

**TRIBHUVAN UNIVERSITY**

**Institute of Science and Technology**

**Report On**

**“ R-Archive: Recommendation system for archived digital documents ”**

In Practical Fulfillments of Requirement of Bachelor of Science in Computer Science and Information Technology(BSc. CSIT)

**Submitted By:**

Nisha Kushum Bhusal (24152/076)

Rupesh Nepal (24163/076)

Yubaraj Karki (24181/076)

**Submitted To:**

Department of Computer Science and Information Technology

Madan Bhandari Memorial College

Binayak Nagar, New Baneshwor, Kathmandu, Nepal

# SUPERVISOR’S RECOMMENDATION

I hereby recommend that this project prepared under my supervision by “Nisha Kushum Bhusal , Rupesh Nepal and Yubaraj Karki” entitled “ **R-Archive: Recommendation system for archived digital documents** ” in partial fulfillment of the requirements for the degree of Bachelor of Science, Computer Science and Information Technology is recommended for the final evaluation.

…………………….

Mr. Parbin Maharjan

**Project Supervisor**

School of Science and Technology

Madan Bhandari Memorial College

# LETTER OF APPROVAL

This is to certify the project prepared by “ Nisha Kushum Bhusal, Rupesh Nepal and Yubaraj Karki” entitled **“ R-Archive: Recommendation system for archived digital documents ”**in partial fulfillment of the requirements for the degree of Bachelor of Science, Computer Science and Information Technology has been evaluated. In our opinion it is satisfactory in the scope and quality as a project for the required degree.

|  |  |
| --- | --- |
| ………………………………………………  Mr. Phul Babu Jha  Coordinator & Supervisor | …………………………………………………  Campus Chief  Madan Bhandari Memorial College |
| …………………………………………  Mr. Parbin Maharjan  Supervisor | …………………………………………………  External  Examiner  Tribhuvan University |

# Acknowledgement:

We are very grateful to the department of Computer Science and Information Technology, Madan Bhandari Memorial College for providing us an opportunity to work on a major project as part of our fourth-year project. We would like to express our deepest appreciation to all those who provided us with the possibility to complete this report. We express our deep gratitude and regards to our project supervisor, **Mr. Prabin Maharjan** monitoring us, providing constant encouragement throughout the completion of this project and guiding towards right direction. We are also thankful to our coordinator **Mr. Phul Babu Jha** for providing us with a intellectual expertise and providing continuously provide help and support.

We would like to also appreciate all the teachers and the correspondence faculty members, seniors, family and friends for their involvement in our project directly and indirectly for their encouragement support and motivation for helping us achieve our goal .

# Abstract:

**“R-Archive”** is a web application designed to maintain to store digital documents form users to archive them and make accessible to any users . Not only archiving the digital documents

Our project mainly aims towards the personalized learning experience and enhancing the user engagement with archived digital library with the tailored resource content-based filtering .

By operating as a central repository, this system gives users easy access the digital interactive educational resources. Our System enables a dynamic and user-friendly experience, allowing users to engage with the vast library of archived documents. Any users are able to upload any informative documents they want to contribute towards the library. Users are not only contributing but are also able to rate the documents as well as users are able to find their resources on the basis of their preferences .While the benefits are good there are some common challenges in implementation and utilization exist. Issues such as user adoption, technical constraints and copyrights .

The main aim of this system is to digitalize and archive the documents for the users to use for free and personalize their learnings.

**Keywords: R-Archive , archive ,personalized learning , content-based filtering**

# Table of Content:

[SUPERVISOR’S RECOMMENDATION i](#_Toc164438926)

[LETTER OF APPROVAL ii](#_Toc164438927)

[Acknowledgement: iii](#_Toc164438928)

[Abstract: iv](#_Toc164438929)

[Table of Content: v](#_Toc164438930)

[List of Figures: vii](#_Toc164438931)

[List of Tables viii](#_Toc164438932)

[List of Abbreviation: ix](#_Toc164438933)

[Chapter 1 : Introduction: 1](#_Toc164438934)

[1.1 Introduction: 1](#_Toc164438935)

[1.2 Problem Statement: 2](#_Toc164438936)

[1.3 Objectives: 2](#_Toc164438937)

[1.4 Scope and Limitations : 3](#_Toc164438938)

[1.4.1 Scope: 3](#_Toc164438939)

[1.4.2 Limitations: 3](#_Toc164438940)

[1.5 Development Methodology: 3](#_Toc164438941)

[1.6 Report Organization: 5](#_Toc164438942)

[Chapter 2 : Literature Review: 6](#_Toc164438943)

[Chapter 3 : System Analysis: 8](#_Toc164438944)

[3.1 Requirement Analysis : 8](#_Toc164438945)

[3.1.1 Functional Requirement: 8](#_Toc164438946)

[3.1.2 Non-Functional Requirement 10](#_Toc164438947)

[3.2 Feasibility Analysis: 11](#_Toc164438948)

[3.3 Analysis 12](#_Toc164438949)

[3.3.1 Object Modeling using Class Diagram: 12](#_Toc164438950)

[3.3.2 Dynamic modelling using Sequence Diagram : 16](#_Toc164438951)

[3.3.3 Process modelling using Activity Diagrams 17](#_Toc164438952)

[Chapter 4 : System Design: 18](#_Toc164438953)

[4.1 Designs 18](#_Toc164438954)

[4.1.1. System Architecture 18](#_Toc164438955)

[4.1.2. Component Diagram 19](#_Toc164438956)

[4.1.3. Model Processing 20](#_Toc164438957)

[4.2 Algorithm Details: 21](#_Toc164438958)

[4.2.1 K-Mean Clustering: 21](#_Toc164438959)

[Chapter 5 : Implementations and Testing: 22](#_Toc164438960)

[5.1 Implementations: 22](#_Toc164438961)

[5.1.1 Tools used : 22](#_Toc164438962)

[5.1.2 Implementation details of modules: 23](#_Toc164438963)

[5.2 Testing : 25](#_Toc164438964)

[5.2.1 Test cases for unit testing: 25](#_Toc164438965)

[5.2.2 Test cases for system testing : 28](#_Toc164438966)

[5.3 Result Analysis 29](#_Toc164438967)

[Chapter 6 : Conclusion and Future Recommendations 30](#_Toc164438968)

[6.1 Conclusion 30](#_Toc164438969)

[6.2 Future Recommendations: 30](#_Toc164438970)

[Reference: 32](#_Toc164438971)

# List of Figures:

Figure 1. 1 Incremental Model 4

[Figure 3. 1 Use Case Diagram 9](#_Toc164436902)

[Figure 3. 2 Gantt Chart 12](#_Toc164436903)

