Chapter 1

Introduction

1. Introduction

1.1 Web based document clustering and retrieval

The main purpose of the web based app for document clustering and retrieval is to implement the document cluster techniques to categorize the documents and documents retrieval. For this purpose of categorization and retrieval this application provide the facility for reading and sharing of most relevant document. User can share their documents and other users can read it. The cluster of documents is also created by program automatically which will depend upon the documents attributes/property on server. Each document is stored in different cluster after the approval of administration. For users who is willing to search the documents of their interest can be able to search the document the cluster technique will retrieve all the related documents from server regarding to user query. Query will base on two types of categories query by example and query without example.

1.2 Existing System

People will have to search the documents regarding to their interest on the different website. They have to visit many sites after their query to search their related documents.

1.3 Problem statement

When some user on web want to search the document of his specific interest or requirement he has to visit many links to search the specific document or articles because they are not categorize on one link and the retrieval of document are also not on one link . Sometime most of the documents don't provide a complete material then user has to visit the other links for the documents this will consumes the time of the user.

1.4 Proposed solution

The documents are categorized in cluster and document sharing time by time this web application will provide readers to all the document of their interests. They web application will provide all the documents to users at the one link. . This web application hold all the documents in different clusters that is share by users and user can easily search all the document regarding to his query.

1.5 Purpose

Purpose of this document is providing a complete description of the features which will be implemented in this web based app for document clustering that based on document clustering technique or document cluster algorithm and project as well as explaining working principles of the system both for users and for developers of the project.

1.6 Scope

The system provide online platform which will based on document cluster and retrieval by using cluster techniques to categorized the documents and documents will save after defining the category of every documents or it will save dynamically as well as document sharing for those people who is looking for the different article /documents on different websites, those people can search many of articles/documents related to their interests on one link the cluster technique will retrieve all the document related to their queries.

1.7 Software Technologies

1.7.1 Eclipse IDE

Eclipse is an integrated development environment (IDE) used in computer programing, and is the most widely used Java IDE. It contains a base workspace and an extensible plug-in system for customizing the environment. Eclipse is written mostly in java and its primary use is for developing Java applications.

1.7.2 JSP Servlet

Java Server Pages (**JSP**) is a technology that helps software developers create dynamically generated web pages based on HTML, XML, or other document types.

Servlet technology is used to create web application (resides at server side and generates dynamic web page). Servlet technology is robust and scalable because of java language.

1.7.3 Ajax

AJAX stands for Asynchronous JavaScript and XML. In a nutshell, it is the use of the XMLHttpRequest object to communicate with server-side scripts. It can send as well as receive information in a variety of formats, including JSON, XML, HTML, and even text files.

1.7.4 Java script

Java Script (is a lightweight, interpreted, programming language with first-class functions. While it is most well-known as the scripting language for Web pages, many non-browser environments also use it, such as node.js and Apache Couch DB. JS is a prototype-based, multi-paradigm, dynamic scripting language, supporting object-oriented, imperative, and declarative (e.g. functional programming) styles. Read more about JavaScript.

1.7.5 CSS Framework (Bootstrap)

Bootstrap is a free and open-source front-end web framework for designing websites and web applications. It contains HTML- and CSS-based design templates for typography, forms, buttons, navigation and other interface components, as well as optional JavaScript extensions. Unlike many web frameworks, it concerns itself with front-end development only.

CHAPTER 2

SYSTEM ANALYISIS

2. System Analysis

Systems analysis is a problem solving technique that decomposes a system into its component pieces for the purpose of the studying how well those component parts work and interact to accomplish their purpose.

2.1 Functional Requirements

- ✓ Web based application is used by user to share their documents
- ✓ Web based application is used by admin to approve the documents to apply clustering techniques for documents categorization
- ✓ Web based application is used by user to search the document by uploading document, writing paragraph and by writing query system will retrieve the documents on the basic of similarity measure.

2.2 Non-Functional Requirement

The System Analysis chapter captures the system requirement for the following areas:

1. **Usability**: Application will be designed in such a way that it is very simple, user friendly and the amount of presented information content is limited and well specified.
2. **Reliability**: Clustering techniques that generate best results are used to make application more reliable.
3. **Performance**: Clustering techniques that are better in performance are used and programming context used in a way to optimize better performance.
4. **Privacy**: This application does not require critical information that hits user's privacy.
5. **Supportability**: This application requires web environment to run in.

2.3 Design Constraints

It has the following constraints:

1. It provides online clustering of documents.

2. It provides online searching of documents.

2.4 User Characteristics

User and admin characteristics are required to use the application.

2.5 User Environment

No special user requirement is required.

2.6Actor Goal List

Actor	Goal
1. User	a. Upload document for online clustering. b. Upload a document for searching relative documents on the basic of similarity measure. c. Search a document by writing a paragraph or writing a query on the basic of similarity measure
2. Admin	a. Approve uploaded document and after apply the clustering technique step by step. b. Admin can create more categories for best match of documents c. Dis approve an uploaded document

