PAPERNERDS: A RESEARCH DIGITAL LIBRARY AND UNDERGRADUATE THESIS PORTAL WITH TEXT MINING ALGORITHMS

An Undergraduate Thesis

Presented to the

Faculty of College of Computer Studies

Laguna State Polytechnic University

Sta. Cruz Campus

In Partial Fulfillment of the requirements for the Degree
BACHELOR OF SCIENCE IN COMPUTER SCIENCE

By:

MIGUEL N. CEMILLER

JIAN CARLO C. CARAYAP

AIRHA M. VILLANUEVA

Under the supervision of:

MIA V. VILLARICA, D.I.T.

FEBRUARY 2022

VISION

The Laguna State Polytechnic University is a center of sustainable development initiatives transforming lives and communities.

MISSION

LSPU provides quality education through responsive instruction, distinctive research, and sustainable extension and production services for improved quality of life.

QUALITY POLICY

LSPU delivers quality education through responsive instruction, distinctive research, sustainable extension, and production services. Thus, we are committed with continual improvement to meet applicable requirements to provide quality, efficient and effective services to the university stakeholders' highest level of satisfaction through an excellent management system imbued with utmost integrity, professionalism and innovation.

College of Computer Studies Goal

The College of Computer Studies graduates are expected to become globally competitive and innovative computing professionals imbued with utmost integrity, contributing to the country's national development goals.

Program Educational Objective

The Bachelor of Science in Computer Science (BSCS) graduates are computing professionals and proficient researchers in designing and developing innovative solutions. It is designed to enable students to achieve the following by the time they graduate:

1. Apply knowledge of computing solutions from fundamentals to complex problems appropriate for the abstraction and conceptualization of computing models.

2. Communicate effectively and recognize the legal, ethical and professional issues governing the utilization of computer technology and to engage in independent learning development as a computing professional.

3. Ability to apply design, develop and evaluate systems' components and processes through mathematical foundations, algorithmic principles and computer science theories.

4. Developed a culture of research for technology advancement.

5. Demonstrated good leadership and a team player that will contribute to nation building and engage in life-long learning as foundation for professional development.

APPROVAL SHEET

The thesis entitled "**PAPERNERDS: A RESEARCH DIGITAL LIBRARY AND UNDERGRADUATE THESIS PORTAL WITH TEXT MINING ALGORITHM**" prepared and submitted by **MIGUEL N. CEMILLER, JIAN CARLO C. CARAYAP**, and **AIRHA M. VILLANUEVA** in partial fulfillment of the requirements for the degree of Bachelor of Science in Computer Science is hereby recommended for approval and acceptance.

Mia V. Villarica, DIT
Thesis Adviser

Approved by the Committee on Oral Examination with a grade of _____.

Jonardo R. Asor
Subject Specialist

Alliana M. Miranda-Ablan
Technical Editor

Ma. Cezzane D. Dimaculangan
Language Critic

Victor A. Estalilla Jr.
Statistician

Reynalen C. Justo, MM-ITM, LPT
Chairman

Accepted and approved in partial fulfillment of the requirement for the degree of Bachelor of Science in Computer Science.

Reynalen C. Justo, MM-ITM, LPT
Dean/Associate Dean

Dr. Rina J. Arcigal
Director/Chairperson
Research and Development

Date Signed: _____

RESEARCH CONTRIBUTION NO.

ACKNOWLEDGEMENTS

The researchers would like to express their heartfelt gratitude to the following individuals who have contributed and supported them in the completion of the study:

MRS. MIA V. VILLARICA, D.I.T, their thesis adviser, for her time and patience in checking the manuscripts over and over again, and for sharing her wonderful ideas and constructive criticisms, which profoundly contributed to the success of this study;

MRS. ALLIANA M. MIRANDA-ABLAN, their technical editor, for correcting the mistakes in the format and contents of the manuscript as well;

MR. JONARDO R. ASOR, their specialization expert, for sharing his much-needed knowledge and suggestions regarding the study;

MR. VICTOR A. ESTALILLA JR., their statistician, for sharing his inputs regarding this study's data sampling design;

MS. MA. CEZZANE D. DIMACULANGAN, their language critic, for her assistance in criticizing the construction and grammar of the manuscript;

MRS. REYNALEN C. JUSTO, MM-ITM, LPT, the Associate Dean of the College of Computer Studies, for allowing the researchers to conduct the study;

And lastly, they are thankful for the effort and hard work of each member of this **RESEARCH TEAM**, who put their heart out in this study.

DEDICATION

The researchers wholeheartedly dedicate this study to **ALMIGHTY GOD**, for His provision of guidance, and for giving the researchers the things they needed in fulfilling His plans. Moreover, they dedicate their study to the following groups of individuals:

FAMILY of the researchers, who have been their source of inspiration and gave them the motivation to keep pushing forward. The study would not have been possible without their constant love and support;

And lastly, their classmates, friends, and relatives who made the researchers feel like they were not alone. This was one hell of a ride, but it was surely a great moment to remember forever. Thank you.

Cemiller, Miguel N.

Carayap, Jian Carlo C.

Villanueva, Airha M.

ABSTRACT

Topic modeling is one of the most powerful approaches in text mining for data mining, latent data discovery, and finding relationships among data and text documents. On the other hand, clustering in text mining has also been proven to be a very effective method for document classification. Moreover, information retrieval is just as important today since IR systems are used on an everyday basis by a wide variety of users. The study sought to develop a research digital library and undergraduate thesis portal integrated with text mining algorithms, but beyond development, the researchers evaluated the use of the LDA topic modeling and K-means clustering on different sets of textual data. Based on the findings, it was discovered that topic models work particularly well with longer texts as opposed to clustering algorithms which work better with shorter texts. LDA topic modeling on short texts returned a coherence score of 0.698 while returning 0.756 for longer texts. K-means on the short texts yielded a perfect V-measure score of 1.00 while yielding a 0.45 V-measure score.

Keywords: *Digital library; Web portal; Undergraduate Thesis; Text mining; Topic Modeling; Clustering; Information Retrieval; Latent Dirichlet Allocation; K-means; Cosine similarity; TF-IDF;*

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

DEFINITION OF TERMS

In designing and developing the system, the researchers were able to identify terminologies enumerated in technical and operational terms that may be useful for better understanding of the study.

Technical Terms

Some terminologies used in the design and development of the developed system were defined in this section.

Bag of Words (BoW)	It refers to a way of extracting features from text for use in modeling, such as with machine learning algorithms.
Clustering	It refers to the task of dividing the data sets into a certain number of clusters in such a manner that the data points belonging to a cluster have similar characteristics.
Data Mining	It refers to the practice of analyzing large databases in order to generate new information.
Digital Libraries	It refers to a collection of documents in organized electronic form, available on the Internet.
Information Retrieval	It refers to the techniques for storing, retrieving, and often disseminating recorded data, especially via the use of a computerized system.
K-Means Clustering	It refers to an unsupervised learning algorithm, which groups the unlabeled dataset into different clusters.
Latent Dirichlet Allocation (LDA)	It refers to a topic model that is used to categorize text in a document to a particular topic.

Lemmatization	It refers to the removal of inflectional endings only and to return the base or dictionary form of a word, which is known as the lemma.
Lowercasing	It refers to the conversion of all characters to lowercase.
Machine Learning	It refers to a type of artificial intelligence (AI) that enables software to grow increasingly accurate at predicting events without having to be explicitly programmed to do so.
Preprocessing	It refers to the preliminary processing of data in order to prepare it for the primary processing or for further analysis.
Python	It refers to an interpreted high-level general-purpose programming language.
Stop Words Removal	It refers to any words in a stop list that are filtered out before or after processing of natural language data.
Text Mining	It refers to the process of transforming unstructured text into a structured format in order to find new insights and significant patterns.
TF-IDF	It refers to Term Frequency – Inverse Document Frequency determines a term's relevance by considering the term's importance in a single document and scaling it by its importance across all documents.
Tokenization	It refers to the process of exchanging sensitive data for non-sensitive data called "tokens" that can be used in a database or internal system without bringing it into scope.
Topic Modeling	It refers to a type of statistical modeling for discovering the abstract “topics” that occur in a collection of documents.

Unsupervised Learning	It refers to a type of machine learning in which the algorithm is not provided with any pre-assigned labels or scores for the training data.
Vectorization	It refers to a step in feature extraction. The idea is to get some distinct features out of the text for the model to train on, by converting text to numerical vectors.

Operational Terms

This section defines any terms or phrases derived from the study operationally, implying the way they were used in the study.

Descriptive research	The term is defined as the collection of information to systematically describe a situation. It answers the what, when, where, and how questions regarding the study. This term is used in the study to imply the interview conducted by the researchers.
Developmental research	The term refers to the systematic study of designing, developing, and evaluating instructional programs, processes, and products. For the purpose of this study, the term is used to refer to the features of the system developed.
Web application	The term web application refers to a collection of web pages delivered over the Internet. It was used to imply the output system of the study.

CHAPTER I

INTRODUCTION AND ITS BACKGROUND

Data grew exponentially every day. Almost all types of institutions and organizations stored data digitally. According to Sagayam (2012), this generally came as textual data in digital libraries and other repositories. In the College of Computer Studies (CCS) of Laguna State Polytechnic University Santa Cruz Campus, the Research Implementing Unit Head (RIUH) or the research coordinators also utilized digital storage to monitor the undergraduate thesis progress of their students. However, as their data grew larger, it became difficult for the research coordinators to track the undergraduate thesis progress of their students. Furthermore, it didn't allow the student researchers to publicly search for relevant documents related to their study through the Internet.

A public access to research papers and a platform for the undergraduate thesis process of their students was figured to overcome the aforementioned limitations. The integration of the said features into a web application would benefit not just those involved in the undergraduate thesis process, but also those researchers outside of the college.

The goals mentioned above were vastly enhanced with the integration of text mining algorithms, a technique for discovering patterns from collections of documents according to Gharehchopogh & Khalifelu (2011). With text mining, the stakeholders were given a better representation of information based on their requirements. Specifically, the integration of Information Retrieval (IR), a method for relevant research papers in a document collection according to a statistical similarity measure based on a user's query according to Hull & Grefenstette (1996), was implemented.

Next, the integration of topic modeling, a technique for finding latent or hidden topics within document collections according to Anupriya & Karpagavalli (2015), recommended related research papers to users. Lastly, the integration of a clustering algorithm, a method for dividing collections of texts into their different clusters according to Yao, et al. (2012), helped suggest the most appropriate group and technical panel members for the students.

Thus, the study aimed to develop Papernerds, a digital library, and an undergraduate thesis portal, that can be used for searching relevant research papers, recommendation of related research papers, and suggestions of group and technical panel members through text mining algorithms for the undergraduate thesis process of students in the College of Computer Studies (CCS) at Laguna State Polytechnic University SCC.

Research Problem

According to Patil & Atique (2013), massive amounts of data were generated every day, resulting in the digitization of the majority of information in government, industry, business, and institutions.

As stated in an interview conducted with the Research Implementing Unit Head (RIUH) of the college, they also utilized the use of online repositories and public libraries for the storage of research papers. The research papers were kept private through a cloud-based file storage service, Google Drive. However, this prevented the general public, particularly college student researchers, from accessing the research papers, especially in the midst of the COVID-19 pandemic, when access to public libraries was limited. Moreover, it was figured that searching relevant resources for a study required great effort, due to the lack of text mining techniques.

On the other hand, the tasks for the college student researchers were distributed through Google Classroom, a web-based learning platform. However, its restrictions adversely affected the undergraduate thesis process of the college. For instance, the service didn't allow the students to form their own groups. By this means, the research coordinators had to track students' group progress manually, and doing that with the current platform was arduous, at least for the research coordinators.

All that being said, the researchers came up with a web application to address the difficulties mentioned above. The solution, "Papernerds: A Research Digital Library and Undergraduate Thesis Portal with Text Mining Algorithms," intended to provide a public digital library for storage and retrieval of research papers, as well as a web portal for dealing with some of the aspects of the undergraduate thesis processes for the College of Computer Studies (CCS) in Laguna State Polytechnic University SCC. Specifically, the study sought to answer the following research problems: (1) How to design and develop a web application that included a public digital library for research papers and a web portal for undergraduate theses for the College of Computer Studies (CCS) at Laguna State Polytechnic University SCC? (2) How to test LDA topic modeling based on coherence and perplexity and K-means clustering based on homogeneity, completeness, and V-measure scores to identify which to integrate to the system for research papers recommendations and student group and technical panel member suggestions? (3) How to examine the performance of the models developed integrated into the system through actual testing?

Research Objectives

The study's goal was to design and develop "Papernerds: A Research Digital Library and Undergraduate Thesis Portal with Text Mining Algorithms" that stored

research papers in a public digital library and dealt with some of the undergraduate thesis processes with research stakeholders for the College of Computer Studies (CCS) in Laguna State Polytechnic University SCC with the use of topic modeling and information retrieval algorithms for the research digital library, and a clustering algorithm for the undergraduate thesis portal, and test the performance of the algorithms integrated in the web application.

Specifically, this study aimed:

1. To design and develop a web application that included a public digital library for research papers and a web portal for undergraduate theses for the College of Computer Studies (CCS) at Laguna State Polytechnic University SCC;
2. To test LDA topic modeling based on coherence and perplexity and K-means clustering based on homogeneity, completeness, and V-measure scores to identify which to integrate to the system for research papers recommendations and student group and technical panel member suggestions.
3. To examine the performance of the models developed integrated into the system through actual testing.

Theoretical Framework

The theoretical framework determined how the algorithms would be utilized in the system development. Sagayam & Srinivasan (2012) stated that text mining, a variation of data mining, was defined as the process of extracting implicit knowledge from textual data, and it was important to be used for vast amounts of unstructured text data. With this, this study was guided by the framework of text mining.