[Figure 3. 3 Class Diagram 15](#_Toc164436904)

[Figure 3. 4 Sequence Diagram 16](#_Toc164436905)

[Figure 3. 5 Activity Diagram 17](#_Toc164436906)

[Figure 4. 1 System Architecture 18](#_Toc164436909)

[Figure 4. 2 Component Diagram 19](#_Toc164436910)

[Figure 4. 3 Model Processing 20](#_Toc164436911)

# List of Tables

[Table 5. 1 Unit Test Case for Registrations Module 25](#_Toc164438758)

[Table 5. 2 Unit Testing for Login Module 26](#_Toc164438759)

[Table 5.3 Unit Testing for the Uploading the documents 26](#_Toc164438760)

[Table 5.4 Unit Testing for Rating the Documents 27](#_Toc164438761)

[Table 5.5 Unit Testing for the Searching 27](#_Toc164438762)

[Table 5.6 Unit Test Case for the recommendations 28](#_Toc164438763)

[Table 5. 7 System testing test case for Usability 28](#_Toc164438764)

# List of Abbreviation:

**CBF** : **C**ontent-**B**ased **F**iltering

**DL : D**igital **L**ibrary

**IDE : I**ntegrated **D**evelopment **E**nvironment

# Chapter 1 : Introduction:

## Introduction:

Exponential growth of digital information and the availability of technologies has increased the need for efficient and accessible means of managing and accessing knowledge and information’s has become norms for the modern society. The concept of digital libraries has become a is solutions offering the users the ability to store, organize and access digital information virtually.

Digital Library is a software-based structure in which the focal point is web expertise. In web, electronic form of data can be accumulated and regained easily by providing rights to privilege the primary data. The primary set of data is focused on digital depiction just like in the form of editorials, published study material, theses, document, e-books and so on.[1].Not only storage of information the scope has been extended to exchanging information, social networking, strong exchanging information between scholars, researchers, developers and data analysts etc.[1]

The development of the R-Archive plays an important role in leveraging the scope of the digital library by introduction of the personalized content-based recommendation system . Not only management of the digital documents but also the ability to make sure that the users are engaged in the learning activities. Users are able to easily access the stored documents as well as upload the document they have to the library to share among themselves

## Problem Statement:

In comparison to the general approach there is central repository stations of the physical documents . The main problem with the libraries is the physical nature of it where the published documents are stored physically and accessible only through physical means.

There exist several digital libraries and digital platform from which the resources can be accessed by digital means but the lack of efficient and personalized platforms for organizing and accessing the resources creates a barrier in learning in existing libraries.

Not everyone has the access and privilege to access the document project aim to create a central repository for effective accessible learning and creating a manageable environment for the archived documents.

## Objectives:

The project aims to provide and enhance the learning opportunity and experience to the users form anywhere however the major objectives of the projects are :

* To personalize the learning experience of the user and increase the engagement of the user to platform by providing recommendation based on the content filtering
* To maintain comprehensive digital library so that the resources is digitally available to the users

## Scope and Limitations :

### Scope:

The system should be able to perform various task and provide and must include various functionalities. The applications can be run through web browser of both desktop and mobile applications . Since the documents are to be easily available to the users when requested if available . The system should be able to recommend the users by analyzing the user activities based on the content they rated . Based the user liked content , resources are clustered into group and labeled and ordered on the basis of the rating and occurrence then the user classified on which cluster label it belongs to. The User interactions and experience with the project should be seamless

### Limitations:

Though the project performs intended task and there are some existing limitations in the project

* The K-mean clustering is not the great approach as the data starts to grow it is best suitable in present due to cold start as we don’t have the required sufficient data
* As the data starts to the clustering size also should increase and not confined to the constructed model labels defined during the initiations
* Copyright is a great limitation to our project as the resources can mislead towards copyright infringement but can be reduced by close observations and report of the resources

## Development Methodology:

To develop the system ,Incremental model of Staged Delivery is taken where service and product are developed and delivered in each stages building or stacking upon another stages

At first project is divided into small, manageable parts, or increments in a sequential manner after the completion of each stage the increments is performed within the SDLC phases