Table 2.1 Actor Goal list

2.7 Use Case Diagram

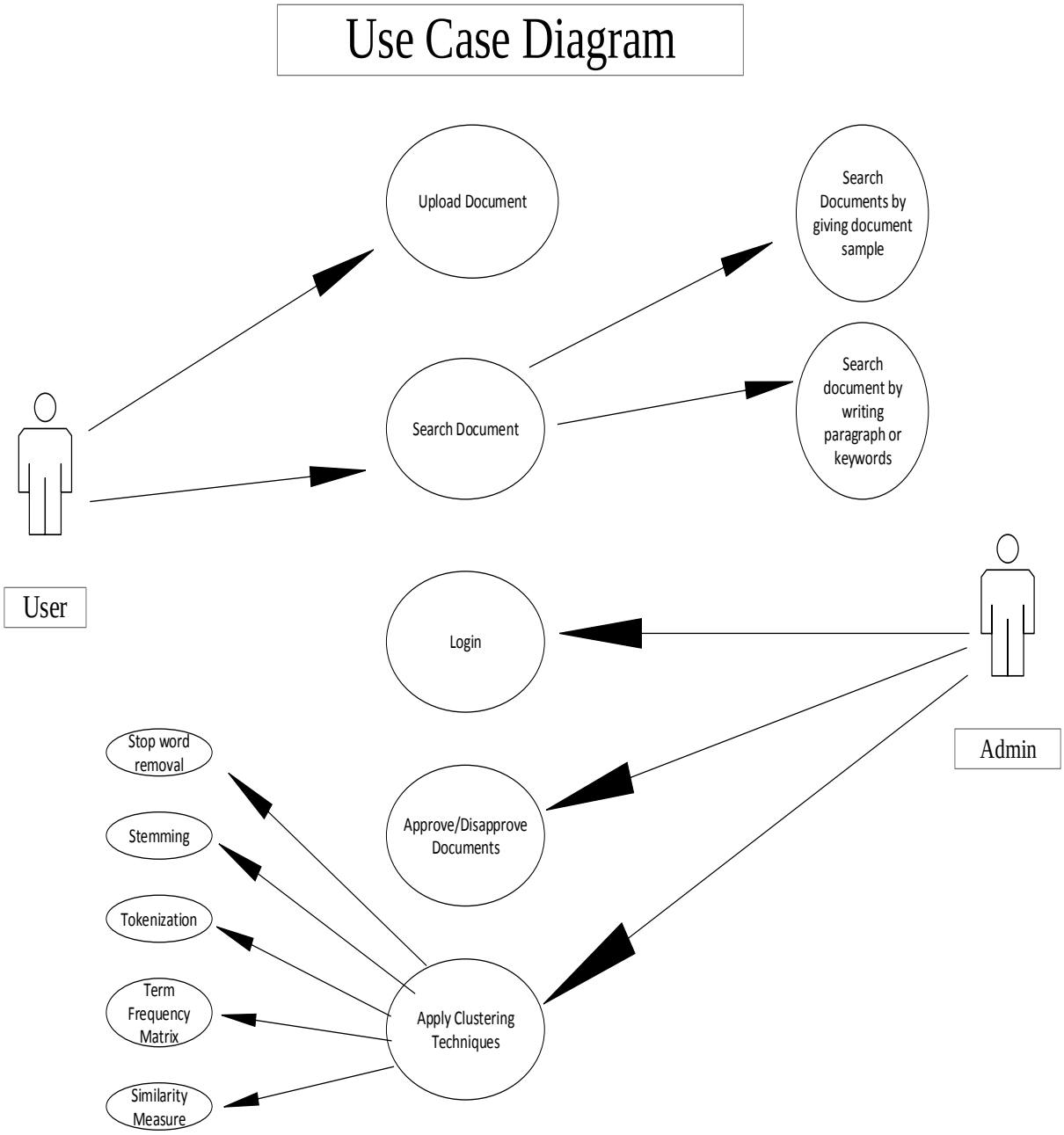


Figure 2.1 Use Case Diagram.

2.8 Use Case Description

Following are some descriptions;

2.8.1 Upload document

User will upload document it will notify admin that a new document is uploaded.

2.8.2 Search documents

User can search documents with two type search by query and search without query

- ✓ Search by query user will give a sample of document or user can write a paragraph and system will retrieve all the documents based on similarity measure
- ✓ Search without query in which user will write some keywords and the system will retrieve documents on the basic of similarity measure

2.8.3 Approve / disapprove documents

When user upload the document it will notify the admin if admin will approve the document it can be further process or if he disapproved the document it will not process further.

2.8.4 Apply clustering techniques

After approval the document is process for further pre-processing clustering steps like stop words removal, stemming and tokenization. After pre-processing activities completed then document will further step up to the frequency matrix and on the basic of frequency matric it will find the similarity between the documents and clustering begins.

2.9. Use case Description (Detailed)

2.9.1 Upload document

Use case ID	1
Use case Name	Upload Document
Actors: User	
Preconditions: document must be .doc or .docx format	
Basic Flow	User will upload the document for clustering

Actor Actions		**System Response**
1.	User will upload word document.	Successfully registered.
Alternative Course of Actions		
1a	The document is not in format of .doc or .docx	
Post Conditions		
Document will be uploaded successfully and it will notify admin for approval		

Table 2.2 Upload Document

2.9.2 Search Document

Use case ID	2	
Use case Name	Search Documents	
Actors: User		
Preconditions: must upload word document, write paragraph or provide some key words.		
Basic Flow	System will retrieve all the documents on the basic of similarity measure after providing sample	
Actor Actions		**System Response**

1.	The user will provide a document paragraph or keywords	The item is uploaded.
Alternative Course of Actions		
1a	Document is not upload must be in .doc or .docx format	
Post Conditions		
The user will have to provide a sample before search		

Table 2.3 Search Documents

2.9.3 Approve/Disapprove Documents

Use case ID	3	
Use case Name	Approve/Disapprove Document	
Actors: Admin		
Preconditions: must be sign in.		
Basic Flow	After signing in admin will approve documents for clustering or disapprove the documents	
Actor Actions		**System Response**
1.	Admin will check the document for approval	Wheatear document is valid or not
Alternative Course of Actions		
1a		

Post Conditions
Document will further process for clustering activities

Table 2.4 Approve/Disapprove Document

2.9.4 Applying Cluster technique

Use case ID		4
Use case Name		Applying Clustering technique
Actors: Admin		
Preconditions: must be sign in and approve the document		
Basic Flow		Admin will apply technique and the system will clustering the document in his relative cluster.
Actor Actions		**System Response**
1.	Admin will apply text preprocessing activities, similarity measure and then save the document in cluster	Document is related to xyz category.
Alternative Course of Actions		
1a		
Post Conditions		
Admin should be apply all the text preprocessing activities		