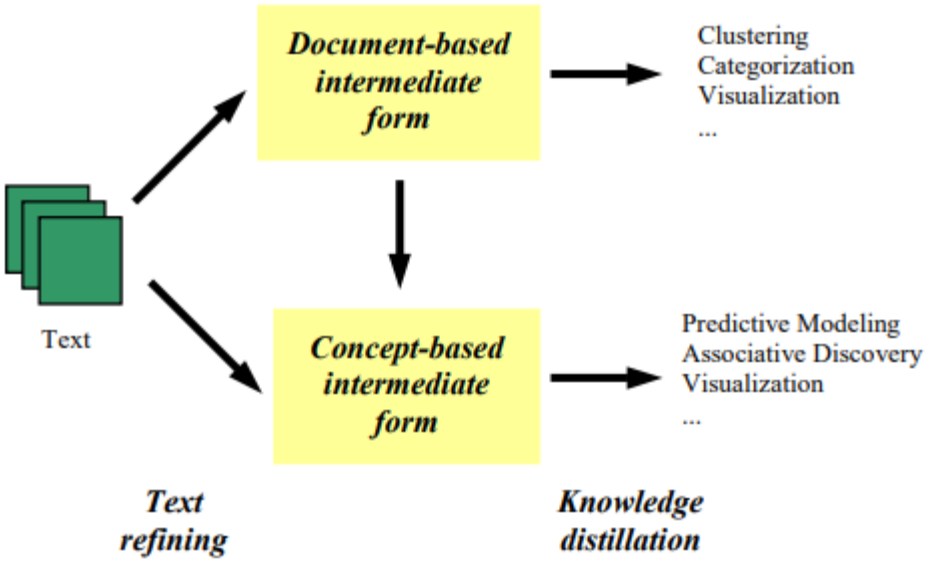


Figure 1. Text mining framework.

The text mining framework, based on Tan (1999), pertained to the methods of text refining and knowledge distillation. Text refining referred to the conversion of unstructured into document-based or concept-based intermediate forms. Knowledge distillation from a document-based intermediate form was used by the researchers in this study to find patterns across collections of documents or texts. For instance, the use of topic modeling, clustering, and information retrieval algorithms were used mainly for grouping or categorizing related textual data. On the other hand, knowledge distillation in a concept-based intermediate form identified patterns or knowledge between objects or concepts.

Conceptual Framework

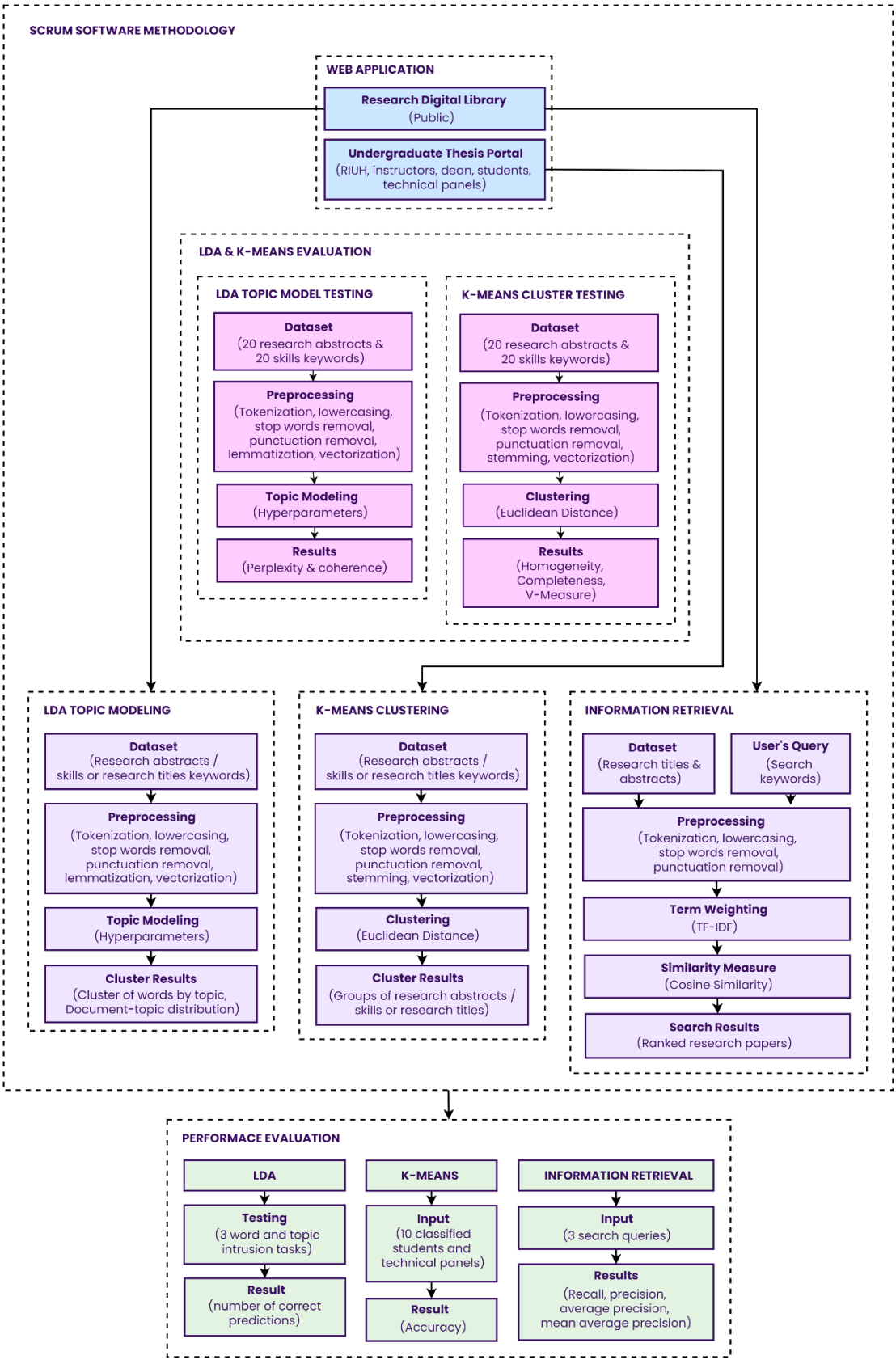


Figure 2. Conceptual framework of the study.

The conceptual framework used the theoretical framework as a reference. It represented the researchers' synthesis of the literature in explaining a phenomenon. Moreover, it mapped out the actions required in the course of the study given their previous knowledge of other researchers' points of view and observations on the subject of study.

The system aimed to provide a public research digital library and an undergraduate thesis portal. It allowed anyone to access the digital library even without an account, however, an account was necessary to be able to view full-text documents. On the other hand, the undergraduate thesis portal was only accessible to certain stakeholders of the study including the Research Implementing Unit Head (RIUH), instructors, dean of the college, students, and the technical panel members.

Within the SCRUM software development methodology, the researchers also integrated three (3) text mining algorithms namely, LDA topic modeling, K-means clustering, and information retrieval. But beforehand, the researchers conducted a series of tests to determine the more appropriate algorithm between LDA and K-means for research paper recommendations and students' group and technical panel member suggestions.

Following this, the researchers integrated the algorithms and models into the web application. First, the LDA topic model was trained on the datasets of either research abstracts or keywords of skills and research titles after being processed through tokenization, lowercasing, stop words removal, punctuation removal, lemmatization, and vectorization for clusters of words by topic and document-topic distributions. Second, the K-means clustering models were trained on preprocessed research abstracts or keywords of skills and research titles through tokenization, lowercasing, stop words

removal, punctuation removal, stemming, and vectorization for clustering based on the Euclidean distance measure. Finally, the information retrieval first subjected data such as titles and abstracts to preprocessing techniques which involves tokenization, lowercasing, stop words removal, and punctuation removal whenever a user conducted a search request. Next, TF-IDF or Term Frequency-Inverse Document Frequency was applied to score the features or the relevant words of the whole corpus. This method allowed the frequent words like stop words and other words which appeared frequently but were not relevant to be scored less, and thus allowed the rare words or words which were usually relevant to be scored higher. In the last step, the cosine similarity utilizing the TF-IDF output measured how similar documents were likely to be in terms of the user's query, resulting in a ranked search result.

For the performance evaluation, the models or algorithms integrated into the web application were examined by the researchers by conducting various tests with sample data. First, the LDA model's performance was evaluated using three (3) word and three (3) topic intrusion tasks, wherein a subject had to identify which doesn't belong for the measurement of coherence of words within a topic and coherence of topics from all the topics. Second, the K-means clustering models were evaluated based on their accuracy from the comparison between true predictions and the prediction of the models from ten (10) classified students based on skills and ten (10) classified technical panel members based on their title works with other students. Finally, the information retrieval was tested using three (3) query sets to calculate its precision, recall, average precision (AP), and mean average precision (MAP).

Scope and Limitations of the Study

The researchers aimed to design and develop a research digital library and an undergraduate thesis portal for the College of Computer Studies (CCS) in Laguna State Polytechnic University SCC, to evaluate certain algorithms, and to examine the performance of the integrated algorithms in the web application. Yet, this study was limited to the following:

- The developed system from the study was only utilized as a public digital library for research papers and an undergraduate thesis portal, accessible only to certain stakeholders which included the RIUH, dean of the college, student researchers, technical panel members, and course instructors.
- The contents of the research digital library of the developed system were only research papers of the College of Computer Studies (CCS) in Laguna State Polytechnic University SCC.
- The research papers uploaded to the system were only limited to PDF files.
- The study only covered the use of descriptive and developmental research methods for the study.
- The study did not cover the deployment or implementation of the system developed.

Significance of the Study

Generally, this study would be useful to certain individuals or groups. Specifically, this study and its system output would be a great benefit to the following:

Research Implementing Unit Head. The RIUH of the College of Computer Studies (CCS) was always tasked to manage numerous undergraduate students of

Computer Science (CS) and Information Technology (IT). The system would help the RIUH by boosting the efficiency of their work. Also, the system would give the RIUH an overview of all the research papers done by students.

Dean of the College. The system would provide the dean of the College of Computer Studies (CCS) an overview of the progress of the undergraduate thesis of students.

Students. The system would be beneficial to the undergraduate students of the College of Computer Studies (CCS). Apart from their ability to be able to gather more research papers relevant to their study through the research digital library, the system could also provide unlimited access to the online research papers repository.

Technical Panel Members. The system would provide the technical panel members with ease of checking the works of the student researchers they were working with.

Course Instructors. The system would provide the course instructors with ease of checking the works of the students they were handling.

Technical Panel Members. The system would provide the technical panel members with ease of checking the works of the student researchers they were working with.

Researchers. This study would be beneficial to present and future researchers who would be conducting a study related to text mining which involves topic modeling, clustering, and information retrieval.

Community. The pandemic adversely affected the nation, and everyone was urged to stay at home. The system would encourage a particular group of researchers

to avoid searching in other places outside for research papers for their own convenience as well. Moreover, institutions, such as the overall campus of Laguna State Polytechnic University SCC, could replicate this study in their own communities to be a part of the growing popularity of digital libraries.

CHAPTER II

REVIEW OF RELATED LITERATURE

This chapter would present the review of related literature and studies; local and foreign, synthesis, and assessment of the different related literature and studies.

Digital Libraries

Digital libraries had been around for a long time. In the study by Bearman (2007) entitled “Digital Libraries,” it was stated that the term “digital libraries” first emerged from the National Information Infrastructure in 1991 and then followed by the US National Political Discourse in 1992. In the pre-WWW context of 1991-1993, the period of popular discovery of the Internet, it was logical that digital libraries focused on fixing the technical issues that people creating multimedia material collections encountered. However, the focus of digital library applications had shifted since the focus on discrete collections during that period.

Borgman (1999), in her study entitled “What are digital libraries? Competing visions,” stated that the term “digital library” was a practical and well-known shorthand for electronic collections, conveying a feeling of deeper material and capabilities than terms like "database" or "information retrieval system." The term implies "the future library" in which the institution was modified to address the new environment in which it exists, as a result of the development of computer networks and digital media.

However, the term evoked a different impression to each person. Fox et al. (1995), in their study entitled “Digital libraries,” pointed out that to some it simply meant that traditional libraries were computerized. To others, who have studied library science, believe it calls for a new approach to library operations, one that includes new

forms of information resources and new storage and preservation technologies. Regardless of the various interpretations of the term, a digital library was a collection of distributed information services, a distributed space of interlinked information, or a networked multimedia information system.

In this regard, Watstein et al. (1999), in their study entitled “Digital library: keywords,” stated that the Word Wide Web (WWW), an interconnected system of public webpages or websites accessible via the Internet, had been referred to as a “digital library” on occasion, with the emphasis on the interconnected system or collection of databases of websites or webpages, rather than the organization, intellectual accessibility, or service attributes.

The arrival of the World Wide Web had enabled the proliferation of new sources of content and information. With the increasing popularity of the Internet, digital libraries had become crucial tools for institutions. In the study by Cabrerizo et al. (2015) entitled “A decision support system to develop a quality management in academic digital libraries,” it was stated that academic digital libraries were taking advantage of the Web's capabilities to assist in teaching, learning, and research. As a result, more people were taking advantage of the services they provide. Students, academics, and researchers all require information, and digital libraries play a critical role in bridging that gap.

In addition to the study stated above where the importance of digital libraries in academic institutions was discussed, Marchionini & Hermann (1995), in their study entitled “The roles of digital libraries in teaching and learning,” expressed the roles of digital libraries in such institutions like serving as a practical role in sharing expensive resources such as books, periodicals, and videos. This solved the problem in which the

students or teachers wouldn't have to travel to other traditional libraries. Further, by making the broad range of information resources available to students and teachers in other schools, digital libraries open new learning opportunities for global rather than just local communities.

Digital libraries were a key technology for developing countries. In the study by Witten et al. (2002) entitled "The promise of digital libraries in developing countries," it was stated that they could assist human development by providing a non-commercial mechanism for distributing humanitarian information on topics such as health, agriculture, nutrition, hygiene, sanitation, and water supply. Perhaps even more important than this was building information collections locally.

In the Philippines, as stated by Garcia et al. (2013) in their study entitled "Open knowledge initiatives in the Philippines: The vibal foundation," Filipiniana.net served as a research portal and digital library on Philippines Studies wherein it collected, organized, and disseminated Filipiniana materials, texts, and books in digital formats. Furthermore, it was organized for Filipino users who desire to study about Philippine Studies through its extensive and comprehensive collections of Filipiniana books and materials spanning the Hispanic era to the present day.

Due to the funding initiatives for digital libraries, attention was drawn to research on digital libraries. Based on the study by Xie (2006) entitled "Users' evaluation of digital libraries (DLs): Their uses, their criteria, and their assessment," an evaluation on two selected digital libraries was conducted to investigate users' use. Nineteen subjects (19) were recruited to participate in the study, and the results of the study yielded some interesting findings. The use of digital libraries showed that the way users interact with them was influenced by their design. In particular, the availability

or lack thereof, as well as the actual design of the features, indicate or lead users on how to use a digital library.