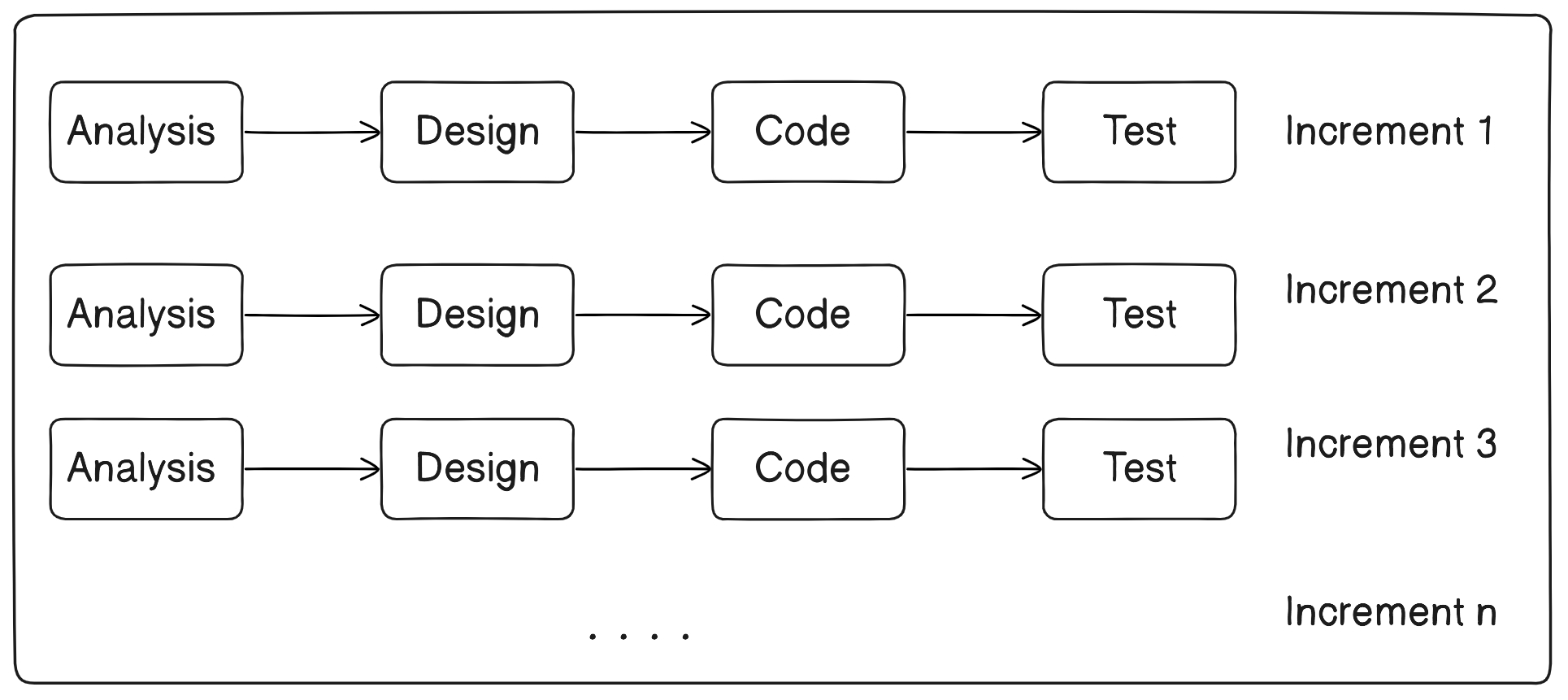


Figure 1. 1 Incremental Model

Incremental Method is consisting of mainly 4 stages

* **Requirement Analysis:** This phase involves analyzing the project's requirements, defining its scope, objectives, and constraints, and creating a project plan. Resources needed for the project are also identified during this phase.
* **Design:** In this phase, the project team develops a detailed plan for implementing the project, including system design, technology selection, and prototyping.
* **Coding and Implementation:** During this phase, the project's code is written according to product requirements. Adherence to coding standards is crucial to improve and update product quality, while physical execution of the designs occurs.
* **Testing:** All system components and additional functionality are tested, integrated, and checked as a whole in this phase. Testing ensures that the product works properly with other pieces of code and is ready for deployment to production.

## Report Organization:

This report is divided into 6 chapters . Each chapter is divided into different headings

The preliminary sections contain overall information about the project. This section includes abstracts, table of content , list of figures and abbreviations

**Chapter 1** provides an overview of the R-archive project introducing its significance and addressing the identified problem, objectives, scopes and limitation of the system.

**Chapter 2** contains literature review section which offers an overview of similar projects,

theories and findings by other researchers.

**Chapter 3** discusses in details about the design includes system analysis which includes requirement analysis, feasibility study and system configurations. It also includes modeling using class diagram, activity diagram and Sequence diagram

**Chapter 4** comprises overall details about the system architecture, component diagram, Model processing diagram which describes the system design as a whole

**Chapter 5** includes the implementation details tools and approaches and modules used in the project along with the test cases and testing performed in the system

**Chapter 6** includes the future scope of the project and necessary future recommendations along with conclusion to the project.

# Chapter 2 : Literature Review:

The exponential growth of information’s and emergence tools , techniques and technologies for the management of those information’s is the norms of the everchanging modern world.

The digitalization of information’s has also revolutionized the way we perceive ,access ,share and utilize those information’s . Digital Library have become an essential tool managing and disseminating vast amount of digital content to diverse userbase. The concept of a "digital library" is mt merely equivalent to a digitized collection with information management tools. It is rather an ENVIRONMENT to bring together collection’s services and people in support of the full life cycle of creation dissemination, use, and preservation of data, information, and knowledge. [2]

The rapid growth of information over internet demands intelligent information agents that can shift through all the available information and find out the most valuable to us. These intelligent systems can be categorized into two classes: Collaborative Filtering (CF) systems and Content-based Filtering (CBF) systems. The difference between them is that **collaborative filtering** is based on collecting and analyzing a large amount of information about the behavior, activities, and preferences of users and the predicting what a user likes based on the similarity of the user to other users, whereas the **content-based filtering** algorithm is based on a description of the element and a profile of a user’s preferences. These algorithms try to recommend items that are similar to those that a user has liked in the past. [3]

To build a CBF recommendation system we need to describe a set of features of an item. To directly compare user and item profiles, CBF heavily relies on similarity metrics—functions that compute how similar or different two feature vectors are.CBF models do not compare users directly; they base their recommendations solely on the user’s past behavior. They derive desired recommendations from the feature-based representation of the items in the database[3]

The similarity between the user can be drawn using the data such as whether the resource is liked by the user and how frequent the resource is liked. Based on the similarity of the user-resource relationship the CBF can be performed . Clustering is the task of grouping a set of objects so that objects in one cluster are more similar to each other than they are to those in other clusters. Clustering is often used as an unsupervised machine-learning tool to find a hidden structure in large datasets. It is based on grouping items in a dataset into several groups, or clusters, such as items in the same group being, on average, more similar than they are to items in different groups[3] . Once the similarity is drawn and resources are clustered based on the user preferences and labeled such that a user belong to a labeled cluster.

Because there are so many options, choosing the best clustering algorithm for your dataset is  challenging .The features of the dataset, the quantity of data items, the number of outliers, and the cluster characteristics are some significant elements that contribute towards  the which clustering algorithm is good for the datasets. There are three popular categories of the  Clustering algorithm which are Partitional clustering, Hierarchical clustering , Density-based clustering. Based on the requirements and constrains of the project K-mean algorithm is one of the best suited for the project. K-Mean algorithm is a partition based algorithm where algorithm works through different iterations to assign the set of data’s into k number of clusters . The first step of k-mean algorithm is to randomly select k centroids, where k is equal to the number of clusters . Centroids are data points representing the center of a cluster. Then by using the Euclidian distance as we can represent each cluster by the point in data space that is the average of the data assigned to it. Since each cluster is represented by an average[4] which means each data point is assigned to the cluster which is closest to the centroid and new centroid is evaluated using mean of datapoint of each cluster. The iteration is performed until the max iterations or until the convergence i.e. iterations yield to same cluster and centroids.

By using the k mean we can construct a clusters of user having a similar preferences to the resources by which recommendations can be constructed using the label of cluster the user belongs to.

# Chapter 3 : System Analysis:

This section describes what a software system does and includes requirements that specify all the fundamentals action of the system. The requirements are the major part in the system development. Once the requirements are collected they determine the structure, functionalities and operational constrains of the system.

## Requirement Analysis :

It is the process we gather the information about the required and existing systems and identifying the user and system requirements from this information. It consists of Functional and Non-functional Requirements.