Table 2.5 Applying cluster technique

CHAPTER 3

Sequence Diagrams & Class Diagram

3. Diagrams

3.1 Document Uploaded Successfully

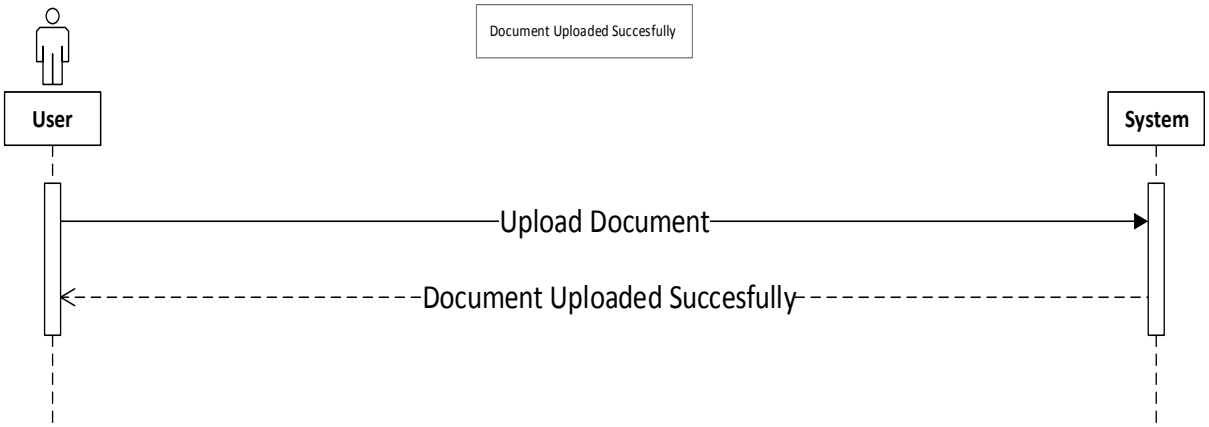


Figure 3.1 Document uploaded successfully

3.2 Document upload unsuccessfully

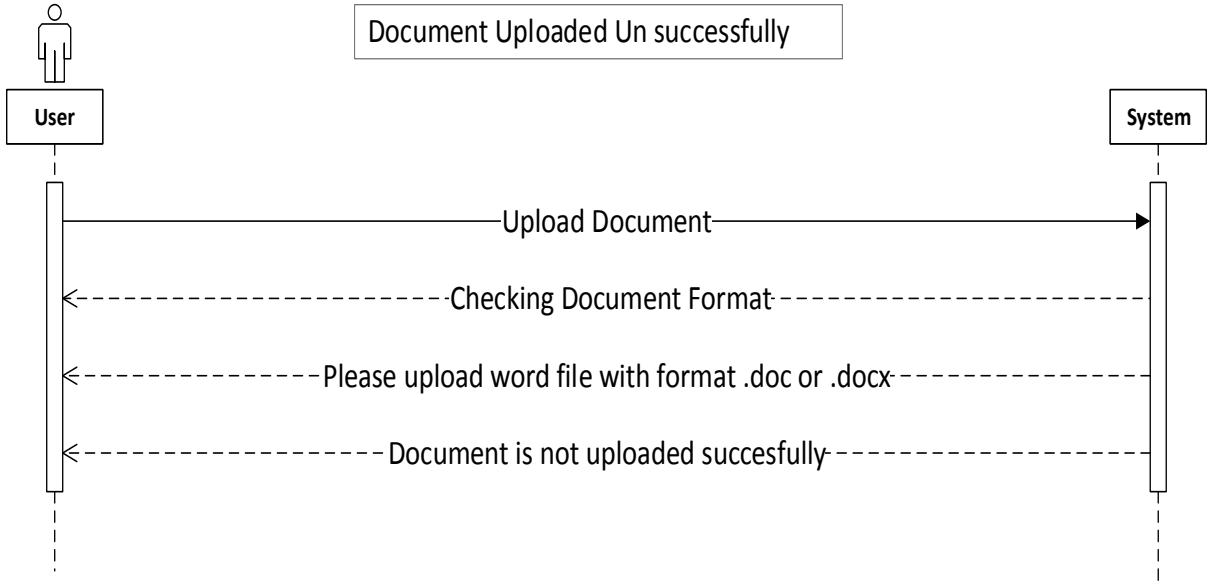


Figure 3.2 Document uploaded unsuccessfully

3.3 Approval of Documents

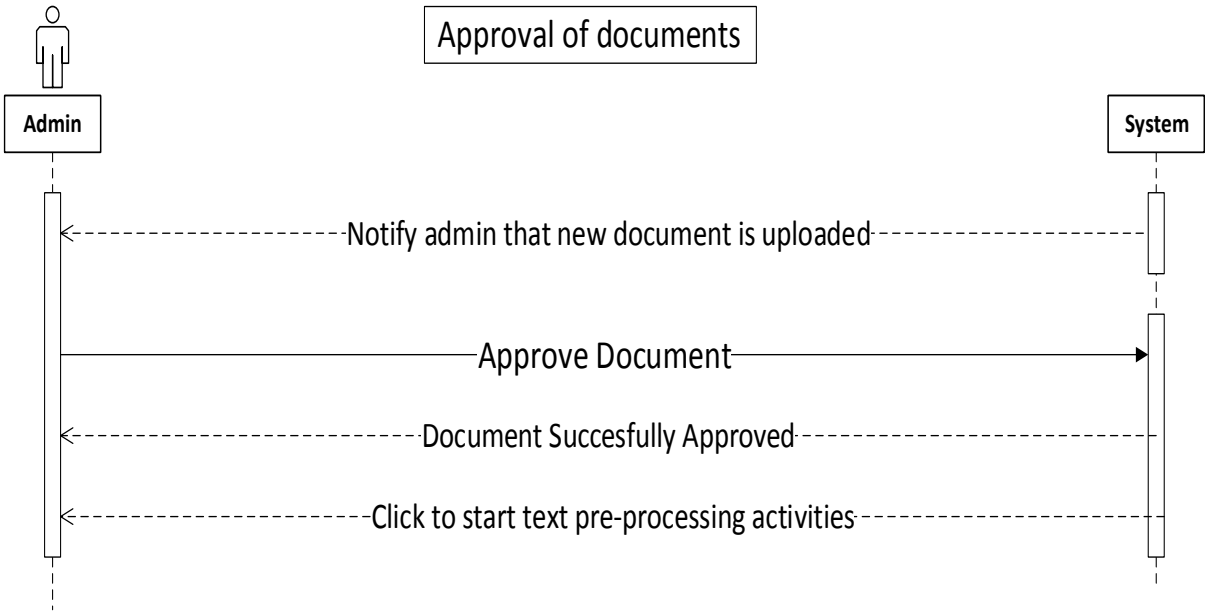


Figure 3.3 Approval of document

3.4 Disapproval of Documents

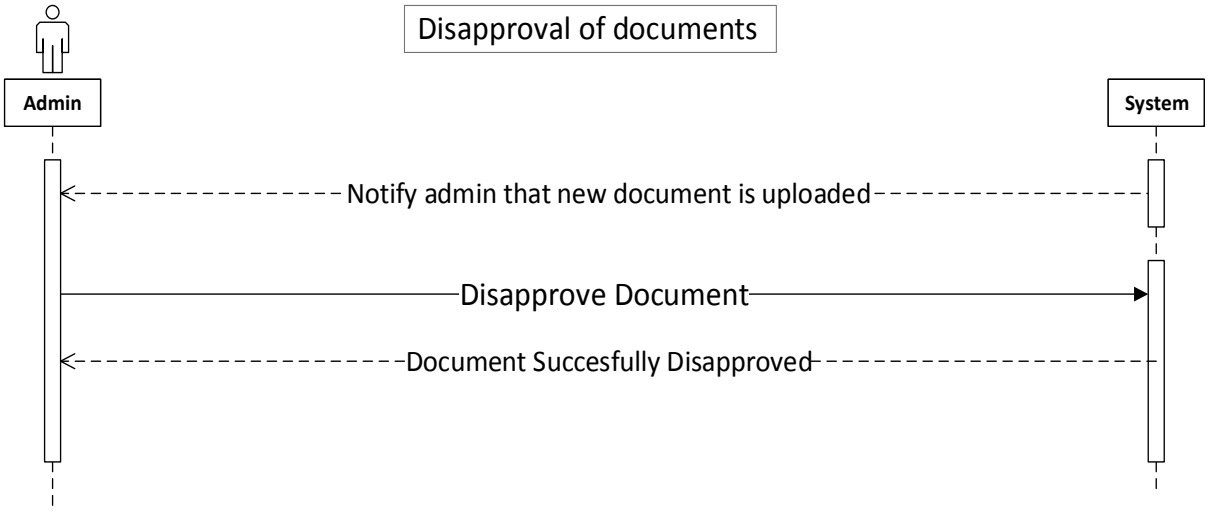