Web Portals

Tatnall (2005), "Portals, portals everywhere," describes a portal in general terms, as a gateway, and a web portal could be seen as a gateway to the information and services on the Web. Thus, a web portal was a system that serves as a centralized place for accessing different resources on the Web.

In the study by Eldrandaly et al. (2009) entitled "Ranking Web Portals Quality Criteria Based on Different Types of Personality," it was stated that a web portal takes information from various sources and puts it all in one place, making it easier for multiple users to access the information. It connects users to a single source of information, data, and services. It could be customized based on the user's function in the organization.

Web portals were one of the most used information products nowadays, as it proves to be very useful. As stated in the study by Nam (2014) entitled "Understanding the motivation to use web portals," web portals, a website or service, offers a broad array of resources and services, such as forums, chats, news, shopping, and links to other sites. It was further stated that they play a role as full size hubs of online communities with shared interests.

Based on the aforementioned studies wherein it was proclaimed that web portals were proclaimed to provide easy access to information from a single location, Dias (2001), in his study entitled "Corporate portals: a literature review of a new concept in Information Management," pointed out that new web-enabled technologies through

web browsers, were increasingly being pushed as a tool for organizational members distributed throughout the organization to gain access to information to overcome information overload.

Thus, they have been implemented in institutions for the benefits it could provide. In the study by Mansourvar & Yasin (2010) entitled "Web portal as a knowledge management system in the universities," it was stated that web portals have become increasingly crucial because of students' need to obtain essential information online. It was also stated that it was vital that universities have a dynamic relationship with the students by sharing the organized knowledge via the web portal.

For instance, Del Rosario et al. (2016), "CollaborateIT: A CCS IT thesis portal with electronic document management system", developed a thesis portal for the IT department in De La Salle University – Manila as part of their study. It covered the entire thesis process as well as the document management of the different thesis documents. The results of the User Acceptance Testing were positive as it was able to solve the problems presented in their study.

With web portals being known as a one-stop website to get the latest updates of the organizations, its usefulness cannot be denied. Nasirun et al. (2012), "Perceived web service quality for students' portal in higher learning institutions," stated that the emergence of the internet encourages organizations to introduce web portals to enhance user's services. As a result, the benefits gained among commercial organizations was unquestionable. However, for non-commercial organizations such as education where the benefits cannot be translated to monetary values, it reflects on the effectiveness and efficiencies of the organizations.

In academic institutions for instance, according to the study by Macrohon et al. (2021) entitled “Design and Development of a System for Managing Student Research: A Case Study for the School of Information Technology, Mapua University,” it was stated that the process of undergraduate student research has been a difficult task, not only for students but also for supervisors and coordinators who supervise various groups of students. With this, the researchers of the study devised a web-based system to streamline the time-consuming process from topic selection through archiving and indexing, reducing, if not eliminating, manual intervention with the involvement of certain stakeholders.

Topic Modeling

These days, understanding large collections of unstructured texts due to the explosion of electronic document archives was a problem. Thus, new techniques were required to automatically organize such large collections. In text mining or machine learning in general, topic modeling was a useful and ubiquitous of such techniques. Based on the study by Wei & Croft (2006) entitled “LDA-based document models for ad-hoc retrieval,” it was stated that topic modeling in the form of topic models offer a formalism for revealing a collection’s topics and have been used to enhance information retrieval.

Kayi et al. (2013), “Topic Modeling Based Classification of Clinical Reports,” described topic modeling as an unsupervised machine learning algorithm that could automatically discover themes or topics in a collection of documents and find topics distributions of each document. Thus, it could be inferred that it does not require a dependent feature in the dataset when building the machine learning model as opposed to supervised classifiers.

This unsupervised nature of topic models was an edge over other supervised classifier. It was considered as their strength as they do not require any priori annotations. The only input they require as stated by Boyd-Graber et al. (2014), “Care and feeding of topic models: Problems, diagnostics, and improvements,” was the text divided into documents and the number of topics you want it to discover.

Topics models have been originated from Latent Semantic Indexing (LSI), most frequently referred to as Latent Semantic Analysis (LSA) as stated in the study by Sendhilkumar et al. (2017) entitled “Generation of Word Clouds using Document Topic Models,” that was a mathematical technique mainly used to index large collections of documents.

Latent Semantic Analysis (LSI) as stated by Deerwester et al. (1990), in their study entitled “Indexing by Latent Semantic Analysis,” uses a technique called Singular Value Decomposition (SVD) to overcome the deficiencies of term-matching retrieval by treating the unreliability of observed term-document association data as a statistical problem. In other words, a “semantic” space was constructed wherein terms and documents that were closely associated were placed near one another in a term-document matrix.

For further discussion of LSI, Rosario (2000) in her study entitled “Latent Semantic Indexing: An overview,” described it as a technique used in information retrieval (IR). It was stated that it projects queries and documents into a space with “latent” semantic dimensions. LSI could be seen as a similarity metric that was an alternative to word overlap measures like TF-IDF where they were represented in vectors of two-dimensional or three-dimensional space. Since it accounts for similarity

measures of words, LSI was not a probabilistic model, therefore it was not considered an authentic topic model.

Probabilistic topic models were the only topic models which were considered authentic. Probabilistic models or probabilistic topic modeling as discussed in the study by Blei et al. (2010), “Probabilistic Topic Models,” were used to summarize a large collection of documents with a smaller number of distributions over words. According to the researchers, such distributions were referred to as "topics" because, when fitted to data, they capture the major themes that run throughout the collection. Basically, unlike LSI/LSA, topics were represented as distributions over words, and documents as probabilistic mixtures of topics.

There were various probabilistic topic models such as Probabilistic Latent Semantic Analysis (pLSA) and Latent Dirichlet Allocation (LDA). Based on the study by Toutanova & Johnson (2007), in their study entitled “A Bayesian LDA-based model for semi-supervised part-of-speech-tagging,” it was pointed out that pLSA was a probabilistic topic model similar to LDA, but without the Bayesian components.

According to the study by Xu (2018) entitled “Topic Modeling with LSA, pLSA, LDA, & lda2vec,” it was stated that pLSA uses a probabilistic method instead of SVD to overcome the drawbacks of LSI/LSA such as the lack of interpretable embeddings wherein we won’t be able to tell what the topics were, the need for a large dataset to get accurate results, and the inefficiency of its representations.

The above-mentioned drawbacks were addressed since as stated by Hofmann (2013) in his study entitled “Probabilistic Latent Semantic Analysis,” pLSA was based on a mixture decomposition obtained from the latent class model, compared to standard LSA/LSI which was based on linear algebra and conducts a Singular Value

Decomposition (SVD) of co-occurrence tables. This results in a more systematic approach with a strong foundation in statistics. Although the approaches seem different, pLSA basically just adds a probabilistic treatment of topics and words on top of LSA.

As stated in the study above, pLSA proves to be a far more flexible model than LSA, however it still runs into a few issues. Levandoski & Lobo (2018), "Document and Topic Models: pLSA and LDA," stated that these issues were caused by pLSA not being a well-defined generative model for new documents as each document was represented simply as a list of numbers which were the mixing of proportions.

The disadvantages of pLSA as pointed out by Cai et al (2008), in their study entitled "Modeling Hidden Topics on Document Manifold," were that it estimates the probability distribution of each document on the hidden topics independently, and the number of parameters in the model grows linearly with the size of the corpus. This results in overfitting which occurs when a model learns the detail and noise in the training data to the point that it lowers the model's performance on new data.

To address the problems with pLSA, LDA which stands for Latent Dirichlet Allocation was developed. According to Wei & Croft (2006), "LDA-Based Document Models for Ad-hoc Retrieval," LDA solves the overfitting problem in pLSA while naturally generating new documents with consistent generative semantics by assuming that the probability distributions of documents over topics were generated from the same Dirichlet distribution with K-parameter hidden variable rather than a large set of individual parameters. As a result, LDA was usually preferable to pLSA since it could quickly generalize to new documents.

Latent Dirichlet Allocation (LDA), the most popular topic modeling technique, was defined further in the study by Blei et al. (2003) entitled "Latent Dirichlet

Allocation,". The authors defined LDA as a generative probabilistic model of a corpus or texts. The basic idea as they stated was that documents were represented as random mixtures over latent topics, each of which was defined by distribution over words. In other words, LDA assumes that every document was a distribution over topics and every topic was a distribution over words.

Furthermore, Chang et al. (2009), "Reading Tea Leaves: How Humans Interpret Topic Models," stated that each document in an LDA model may have multiple topics. Each of the topics has its own internal consistency, with words from that topic frequently appearing together in papers and/or appearing infrequently outside of that topic. It was up to the creator of the model to determine what the consistency refers to, and thus how the topic could be interpreted.

In evaluating LDA models, perplexity and topic coherence were the metrics used to evaluate model fit. According to the study by Jacobi et al. (2016) entitled "Quantitative Analysis of Large Amounts of Journalistic Texts Using Topic Modeling," perplexity was how well a model predicts a sample. It was stated that the number with which the LDA model predicts the data was a good indication of the right number of topics from a computational standpoint, with a lower perplexity indicating a better prediction. To calculate the perplexity, a part of the data to train an LDA model was used. The model was then tested using the data that was left out. This procedure was repeated for models with varying numbers of topics, until it was clear which number results in the lowest amount of perplexity.

On the other hand, the topic coherence metric as stated by Stevens et al. (2012), in their study entitled "Exploring Topic Coherence over many models and many topics," scores a single topic by measuring the degree of semantic similarity between

high scoring words in the topic. This metric helps in distinguishing between topics that were semantically interpretable and topics that were statistical inference artifacts. The topic coherence score could also be calculated in the same way as the perplexity metric in the study above.

Several studies showed most of the application of the LDA generative probabilistic model in recommendation systems. For instance, the study by Tu et al. (2014) entitled "Online Dating Recommendations: Matching Markets and Learning Preferences," developed an online dating recommendation system and designed the LDA model to learn the user preferences from the observed user messaging behavior and user profile features. Experimental results showed that their LDA model could successfully classify similar users and learn their preference.

Moreover, Nagori & Aghila (2011), in their study entitled "LDA Based Integrated Document Recommendation Model for e-Learning Systems," integrated a personalized document recommendation model based on LDA to their e-learning recommender system to help learners find the highly relevant documents to their fields. The integrated model proved to be an essential component for their e-learning system.

Next, LDA was also applied on a recommender system in the study of Jiang et al. (2012), "Recommending Academic Papers via Users' Reading Purposes." The researchers proposed the LDA method that outperforms the traditional TF-IDF method in effectively capturing the similarity between short texts given different reading purposes to recommend academic papers to users from problem-related and solution related aspects.

Another study by Kavvadias et al. (2020) entitled "Supporting topic modeling and trends analysis in biomedical literature" implements the LDA model to allow

biomedical professionals with no experience in programming to execute topic modeling and trend analysis experiments on their own from biomedical literature, integrated in a web-based application. This was proposed and developed to address the problem of biomedical professionals manually reading and analyzing publications and classifying information into research domains and subdomains as the number of biomedical scientific publications grows exponentially every year, complicating literature analysis. The study revealed that the system functionalities were well received, with an overall usability score of 76/100 on the System Usability Score (SUS) scale.

Lastly, in the Philippines, as stated in the study by Cuaton et al. (2021) entitled "A topic modeling analysis on the early phase of COVID-19 response in the Philippines," the early phase of the government's pandemic response was analyzed using the LDA based on the official press releases issued by the Department of Health (DOH) from early January to mid-April 2020. The researchers' results revealed five latent themes or topics with the government's highest effort of 40% centered on the topic, "Nationwide Reporting of COVID-19 Status", while the topic, "Contact Tracing of Suspected and Infected Individuals" received the least attention, with only 11.68%, indicating a lack of priority in this area. As a result, the researchers concluded that the government was unprepared in the early stages of the pandemic, but it made efforts to reorganize its financial and operational objectives to focus on disease control.

Term Weighting Schemes

Deng et al. (2014), in their study entitled "A study of supervised term weighting scheme for sentiment analysis," defined term weighting as an approach that assigns weights or values to terms to improve the performance of text mining tasks. These weights measure the relevance of terms and denote how much these terms contribute to

a variety of document-related tasks, such as information retrieval, text classification, and text clustering. It has been shown to improve the performance of a variety of tasks.

According to the study by Lan et al. (2005) entitled "A Comprehensive Comparative Study on Term Weighting Schemes for Text Categorization with Support Vector Machines," it was stated that a term weighting scheme, an important step in automatic text classification, was used to convert documents as vectors in the term space.

These vectors in the term space were represented in a vector space model (VSM) as stated in the study by Ko (2012) "A Study of Term Weighting Schemes Using Class Information for Text Classification."

In VSM, according to the study by Qimin et al. (2014) entitled "Text clustering using VSM with feature clusters," each document was represented as a numerical feature vector, which consists of weights of many terms extracted from the text corpus. The vector was often high-dimensional due to the huge vocabulary of words in documents.

Term weighting schemes were divided into two categories. Samant et al. (2019), "Improving Term Weighting Schemes for Short Text Classification in Vector Space Model," stated the categories of term weighting schemes. The categories of schemes as stated by the researchers were supervised and unsupervised weighting schemes. Supervised term weighting schemes use category information, which was the number of positive category documents in which the term was present while unsupervised term weighting schemes, such as Term Frequency-Inverse Document Frequency (TF-IDF), do not consider the category information of a term for weight assignment.

Supervised term weighting schemes for general text classification includes Information Gain (IG), Gain Ratio (GR), and chi-square ($\chi 2$), which were then multiplied with Term Frequency (TF) to incorporate the "local" relevance of the term in documents. However, these schemes as stated by Xuan et al. (2014), "A New Improved Term Weighting Scheme for Text Categorization," and the study stated above make use of the prior information about the membership of training documents in predefined categories to assign weights to terms. Thus, these supervised weighting schemes were considered unsuitable for unsupervised text mining tasks.