### Functional Requirement:

Functional requirements are the requirements that the end user specifically demands as basic facilities that the system should offer. These requirements describe what the system should do and sometimes not do. Users have to register themselves to fully access and utilized the services offered by the applications. The functional requirements for the R-Archive are as follows :

* **User Registrations, login and authentications :**

Users must be able to register account securely providing necessary information’s not only registrations user are also securely authenticated by the system

* **Update and Delete Details :**

Users are able to update and delete account details and make changes to their account . whereas admin has the privilege to access every information’s to update and delete

* **File Upload :**

Both users and admin are able to upload digital doc to the database and server

* **Clustering Users and Recommendations :**

System is able to cluster the users based on their preferences of the book they liked and based on the clusters recommend the book to the users

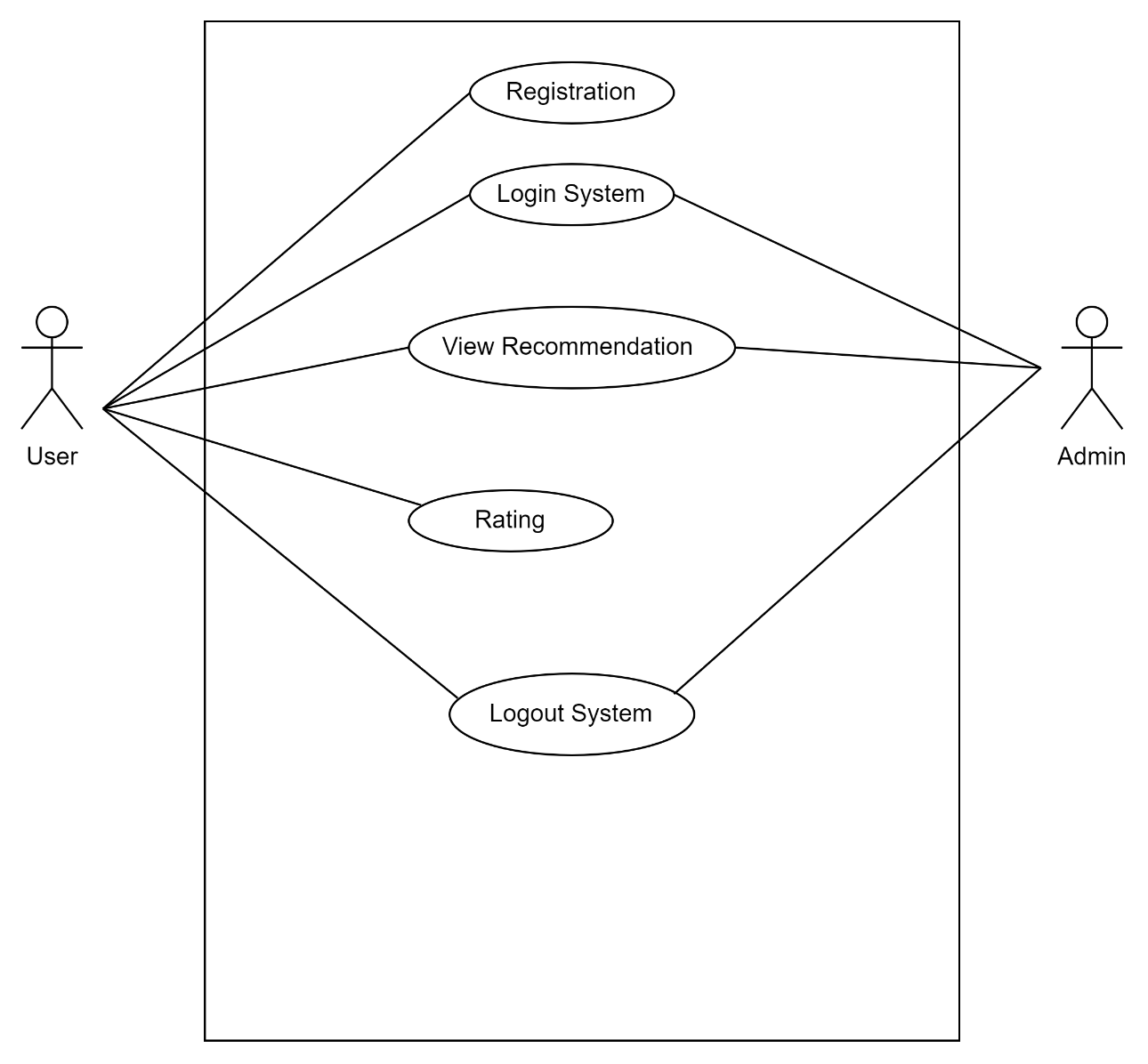


Figure 3. 1 Use Case Diagram

### Non-Functional Requirement

These requirements focus on the operational aspects and expected behavior of the system. Nonfunctional requirements are crucial as they are rooted in user expectations and guide the system's functionality. The reliability of the system is depending upon meeting these nonfunctional requirements. From the observation can state that the following are must have Nonfunctional requirement to our system.

* **Security:**

Since the information about the users and resources invested into the benefit of both the users and the organizations are valuable the system must be secured using various authentication approached such as JWT 0auth basic authentication and so on.

The user experience must also be secure to the users as system must isolate different variabilities form the users.

* **Performance:**

The system should be able to respond to user requests within a second as the site will be well optimized with the reliability target of 99%.

* **Quality Of Service(QOS):**

The system should be configured to present the user with best quality services to the users

* **Portability :**

The system must be accessible form any devices and should run in almost every system environment giving user access form anywhere.

* **Easy To Use:**

The system is naturally designed and user-friendly, giving them good experience and can navigate with ease.

## Feasibility Analysis:

A feasibility Analys is a study to evaluate the feasibility of the project in terms of how beneficial product development will be for the organization from a practical point of view and whether to move forward with the project or not.

* **Technical :**

As the current system is not so heavy, it will run with the minimum resources and the available technology is feasible for the project development. The only technical aspects needed are Operating system i.e. windows, IOS, Linux,Python,JS & Database Postgres SQL Server, Internet browser, internet connection.

* **Operational:**

Operational Feasibility is the measure of how well the project will perform. All the functionality of the system are possible to create .The calculations and database queries are possible to execute without any errors and extra requirements. The software configurations used by the system are possible to establish. The system will operate over internet thus making the user available with the places.

* **Economical:**

In the case of economic feasibility, there is no or minimal expense as we have been using free resources for its development as the only expense being that hosting and maintenance. For the users to access the application, the only cost involved will be in getting access to the Internet and the hosting cost of the application .

* **Schedule :**

The requirement can be meet and fulfilled in the time period of the 12th week

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Activity | 1st week | 2st week | 3st week | 4st week | 5st week | 6st week | 7st week | 8st week | 9st week | 10st week | 11st week | 12st week |
| Requirement  Gathering |  |  |  |  |  |  |  |  |  |  |  |  |
| Planning and  Designing |  |  |  |  |  |  |  |  |  |  |  |  |
| Development |  |  |  |  |  |  |  |  |  |  |  |  |
| Testing |  |  |  |  |  |  |  |  |  |  |  |  |
| Documentation |  |  |  |  |  |  |  |  |  |  |  |  |

Figure 3. 2 Gantt Chart

## Analysis

### Object Modeling using Class Diagram:

In the given class diagram, we have all together three classes model which are   
**User , Resources , Tag , UserResourceInteraction.** The attributes and value of the object constructed with the given classes is stored in the database using **ORM**(**O**bject **R**elational **M**odel)

The **User**  class describe the overall user properties form the basic information’s to the privilege they have.