Figure 3.4 Disapproval of Document

3.5 Applying Cluster Technique

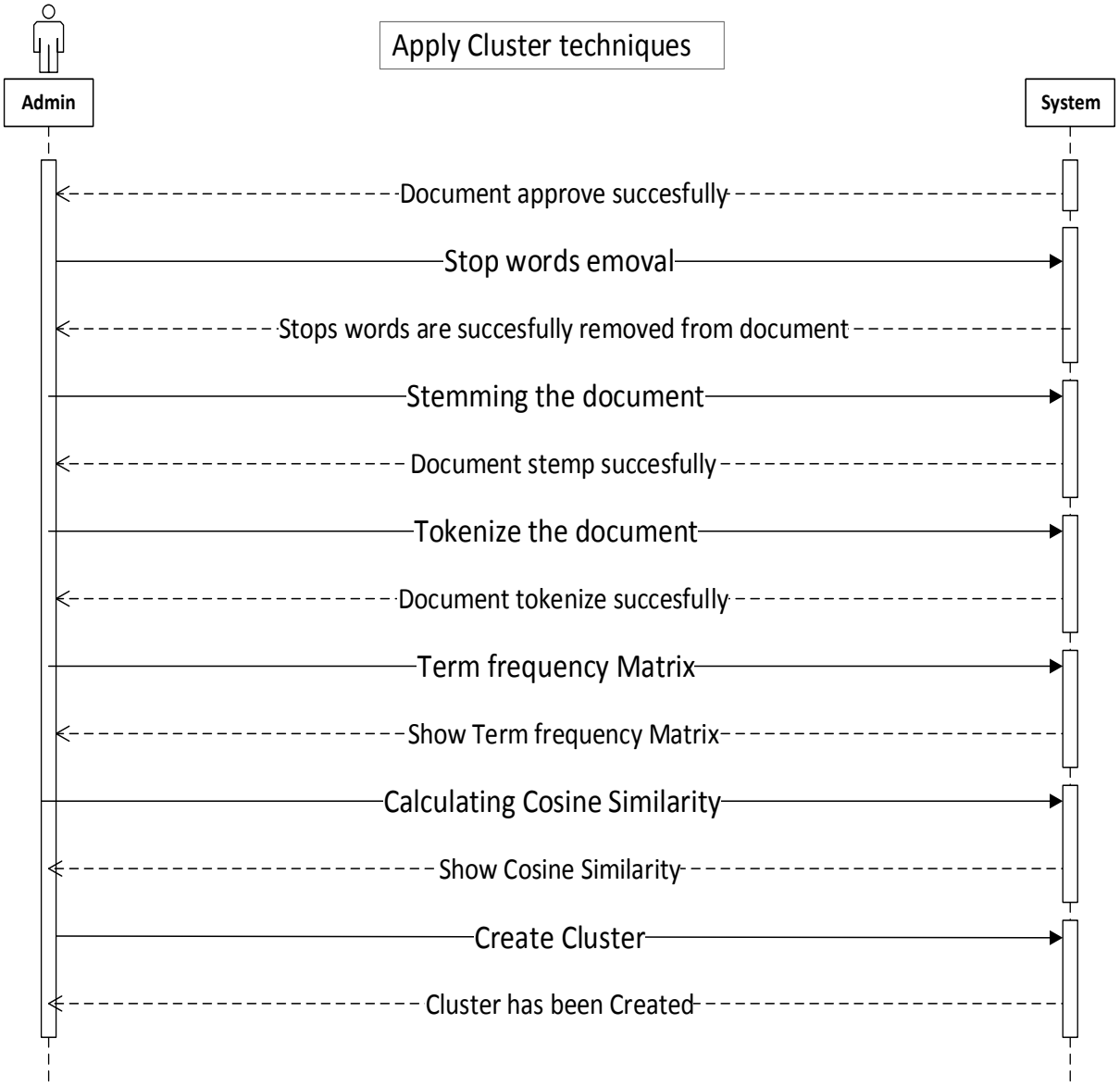


Figure 3.5 Apply cluster technique

3.6 Search Document By Example

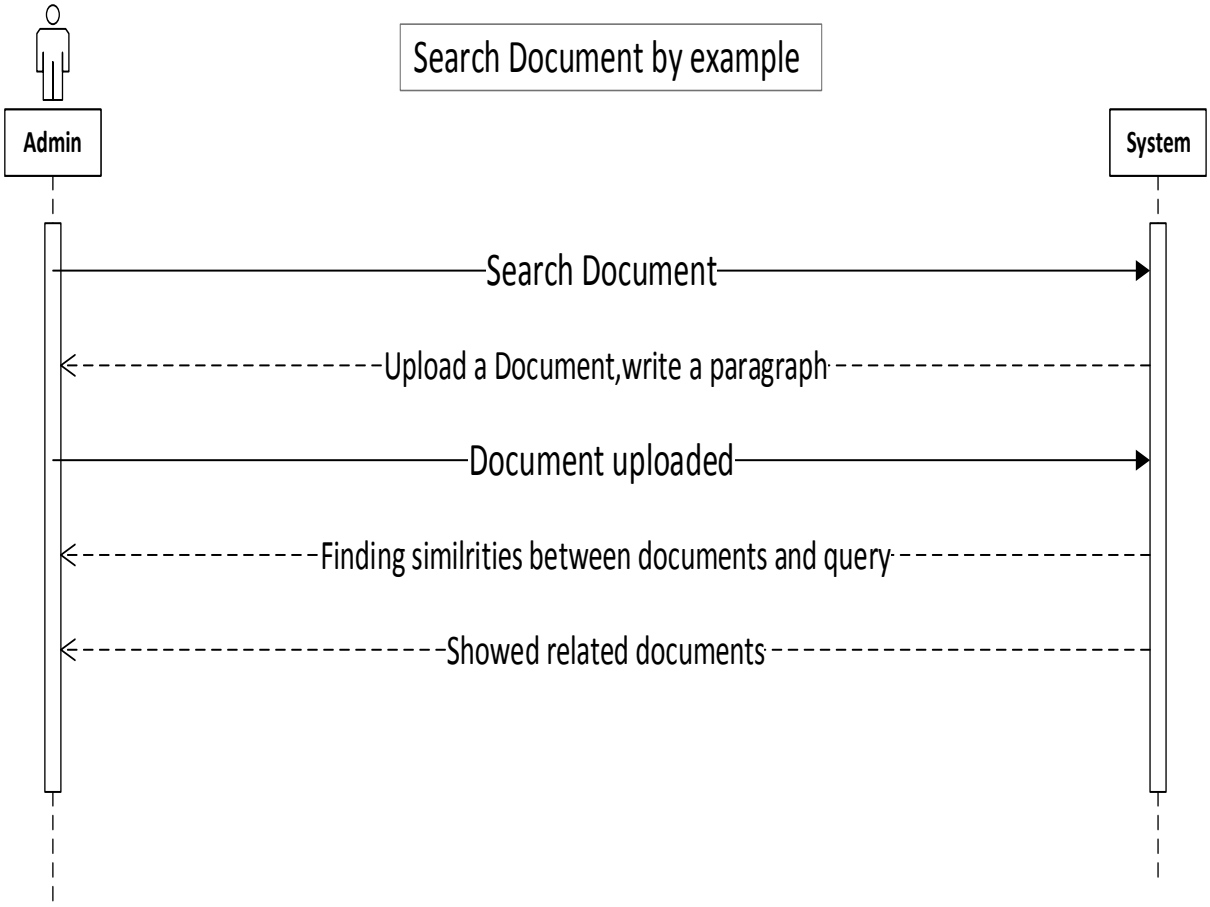


Figure 3.6 Search document by example

3.7 Search Document Without Example

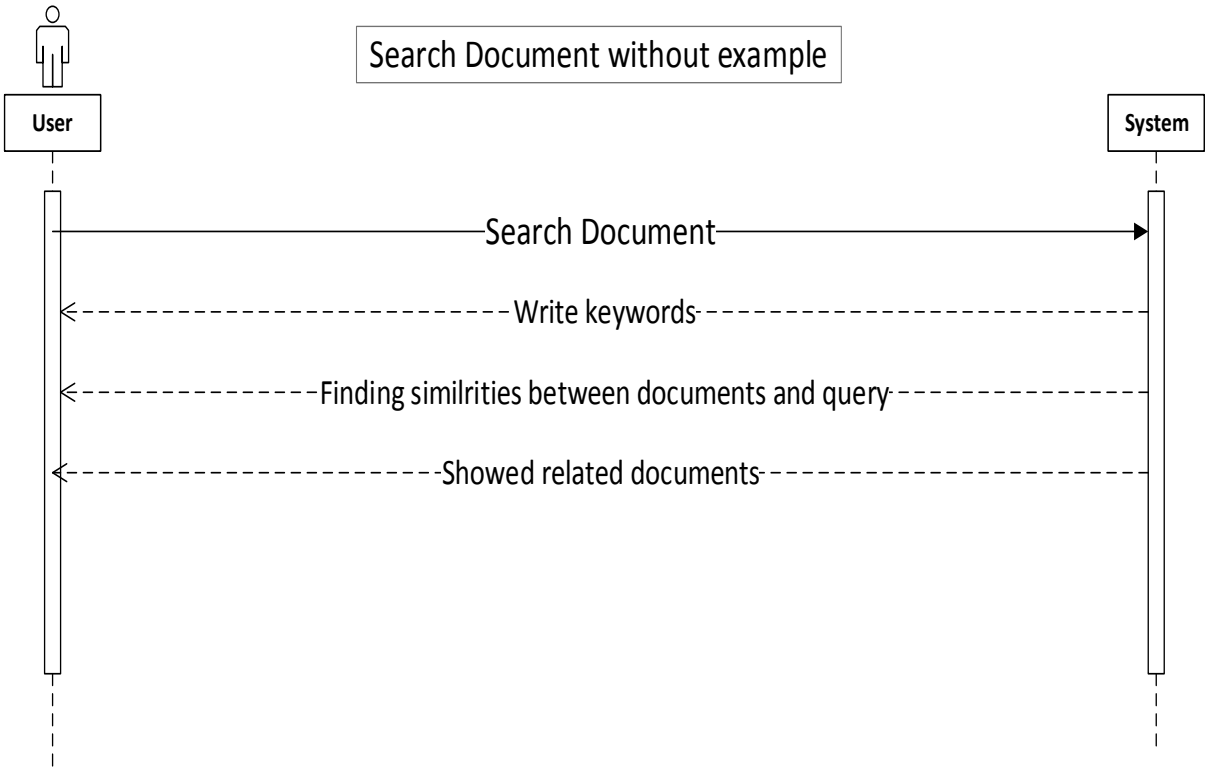


Figure 3.7 Search document by example

3.8 Class Diagram

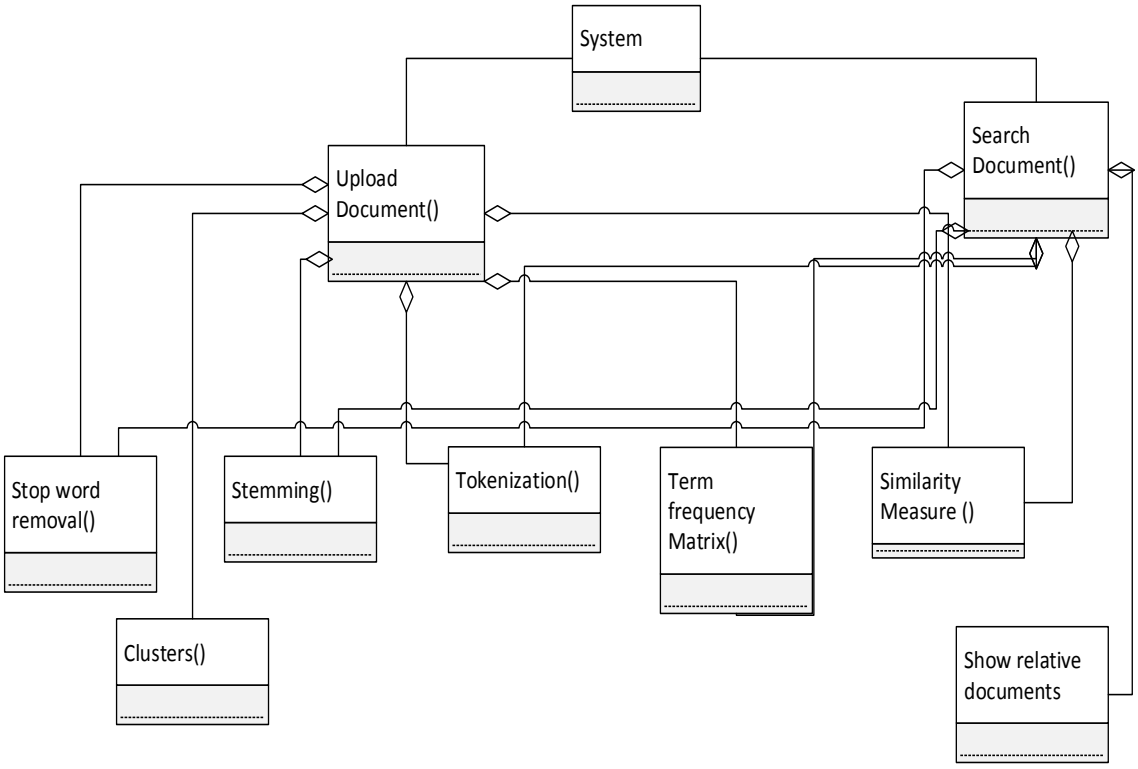


Figure 3.8 Class Diagram

CHAPTER 4

SYSTEM DESIGN

4. System Design

Systems design is the process of defining the architecture, components, modules, interfaces, and data for a system to satisfy specified requirements. Systems design could be seen as the application of systems theory to product development.

4.1 Activity Diagram

Are graphical representations of workflows of stepwise activities and actions with support for choice, iteration and concurrency, In the Unified Modeling Language, activity diagrams are intended to model both computational and organizational processes (i.e. workflows).