On the other hand, unsupervised term weighting schemes do not rely on prior information about membership of training documents to categories. Quan et al. (2011), in their study entitled "Term Weighting Schemes for Question Categorization," includes binary representation, Term Frequency (TF), and Term Frequency-Inverse Document Frequency (TF-IDF). These unsupervised weighting schemes were suitable for unsupervised text mining tasks.

The simplest binary representation was used to represent 1 as presence & 0 as absence of the term in document, the term frequency (TF) was one of the most used method where each word was associated with the frequency of occurrence of that term in the document, and TF-IDF was the most widely used by researchers in text classification according to the study by Motwani & Tiwari (2013) entitled "Comparative study and analysis of supervised and unsupervised term weighting methods on text classification."

TF-IDF as defined in the study by Kuang & Xu (2010) entitled "Improvement and Application of TF • IDF Method Based on Text Classification," was simply TF multiplied by IDF, where TF stands for term frequency, which was used to determine

the term's describing ability, and IDF stands for inverse document frequency, which was used to determine the word's differentiating ability.

Basically, according to the study of Qaiser & Ali (2018) entitled “Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents,” it was stated that TF was simply the number of times a term appears in a document while Inverse IDF was a measure of whether a term occurs too frequently or infrequently in a collection of documents. IDF was calculated to scale down the words that occur too frequently such as stop words.

According to the study by Forman (2008) entitled “BNS Feature Scaling: An Improved Representation over TF-IDF for SVM Text Classification,” TF-IDF was the most extensively used representation for real-valued feature vectors in machine learning for text classification.

Vijayarani et al. (2015), in their study entitled “Preprocessing Techniques for Text Mining – An Overview,” defined TF-IDF as a numerical statistic that reveals how important a word was to a document in the collection. The researchers also further stated that was often used as a weighting factor in information retrieval and text mining.

TF-IDF according to Ramos (2003), “Using TF-IDF to Determine Word Relevance in Document Queries,” calculates values for each word in a document using an inverse proportion of the frequency of the word in a specific document to the percentage of documents in which the word appears, as the name implies. The researcher also stated that words with high TF-IDF numbers indicate a strong relationship with the document they appear in, implying that if the word appears in a query, the document may be of interest to the user.

Information Retrieval (IR)

Based on the study by Sinhal (2001) entitled "Modern Information Retrieval: A Brief Overview," people have recognized the significance of archiving and finding information for thousands of years. With the introduction of computers, it became possible to store vast volumes of data, and it became necessary to extract meaningful information from these collections. The field of Information Retrieval (IR) sprang from this need in the 1950s.

The term "information retrieval" (IR) has been around long before the Internet was born. According to Swanson (1988), in "Historical Note: Information Retrieval and the Future of an Illusion," the term "information retrieval" was first introduced into the literature of documentation by Calvin Mooers in 1950. This developed out of a problem with subject access to information and how to index scientific journals and reports.

When the Internet was born, it led to the rapid growth of the World Wide Web (WWW). Thus, the difficulty in finding needed information, effective and efficient information retrieval systems have become necessary as stated in the study by Liu (2011) entitled "Learning to Rank for Information Retrieval.

Due to the rapid growth of the Internet and the Web, as well as its massive popularity in communities and institutions, it resulted in a great amount of information on them. Liaw & Huang (2003), "An investigation of user attitudes toward search engines as an information retrieval tool," stated that the primary use of the Internet was for information retrieval from search engines, such as Google, Bing, and Yahoo!

Search engines, especially the ones in the Web, were the most well-known examples of information retrieval systems. Voorhees (1999), "Natural Language

Processing and Information Retrieval," stated that Information retrieval (IR) systems in search engines, search a collection of natural language documents with the purpose of returning only the documents that were relevant to a user's query. Moreover, the researcher stated the difference of IR systems from database and expert systems. IR systems, according to the researcher, work with unstructured natural language text, as opposed to database systems, which require highly-structured data and formal semantics. Also, IR systems, unlike expert systems, do not try to derive or synthesize specific solutions, instead of returning documents with information that was similar to the query.

Today, IR systems as stated by Baeza-Yates & Ribeiro-Neto (1999), "Modern Information Retrieval," have changed considerably in the last years with the expansion of the Internet and the emergence of modern and inexpensive graphical user interfaces and mass storage devices. Multimedia devices must already provide excellent and intelligent search functionality due to the massive increase in stored digital content.

As aforementioned, an IR system retrieves information that might be useful to the user. With this, according to the study by Mahmud (2013) entitled "A Simple Information Retrieval Technique," it was stated that to be effective in its attempt to meet information the user needs, the IR system must somehow interpret the contents of the documents in a collection and rank them according to their degree of relevance to the user's query. Thus, it could be deduced that an IR system's primary goal was to return all relevant documents to a user's query while retrieving as few non-related documents as possible.

The degree of relevance was calculated with similarity measures. Similarity measures, as defined in the study by Lesot et al. (2009) entitled "Similarity measures

for binary and numerical data: a survey," were functions that measure the degree to which two objects were alike. Further, the researcher stated that text mining tasks or information retrieval require a similarity measure, and that the choice of a similarity measure for a task was critical as there was no universal similarity measure for all tasks.

There were various similarity measures, with the Euclidean distance and the cosine similarity measures being the most popularly used. Euclidean distance as stated in the study by Huang (2008) entitled "Similarity Measures for Text Document Clustering," was a standard metric for geometrical problems. It was the ordinary distance between two points and could be easily measured with a ruler in two- or three-dimensional space. Euclidean distance was widely used in clustering problems, such as text clustering. The K-Means algorithm uses it as the default distance metric.

On the other hand, cosine similarity as stated by Rahutomo et al. (2012), "Semantic Cosine Similarity," was a widely implemented metric in information retrieval. Using this model, the similarity between two documents could be measured by calculating the cosine value or angle between the term vectors of two documents. In search engines, the similarity values between the user's query and documents were sorted from the highest one to the lowest one, which serves as the ranking of the search results. The higher similarity score between the document's term vector and query's term vector means more relevancy between document and query.

But, before cosine similarity could be calculated, documents were first represented as term vectors in a Vector Space Model (VSM), using TF-IDF where terms were weighted based on their importance in a document. As pointed out by Alodadi & Janeja (2015), in their study entitled "Similarity in Patient Support Forums," documents were converted from textual format to vectors, with each document corresponding to a

point in space or a vector in vector space. To create the vector space, the TF-IDF weighting scheme was used wherein each document was represented as a list of terms with a given weight.

In evaluating information retrieval (IR) systems, according to the study by Arora et al. (2016) entitled “Evaluation of information retrieval: precision and recall,” the recall and precision metrics, which were basic measures, were used to test the efficacy of information retrieval (IR) systems. Further, the researchers stated that basically, the evaluation of IR systems revolves around the notion of relevant and non-relevant documents. With this, human decision-making was used to decide if a document was relevant or not to a user's query. However, the measures, recall, and precision, need to be extended with other metrics to evaluate ranked information retrieval systems.

Precision and recall as further discussed in the study by Sathya Bama et al. (2015) entitled “A Survey on Performance Evaluation Measures for Information Retrieval System,” were the two main evaluation metrics for unranked information retrieval systems. Recall, the proportion of real positive cases that were correctly predicted positive, was the ratio of the number of relevant records retrieved to the total number of relevant records in the database. Recall alone was not enough since it does not care about the irrelevant documents retrieved. Thus, precision was also used. Precision, the proportion of predicted positive cases that were correctly real positives, was the ratio of the number of relevant records retrieved to the total number of irrelevant and relevant records retrieved, with relevant documents that were not retrieved receiving a precision score of zero.

In order to really evaluate ranked information retrieval systems and take into account the ranking or relative order of the documents retrieved by search engines, the aforementioned metrics were used with Average Precision (AP) and Mean Average Precision (MAP). Trotman (2005), "Learning to Rank," stated that AP for a single query was computed as the sum of precisions divided by the number of relevant documents in the ranked list.

MAP was then, as stated in the study Turpin & Scholer (2006) entitled "User Performance versus Precision Measures for Simple Search Tasks," the mean of average precision scores over a set of queries. Moreover, it assumes users were interested in finding many documents for each query. Thus, it could be inferred that this was the most commonly used measure in research papers.

Clustering

Clustering, or unsupervised classification, as stated in the study by Abbas (2008) entitled "Comparisons Between Data Clustering Algorithms," was a division of data into groups of objects that were similar. Each group, called a cluster, consists of objects that were similar between themselves and dissimilar compared to objects of others from other groups or clusters. Moreover, in clustering or unsupervised classification, the data provided were not labeled, and the task of the clustering algorithm was to create labels with clusters, but the category labels were data-driven, as opposed to supervised classification, such as Naïve Bayes, where the data provided were labeled, and the algorithm was tasked to label a newly encountered, yet unlabeled, pattern.

There were two very basic clustering algorithms, which were Hierarchical Clustering and K-Means algorithms, as stated in the study by Sonagara & Badheka (2014) entitled “Comparison of Basic Clustering Algorithms.”

Hierarchical Clustering (HC), as defined in the study by Xu & Wunsch (2005) entitled “Survey of Clustering Algorithms,” organizes data into a hierarchical structure according to the proximity matrix. The results of HC were usually depicted by a binary tree or dendrogram. The root node of the dendrogram represents the whole data set and each leaf node was regarded as a data object.

Hierarchical Clustering (HC), based on the study by Kameshawaran & Malarvizhi (2014) entitled “Survey on Clustering Techniques in Data Mining,” were categorized into agglomerative (bottom-up) and divisive (top-down). Agglomerative clustering starts with one-point (singleton) clusters and merges two or more of the most comparable clusters in a recursive manner, while divisive clustering begins with a single cluster containing all data points and splits the most appropriate cluster recursively. The process was repeated until a criterion for stopping was met.

However, Hierarchical Clustering has its own weaknesses. Kaushik, M., & Mathur (2014), “Comparative Study of K-Means and Hierarchical Clustering Techniques,” pointed out that the performance of HC was less as compared to K-Means, and that despite the weaknesses K-Means may present, K-Means was still the most popular algorithm of the two due to its simplicity and efficiency.

K-Means, on the other hand, as defined in the study by Wagstaff et al. (2001) entitled “Constrained K-Means Clustering with Background Knowledge,” was a method used commonly to automatically partition data into *k* groups. However, the *k* number of groups or clusters was selected initially before the algorithm was applied.

The algorithm of K-Means was composed of a series of steps. According to the study by Dehariya et al. (2010) entitled "Clustering of Image Data Set Using K-Means and Fuzzy K-Means Algorithms," first, *k* points were selected as initial group centroids in the space the objects being clustered. Second, each object was assigned to the group that has the closest centroid. Third, when all objects have been assigned, the positions of the centroids were recalculated. Last, steps 2 and 3 were repeated until the centroids no longer move. This results in the objects being divided into groups.

In calculating the distance between objects and the closest centroid, a distance or similarity measure was used. As stated in the study by Singh & Singh (2012) entitled "Performance Evaluation of K-Means and Hierarchical Clustering in Terms of Accuracy and Running Time," K-Means clustering uses the Euclidean distance method to calculate to the distance between each data object and the cluster centers or centroids.

As aforementioned, the number of clusters for K-Means was required to be specified. According to Pham et al. (2005), "Selection of K in K-Means Clustering," typically, a number of iterations were needed to acquire a decent clustering result, in which the user executes the algorithm with different values of *k.*

In evaluating the K-Means algorithm and determining the most suitable *k* number of clusters, a method called the Silhouette Coefficient could be used. According to the study by Kumar (2020) entitled "Silhouette Method — Better than Elbow Method to find Optimal Clusters," it was stated that to get the silhouette score, the Silhouette Coefficient method computes the silhouette coefficients for each point and averages them across all samples. The silhouette score, according to the researcher, was a measure of how similar an object was to its own cluster (cohesion) compared to other clusters (separation). The value of the silhouette ranges between [1, -1], where a high

value indicates that the object was well matched to its own cluster and poorly matched to neighboring clusters.

K-Means clustering works well with text documents. In the study by Jacob & John (2016) entitled “Improved Clustering of Documents using K-Means Algorithm,” it was shown that K-Means could be used in clustering texts or documents into their relevant clusters. The texts were preprocessed, and then the terms were weighted using TF-IDF from the Vector Space Model (VSM) for each document and give more weight to the term with less frequency and high importance in the clusters. K-Means provides a faster way to create clusters from a set of random documents. It calculates the vector value for each document from the vector space model and based on their values, new clusters were formed.

In further application of the K-Means algorithm, Himel et al. (2017), “Weight Based Movie Recommendation System Using K-Means Algorithm,” developed a recommendation system that recommends movies based on the data of the users with the use of the K-means algorithm. Basically, it takes the user’s activity as the data and based on that data, it recommends the movies to the user. The total weighted value of each movie was first calculated before the K-Means algorithm was applied on those weighted values.

Text Mining

Text mining, as defined in the study by Waegel (2006) entitled “The Development of Text Mining Tools and Algorithms,” was a form of information retrieval (IR). It was defined as the extraction of implicit, previously unknown, and potentially useful information from textual data. Also, text mining uses computational

linguistics and natural language processing (NLP), as well as typical data mining, machine learning (ML), and statistics.

In clarification, as stated by Gupta & Lehal (2009), “A Survey of Text Mining Techniques and Applications,” text mining was not the same as what we're used to in online search. In most cases, when someone searches for anything, they were seeking for something that was already well-known and has been written by someone else. It was the automated extraction of information from various textual resources by a computer to uncover new, previously undiscovered information.

Baharudin et al. (2010), in their study entitled “A review of machine learning algorithms for text documents classification,” they stated that text mining studies have lately gained more importance as a result of the increased availability of electronic documents from a range of sources. The internet, governmental electronic repositories, news articles, biological databases, chat rooms, digital libraries, online forums, electronic mail, and blog repositories were all examples of unstructured and semi-structured information resources. Therefore, accurate categorization and knowledge discovery from these resources was an important area for research.

In the Philippines, according to the study by Cheng & Sharmayne (2020) entitled “Analysing digital banking reviews using text mining,” it was stated that the goal of the researcher’s work was to analyze digital banking application reviews using text mining approaches. With more consumers becoming technologically literate, financial institutions must design a digital banking application that would set them apart from the competitors. Topic modeling with LDA was used to investigate customer issues, and association rules between digital banking features and review score would be mined.

