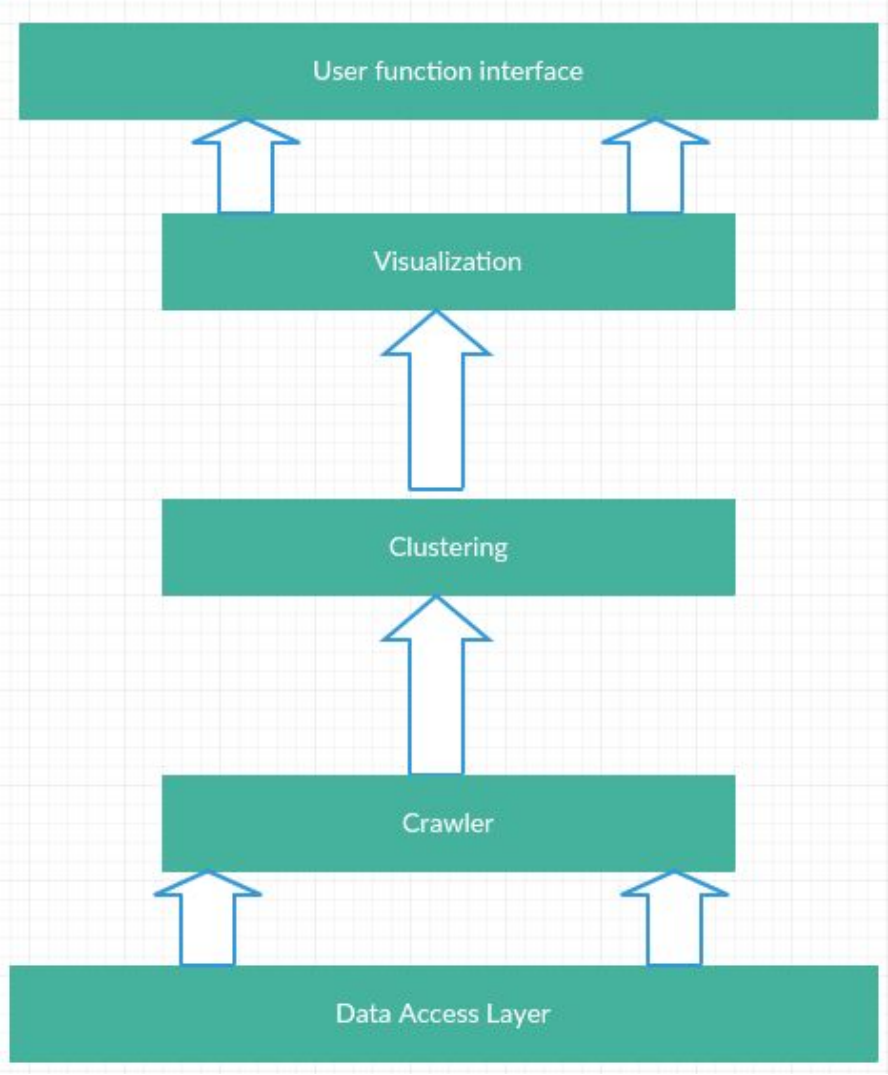

Figure 3.1 Architecture Diagram for this Project

3.2 Class diagram

Class Diagram as shown in Fig. 3.2 provides an overview of the target system by describing the objects and classes inside the system and the relationships between them.