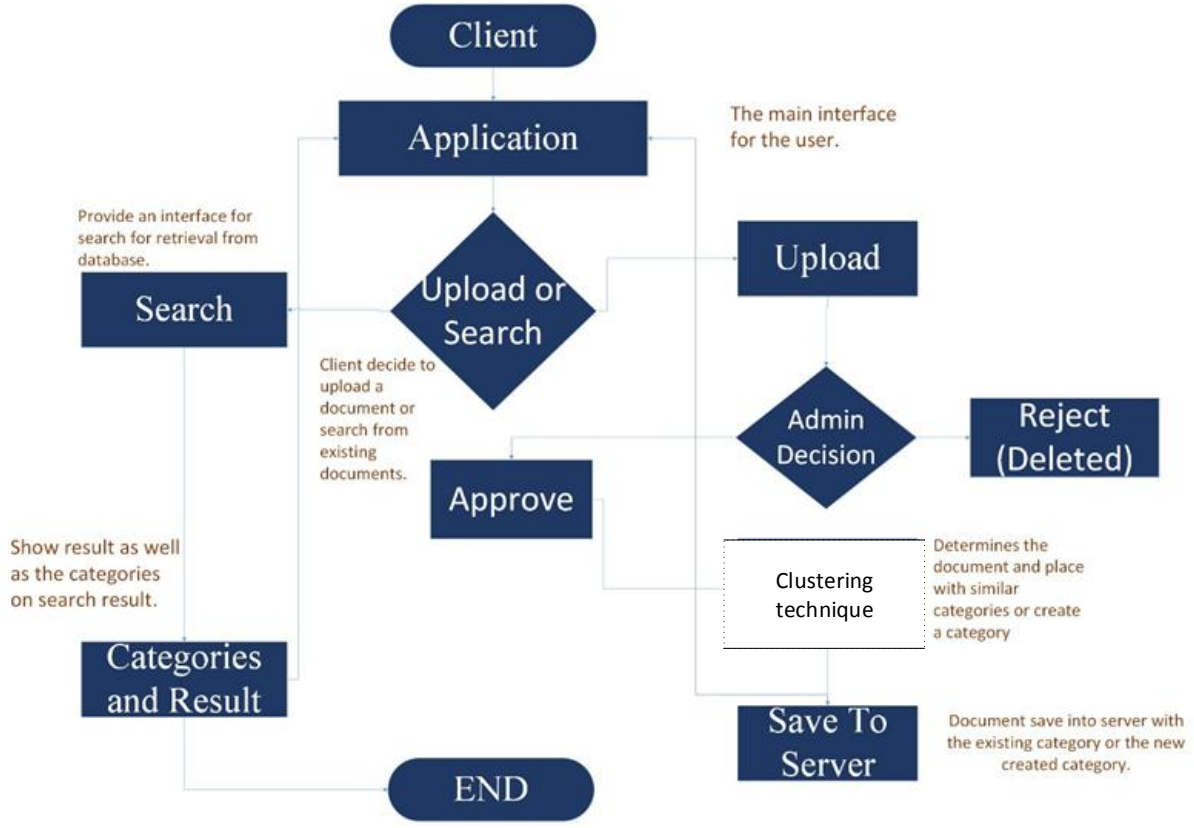


Figure 4.1 Activity Diagram

CHAPTER 5

IMPLEMENTATION

5. Implementation

Implementation is the carrying out, execution, or practice of a plan, a method, or any design, idea, model, specification, standard or policy for doing something. As such, implementation is the action that must follow any preliminary thinking in order for something to actually happen.

5.1 Tools and Technologies

Eclipse

Data base (My Sql)

Server (Xampp)

CHAPTER 6

SYSTEM TESTING

6. System Testing

System testing is defined as "testing the behavior of a system as per software requirement and specifications".

Testing is the fully integrated applications including external peripherals in order to check how components interact with one another. System testing enables us to test, Verify and validate both the business requirements as well as the applications architecture.

The application is tested thoroughly to verify that it meets the technical and functional specifications.

6.1 Why System Testing is Required

It is the first level of software testing where the application is tested as a whole. It is done to verify, validate the functional, business, technical requirements of the software.

It also includes the verification and validation of software application architecture.

6.1.1 Black Box Testing

Black-box testing is a method of software testing that examines the functionality of an application without peering into its internal structures or workings.

6.1.2 White Box Testing

A software testing technique whereby explicit knowledge of internal workings of item being tested are used to select the test data.

6.2 Test Cases

6.2.1 Verify, that user upload a valid format of a document

Test Case Id:	001	Test Engineer:	Majid Liaquit
Test Date:	10-Aug-2016	Test Case version:	1.0
Reviewed by:	Testing Team Leader	Use Case reference(s):	N/A
Objective:	**Verify that the user upload a valid format of a document**		

Product/Version/Module Environment:	The user currently using the app on browser	
Pre-Requisite:	the user must be visited to clustering app	
Step No.	Execution Description The user must be uploaded a valid format of document.	procedure Result The document uploaded successfully.
Comments:	The test passed successfully	
▪ Passed	Failed	Not Executed

Table 6.1 Verify that the user uploaded valid format of document

6.2.2 Verify, that admin must be login before applying clustering on documents

Test Case Id:	002	Test Engineer:	Majid Liaquit
Test Date:	10-Aug-2016	Test Case version:	1.0
Reviewed by:	Testing Team Leader	Use Case reference(s):	N/A
Objective:	**Verify that admin must be login before applying clustering on documents**		

Product/Version/Module Environment:	The admin currently using the app on browser	
Pre-Requisite:	The admin must be sign up.	
Step No.	Execution Description The admin can apply clustering on document	procedure Result Admin sign in successfully
Comments:	The test passed successfully	
▪ Passed	Failed	Not Executed

Table 6.2 verify that admin is login before applying clustering on document

6.2.3 Verify, that document is store in his relevant cluster

Test Case Id: 003	Test Engineer: Majid Liaquat	
Test Date: 10-Aug-2016	Test Case version: 1.0	
Reviewed by: Testing Team Leader	Use Case reference(s): N/A	
Objective:	**Verify that document is store in relevant cluster.**	
Product/Version/Module Environment:	The admin currently using the app on browser	
Pre-Requisite:	The admin must be sign in.	
Step No.	Execution Description The admin will approve the	procedure Result The document is store in his

	document and apply all the clustering techniques to store a document in his relevant cluster	relevant cluster
Comments:	The test passed successfully	
▪ Passed	Failed	Not Executed

Table 6.3 Verify, that document is store in his relevant cluster

6.2.4 Verify, that user search the documents by giving query

Test Case Id: 004	Test Engineer: Majid Liaquit	
Test Date: 10-Aug-2016	Test Case version: 1.0	
Reviewed by: Testing Team Leader	Use Case reference(s): N/A	
Objective:	**Verify, that user search the documents by giving query**	
Product/Version/Module Environment:	The user currently using the app on browser on pc.	
Pre-Requisite:	The user must have document with valid format or query to search the document	
Step No.	Execution Description	procedure Result
	The user provide the document and system will retrieve all the documents regarding to the query	System will retrieve all the documents which is relevant to the system.
Comments:	The test passed successfully	

▪ Passed	Failed	Not Executed

Table 6.4 Verify, that user search the documents by giving query

CHAPTER 7

CONCLUSION

7. Conclusion

7.1 Good Features of the System

- User Friendly.
- Fast Searching.
- Documents sharing
- Document reading
- Stored documents in cluster
- Storing of documents is based on document clustering.
- Retrieval of documents is based on document clustering
- Admin will apply all the clustering techniques to the documents after approval

7.2 Limitations of the system

- Compatible with windows operating system.