* **id (int pk) :** It is an attribute representing the unique identifier for the user with the primary key constrains
* **password (varchar) :** It is an attribute representing the password for the user with is hashed by default
* **date\_joined(timestamp) :** It consist of the timestamp for the creation of the account in the form of “YYYY-MM-DD HH:MM:SS”
* **last\_login(timestamp) :** It consist of the timestamp of the last login date and time of the account in “YYYY-MM-DD HH:MM:SS”
* **email(varchar) :** It consist the basic email information
* **first name(varchar) :** It consist of the user first name
* **last name(varchar) :** It consist of the user last name
* **age(int) :** It consist of the user age
* **address(varchar) :** It consist of the user address
* **phone\_number(varchar) :** It consist of the user phone number
* **is\_admin(bool) :**  It determines the permission the user has over the system if yes the user has an admin privilege if no the user is just the normal user
* **is\_active(bool) :** It is to determine the user is active or not if not active user is not able to login into the system
* **is\_staff(bool) :**  similar to the is\_admin it gives the additional privilege to the admin to edit the data in the database form the admin panel
* **profile\_pic(varchar):** It consist of the URL route where the profile pic of the user is stored

The **Resources** class describe the overall properties of the resources which is the basic information’s of the resources :

* **resource id(int pk) :** It is a unique resource identifier for the resources with the primary key constrains
* **name(varchar) :** It consist of the resource name
* **description(varchar) :** It consist of the resource description such as short summary about the resource
* **tag(int fk ) :** It consist of the one-to-many relationship with the tag id of the **Tag class** following the foreign key constrains
* **resource\_image(varchar) :** It consist of the URL route where the image of the resource is stored mostly being the cover page of documents
* **resource file(varchar) :** It consist of the URL route of the resource path where the resources is stored
* **created\_at(varchar) :** It consist of the timestamp of the date and time of the resources uploaded in “YYYY-MM-DD HH:MM:SS”
* **uploaded\_by(int fk) :** It consist of the one-to-one relationship with the user id of the **User** **class** following the foreign key constrains

The **Tag** class describe the overall properties of the resources tags:

* **tag\_id(int pk) :** It is a unique Tag identifier for the resources with the primary key constrains
* **tag\_name(varchar) :** It consist of the name of the tag associated with to the tag\_id

The **UserResourceInteractions** class describe the nature of the user and resource and how the interactions is stored using rating :

* **id(int pk ) :** It is a unique identifier of the interaction data stored in the database following primary key constrains
* **resource id(int fk) :** It consist of the one-to-one relationship with the resource id of the **Resource class** following the foreign key constrains
* **rating(float) :**
* **user id(int fk) :** It consist of the one-to-one relationship with the user id of the **User class** following the foreign key constrains

