A

PROJECT REPORT

ON

AUTO ASSIGN ARTICLES

USING DATA SCIENCE, NLP, and CLUSTERING

IN ASSOCIATION WITH GEEKSFORGEEKS

Submitted

By

AMIYA RANJAN ROUT (1805106002)

Under the Esteemed Guidance of

Mrs. Swarnalata Pati

ASSISTANT PROFESSOR

COLLEGE OF ENGINEERING AND TECHNOLOGY

BHUBANESWAR

GEEKSFORGEEKS, A – 118, SECTOR-136,

NOIDA, UTTAR PRADESH – 201305

COLLEGE OF ENGINEERING AND TECHNOLOGY

(Techno Campus, PO- Ghatikia, Mahalaxmi Vihar, Bhubaneswar-751029)

DEPARTMENT OF COMPUTER SCIENCE & APPLICATION

CERTIFICATE

This is to certify that the Thesis/Dissertation entitled **Auto Assign Articles** is being submitted by **AMIYA RANJAN ROUT** with regd.no **1805106002** in fulfilment of the requirement for the award of the degree of MCA. In to the College of Engineering & Technology is a record of bonafide work carried out by us under our guidance and supervision from February to May 2021.

The results presented in this thesis have been verified and are found to be satisfactory. The results embodied in this thesis have not been submitted to any other University for the award of any other degree.

INTERNAL GUIDE	HOD
Mrs. Swarnalata pati	**Dr. Jibitesh Mishra**
(Assistant Professor)	**(Associate Professor)**

COLLEGE OF ENGINEERING & TECHNOLOGY

(Techno Campus, PO- Ghatikia, Mahalaxmi Vihar, Bhubaneswar-751029)

DEPARTMENT OF COMPUTER SCIENCE & APPLICATION

ACKNOWLEDGMENT

I am extremely grateful to **Professor Dr. Jibitesh Mishra, H.O.D** and **Mrs. Swarnalata Pati** as my Internal Guide, Department of CSA, College of Engineering & Technology for their encouragement and valuable guidance in bringing shape to this dissertation.

I will be failing in duty if I do not acknowledge with grateful thanks to the author of the references and other literature referred in this Project.

I express my thanks to all staff members and friends for all the help and coordination extended in bringing out this Project successfully in time.

Finally, I am very much thankful to my parents who guided me for every step.

Date: 01.05.21

Amiya Ranjan Rout

(1805106002)

Place: Balasore

COLLEGE OF ENGINEERING & TECHNOLOGY

(Techno Campus, PO- Ghatikia, Mahalaxmi Vihar, Bhubaneswar-751029)

DEPARTMENT OF COMPUTER SCIENCE & APPLICATION

DECLARATION

I **Amiya Ranjan Rout** bearing **Roll No: 1805106002** a Bonafede student of **College of Engineering and Technology,** would like to declare that the project titled **"AUTO ASSIGN ARTICLES"** a fulfilment of MCA Degree course of College of Engineering & Technology is my original work in the year 2021 under the guidance of **Mrs. Swarnalata Pati,** Assistant Professor, Computer Science & Application Department and it has not previously formed the basis for any degree or diploma or other any similar title submitted to any university.

Date: 01.05.21

Amiya Ranjan Rout

(1805106002)

Place: Balasore

ABSTRACT

GeeksforGeeks is the largest computer science portal not only in India but also all over the world. According to the Alexa Traffic Rank, the most visitors of GeeksforGeeks are from India (44.8%) and the USA (19.6%). In 2020, GeeksforGeeks ranked amongst the top 50 websites in India. GFG got more than 500 technical articles per day from various contributors from all over the world (mostly from India). And currently, almost 60 (30 internals & 30 externals) reviewers in various computer science domains are working in GFG Content Team. So, it's a very tedious task for the manager to assign those articles among the reviewers daily, and it consumes a lot of time for the manager. So, we are building such a tool that can process the title of the article in such a way that it can assign that article automatically to the corresponding reviewer and it will save lots of time for the manager. For example, an article named "Variables in C Programming Language" will assign automatically to that reviewer who is reviewing the article in the C Programming domain. So, we are going to use the Natural Language Processing and Classification Algorithm for this task.