Machine Learning (ML)

Machine learning (ML), as defined in the study by Sah (2020) entitled "Machine Learning: A Review of Learning Types," was the study of computer algorithms that provides systems the ability to automatically learn and improve from experience. It was seen as a subset of artificial intelligence (AI). Further, based on the learning approach, the sort of data they input and produce, and the type of problem they solve, machine learning algorithms can be classified into primary categories, which include supervised, unsupervised, and reinforcement learning.

The difference in the ML categories mentioned above were stated in the study by Kotsiantis (2007) entitled "Supervised Machine Learning: A Review of Classification Techniques." In their study, it was stated that the learning was called supervised when the instances or dataset given contains known labels or corresponding correct outputs, while the learning was called unsupervised when the instances were unlabeled. Moreover, in reinforcement learning, the learner was not told which actions to take, but the learner must discover which actions provide the best reward by trying each action in turn.

In the study by Portugal et al. (2018) entitled "The Use of Machine Learning Algorithms in Recommender Systems: A Systematic Review," it was discussed that when supervised learning were provided with training data and correct predictions, the ML algorithm was tasked to learn based on the training data, and to apply the knowledge that was gained using real data. For instance, for book classification in a bookstore, a training set of the information about the books including their correct classifications was learned by the ML algorithm. When a new book arrives at the

bookshop, the algorithm could classify it using the knowledge it has learned about book classification.

Unsupervised learning, on the other hand, as stated by Celebi & Aydin (2016), in “Unsupervised Learning Algorithms,” the ML algorithms do not have a training set. They were given data about the real world and were expected to learn from it on their own. The goal of unsupervised learning algorithms was to discover hidden patterns in data. The most well-known subfield of unsupervised learning was clustering.

Lastly, reinforcement learning, as stated in the study by Bhatt (2018) entitled “Reinforcement Learning 101,” was a type of machine learning technique that allows an agent to learn by trial and error in an interactive environment using feedback from its own actions and experiences. Unlike supervised learning, where the agent receives feedback in the form of a correct set of actions for completing a task, reinforcement learning uses rewards and punishments to signify positive and negative behavior.

Machine learning, according to the study by Wellers et al. (2017) entitled “8 Ways Machine Learning Is Improving Companies Work Processes,” was capable of processing huge amounts of data that’s why many of today's leading companies, such as Facebook, Twitter, and Google make machine learning a central part of their operations in automating decision processes. Machine learning has become a crucial competitive differentiator for many companies and organizations.

In the field of education, Kučak et al. (2018), in their study entitled “Machine Learning in education a survey of current research trends,” stated that researchers and scientists were now interested in the use of machine learning since the majority of today's classrooms employ digital resources and were investing in machine learning, which might be used to support teachers and guide students.

There were various programming languages used for machine learning, but the most popular among data scientists, as stated in the study by Kumar (2019), "Python Vs R: What's Best for Machine Learning," were Python and R. Although they both share similar features and open-source, Python was designed to be a widely-used programming language, whereas R was designed for statistical analysis.

The two languages were further compared in the study by Odhiambo et al. (2020) entitled "An Analytical Comparison Between Python Vs R Programming Languages." In their study, it was stated that around 65% of all developers worldwide use Python for machine learning, compared to 25% of developers who use the R programming language. More importantly, Python was said to have a smooth learning curve in comparison to R, which was difficult to learn mainly because of its syntax was hard to read.

Software Development Model

Software Development Life Cycle (SLDC), as stated in the study by Leau et al (2012) entitled "Software Development Life Cycle AGILE vs Traditional Approaches," was a method for developing and managing software. It usually covers several steps, ranging from preliminary development analysis to software testing and evaluation after development. It also included the models and methodologies used by software development teams to create software systems, with the methodologies serving as a framework for planning and controlling the entire development process.

As aforementioned, software development teams used SDLC models. According to Mishra & Dubey (2013), "A Comparative Study of Different Software Development Life Cycle Models in Different Scenarios," there were various SDLC models widely used for developing software, with each having its own advantages and

disadvantages. For instance, traditional models, such as the most popular Waterfall model, was advantageous if the requirements were known beforehand and a full control of the project was necessary. The phase was not repeated and the development does not proceed to the next phase unless and until the previous phase was fully accomplished. As a result, it wasn't particularly useful when the project requirements were dynamic.

Next, in incremental models, according to Stoica et al. (2013), "Software Development: Agile vs Traditional," requirements were divided into subsets, and there were multiple development cycles involved, with each increment involving the building of a module and the addition of additional features in a new version until the system was complete. However, it requires good planning and design. Thus, this model was only recommended to be used when system requirements were clearly defined and understood, before the system was built incrementally.

Moreover, iterative models, based on the study by Bhuvaneswari & Prabaharan (2013) entitled "A Survey on Software Development Life Cycle Models," was a model in which development begins with the specification and implementation of a part of the software, which may then be reviewed to identify additional requirements. Less time was spent documenting in this model, and more time was allocated to designing. The drawback, therefore, was that design issues may arise because not all requirements were gathered upfront for the entire life cycle.

Agile software development methodology, as opposed to the traditional software development methodologies stated above, was defined in the study by Szalvay (2004) entitled "An Introduction to Agile Software Development," as a software development methodology based on the idea of incremental and iterative development, in which different phases of a development life cycle were revisited over and over

again. It converges on solutions iteratively by using consumer feedback to enhance the software.

Agile development consists of various methodologies, but SCRUM, as stated by Mahalakshmi & Sundararajan (2013), “Traditional SDLC vs Scrum Methodology – A Comparative Study,” was the most famous and the most powerful methodology which provides benefits to software development companies, since traditional SDLC was not able to meet the requirements of the market.

In addition to the study above, Ashraf & Aftab (2017), “Latest transformation in SCRUM: A state of the art review,” pointed out that in the software industry, SCRUM has been acknowledged as one of the most promising and widely used agile process models. Its contribution to enhanced productivity, improved collaboration, quick reaction to altering market needs, and faster delivery of quality products was the reason for its widespread recognition.

SCRUM, according to the study by Schwaber & Sutherland (2017) entitled “The scrum guide,” was a process framework that has been used to manage work on complex products since the early 1990s. It was a framework within which various processes and techniques were employed. SCRUM makes clear the relative efficacy of the product management and work techniques so that the product, the team, and the working environment could continuously improve.

Furthermore, SCRUM, according to the study by Sharma & Hasteer (2016) entitled “A Comprehensive Study on State of Scrum Development,” was a lightweight framework that was easy to understand and capable of managing complex product development under time constraints. The software product was produced in SCRUM in a series of iterations or increments over the course of a sprint (usually 30 days).

Since SCRUM allows time constraints and collaboration between stakeholders, Srivastava et al. (2017), "SCRUM Model for Agile Methodology," stated that it increases the speed of development with the successive and incremental features to be developed, and it aligns the mottos of individual and organizations and supports stakeholders value creation.

Related Systems

Table 1. Features comparison of existing systems.

FEATURES	PAPERNERDS	GOOGLE CLASSROOM	GOOGLE SCHOLAR	SCIENCE DIRECT
Accessible online	✓	✓	✓	✓
Security recovery	✓	✓	✗	✗
Documents/articles /books library	✓	✗	✓	✓
Student group members suggestion	✓	✗	✗	✗
Technical panel members suggestion	✓	✗	✗	✗
Related documents recommendation	✓	✗	✓	✓
Undergraduate thesis process	✓	✗	✗	✗
Download/upload documents	✓	✓	✗	✓
Semantic search engine	✓	✗	✓	✓
Highlight specific words in search results	✓	✗	✓	✓
Responsive display for different screen sizes	✓	✓	✓	✓

Google Classroom by Google Inc. was an online classroom that allowed participants to communicate with one another, view presentations or videos, interact with other participants, and engage with resources in workgroups. It was announced by

Google in May 2014 as a new tool in Google Apps for Education. In its initial release, more than 30 teachers at Daffodil International University started using Google Classroom in September 2014.

Google Scholar by Google Inc. was a subset of Google's larger search index that includes full-text journal articles, technical reports, preprints, theses, books, and other scholarly resources, as well as scholarly Web pages. Science, particularly medical science, looks to be the strongest, whereas social science tends to be the weakest. The company claimed to have full-text literature from all major publishers, as well as hosting companies like Highwire and Ingenta, with the exclusion of Elsevier and the American Chemical Society.

ScienceDirect, operated by the British-Dutch publisher Elsevier and launched in March 1997 to provide a database and web host environment for its journal and those of other participating publishers. The website provides access to a large bibliographic database of scientific and medical publications. It hosts over 18 million pieces of content from more than 4,000 academic journals and 30,000 e-books of this publisher. It also aimed to create an integrated environment, which would permit efficient and transparent access to the scientific literature by integrating abstracting and indexing searching and by linking to other publishers' sites.

Synthesis

Digital libraries had been around for a long time. Cabrerizo et al. (2015) stated that it had proven to be beneficial in providing information between students, researchers, etc., especially with the use of the World Wide Web (WWW) through the Internet. In fact, Watstein et al. (1999) stated that even the WWW was a digital library since it was a collection of web pages. It was essential to be developed, because for

instance as stated by Marchionini (1995), digital libraries served as a sharing platform in institutions to serve books, periodicals, and even videos.

The implementation of a web portal was deemed necessary for the researchers' study since according to Eldrandaly (2009), it made it easier for multiple users to access information while connecting them to a single source of information, data, and services. Further, Mansourvar & Yasin (2010) stated that web portals had become increasingly crucial in institutions because of students' management of information online. For instance, Del Rosario et al. (2016) developed a thesis portal for their university, which covered the thesis process of different thesis documents.

Beyond system development, the researchers wanted to integrate the use of text mining with machine learning (ML) algorithms. Waegel (2006) defined text mining as the extraction of potentially useful information from textual data. Baharudin et al. (2010) stated that it had gained more importance as a result of the growth of electronic documents from various sources. Machine learning, as defined by Sah (2020) was the study of computer algorithms that provided systems the ability to automatically learn and improve from experience, and it was classified into primary categories: supervised, unsupervised, and reinforcement learning. Unsupervised learning, in which the data given were usually about the real world according to Celebi & Aydin (2016), was used by the researchers because supervised learning, according to Kotsiantis et al. (2007), required a dataset with known labels and reinforcement learning, as stated by Bhatt (2018), learns by trial and error.

Table 2. Findings in selecting the best topic model.

NAME	FINDINGS
LSI / LSA	Not an authentic topic model
pLSI / pLSA	Runs into overfitting
LDA	A generative probabilistic topic model

Several topic models were considered by the researchers from various studies. Rosario (2000) stated that LSI, the origin of topic models, was not an authentic topic model since it was not a probabilistic topic model. Probabilistic topic models include pLSA and LDA. However, Cai et al. (2008) stated that pLSA ran into overfitting which occurs when a model learned the detail and noise in the training data to the point that it would lower the model's performance on new data. Therefore, LDA, as defined by Blei et al. (2003), as a generative probabilistic topic model, which would the issues encountered with pLSA, was the chosen by the researchers.

Table 3. Findings in selecting the best similarity measure.

NAME	FINDINGS
Euclidean distance	Only used for clustering
Cosine Similarity	Usually used for IR systems

There were two most commonly used similarity measures used for different tasks according to various studies. Lesot et al. (2009) pointed out that IR systems require similarity measures usually for search results ranking. Euclidean distance, as defined in the study by Huang (2008), was a similarity measure usually only used in clustering problems, where distance between two points was calculated. On the other hand, cosine similarity as stated by Rahutomo et al. (2012), was the metric usually used in IR systems, where the cosine value or angle between the term vectors of two documents was calculated. Therefore, cosine similarity was chosen.

Table 4. Findings in selecting the best clustering algorithm.

NAME	FINDINGS
Hierarchical Clustering	Represents results in a dendrogram
K-Means	Better performance than Hierarchical Clustering

Studies of different clustering algorithms reveal which was best for clustering tasks. Hierarchical Clustering, as pointed out by Xu & Wunsch (2005), represented results in a binary tree or dendrogram. On the other hand, K-means, the most popular clustering algorithm, was discussed by Wagstaff et al. (2001) as a method used to divide data into *k* groups. In comparison in terms of performance, Kaushik & Mathur (2014) pointed out that the performance of K-means was better than Hierarchical Clustering.

In evaluating the algorithms, several studies regarding the topic modeling, clustering, information retrieval, were discussed. For the LDA topic modeling, it was evaluated based on perplexity and topic coherence. Jacobi et al. (2016) defined perplexity as a metric for how well the model predicts a sample, with a lower perplexity indicating a better prediction. Topic coherence, on the other hand, was a metric that helped in distinguishing between topics that were semantically interpretable, and thus be able to identify the optimal number of *k* clusters.

Moreover, information retrieval was evaluated based on recall, precision, average precision (AP), and mean average precision (MAP). Sathya Bama et al. (2015) stated that in calculating recall, the number of relevant records retrieved was divided by the total number of relevant records. Further, in calculating precision, the number of relevant records retrieved was divided by the total number of irrelevant and relevant records retrieved. AP and MAP metrics were used to consider the ranking of the search results. Trotman (2005) defined AP for a single query as the sum of precisions divided

by the number of relevant documents in the ranked list. On the other hand, Turpin & Scholer (2006) defined MAP as the mean average scores over a set of queries.

Next, the K-means clustering algorithm was evaluated based on the silhouette coefficient method. Kumar (2020) stated that the method would determine the suitable number of *k* number of clusters since it was a measure of how similar an object is to its own cluster compared to other clusters.

Table 5. Findings in selecting the appropriate software development methodology.

TYPE	NAME	FINDINGS
Traditional Software Development	Waterfall	Phases cannot be repeated
	Incremental	Each increment involves the building of a module
	Iterative	Starts the implementation, and identifies additional requirements after
Agile Software Development	SCRUM	Combination of incremental and iterative development

In selecting the software development model, the SCRUM software methodology was found to be the most appropriate for the study of the researchers. This development model, as stated by Szalvay (2004), was a type of agile software development model, which was based on the idea of incremental and iterative development. The Waterfall model wasn't chosen even though it's relatively popular since according to Mishra & Dubey (2013), the development could not proceed to the next phase unless and until the previous phase was fully accomplished.

CHAPTER III

RESEARCH METHODOLOGY

This chapter would present the methodologies used to accomplish the objectives of the study. It included the research design, population of the study, fact-finding techniques, design and development of the system, and evaluation and examination of the algorithms and their performance respectively.