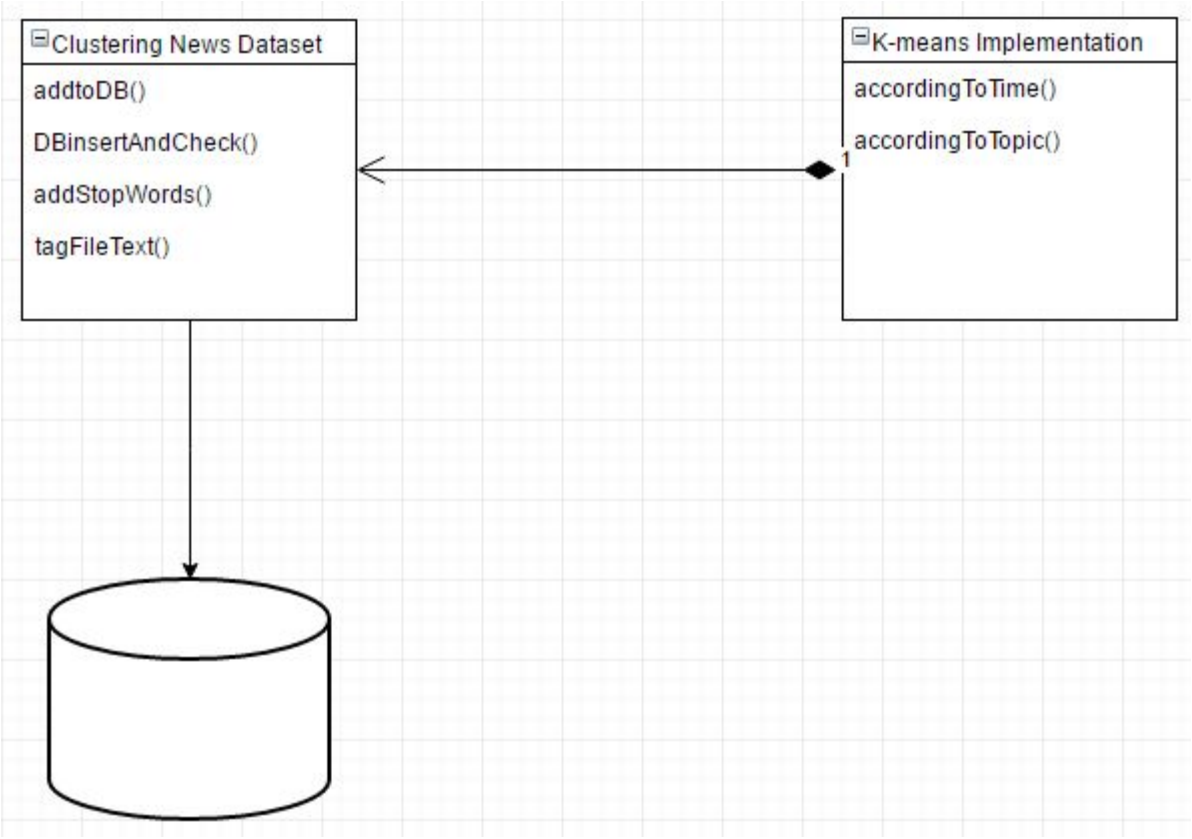

Figure 3.2 Class Diagram for this project

3.3 Sequence diagram

In this sequence diagram as shown in Figure 3.3 user first press the button then enter the URL address (dawn.com/newspaper/2012-01-01/) randomly and then our software crawled the news from the website and stored it to the given folder path.