7.3 Future Enhancement

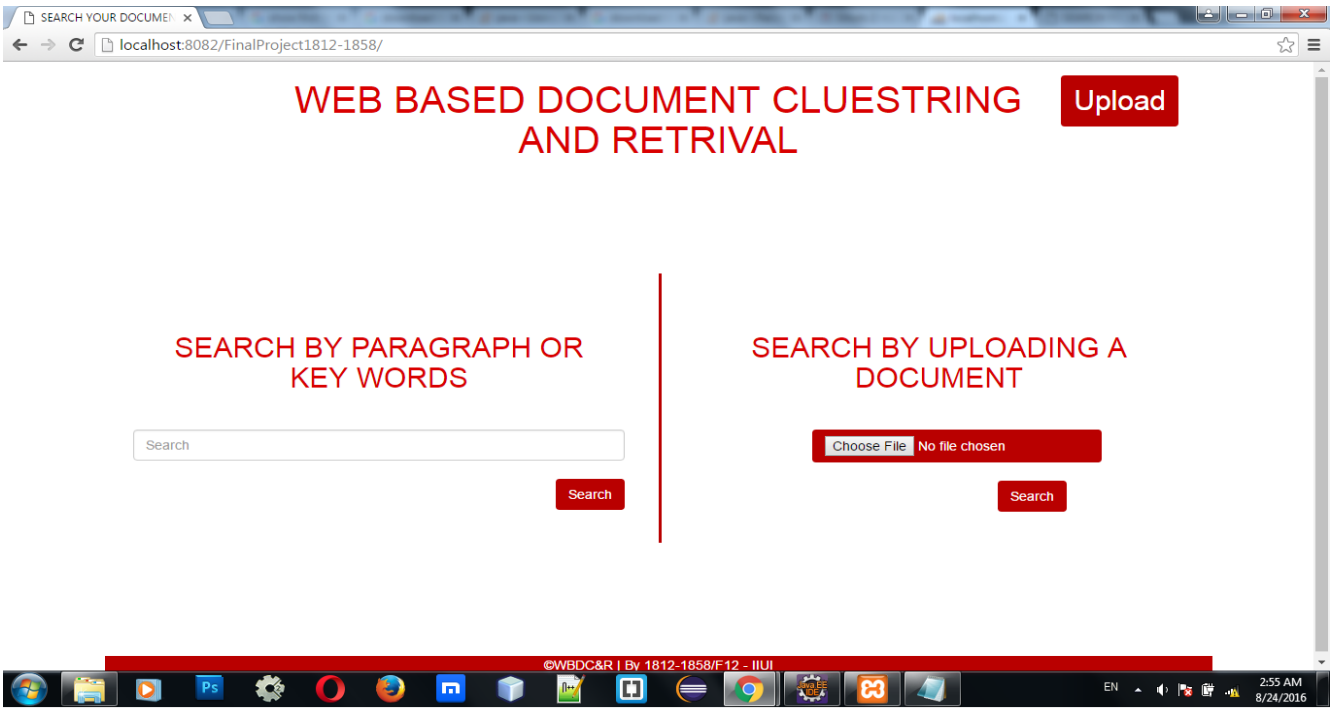
- Make an android application for android mobiles.

CHAPTER 8

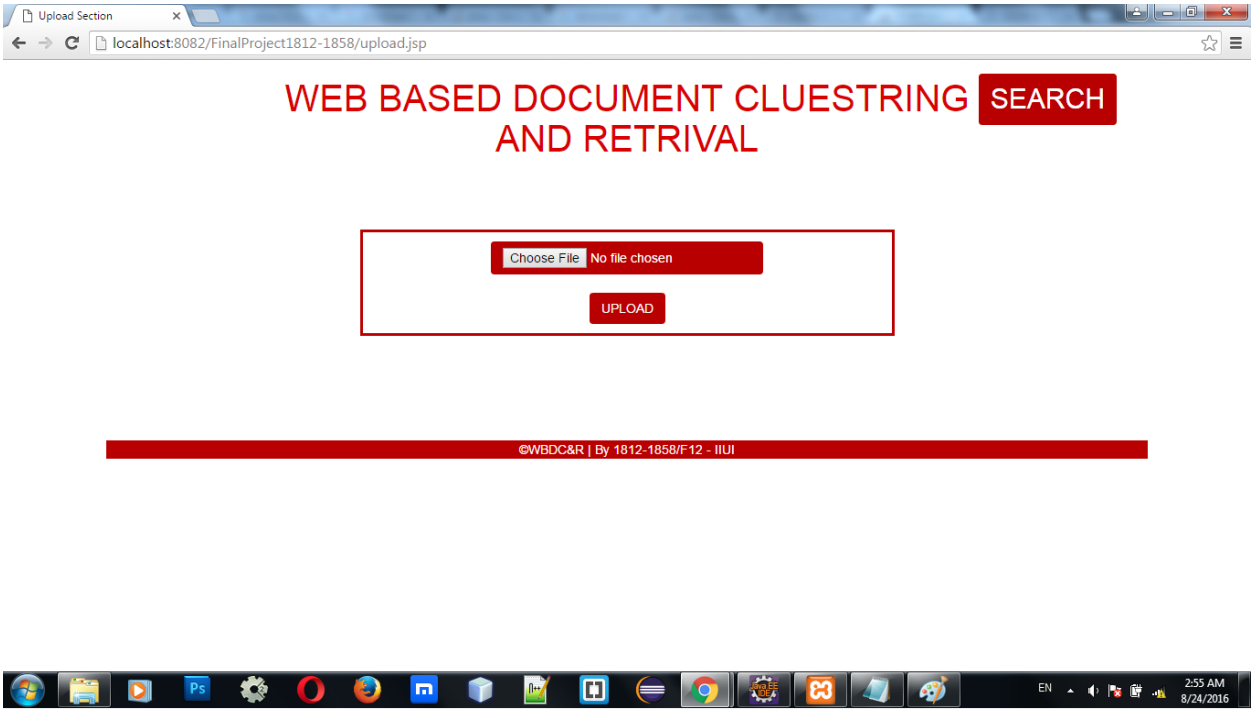
USER MANUAL

8. User Manual

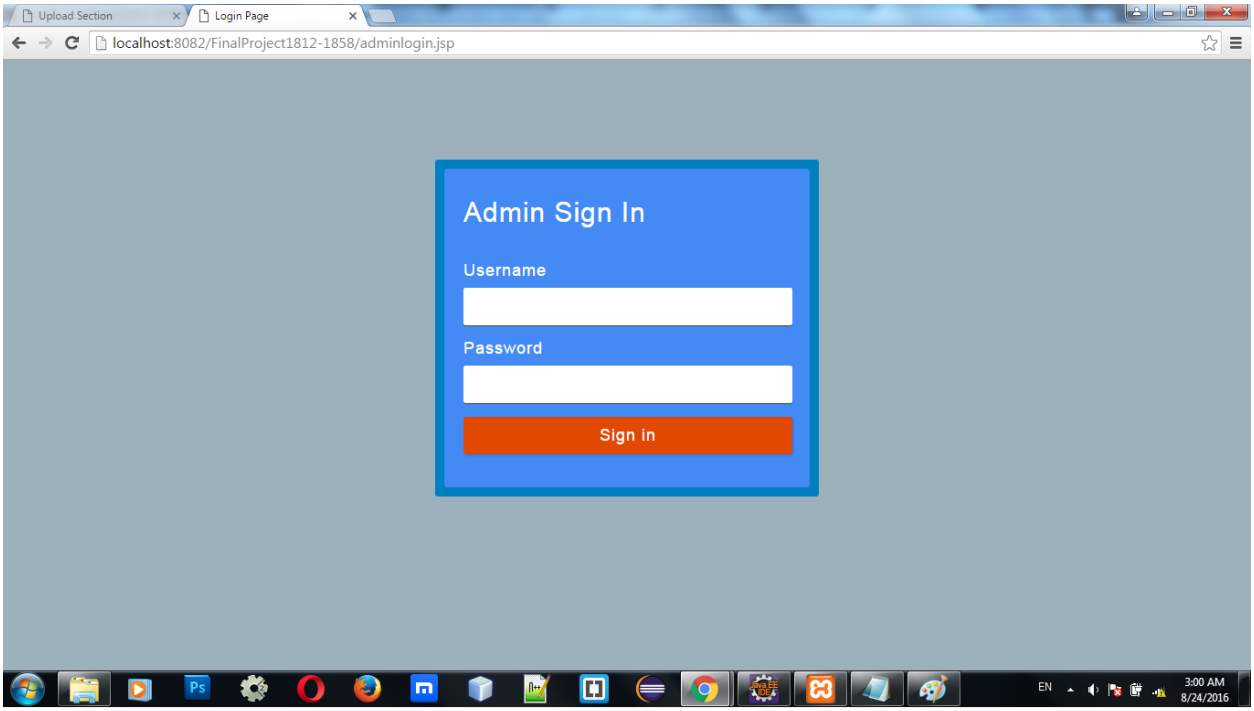
8.1. Main Page (Search document by query & upload for clustering)