Research Design

The researchers of this study used descriptive research and developmental research design methods. As stated by Fireman Kramer (1985), descriptive research was regarded as the initial step in the development of new knowledge that described the characteristics of variables under the study that leads to a tentative hypothesis. This method was used by the researchers to gather the information that systematically described the study through interviews. The researchers, therefore, were able to thoroughly investigate the background of the research problem before further research was carried out.

On the other hand, developmental research, as defined by Richey (1994) as the systematic study of designing, developing, and evaluating instructional programs, processes, and products that must meet the criteria of internal consistency and effectiveness, was used by the researchers to determine the features to be developed for the system. Also, it was used to examine the consistency and efficacy of the development of the system.

Population of the Study

The target users of the study were composed of various stakeholders involved in the undergraduate thesis of the College of Computer Studies (CCS) department and other departments in Laguna State Polytechnic University SCC.

Specifically, the stakeholders of the study included the Research Implementing Unit Head (RIUH), the Dean/Associate Dean, the student researchers, the course instructors, the research advisers, and the subject specialists of the CCS. Additionally, the language critics, the technical editors, and the statisticians of the College of Arts and Sciences (CAS), the College of Teacher Education (CTE), the College of Engineering (COE), and the CSS, were also included. Lastly, the stakeholders of the study were also composed of guest users.

Fact-Finding Techniques

The researchers used various fact-finding techniques, such as interviews, and online research for the collection of data and information which were critical for the study.

Designing the Application

This section of designing the application discusses the similar methods used by other researchers in carrying out tasks with the algorithms used in this study.

LDA Topic Modeling

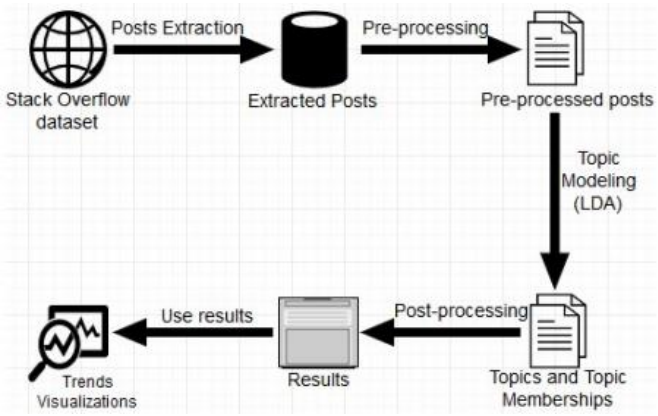


Figure 3. LDA to examine linguistic trends.

Based on the study by Johri & Bansal (2018), textual data from a famous question and answer website called StackOverflow was analyzed using the Latent Dirichlet Allocation (LDA) topic modeling algorithm. In their design, it was seen that data needed to be extracted first and then preprocessed before being subjected to a topic model. Next, the output of the LDA was processed in order to better visualize the trends.

K-Means Clustering

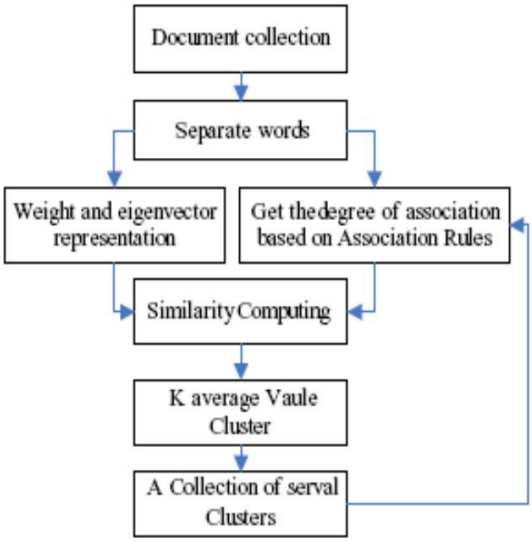


Figure 4. K-means to cluster text documents.

According to Gupta & Lehal (2009), the K-means clustering algorithm for the calculation of the similarity between text documents, took into account not only the eigenvector based on the algorithm of term frequency statistics, but also the degree of association between words, and then the relationship between keywords. This reduced the sensitivity of the input sequence and frequency, and to some extent, it considered semantic understanding, effectively raising the similarity accuracy of small text documents.

Information Retrieval

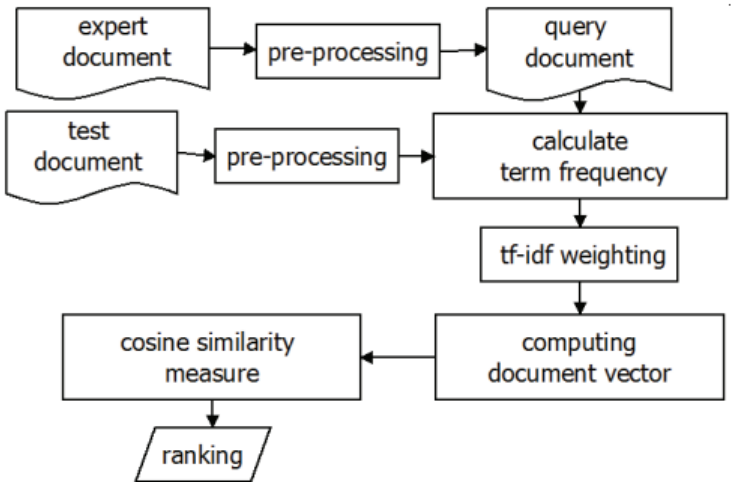


Figure 5. Information retrieval for ranked search results.

According to Lahitani et al. (2016), to test the similarity of a document, text-based documents were first pre-processed and then generated terms used as keywords. Along with the query document (q), the number of terms contained in each document (TF) and the total number of times each term contained in the document throughout (DF) were calculated. Each document's whole term was weighted using the TF-IDF method. After each document test was weighted, the next step was to calculate the vector lengths of each document, and then the weight of the term lock on the query document (q) with the weight terms

contained in the document were weighted to determine the level of similarity performed on each document.

Developing the Application

This method of application development enumerated the data collection that covers the fact-finding techniques, the algorithm analysis, the data model generation, and the development methodology.

Data Collection

Interview

The researchers conducted short interviews with the Research Implementing Unit Head (RIUH) of the college through online video and chat consultations whenever classes were held. The researchers were able to collect the necessary information needed for this stud, such as the undergraduate thesis process and stakeholders involved. Furthermore, the researchers were also able to identify the needs of the client and the potential benefits of this study to undergraduate students. Based on the identified needs and wants of the client, the researchers were able to determine the key features and functionalities of the developed system.

Moreover, the researchers were able to collect critical datasets for machine learning models, such as research papers and list of titles of the alumni BSIT and BSCS students of the College of Computer Studies (CCS). With this, the researchers were able to extract research abstracts from 2021 and titles from 2018-2021 which were used for training. However, the number of abstracts collected from the research papers wasn’t enough for effective topic modeling

as lack other topics related to the undergraduate programs for Computer Science (CS) and Information Technology (IT).

Table 6. Datasets of abstracts and titles from interviews.

PROGRAM	ABSTRACTS	TITLES
IT	67	104
CS	12	19
TOTAL	**79**	**123**

The total abstracts collected from the research papers of IT and CS were seventy-nine (79), while the total titles collected were one hundred twenty-three (123).

Online Research

The researchers made use of the vast number of resources available on the Internet to learn and acquire the necessary resources and information needed in the development of the web application. Furthermore, the Internet was also utilized by the researchers to gather ideas, concepts, information, and processes of pre-existing software and applications which were related to this study. Since the Internet had a vast amount of information, the researchers intensively selected and filtered the information to extract the needed information to formulate new ideas and concepts in the development process of Papernerds and integrated them with the existing ones to determine the features and functionalities of the system.

Moreover, as stated in the fact-finding technique above, the abstracts gathered weren't sufficient for effective topic modeling. Therefore, the researchers further collected nine hundred twenty-one (921) abstracts online

which were related to the programs offered by the college in addition to the abstracts already collected for a total of one thousand (1000) abstracts. An abstract from a research paper was deemed sufficient for an abstract was simply a short summary of a completed research manuscript. It was intended to describe the major aspects of the entire paper without going into great detail, and it contained the introduction, significance, methodology, and the results of the study which were enough to cluster the research papers of undergraduate students into different topics.

Lastly, the researchers utilized online research and created a training set of around fifty (50) keywords regarding the skills of students in an undergraduate thesis in order for the clustering algorithm to provide a recommendation of possible group members of students.

Table 7. Datasets of abstracts and skills keywords from online research.

ABSTRACTS	SKILLS KEYWORDS
921	50

The total abstracts gathered online from various sources, such as IEEE Xplore and Google Scholar, were nine hundred twenty-one (921), while the total keywords of skills gathered from various articles were fifty (50).

Algorithm Analysis

This section of algorithm analysis presented the methodologies of the algorithms and techniques used by the researchers, enumerating the LDA topic modeling, the TF-IDF term weighting scheme, the cosine similarity measure, and the

K-means clustering algorithm. It included their architecture and their advantages and disadvantages for a better representation of their general behavior.

LDA Topic Modeling

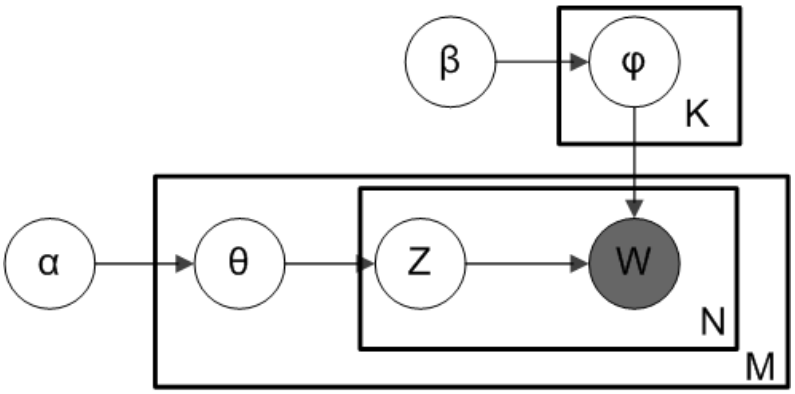


Figure 6. LDA plate diagram.

Where: α was the per-document topic distributions, β was the per-topic word distribution, θ was the topic distribution for document *m*, φ was the word distribution for topic *k*, z was the topic for the *n*th word in document *m*, and was the specific word.

LDA, the most popular topic modeling technique, as defined by Blei et al. (2003), was a generative probabilistic model of a corpus or text. The basic idea as they stated was that documents were represented as random mixtures over latent topics, each of which was defined by distribution over words. In other words, LDA assumed that every document was a distribution over topics and every topic was a distribution over words.

The advantages of LDA included being a probabilistic model with interpretable topics, and it didn't need to know topics in advance. On the other hand, its disadvantages included the inability for representation of the

relationship among topics, and it required the number of topics to be specified in advance.

TF-IDF Term Weighting Scheme

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

Figure 7. TF-IDF formula.

Where, according to Barysevich (2019), $tf_{i,j}$ is the number of occurrences of *i* in *j*, df_i was the number of documents containing *i*, and *N* was the total number of documents.

TF-IDF, or Term Frequency–Inverse Document Frequency, as defined by Soucy & Mineau (2005), was a numerical statistic in information retrieval that was meant to reflect how essential a term was to a document in a collection or corpus. It was the most common weighting method used to describe documents for the Vector Space Model (VSM) to get two-real valued vectors, particularly in Information Retrieval tasks.

Basically, TF-IDF, according to Van Zaanen & Kanters (2010), was the product of Term Frequency (TF) and Inverse Document Frequency (IDF).

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{i,j}}$$

Figure 8. Term Frequency (TF) formula.

TF which according to Agrawal & Gupta (2014), measured how frequently a term occurs in a document. Since every document was different in length, a term may appear much more frequently in large documents than in

shorter ones. Thus, the term frequency was often divided by the total number of terms in the document as a way of normalization.

$$idf_j = log\left[\frac{n}{df_j}\right]$$

Figure 9. Inverse Document Frequency (IDF) formula.

On the other hand, according to Wynne & Wint (2019), IDF measures how important a term is. While computing TF, all terms were considered equally important. However, it was known that certain terms, such as “is”, “of”, and “that”, may appear a lot of times but have little importance. Thus, it was needed for the weight of rare words across all documents in the corpus to be calculated. The words that occurred rarely in the corpus received a high IDF score.

The advantages of the TF-IDF term weighting scheme included being a basic metric in extracting the most descriptive terms in a document, the similarity between documents could be computed easily, and it could be used for stop words removal. Its disadvantages, however, include making no use of semantic similarities between words, and it was only useful as a lexical level feature.

Cosine Similarity Measure

$$cos(\theta) = \frac{A \cdot B}{\|A\| \, \|B\|}$$

Figure 10. Cosine similarity formula.

According to Asenova & Chrysoulas (2019), θ was the angle between vectors *A* and *B.* As θ ranged from 0 to 90 degrees, cos θ ranged from 1 to 0. Since TF-IDF vectors were non-negative, θ ranged from 0 to 90 degrees only.

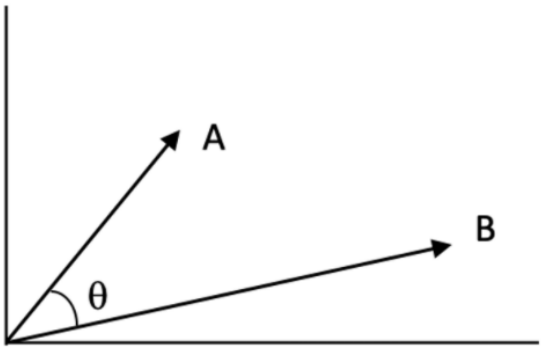


Figure 11. Angle of vectors A and B.

Cosine similarity, as defined by Rohilla et al. (2021), was a metric used regardless of their size to measure how similar the documents or texts were. Mathematically, it measured the cosine of the angle between two vectors projected in a multi-dimensional space. The lower the angle, the greater the resemblance to cosine.

The documents or texts, according to Huang (2008), were represented as term vectors. Further, the similarity of the two documents corresponded to the correlation between the vectors. The cosine of the angle between vectors, or cosine similarity, was used to quantify this. Cosine similarity was one of the most popular similarity measures applied to text documents, such as in numerous information retrieval applications and clustering.