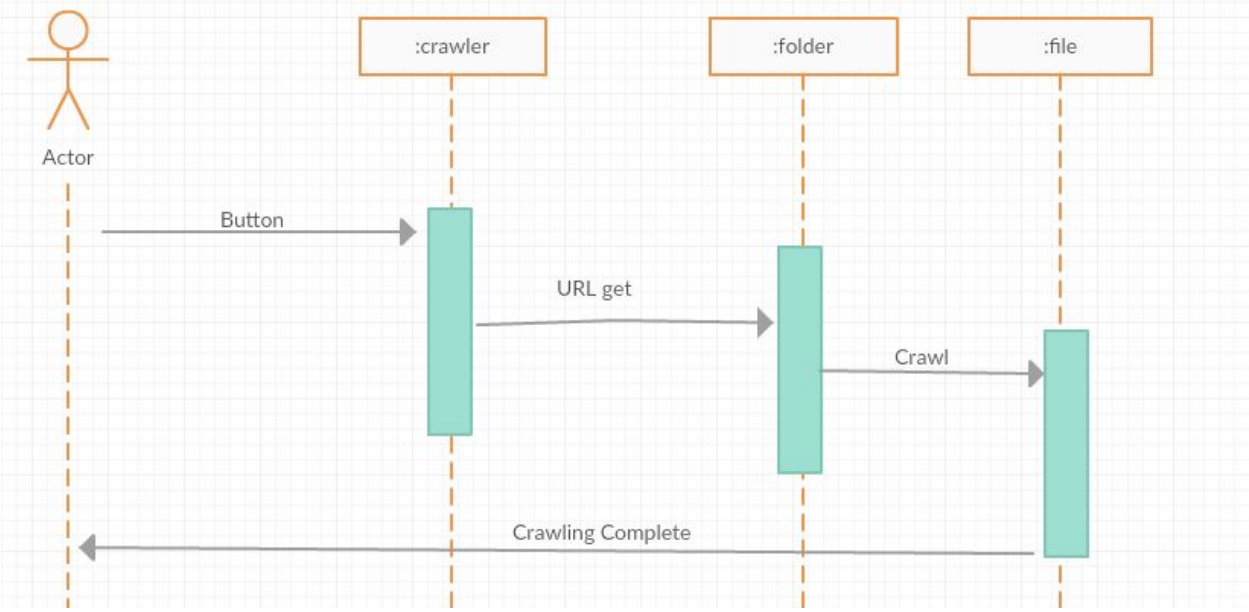

Figure 3.3 Sequence Diagram for web Crawler

In this sequence diagram as shown in Figure 3.4, user press the button cluster user will get the option to chose from get time base cluster or topic base cluster and then pre processing tokenization will occur and remove stop word from file and display result to the user.