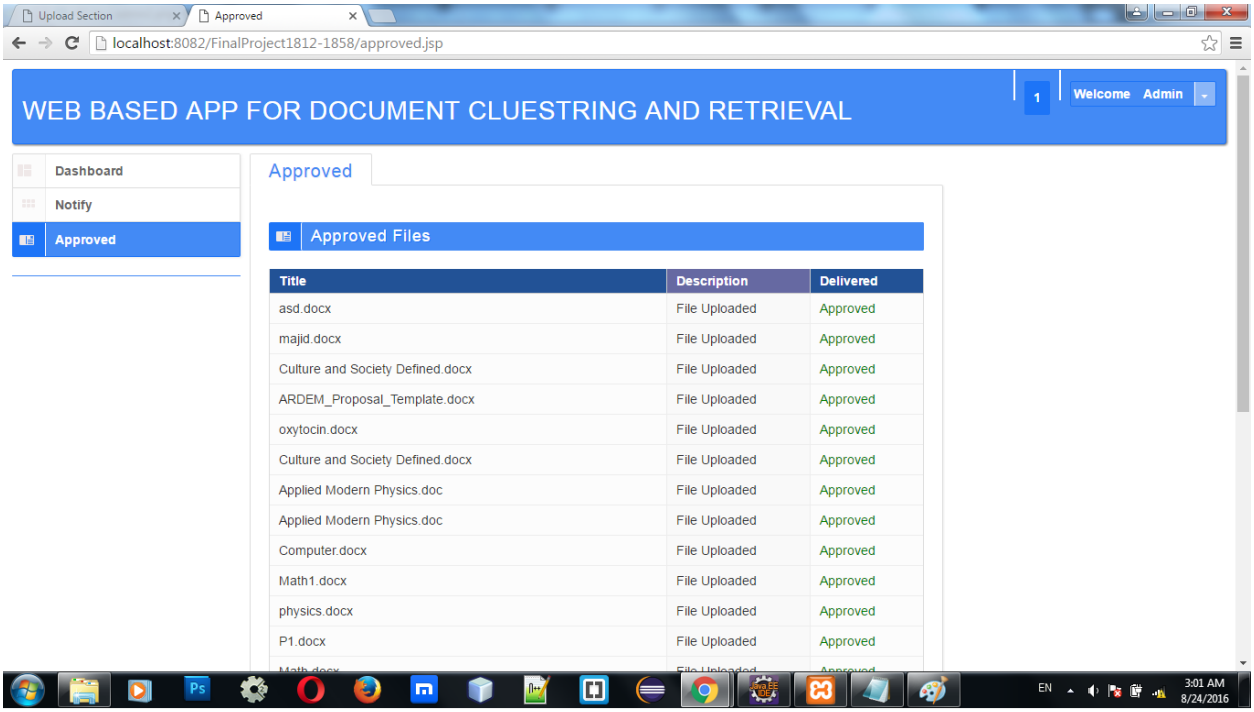
8.2 Upload Document



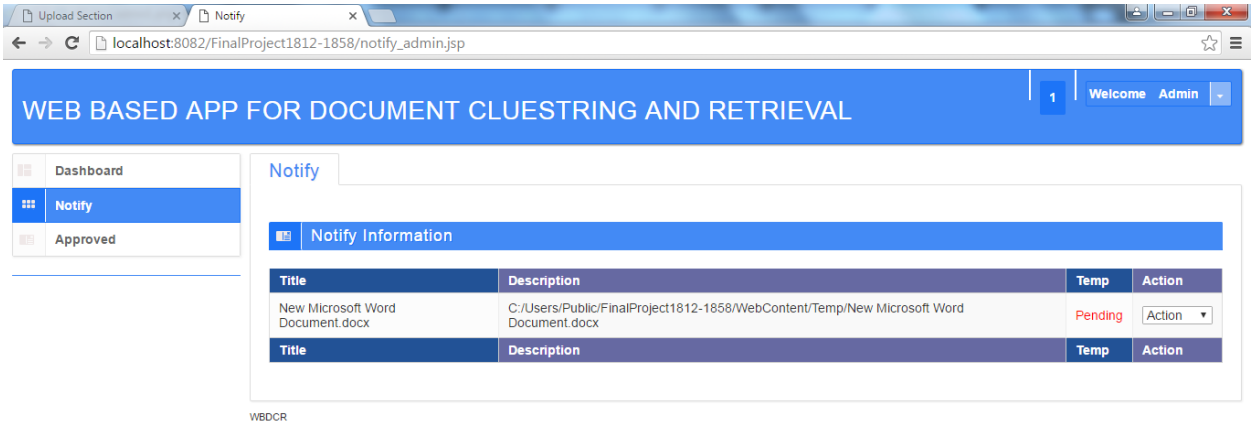
8.3 Admin Login



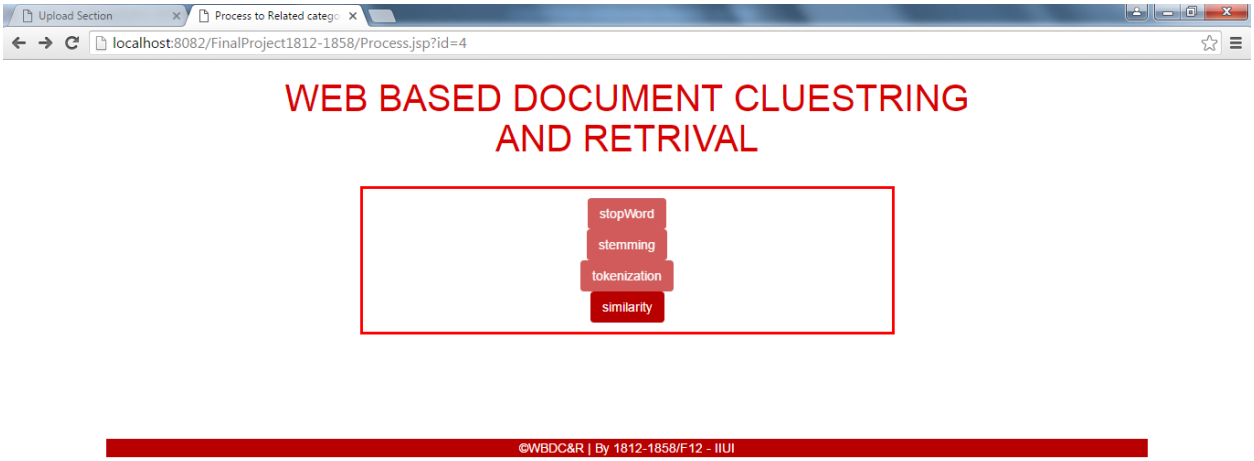
8.4 Admin Notification



8.5 Document Approval



8.6 Applying Text Clustering



References

[1] Fabrizio Sebastiani “Machine Learning in Automated Text Categorization”.

[2] Zdravko Markov. MDL-based Unsupervised Attribute Ranking.

[3] Peter Náther N-gram based Text Categorization.

[4] N-Gram-Based Text Categorization William B. Cavnar and John M. Trenkl.

[5] Cavnar, William B. and Vayda, Alan J., “Using superimposed coding of N-gram.