INDEX

Chapter 1: Introduction

1.1 About GeeksForGeeks

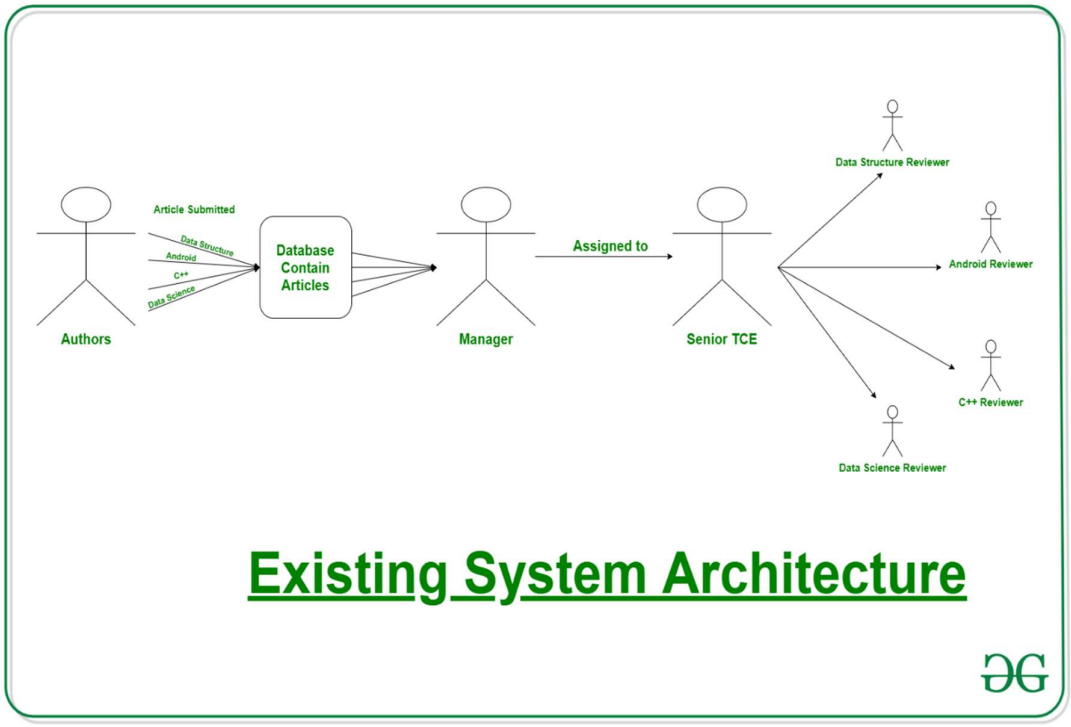
With the idea of imparting programming knowledge, Mr. Sandeep Jain, an IIT Roorkee alumnus started a dream, GeeksforGeeks. Whether programming excites you or you feel stifled, wondering how to prepare for interview questions or how to ace data structures and algorithms, GeeksforGeeks is a one-stop solution.
With every tick of time, we are adding arrows in our quiver. From articles on various computer science subjects to programming problems for practice, from basic to premium courses, from technologies to entrance examinations, we have been building ample content with superior quality. In a short span, GeeksforGeeks has built a community of 1 Million+ Geek around the world, 20,000+ Contributors, and 500+ Campus Ambassadors in various colleges across the nation. GeeksforGeeks is the largest computer science portal not only in India but also all over the world and used by millions daily. According to the Alexa Traffic Rank, the most visitors of GeeksforGeeks are from India (44.8%) and the USA (19.6%). In 2020, GeeksforGeeks ranked amongst the top 50 websites in India.