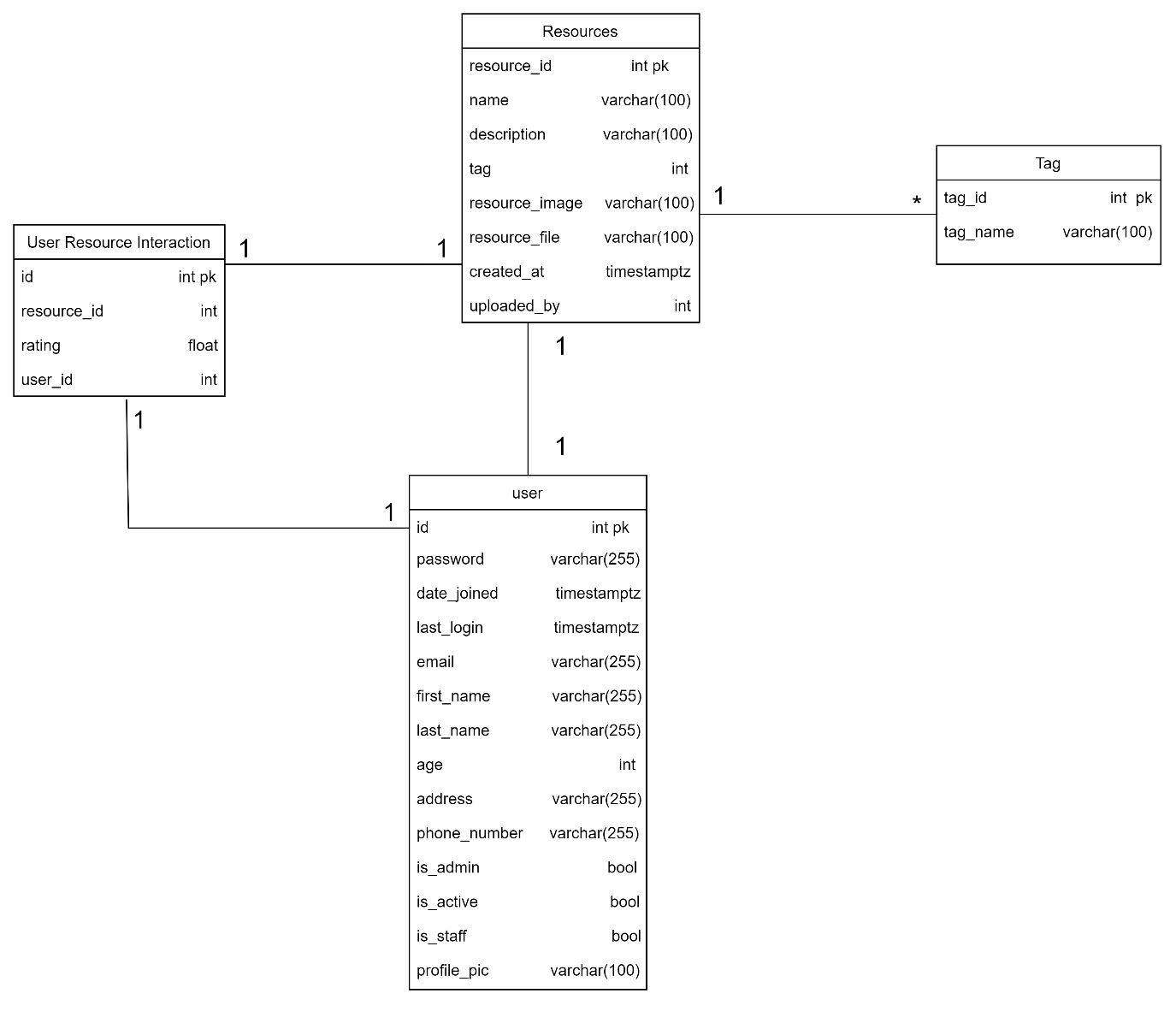


Figure 3. 3 Class Diagram

### Dynamic modelling using Sequence Diagram :

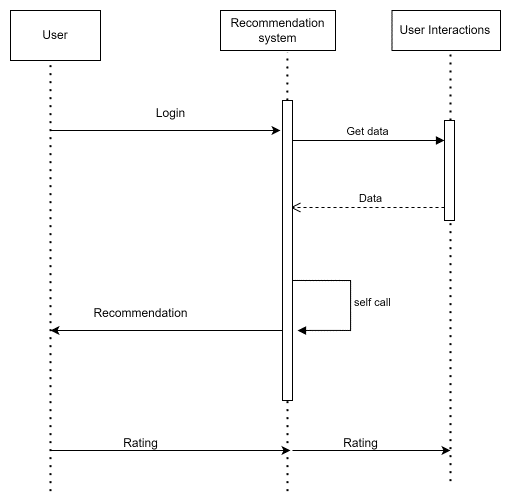


Figure 3. 4 Sequence Diagram

The Sequence diagram mainly illustrate the way how the user interacts with the Recommendation System and the recommendation is generated and given to the users

At first the user is authenticated by the system once authenticated the system starts to look for the stored information about the user and the resources the user interacted with based on the fetched information the system calls the K-MEAN algorithm and based on the cluster recommend the resources to the user . Then the user visits the resource rate the resource the rating is stored as the user interaction which is then again used for creating recommendation for other or same user.

### Process modelling using Activity Diagrams

Figure 3. 5 Activity Diagram

Activity diagram illustrate the step-by-step activity the user demonstrates for the application where the user is authenticated if exist login into the homepage if not register the user. The user can view the recommendations as well as search for the required resources than rate the visited resources. The rating is used for the recommendations as illustrated by sequence diagram.

# Chapter 4 : System Design:

The essence of system design is making decisions about the logical organization of the software. The system design not only includes the logical organization but also logical implementations and functionality of the system.

## Designs

### System Architecture

The system architecture for project **R-archive** is client-server architecture. In client-server

client interacts with the browsers consisting of the frontend of the project . The frontend then interacts with the Backend server which consist of implementation logic.