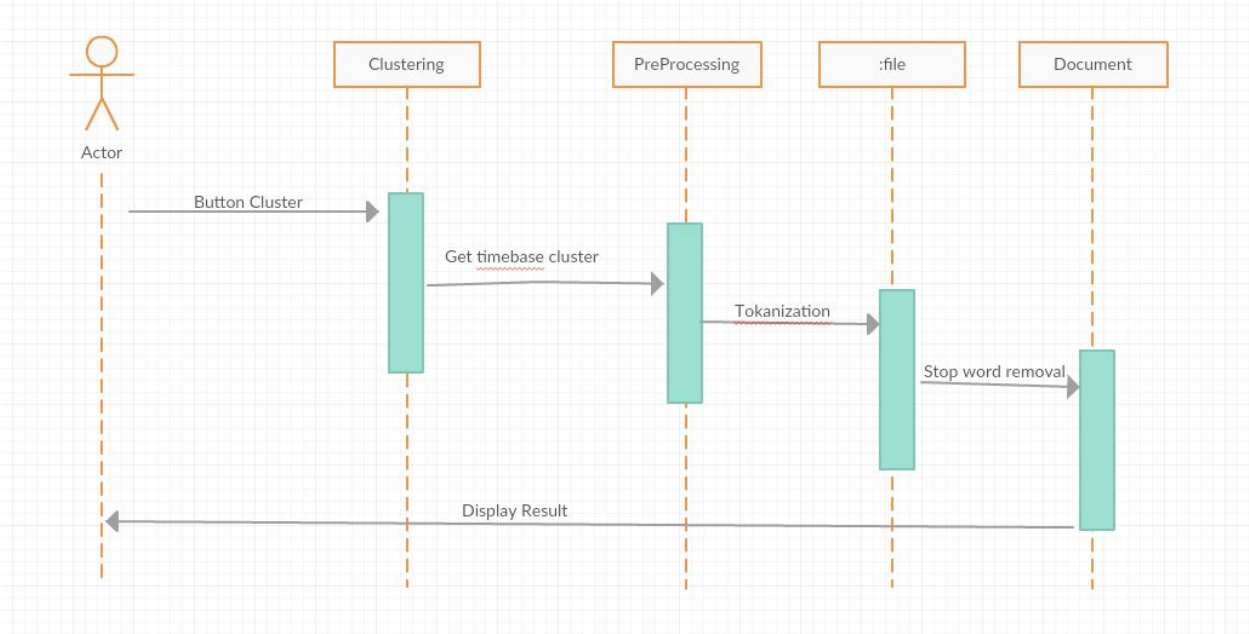

Figure 3.4 Sequence Diagram for displaying clusters on timeline

3.4 User Interface:

Figure 3.5 User Interface for this project

Figure 3.6 User Interface for this project

3.5 Software COTS:

Short for commercial off-the-shelf, an adjective that describes software or hardware products that are ready-made and available for sale to the general public. In our project the software cots which we are using are given below.

3.5.1 Jsoup:

Jsoup is a Java library for working with real-world HTML. It provides a very convenient API for extracting and manipulating data, using the best of DOM, CSS, and jquery-like methods.
jsoup implements the WHATWG HTML5 specification, and parses HTML to the same DOM as modern browsers do.

- scrape and parse HTML from a URL, file, or string
- find and extract data, using DOM traversal or CSS selectors
- manipulate the HTML elements, attributes, and text
- clean user-submitted content against a safe white-list, to prevent XSS attacks
- output tidy HTML

jsoup is designed to deal with all varieties of HTML found in the wild; from pristine and validating, to invalid tag-soup; jsoup will create a sensible parse tree.

3.5.2 Pos tagger:

The library provided lets you “tag” the words in your string. That is, for each word, the “tagger” gets whether it’s a noun, a verb ..etc. and then assigns the result to the word. For example: “This is a sample sentence” will be output as

This/DT is/VBZ a/DT sample/NN sentence/NN

To do this, the tagger has to load a “trained” file that contains the necessary information for the tagger to tag the string. This “trained” file is called a model and has the extension “.tagger”. There are several trained models provided by Stanford NLP group for different languages.

3.5.3 JfreeChart :

JFreeChart is a free 100% Java chart library that makes it easy for developers to display professional quality charts in their applications. JFreeChart's extensive feature set includes:

- a consistent and well-documented API, supporting a wide range of chart types.
- a flexible design that is easy to extend, and targets both server-side and client-side applications.
- support for many output types, including Swing and JavaFX components, image files (including PNG and JPEG), and vector graphics file formats (including PDF, EPS and SVG).

3.6 References:

Jsoup: jsoup Java HTML Parser, with best of DOM, CSS, and jquery

jsoup.org

Pos Tagger:

Java example for using Stanford POSTagger

www.programcreek.com

JfreeChart:

http://www.jfree.org/jfreechart/

Chapter 4

Implementation Issues

4.1. Tools and Technologies Used:

- Net beans IDE.
- Java APIs.
- News archives.

4.1.1 Net beans IDE:

For java coding we use Net Bean for it because it provide different type of features like provide facility of connectivity with the database, filling and support different type of libraries. In this section you will provide the reason behind using all the existing tools and technologies that you may have used during the development of your project. This includes development environment that you have used. How have you deployed the development environment? What different kind of packages you have used? Are there any third party libraries involved etc.

4.1.2 Java APIs:

The Java language defines the syntax and semantics of the Java programming language. It defines basic vocabulary such as primitive types, if/else blocks, the syntax of class declaration, exception syntax, variable scoping rules, and everything else necessary for the language to function

1 Jsoup:

jsoup is a Java library for working with real-world HTML. It provides a very convenient API for extracting and manipulating data, using the best of DOM, CSS, and jquery-like methods.
jsoup implements the WHATWG HTML5 specification, and parses HTML to the same DOM as modern browsers do.