1.2. Problem Statement

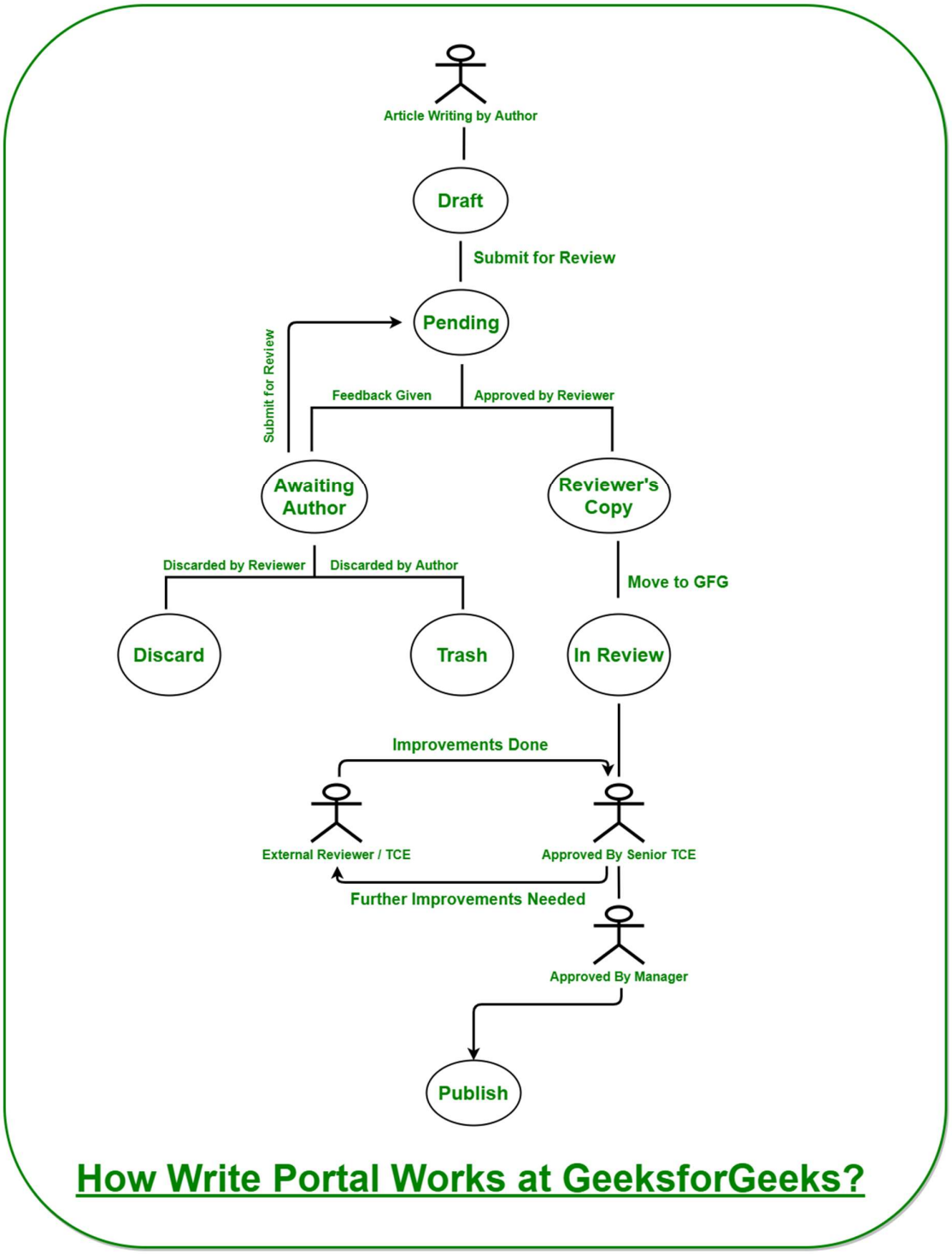
GFG got more than 500 technical articles per day from various contributors from all over the world (mostly from India). And currently, almost 60 (30 internals & 30 externals) reviewers in various computer science domains are working in GFG Content Team. So, it's a very tedious task for the manager to assign those articles among the reviewers daily, and it consumes a lot of time for the manager. So, we are building such a tool that can process the title of the article in such a way that it can assign that article automatically to the corresponding reviewer and it will save lots of time for the manager. For example, an article named "Variables in C Programming Language" will assign automatically to that reviewer who is reviewing the article in the C Programming domain. So, we are going to use the Natural Language Processing and Classification Algorithm for this task.

1.3. Existing System

In the existing system, we do the assignment task manually. We assign the articles to the reviewer by inspecting the title and assign it to the corresponding reviewer which takes a lot of time. At first, all the articles come to the manager then he distributes them with the teams. Total 4 teams are working at the GFG content team (Team DSA, Team Web Technologies, Team GATE, and NET, Team Other Computer Science Subjects). The manager assigns the article to these 4 mentors then he/she assign them to the reviewers. So, this system is very tedious and time-consuming.