The advantages of cosine similarity included being a good method for calculating the similarity between two vectors. Also, it had meaningful

semantics for ranking similar documents. However, the only disadvantage was that it didn't account for the fact that various users had different rating scales.

K-Means Clustering

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2$$

Figure 12. K-means formula.

Where, according to Yuan & Yang (2019), *k* represents the number of cluster centers, c_j represents the k^{th} center, and x_i represents the i^{th} point in the dataset.

Moreover, the algorithm, as stated by Yi et al. (2010), aimed at minimizing an objective function, in this case, a squared-error function, where $\| x_i - c_i \|^2$ was a chosen distance measurement (Euclidean distance) between data point x_i and the cluster center c_i. This criterion tried to make the resulting *k* clusters as compact and as separate as possible

K-means, according to Sahu et al. (2009), was shown to be effective in producing good clustering results for many practical applications. It was attractive in practice because it was simple and it was generally very fast. It partitions the input dataset into *k* clusters. Each cluster was represented by an adaptively-changing centroid, starting from some initial value. The squared distances between the input data points and centroids are computed, and the inputs to the nearest centroid were assigned.

Since the algorithm's proposal three decades ago, Zhao et al. (2018) stated that it still remained popular due to its simplicity and versatility.

The advantages of K-means clustering were that it was easy to implement, it could warm-start the position of centroids, and it could easily adapt to new examples. On the other hand, its disadvantages were that it required the number of clusters to be specified in advance, it couldn't handle data and outliers, and it was not suitable to identify clusters with non-convex shapes.

Data Model Generation

This section of data model generation enumerated the methodologies in generating the data models of the algorithms starting with data preprocessing and then followed by evaluation of LDA and K-means, topic modeling, information retrieval, and clustering.

Data Preprocessing

Preprocessing of the text data was an essential step for text mining tasks. It allowed text to be transformed into a more digestible form so that machine learning algorithms could perform better. It helped to get rid of unhelpful parts of the data, or noise, through different techniques such as the conversion of all characters to lowercase, removal of stop words and punctuations, tokenization, stemming, lemmatization, and vectorization.

Lowercasing

Although often ignored, Tuncer (2020), stated that lowercasing text data was one of the simplest and most effective forms of text preprocessing. It could

be used to solve most text mining problems, and it could be especially useful in significantly helping with the consistency of expected output.

Stop Words Removal

Stop words, according to Teha (2020), were words in English that didn't contribute much meaning to a statement, such as "a", "the", "is", and "are". They could be safely ignored without compromising the sentence's meaning.

Punctuation Removal

Punctuation removal was the removal of punctuations in a string. According to Ebrahim (2021), eliminating punctuation was a typical preprocessing step in many data analysis and machine learning tasks.

Stemming

Sawhney (2021) defined stemming as the process of reducing a word to its stem or root format. For instance, the terms, "branched," "branching," and "branches" could all be reduced to the same word "branch."

Tokenization

It is defined by Kannan et al. (2014) as the process of breaking a stream of text into words, phrases, symbols, or other meaningful elements called tokens. The aim of tokenization was the exploration of the words in a sentence. The list of tokens would become the input for further processing such as parsing or text mining.

Lemmatization

According to Kotov (2017), lemmatization usually referred to doing things properly with the use of a vocabulary and morphological analysis of words, normally aimed to remove inflectional endings only and return the base or dictionary form of a word, which was known as the lemma.

Vectorization

In vectorization, according to Brownlee (2017), text data were encoded as integers or floating-point values for use as input to a machine learning algorithm.

LDA and K-Means Evaluation

In accomplishing the second specific objective of the study, an evaluation of the algorithms was carried out in order to identify which algorithm from LDA and K-means was more appropriate to integrate into the system for research papers recommendation and student group and technical panel member suggestions.

	abstract
0	Government is one of the many who ought to adv...
1	Documents are one of the most important things...
2	The study entitled “the GEARS Online Publicati...
3	The study entitled “Blue Card Management Infor...
4	The study of Daily Diagnosis of Calauan Rural ...

Figure 13. First five (5) data of the twenty (20) abstracts dataset.

A sample dataset of twenty (20) abstract paragraphs that would be used to train the models was prepared by the researchers in order to carry out the evaluation. The above figure showed the first five (5) abstracts data from index 0 to 4.

	skills_keywords
0	programming code
1	research paper literature
2	search literature
3	write literature
4	code frontend design

Figure 14. First five (5) of the twenty (20) skills dataset.

A sample dataset of twenty (20) skills keywords that would be used to train the models was prepared by the researchers in order to carry out the evaluation. The above figure showed the first five (5) skills data from index 0 to 4.

LDA Topic Model Testing

The researchers tested the LDA topic model on the datasets prepared based on the perplexity and coherence scores on each of them.

Perplexity

$$per(D_{test}) = exp\{-\frac{\sum_{d=1}^{M} \log p(\mathrm{w}_d)}{\sum_{d=1}^{M} N_d}\}$$

Figure 15. Perplexity score formula.

In the figure above, *M* was the number of documents (in the test sample, presumably), *wd* represented the words in document *d, and* N_d was the number of words in document *d*.

It was seen as a good measure of performance for LDA. It captured how surprised a model was of new data it has not seen before,

and it was measured as the normalized log-likelihood of a held-out test set.

Coherence

$$CoherenceScore = \sum_{i<j} score(w_i, w_j)$$

Figure 16. Coherence score formula.

Where *wi,wj* are the top words of the topic.

It measured the score of a single topic by measuring the degree of semantic similarity between high-scoring words in the topic. These measurements helped distinguish between topics that were semantically interpretable topics and topics that were artifacts of statistical inference. A set of statements or facts was said to be coherent if they support each other. To decide the optimum number of topics to be extracted using LDA, a topic coherence score was always used to measure how well the topics are extracted.

According to Sara (2019), a coherence score of 0.30 was bad, 0.40 was low, 0.55, was okay, 0.65 was good, 0.70 was nice, 0.80 was unlikely, and 0.90 was probably wrong.

K-Means Cluster Testing

The clustering algorithm was evaluated based on homogeneity, completeness, and V-measure in order to determine how well it would work on the two types of datasets prepared by the researchers.

Homogeneity

$$h = 1 - \frac{H(C|K)}{H(C)}$$

Figure 17. Homogeneity score formula.

It was a measure of the ratio of samples of a single class pertaining to a single cluster. The fewer different classes included in one cluster, the better. The lower bound should be 0.0 and the upper bound should be 1.0, with 1.0 representing a perfectly homogeneous clustering.

Completeness

$$c = 1 - \frac{H(K|C)}{H(K)}$$

Figure 18. Completeness score formula.

It was a measure of how much similar samples were put together by the clustering algorithm. A perfectly complete clustering was one where all data points belonging to the same class were clustered into the same cluster. The completeness score ranged between 0.0 and 1.0. A score of 1.0 stood for perfectly completeness labeling.

V-Measure

$$NMI = 2 * \frac{h * c}{h + c}$$

Figure 19. V-measure score formula.

It was the harmonic mean between homogeneity and completeness. This score was a measure between 0–1 that actually quantified the goodness of the clustering partition. In fact, it required that both homogeneity *h* and completeness *c* were maximized. V-measure or NMI was 1 when both *h* and *c* were 1. Moreover, if the clustering doesn't satisfy any of the two conditions V-measure would be zero.

Topic Modeling

According to Kayi et al. (2013), topic modeling was an unsupervised machine learning technique that could scan a series of documents, find word and phrase patterns within them, and automatically cluster word groups that best characterized the set. It was great for document clustering and organizing large blocks of textual data. Latent Dirichlet Allocation (LDA) was currently the most popular technique for topic modeling.

LDA, as stated by Niu et al. (2015), represented each document as a probability distribution over topics, and each topic as a probability distribution over words. Thus, according to Schwarz (2018), it was an alternative to predefined document classifications for analyzing the content of vast amounts of unclassified data. This major upside of the LDA technique was that one does not have to know in advance what the topics would look like. Basically, documents were represented by an LDA topic model as mixtures of topics that contain words with varying probability.

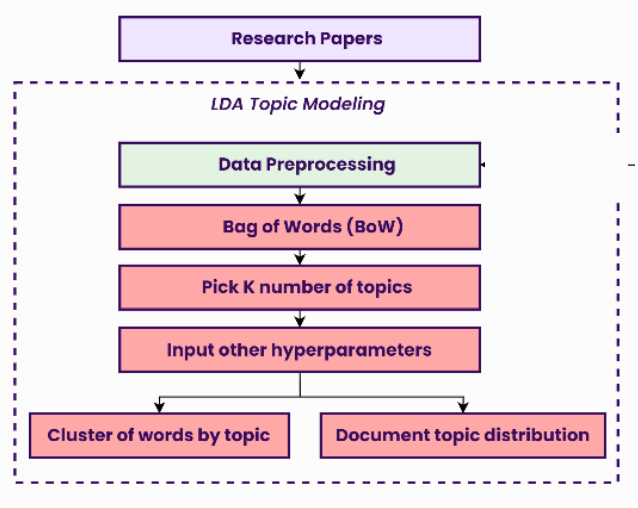


Figure 20. Topic model generation with LDA.

The methodology for topic modeling in the study started with the training of data which are one thousand (1000) research papers' abstracts as mentioned previously.

Data was then preprocessed with stop words removal and lemmatization. A Bag-of-Words (BoW) model was then created in order to represent the texts as vectors in a Vector Space Model (VSM). Next, generating the topic model, the *k* number of topics was specified, along with other hyperparameters of LDA. The output of the LDA model included the creation of topics and topic probability distributions over documents which both were represented in matrices.

```
[(0,
  '0.039*"be" + 0.032*"topic" + 0.025*"test" + 0.025*"model" + 0.021*"ordering" + 0.018*"chest" + 0.014*"segmentation" + 0.014
*"document" + 0.011*"system" + 0.011*"diagnosis"'),
 (1,
  '0.074*"be" + 0.035*"system" + 0.023*"model" + 0.020*"expert" + 0.019*"use" + 0.017*"month" + 0.017*"topic" + 0.010*"where" +
0.010*"work" + 0.009*"first"'),
 (2,
  '0.026*"method" + 0.026*"propose" + 0.017*"be" + 0.017*"use" + 0.017*"base" + 0.017*"image" + 0.017*"proposal" + 0.017*"huma
n" + 0.017*"tissue" + 0.017*"perform"'),
 (3,
  '0.061*"be" + 0.030*"base" + 0.027*"model" + 0.027*"service" + 0.025*"system" + 0.015*"topic" + 0.012*"present" + 0.012*"web"
+ 0.012*"method" + 0.012*"component"'),
 (4,
  '0.046*"be" + 0.045*"emergency" + 0.045*"decision" + 0.018*"model" + 0.018*"system" + 0.018*"problem" + 0.018*"propose" + 0.0
18*"general" + 0.018*"process" + 0.018*"method"'),
 (5,
  '0.037*"be" + 0.019*"fault" + 0.018*"time" + 0.017*"system" + 0.015*"period" + 0.015*"attention" + 0.012*"series" + 0.012*"sp
ecial" + 0.010*"model" + 0.009*"have"'),
```

Figure 21. LDA topics example.

An example of generated topics of an LDA model with word probabilities of the terms from most relevant to least was represented in the figure above from topic “0” to topic “5.”

```
[(0, 0.001176592), (1, 0.0011766074), (2, 0.0011765956), (3, 0.0011765879), (4, 0.0011765622), (5, 0.001176617),
0.0011765322)]
```

Figure 22. Document topic probability distribution.

Moreover, an example of topic probabilities within a document was shown in the figure above. Similar to the probabilities of keywords in LDA topics, documents were given a topic probability distribution.

Information Retrieval

Information retrieval (IR), as stated by Strzalkowski (1995), would search a collection of natural language documents with the purpose of returning only the documents that were relevant to a user's query.

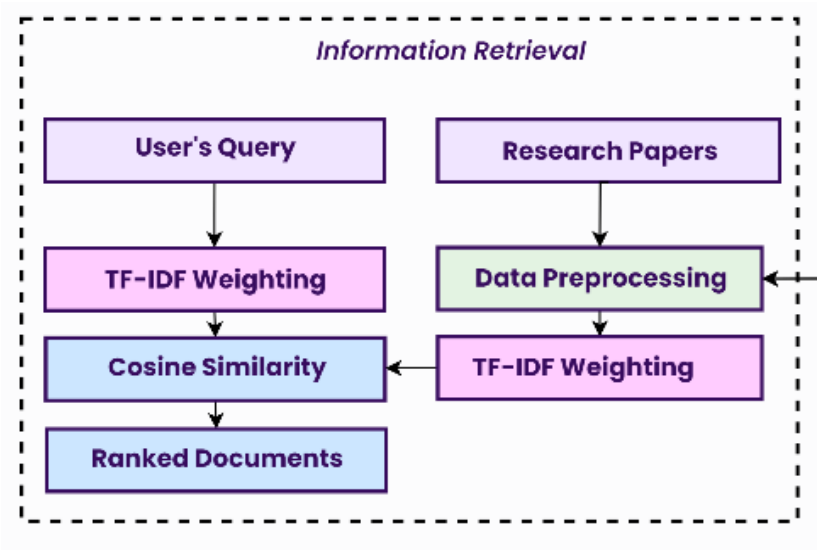


Figure 23. Information retrieval with TF-IDF and cosine similarity.

It usually comes in the form of search engines. But before returning search results with ranking to a user, documents and queries were first represented as vectors using a Vector Space Model (VSM) where the cosine similarity was calculated.

Relevance rankings of documents, as stated by Sidorov (2003), could be calculated using the cosine similarity function, and among many different ways to turn a document into a vector, TF-IDF weighting was often used.

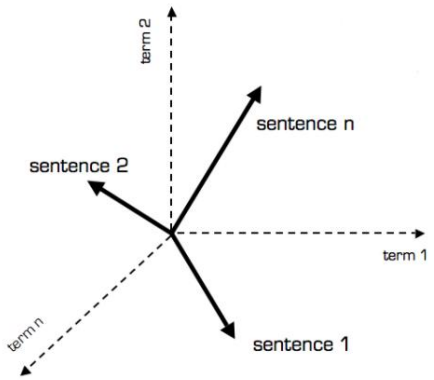


Figure 24. Vector Space Model (VSM) for documents and queries.