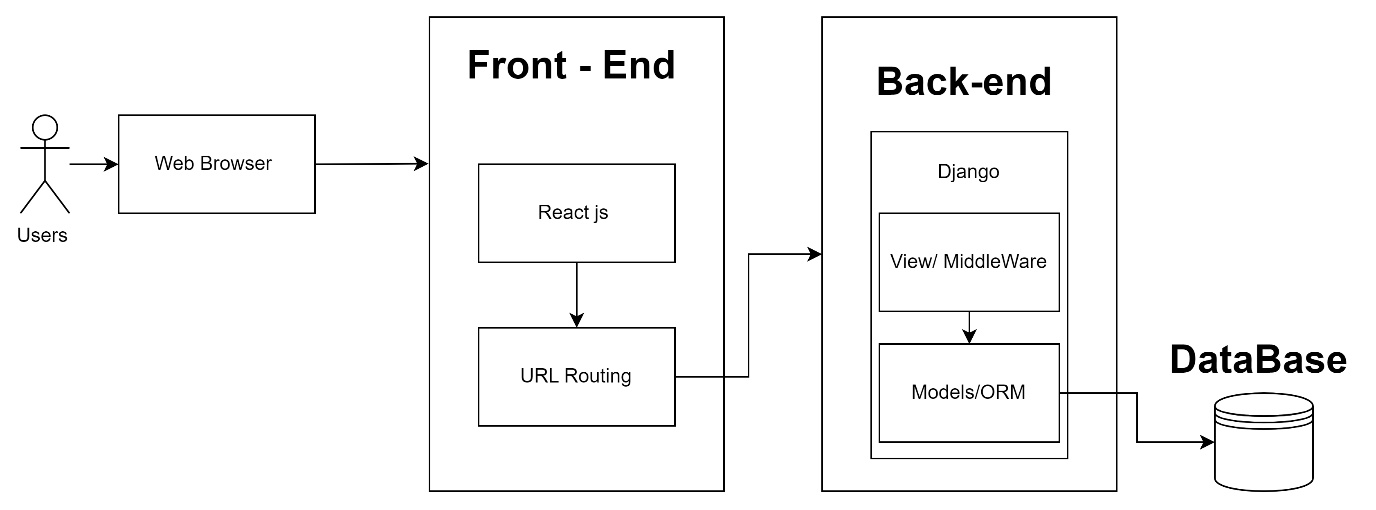


Figure 4. 1 System Architecture

### Component Diagram

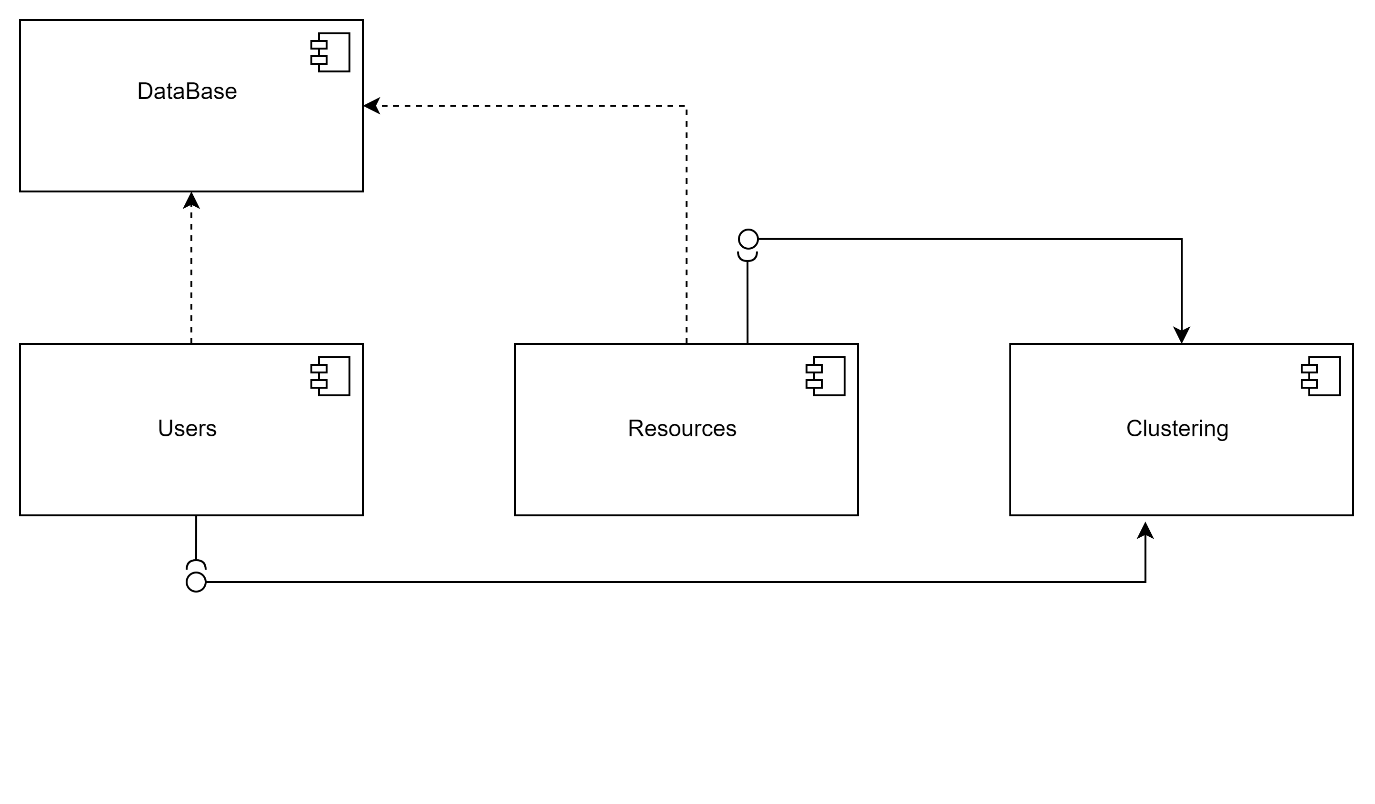


Figure 4. 2 Component Diagram

The component diagram describes the components which construct our system mainly our application is composed of the 3 components one which are users , resources and clustering. The clustering component uses both dependent over the users and resources.

### Model Processing

Figure 4. 3 Model Processing

The models processing diagram depicts the overall process of modeling our recommendation algorithm where the cylindrical shape represent the database data and the rectangle shape represent the steps taken for the processing.

K-mean algorithm in the recommendation system uses the user and user resource interaction data form the database using those data the k mean algorithm construct cluster and label the users . The ranking of the resources is done based on the count occurrence of the rating and the average rating of the resources. The ranking the resources are arranged with the highest rank in the first order and the resource is displayed to the user

## Algorithm Details:

### K-Mean Clustering:

Clustering is the process of dividing a dataset into groups of similar data points, called clusters, where the data points within each cluster are more similar to each other than they are to data points in other clusters. The goal of clustering is to identify meaningful patterns or structures in the data.

The most clustering algorithms, the main thing that we need is a concept of distance between our data. a. If our data live in some space X then to find the distance between the data points we use cluster center the cluster centers are shown as x and x′ and computes a distance between them. We’ll write such a distance as ||x − x′||. If X is ℝD, then a natural choice would be the Euclidean (L2)[4]

||x − x′ || = [4]

When the data space X is ℝD and we are using Euclidian distance we can represent each cluster by the point in data space that is the average of the data assigned to it. Since each cluster is represented by an average, this approach is called K-Means[4]

**Steps involve working on K-mean algorithm**:

* Input a data vector with the number of clusters(N)
* Initialize with the random values
* For the given number of the iterations , iterate through the items
* Find the mean closest to the item by calculating distance of the items with the each of the mean centroid of the data
* Assign label to each mean items
* Update mean by shifting average to the items in that cluster.

# Chapter 5 : Implementations and Testing:

## Implementations:

R-Archive is a web application which provide a central repository to the users to upload and maintain digital documents along with personalizing the learning experience of the users

The web applications use the unsupervised learning methodology for recommendations of the resources to the user . It uses the K mean algorithm to cluster and label the users based on the past resource interactions of the users. The application is developed using python ,Django and react as a programming language and a Postgres SQL as a database

### Tools used :

* **Frontend:**

**React:** React is a free and open-source front-end JavaScript library for building user interfaces based on components. It is maintained by Meta and a community of individual developers and companies.

**Html:** Hypertext Markup Language or HTML is the standard markup language for documents designed to be displayed in a web browser. It defines the content and structure of web content. It is often assisted by technologies such as Cascading Style Sheets and scripting languages such as JavaScript.

**Tailwind:** Tailwind CSS is an open-source CSS framework. The main feature of this library is that, unlike other CSS frameworks like Bootstrap, it does not provide a series of predefined classes for elements such as buttons or tables.

* **Backend Tools:**

**Django:** Django is a free and open-source python-based web framework that runs on a web server. Django Rest Framework is used to build the RESTful APIs with Python and Django.

**Database**: PostgreSQL: PostgreSQL is a free and open-source relational database management system emphasizing extensibility and SQL compliance. PostgreSQL supports both non-relational and relational data types.

* **IDE Used :**

**Vs code :** VS code is a source-code editor developed by Microsoft for Windows,

Linux and MacOS. It supports different programming languages with their own

extensions as well as provide various features such as debugging, IntelliSense , Lsp,etc

**Google Collab** : Google Collab is a web-based cloud ide provided by the google with it preexisting python env and libraries

* **Diagram Tool Used**

**Draw.io** : A free and open-source online diagramming tool which enables users to generate many sorts of diagrams and structures.

### Implementation details of modules:

**K-Mean Algorithm:**

* **Data sets and data collections:**

The data stored in the database is utilized by the algorithm by converting the data of the both users, resources and their interactions into data frames i.e. list of list in tabular formatting.

* **Data Pre-processing :**

After fetching the data form the database the most imp data is the User Resource interactions where user rating for the resources is stored. The most important part of the data for the resources is the Tag . We expand the resources into the tag they poses then we combine both the expanded tags and the resources into the single datasets where the tags are assigned with respective rating form the resources and grouped by user mean average value.

* **Elbow Method :**

Elbow method is a useful tool for determining the appropriate number of clusters in clustering algorithms. The elbow Method compromises of the iterations through the increment where the datasets are fit over the different value of K(i.e. clusters ) and inertia(sum of square) over K-mean algorithm. We construct a graph and evaluate the number cluster against the inertia by visual inspections we determine the stability in the plots as the cluster increases the significant changes in the inertia decrease which can create a elbow like structure in the graph. The Optimal value of K is taken such that the increase in the cluster doesn’t make significant changes in the sum of square of the

* **Labeling :**

For labeling we identify the optimal value of **k** i.e. for number of clusters . By using **k** labeling is done based on the which cluster does the datasets belongs to using K-Mean algorithm. The users are labeled based on the tags and their rating as we starts to cluster them together.