Data Structure Reviewer
Article Submitted
Data Structure
Android
C++
Data Science
Database
Contain
Articles
Assigned to
Android Reviewer
Authors
Manager
Senior TCE
C++ Reviewer
Data Science Reviewer
Existing System Architecture

1.4 How Write Portal Works at GFG?



1.4. Proposed solution

In the proposed system our main focus is:

1. Gathered all the raw data from the various sources (database, webpages) and clean the data using the data science technique.
2. Text pre-processing and vectorization using Natural language Processing Technique.
3. Text classification and Grouping texts in threads with the nearest neighbour's search algorithm controlled by the Levenshtein distance within each group using the clustering technique.
4. Map each cluster to the corresponding reviewers.

Some FAQs

1. **How do we get the text inside the article?**

   Articles have text embedded inside along with a lot of unwanted boilerplate, ads, copyright messages etc. How can you segment the article to just get the article text and throw away the rest? This is an information extraction problem. Most sites use HTML DOM/SAX parsing along with a lot of other heuristics.

2. **How do we get definitive terms inside the article?**

   Articles are a lot of text, that include all kinds of conjunctions, connectives, pronouns, nouns, numbers, etc. What is most important to an article? Techniques like TF-IDF exist to get to a good distance. Some types of features are more important than others - especially when the objective is to group related articles together. When you are considering technical articles, you are more interested in putting articles from a technical term together. For example, we need the term like Android, HTML, Java, Machine Learning, LinkedList etc.

3. **How do you know two documents are similar?**
   Till now, you've brought an article to its vector form. A vector of keywords and weights - that depict importance to the article. Given that you have

two such vectors, how do we figure the similarity between them? Measures like cosine similarity exist here. There are several similarity measures, and each one has pros and cons. For example, cosine similarity also gives importance to terms found in one article, and not found in the other. So, if one document is a superset of another, the cosine similarity may still be low. If you don't want this property - you may look at other distance measures.

4. **How do you group a set of related documents together?**
   Document clustering is an extremely well researched topic. To start with, you'll get algorithms off the books - like hierarchical agglomerative clustering, k-means clustering, and top-down clustering. k-means clustering may not be best suited when you don't know how many clusters you have to group articles into. So, you'd have to inspect HAC and the top-down methods.

Chapter 2. System Requirement

Software Requirement

- IDE:
  - Jupyter Notebook
- Language:
  - Python
- Technique:
  - Data Science
  - Natural Language Processing
  - Clustering Algorithm

Hardware Requirement

- Intel i5 10^{th} Generation Processor
- Windows 10 / UBUNTU 14.0.4
- Jupyter Notebook Installed

Chapter 3: Software Requirement Analysis

3.Software Requirement Analysis

Analysis is a detailed study of the various operations performed by a system and their relation within and outside the system. A key question is what must be done to solve the problem. One aspect of the analysis defining the boundaries of the system and determining whether or not a candidate system should consider other related system. During analysis data are collected on the available files decision points and transaction handled by the parent system. Some logical system models and tools are used in the analysis. Data flow diagrams, interviews, onsite observation and questionnaires are examples, the interview is commonly used in analysis. It requires special skill and sensitivity to the subject being interview bias in data collection and interpretation can be problem.

Training experience and common sense are required for collections of the information are needed to do analysis. Once analysis is completed, the analyst has firm understanding of what is to be done. The next step is to decide how the problem might be solved. Thus, in system design, we be move from the logical to the physical of the life cycle.

3.1. System Planning and the initial investigation

The most critical phase of managing system projects is planning to launch a system investigation, we need plan detailing the steps to be taken, the people to be questioned and they outcome expected. The initial investigation has the objective of determining whether the users request has potential merit. The major steps are defining user requirements. When the initial investigation is completed. The user receives a proposal summarizing the finding the recommendation of the analyst.

3.2. Information Gathering

A key part of feasibility analysis is gathering information about the present system. The analyst knows what information to gather, where to find it, how to collect it and what to make of it. The proper use of tools for gathering information is the key to successful analysis. The tools are the traditional interview, questionnaire, and on-site observation. We need to Know, for example how to structure an interview, what makes up a questionnaire, and what to look for on-

site observations. These tools when learned help analysis assess the effectiveness of the present system and provide the groundwork for recommending a billing system.

3.3. Feasibility Study

The main objective of the preliminary analysis is to identify the problem, evaluate the system concept of feasibility, and perform the economic and technical analyses perform the cost benefit analysis. After the clarification analysis the solution proposed it is checked that it is practical to implement that solutions. This is done through the feasibility study. It is checked for various aspect whether the proposed solution is technically or economically feasible or not. On the basis of which it has been categorized into four classes viz

1.) Technical
2.) Economic
3.) Legal
4.) Operational

The outcome of the preliminary analysis should be clear so that an alternate way to do the job can be found out.