- scrape and parse HTML from a URL, file, or string
- find and extract data, using DOM traversal or CSS selectors
- manipulate the HTML elements, attributes, and text
- clean user-submitted content against a safe white-list, to prevent XSS attacks
- output tidy HTML

jsoup is designed to deal with all varieties of HTML found in the wild; from pristine and validating, to invalid tag-soup; jsoup will create a sensible parse tree.

2 Pos tagger:

The library provided lets you “tag” the words in your string. That is, for each word, the “tagger” gets whether it’s a noun, a verb ..etc. and then assigns the result to the word. For example: “This is a sample sentence” will be output as

This/DT is/VBZ a/DT sample/NN sentence/NN

To do this, the tagger has to load a “trained” file that contains the necessary information for the tagger to tag the string. This “trained” file is called a model and has the extension “.tagger”. There are several trained models provided by Stanford NLP group for different languages.

3 JfreeChart :

JFreeChart is a free 100% Java chart library that makes it easy for developers to display professional quality charts in their applications. JFreeChart's extensive feature set includes:

- a consistent and well-documented API, supporting a wide range of chart types.
- a flexible design that is easy to extend, and targets both server-side and client-side applications.
- support for many output types, including Swing and JavaFX components, image files (including PNG and JPEG), and vector graphics file formats (including PDF, EPS and SVG).

4.1.3 News Archives:

In our software we use Don news archives from their website of two years. The first step in this project is to crawl documents from a news archive for his we have made our own web crowler which crowls only the texts from that respective article after getting almost 25 to 30 news documents. In our next step we will find out the term frequency for each saved document and also indentify the time occurance in that respective documents.

4.2 Software Description

4.2.1 Introduction

Time plays an important role in any information space and it has been studied in several areas like information extraction topic detection, query log analysis, and summarization. Time and temporal measurements can help recreating a particular historical period or a

collection of documents. Content clustering is very useful for this purpose but there are several improvements that can be made.

The purpose of this work is to develop extract lifespan of news in different newspapers on given timeline and place them in different clusters. Clustering is the process of grouping similar things or people or documents on the bases of their any similarity. In this work we will use text clustering that is the application of cluster analysis for temporal information in textual documents. This work can be used for automatic document organization, topic extraction and fast information retrieval or filtering . Example of working mechanism of our proposed work is shown in Fig 1.

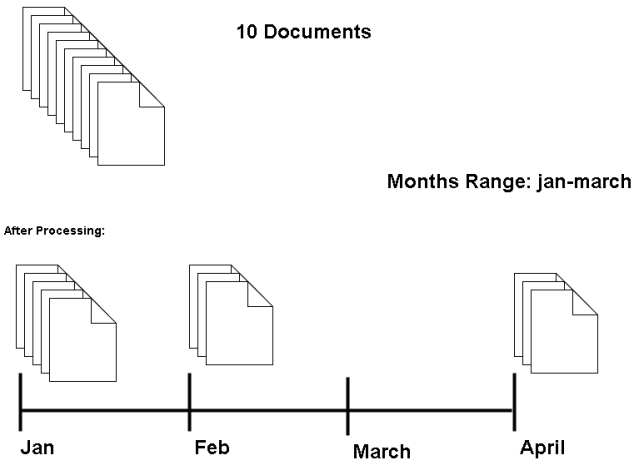

Figure 4.1: Working mechanism

The scope of this project is to survey the lifespan of news in different newspapers within the given time, which is useful in journalism and in other public surveys.

4.2.2 Methodology

The first step in this project is to crawl documents from a news archive for his we have made our own web crowler which crowls only the texts from that respective article after getting almost 25 to 30 news documents. In our next step we will find out the term frequency for each saved document and also indentify the time occurance in that respective documents. Now using this data we will apply K-means[4] Algorithm to create clusters based on thier time and thier topics.Clustering is the task of grouping a set of objects in such a way that objects in the same group are more similar to each other than to those in others groups[2]. Now we have our require clusters and the last step of this

project is to visualize in a proper timeline which will show the intensity of the time those clusters represent.