* **Ranking:**

The rank is identified by using the occurrence of the resources and the rating they received

Raking is done by identifying the fitness of good will for our cluster by multiplying the average rating and the occurrence count for the given resources each resource are grouped by based on the labeled users associated with the given resources

**Objective(goodness)= Rating ["mean"]\*Rating ["count"]**

Greater the objective(goodness) higher will the resources be places in the ranking and the **highest-ranking resources is recommended first to the user**.

## Testing :

### Test cases for unit testing:

**For Registration Module:**

**Table 5. 1 Unit Test Case for Registrations Module**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Test Case | Precondition | Steps | Expected Result | Actual Result | Status |
| Register with correct data. | User must enter all the details according to format and precautions. | 1. Open Signup Page  2. Enter Valid data  3. Click Signup button Example: Email: admin@mail.com Username: admin Password: hello123@ | The registration is successful. | The new user is successfully registered in the system. | Pass |
| Register with empty form attributes. | User must enter all the details according to format and precautions. | 1. Open Signup page.  2. Enter valid data but leave one field empty.  3. Click signup button. Example: Email: user@mail.com Username :user  Password: user123 | The registration is unsuccessful. | The field may not be blank | Pass |

**For Login Module :**

**Table 5. 2 Unit Testing for Login Module**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Test Case | Precondition | Steps | Excepted Result | Actual Result | Status |
| Login with correct data | The user should be registered. | 1. Open Login page.  2. Enter Valid Data  3. Click Login Button. Example: Username: Admin Password: hello123@ | The user must successfully login to the app. | The user is successfully logged in | Pass |
| Login with incorrect data | The user should be registered. | 1. Open Login page.  2. Enter correct username but incorrect password.  3. Click login button.  Example: Username: admin Password: hello123 | The user must not be able to login to the app. | Invalid Credentials. | Pass |

**For Uploading the Documents :**

**Table 5.3 Unit Testing for the Uploading the documents**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Test Case | Precondition | Steps | Expected Result | Actual result | Status |
| Upload documents | The user must be registered. | 1.Open login page.  2.Click on upload documents :- Document Lecture\_Notes.pdf | Document is successfully. | Document is uploaded. | Pass |

**For Rating the Documents :**

**Table 5.4 Unit Testing for Rating the Documents**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Test Case | Precondition | Steps | Expected Result | Actual Result | Status |
| Rate the document | The user must be register. | 1.Login to system  2.Click any sort of document .  3.Go to rating area.  4.Rate according to your importance | Document get rating in the form of star. | The document is rated | Pass |

**For Searching Module :**

**Table 5.5 Unit Testing for the Searching**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Test Case | Precondition | Steps | Expected Result | Actual Result | Status |
| Searching the document | The user must be register. | 1.Login to system  2.Go to search bar and search document by its name .  For example:- Neural Network | The document is found. | The document related with the name of the document is found. | Pass |

**For recommendation Module :**

**Table 5.6 Unit Test Case for the recommendations**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Test Case | Precondition | Steps | Expected Result | Actual Result | Status |
| Recommend document | The user must be register. | 1.Login to system  2.Based on the cluster of the user the documents are recommended to the users | We can see a recommendation in home page. | Recommended  document is in home page. | Pass |

### Test cases for system testing :

System testing is defined as testing of a complete and fully integrated software product.

System testing of software or hardware is testing conducted on a complete, integrated

system to evaluate the system's compliance with its specified requirements. system testing

falls within the scope of black box testing, and as such should require no knowledge of

inner design of the code or logic. One of the types of system testing is the usability testing

which is performed in the system.

**For Usability**

**Table 5. 7 System testing test case for Usability**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Test Case | Precondition | Steps | Expected Result | Actual Result | Status |
| Click on various  links on the  system. | The user must login to the integrated website | 1.Login to system  2. User must Navigate through the websites | Link should redirect users  to functional web pages | The user is able to easily redirect and navigate to the webpages and use the well-defined functionality of the webpage | Pass |

## Result Analysis

All the modules of the system are working as expected and the output of the system is up

to expectations. The developed web application provides us with the feature of uploading the documents , rate them and construct a recommendation based on the user liked content (i.e. rating).

This application uses a K-mean algorithm to find the optimal distance to construct a cluster between the users and cluster users and recommendation is given to the users based on the clusters. The cluster is constructed for users based on the documents they rated.

Main focus is given on creating the user-friendly interface so that user can navigate

throughout the system very easily. Any users can upload the documents so that the documents can be archived into the central repository. The archived documents are then recommended to the users based on the Users clusters

Overall, the web application is working according to expectations and without any problem.

# Chapter 6 : Conclusion and Future Recommendations

## Conclusion

In conclusion, content-based recommendation with rating and clustering techniques

represents a powerful approach to enhancing the user experience within a digital library

platform. By leveraging user ratings to understand preferences and clustering techniques to

group similar items, content-based recommendation system can provide highly personalized and relevant resource suggestions present in digital library. By analyzing user ratings, the system gains valuable insights into individual preferences, allowing it to recommend items that align closely with each users interests and tastes.

Overall, using content-based recommendation with rating and clustering techniques in a

digital library offers a smart way to help users to personalize the learning digital journey of the users.

## Future Recommendations:

R-archive project is just the solution to personalize the digital libraries according to the user’s taste. The content-based recommendation is the first stage for the project as the user increases so does the dataset . as the dataset increases the project will have enough capacity and data to further personalized the user experience not only based on the content based but also collaborative based . The clustering has a major disadvantage for selecting optimal value of K i.e. number of clusters for the project. The major advantage the clustering gave was that the cold start or the dead start where there is not enough sufficient data for the other implementation of personalization.

The system can be improved as taking following into considerations.

* By personalizing the collaborative filtering technique into the system
* By using advance machine learning algorithm to make the recommendations more accurate and more efficient
* By optimizing the backend and frontend system into the standardized industrial level
* By implementing the proper cloud storage service into the system
* By implementing the compression algorithms and techniques to improve storage and efficiency of the library.

# Reference:

[1] A. Ur Rahman, A. Fahd, and Alhaidari, “The Digital Library and the Archiving System for Educational Institutes,” vol. 20, 2018.

[2] Singh and Sukhdev, “Digital Library: Definition to Implementation Item Type Presentation.”

[3] M. Koroteev, “Review of Clustering-Based Recommender Systems,” doi: https://doi.org/10.13140/RG.2.2.34745.90725.

[4] R. Adams, “K-Means Clustering and Related Algorithms COS 324 -Elements of Machine Learning Princeton University.”