3.3.1. Technical Feasibility

During the technical feasibility studies following issues are taken into consideration

1.Whether the required technology is available or not?

2. Required resources are available or not? (Manpower, programmer, software and hardware etc)

Once the technical feasibility is established, it is important to consider the monetary factors also. Since it might happen that developing a particular system may be technically possible but it may be require huge investments and benefits may be less. For evaluating this, economic feasibility of the proposed system is carried out.

As in our proposed system our team has technically trained manpower with knowledge of developing the system. We are going to use web technology in our system, which is readily available. Software to be used is also available easily. So technically the project is feasible.

3.3.2. Economic Feasibility

For any system if the expected benefits equal or exceed the expected costs, the system can be judged to be economically feasible. In economic feasibility, cost benefit analysis is done in which expected costs and benefits are evaluated.

Economic analysis is used for evaluating the effectiveness of the proposed system. In economic feasibility, the most important is cost benefit analysis. As the name suggests, it is an analysis of the cost to be incurred in the system and benefits derivable out of the system.

As in our institute the hardware and software required for this type of system is already available so economically our project is feasible.

3.3.3. Cost Benefit Analysis

Developing an IT application is an investment. Since after developing that application, it provides the organization with profits. Profits can be monetary or in the form of the improved working environment. However, it carries risks, because in some cases an estimate can be wrong. And the project might not actually turn out to be beneficial.

Cost Benefit Analysis helps to give management a picture of cost, benefits and risks. It usually involves comparing alternate investments.

Cost Benefit determines the benefits and saving that are expected form the system and compare them with the expected costs.

The cost of an information system involves the development cost and maintenance cost. The development costs are one-time investments whereas the maintenance cost is recurring. The development cost is basically the cost incurred during the various stages of the system development.

Each phase of the life cycle has a cost. Some examples are:

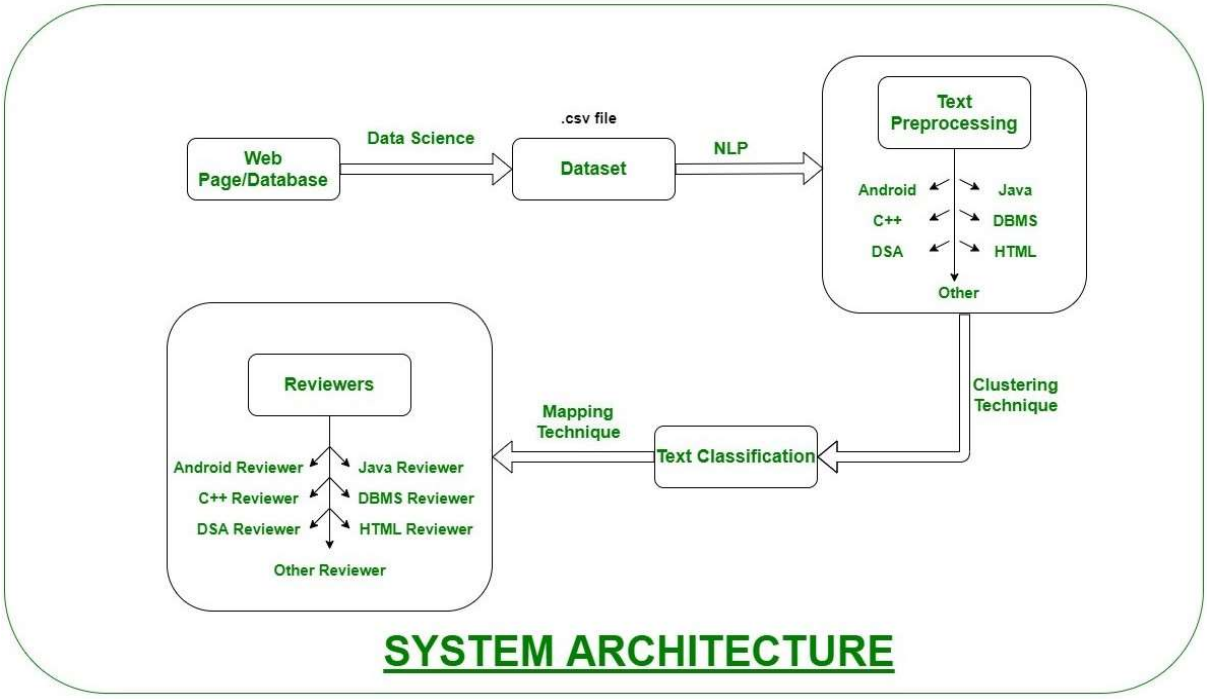
1. Personal
2. Equipment
3. Supplies
4. Overheads etc.