4.2.3 Findings

The data on timeline will be represented as on y-axis we have tterm frequency which will show the intesity of that particular cluster and on x-axis we have timeline which is represting the duration of the documents. Each graph will show the multiple cluster and which cluster will be represented is repsented by line and each line will be the represented a certain document in that cluster.

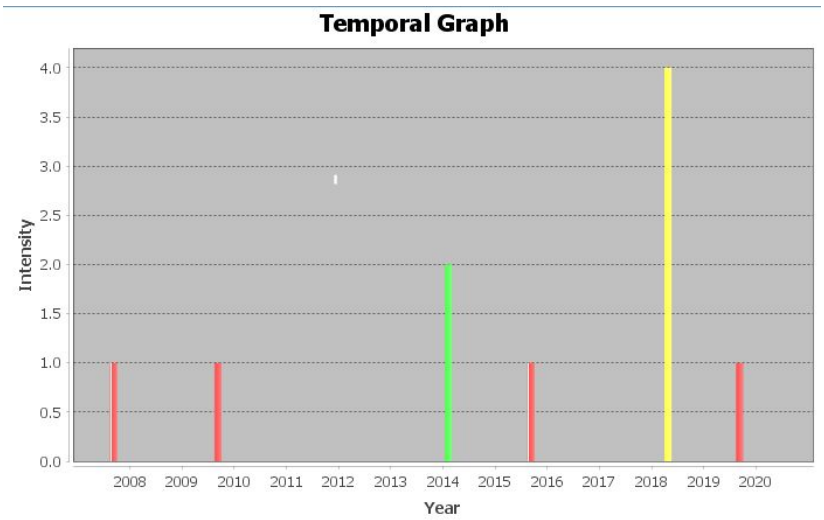

Figure 4.2: Sample output for the given data

4.2.4 Conclusion

After getting data set or website link as a input. The data set will be reviewed and the website link will be cawled and the result will display the timeline which shows the duration from which documents belong or will represent it. This timeline will be in hierarchical from where years will be first layer and month will be the second layer so that the cluster can be further classified in months of the same year.

Chapter 5

Testing and Deployment

This section of the document will cover the testing phase implemented on the project. This section identifies gaps, errors, or missing requirements in contrary to the actual requirements. It also gives a description how to access the website and courses.

5.1. Testing Methodology

Testing plan plays a very important role to ensure that the product delivered is stable and performing all its functional tasks correctly.

The testing methodology that we used for this product is Black Box Testing. This kind of testing exams the functionality of the project without knowing that function is working in the back. The reason for adapting this methodology is it's commonly used in industry and to check whether the product is working as it was expected to work.

5.2. Test Cases

5.2.1 Test Case 1: Creating Clusters

Table 5.1: Test Case 1

Description: To test whether the system is creating the clusters correctly or not.
How test case generated: We input the number of clusters for multiple test documents and check whether the system create the clusters right or not.
Expected Results: Successful creation of user account.
Actual Results: Passed

5.3 Installation Process Description

.EXE file will be given of this project. We will deploy our software on pc it is just an .exe file so that user can easily use it by just clicking on it.

Chapter 6

Conclusion and Future Work

In this chapter, we have focused on our objectives that are achieved in the development of the project and the ideas that are grown during the development.

6.1. Evaluation of Objectives and Aims

We are proud to say that we have successfully achieved our objectives that we had planned at the start of the project.

6.2. Evaluation of Project Management

Our supervisor Sir Azhar Iqbal played an important role to motivate us throughout the development of the project and helped us in generating course content. He always used to say, “Motivation is what gets you started. Habit is what keep you going”. Both project partner work on each module as assigned by a supervisor.

6.3 Discussion

In the beginning of the project, we are not clear that how would we able to meet the planned objectives. We start our project development in java. We discuss the project with supervisor on regular basis in the meeting and develop the discussed modules of project. We started from the basic requirements of the project and developed one by one to meet our aims of the project resulting in successfully completion of our project.