In proposed system all hardware and Software are available in the company so cost is 0%.

3.3.4. Legal Feasibility

It includes study concerning contracts, liability, violation and legal other traps frequently unknown to the technical staff.

Chapter 4: System Architecture



SYSTEM ARCHITECTURE

Chapter 5: What are my Responsibilities?

In this project I have assigned the following tasks:

1. Prepare the proper documentation for this project.
2. Gathering all the data from the web pages using the web scraping technique.
3. Clean the raw data and prepare a complete table using the data science technique.
4. Text pre-processing and vectorization using NLP technique.

Chapter 6: Project Plan & Approach

6. Project Plan & Approach

Software development team consists of 3 members.

Phase-1: Requirement Gathering and Analysis

Gathering system requirement and prepare a System Requirement Specification document.

After collect the information it is analysed that the available resources can fulfil all the requirements. And it is also be examined that what resource will be used.

Phase-2: System Design

Make a detailed analysis of the system and prepare a System Design Document on the basis of SRS.

Phase-3: Prepare UTC/STC for testing the software

Prepare Unit Test Cases document. This document will be used to verify whether the functional requirements of the system have been met.

Phase-4: Develop the software

Develop the planned system

Phase-5: Test the software using the prepared UTC/STC and Rework if needed

Run your software programs using the respective UTC to verify & test the software.

Phase-6: Demonstrate the software to users & Implement it

Chapter 7: Project Schedule

PROJECT SCHEDULE

Step #	*Time Frame*
Phase 1. Requirement Gathering	**18 Days**
Phase 2. System Design	**25 Days**
Phase 3. Prepare UTC & STC for Testing the software	**15 Days**
Phase 4. Develop the software	**32 Days**
Phase 5. Test the software	**13 Days**
Phase 6. Demonstrate the software to users & implement it	**12 Days**

Chapter 8: Software Design

8.1. Scope

In this section we define the scope of the design effort. The design phase is an important part of the system development phase. A good design of the system needs creativity and flair from the designer and is the key to effective and successful engineering. The following are the basic objectives of the software design process:

- To describe the process of software design where informal ideas are transformed into detailed implementation description.
- Introduction of different stages in the design process.
- Understanding whether an Object Oriented or a Functional Oriented approach or both should be applied to the software.
- Determining and improving, cohesion control and coupling within subsystems.

In the project, the design phase has been identified as one of the most crucial documents. In this phase, we have identified the various aspect of the “USER MANAGEMENT”, which have to be implemented as subsystems and their further components.

8.2. System model specification

The system model chosen for the software project is the prototyping model. It begins with requirement gathering. The overall objectives of the software are well defined, known requirements are identified and an area where further definition is mandatory has been outlined. After this the software process is initiated and a prototype is built and then it is evaluated by the customer and used to refine requirements for the software to be developed. A prototype is a

working system that is developed to test the ideas about the new system and prototyping is a process of building a model of the system to be developed.

This approach is used in our project because it is difficult to know all the requirements in the beginning of the project. Such situation arises when no other system like proposed one is built earlier.

The complete system of our project is based on iterative model where the user keeps specifying the requirements according to the model made available and the changes are incorporated into the system to build a better model. In this type of approach, we will develop a prototype and show it to the user. The user verifies the prototype. In case if there are some suggestions from user then again that functionality is added to the system and again the user evaluates it. This cycle gets repeated till the time the user is completely satisfied with the prototype. Then the actual project is built and implemented thereafter.

Chapter 9: References/Bibliography

- https://towardsdatascience.com/news-aggregator-in-2-weeks-5b38783b95e3
- https://towardsdatascience.com/building-a-news-aggregator-from-scratch-news-filtering-classification-grouping-in-threads-and-7b0bbf619b68
- https://ilyagusev.github.io/tgcontest/en/science.html
- https://contest.com/docs/data_clustering
- http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.59.7746&rep=rep1&type=pdf
- https://www.aclweb.org/anthology/D09-1001.pdf

THE END