Where N was the number of distinct terms over all the document and queries. The i^{th} index of a vector contained the score of the i^{th} term for that vector. The vector outputs of the TF-IDF weighting would then be used to obtain the cosine similarity of the documents and the query by taking their dot product and dividing that by the product of their norms.

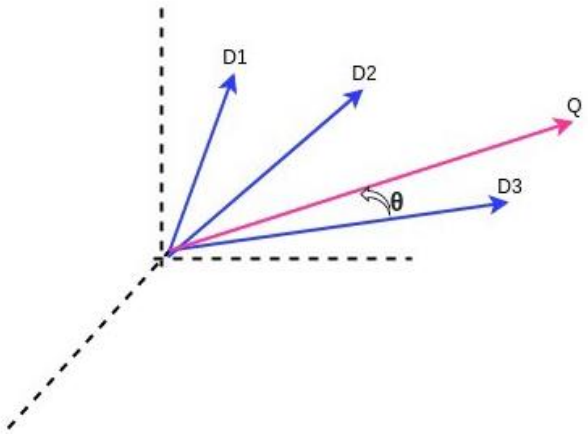


Figure 25. Cosine similarity for documents and query.

The cosine of the angle between the vectors would then be returned. The above figure showed the calculation of the cosine angle between three (3) documents (*D1*, *D2*, *D3*) and a query (*Q).*

Clustering

It was a type of unsupervised learning method for data analysis. One of the most popular techniques was K-means clustering. It was a simple approach for partitioning a dataset into *k* distinct, non-overlapping clusters. It used the Euclidean distance method to find out the distance between the points. To perform K-Means, the desired number of clusters *k* was first specified.

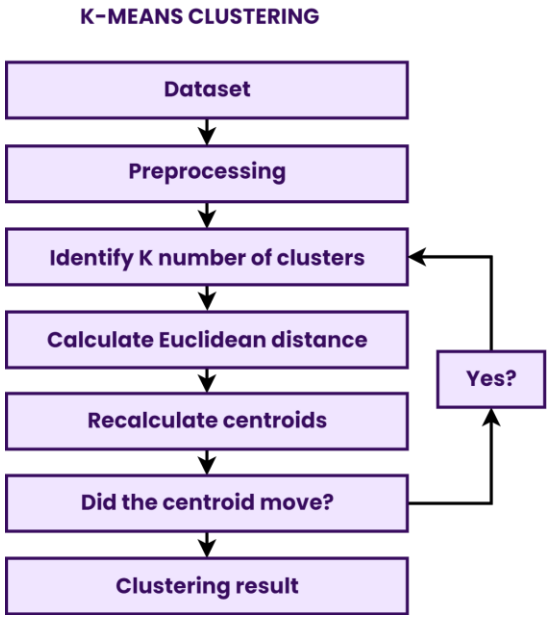


Figure 26. K-means clustering model generation.

It showed the steps in creating the K-means algorithm models. The dataset of fifty (50) skills keywords or one hundred twenty-three (123) research title keywords were all both used in training two different K-means models for two different tasks.

Table 8. Training set of the K-means models.

DATASET	TRAIN SET
Students skills keywords	50
Research titles keywords	123

The above table showed the training set for the K-Means clustering models. Since K-means was an unsupervised clustering technique, there were no tests and validation sets, unlike classification techniques.

Before training the models, data were first preprocessed with tokenization, lowercasing, stop words removal, punctuation removal, stemming, and vectorization.

The algorithm of K-means, according to Dehariya (2010), was composed of a series of steps. First, *k* points were selected as initial group centroids in the space the

objects being clustered. Second, each object was assigned to the group that has the closest centroid. Third, when all objects had been assigned, the positions of the centroids were recalculated. Last, steps 2 and 3 were repeated until the centroids no longer move. This results in the objects being divided into groups.

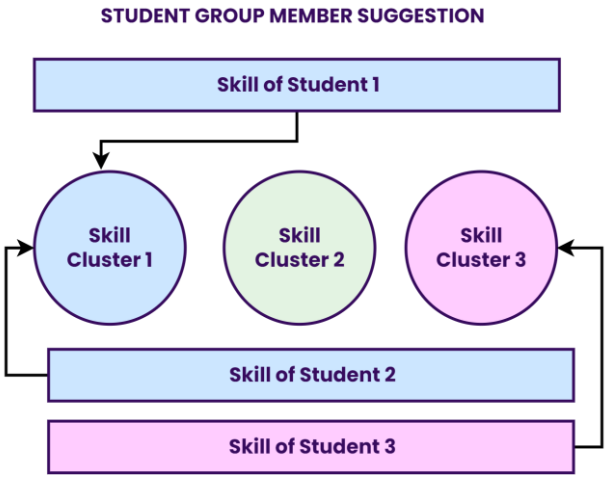


Figure 27. K-means model for student group member suggestion.

The figure above showed the utilization of one of the K-means models in the web application for the purpose of student group member suggestions. The algorithm, integrated into the web application, would first cluster the student by his/her skill. Then the web application would then return a suggestion of a student in the same cluster.

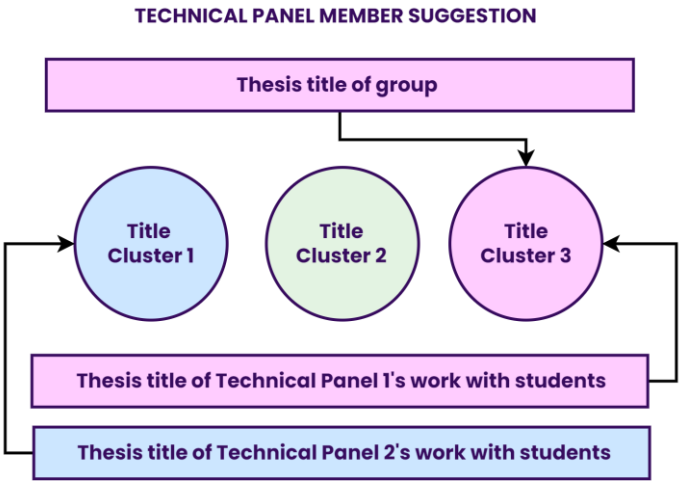


Figure 28. K-means model for technical panel member suggestion.

It depicted how the other K-means model was used in the web application. The cluster to which the thesis title of a group of students belongs was compared with the cluster of the research title works of each technical panel member with other students. Since a technical panel member may have multiple works with other students, the algorithm would just summarize everything into one text data.

Development Methodology

The system was developed with the use of a Software Development Life Cycle (SDLC) model. According to Sree & Priya (2016), it was said that the Software Development Life Cycle was the most important aspect of software system development, and there were many various development methodologies used currently.

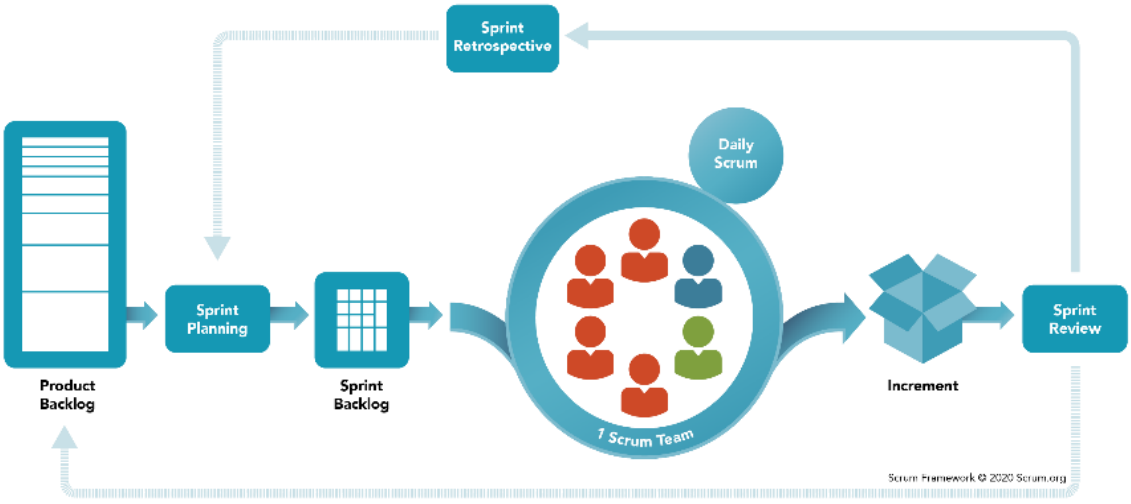


Figure 29. SCRUM software development methodology.

Out of all the software development methodologies, SCRUM was chosen by the researchers. As stated in the study by Srivastava et al. (2017), SCRUM was basically a lightweight, agile framework that outlines steps for managing and controlling the software and product development process. It was the combination of the Iterative model and the Incremental as builds were sequential and incremental in terms of the

features to be developed with constant communication between the stakeholders. Moreover, SCRUM, according to the researcher, was created to speed up the development process. The SCRUM workflow consists of the product backlog, spring planning, sprint backlog, sprint with daily scrum, increment, sprint review, and sprint retrospective.

Product Backlog

The SCRUM workflow starts with the construction of a product backlog. The researchers first constructed a product backlog, which was simply a list of features or functionalities for the system or product.

PRODUCT BACKLOG						
ID	As a...	I want to be able to...	So that...	Priority	Sprint	Status
1	User	search reseach papers	I can find relevant documents to my query	High	1	Done
2	User	view related research papers	I can find more relevant research papers	High	1	Done
3	Student	access the undergraduate thesis portal	I can go through the undergraduate thesis process	High	1	Done
4	Student	be recommended group members	I can have an idea of the perfect group	High	1	Done
5	Student	be recommended technical panel members	I can have the most suitable technical panel members	High	2	In Progress
6	Guest	create an account	I can have full access to the digital library	Medium	2	Done
7	Guest	login	I can view full text research papers	Medium	2	Done
8	User	view full text reserch papers	I can be able to see the full context	Medium	2	Done
9	Student	upload deliverables	I can complete my requirements	Low		To be started
10	RIUH/Dean	view all users	I can track the users registered	Low		To be started
11	RIUH	monitor the groups' progress	I can see who is working and not working	Low		To be started
12	RIUH	upload final manuscripts	I can store the graduate students' papers in the system	Low		To be started
13	Technical Panel	check the group's progress	I can check their research papers	Low		To be started

Figure 30. Product backlog of the researchers.

The items on the list, which could be thought of as a "To-Do" list, were chosen based on suggestions from the Research Implementing Unit Head (RIUH) and ideas from the researchers. Also, besides listing the features, the researchers also prioritized the items. High-priority items were prioritized at the top of the list, while non-priority items were pushed to the bottom. Once the product backlog was built, it was regularly maintained by the researchers to make sure prioritization is correct and new items were added or removed.

Sprint Planning

In sprint planning, the researchers then addressed which product backlog items would be worked on. The researchers set up a meeting with one another in less than an hour to identify what could be delivered and how that could be achieved. Also, the duration of the sprint was addressed during the meeting. By the end of the sprint planning, the researchers were then ready to start working on the sprint backlog.

Sprint Backlog

PRODUCT BACKLOG						
ID	As a...	I want to be able to...	So that...	Priority	Sprint	Status
1	User	search reseach papers	I can find relevant documents to my query	High	1	Done
2	User	view related research papers	I can find more relevant research papers	High	1	Done
3	Student	access the undergraduate thesis portal	I can go through the undergraduate thesis process	High	1	Done
4	Student	be recommended group members	I can have an idea of the perfect group	High	1	Done

Figure 31. Sprint backlog for the first sprint.

The output from sprint planning was a sprint backlog. It was a list of product backlog items identified by the researchers for the sprint. For instance, the researchers included the first four (4) product backlog items in their first sprint, such as “search research papers”, “view related research papers”, etc.

Sprint

Following the creation of the sprint backlog, the researchers then began the sprint. Whenever a sprint begins, the researchers or the developer start their work. The developer built the features identified with the help of the sprint backlog with a limit to a set period of time during which the work had to be completed. During the sprint, the researchers held a daily scrum, or a short daily standup meeting in less than 15 minutes and discussed the progress the

researchers' developer was working on. Also, daily scrums were held by the researchers to ensure everything was on the right track.

Increment

The outcome of a sprint was a usable product called an increment or product increment. The increment was the integration of all the completed items done by the researchers or developer based on the sprint backlog. The researchers then marked the completed items as "Done" to show that it was in a usable condition which could be made available to the end-user in the form of a system.

Sprint Review

At the end of a sprint, the researchers and the Research Implementing Unit Head (RIUH), a key stakeholder, reviewed the outcome of a sprint in an event called sprint review. In sprint review, the researchers and the RIUH met to review the completed work and addressed any concerns.

Sprint Retrospective

After the sprint review, the researchers convened for another meeting in less than an hour to wrap up the sprint in a phase called sprint retrospective. In this phase, the researchers addressed what went well, what could be improved, and what they would be committed to improving in the next sprint. For the next sprint, the processes in the SCRUM workflow, starting with sprint planning, were repeated.

Performance Evaluation of Algorithms

In this phase, the researchers performed an examination of the performance of the algorithms or models integrated into the web application enumerating the LDA topic model, the K-means clustering models, and the information retrieval algorithm through actual testing of data.

LDA Topic Model

Interpretation-based approaches, particularly word intrusion, and topic intrusion, developed by Chang et al. (2009), were used to evaluate the performance of the LDA model. These approaches were considered a "gold" standard for evaluating topic models since they use human judgment to maximum effect.

In word intrusion, subjects were presented with 6 randomly ordered words, 5 of which belong to a given topic and one which does not - the 'intruder' word. Subjects were asked to identify the intruder word.

#1 - [dog, cat, horse, apple, pig, cow]

#2 - [car, teacher, platypus, agile, blue, Zaire]

Figure 32. A word intrusion task.

In the first set, most people would readily identify *apple* as the intruding word in the set because the remaining words, *dog, cat, horse, pig, cow* made sense together as they were all animals. On the other hand, for the second set which lacks such coherence, identifying the intruder was difficult. People would usually pick an intruder at random, implying a topic with poor coherence.

Table 9. Word intrusion tasks prepared by the researchers.

SET	WORDS					
1	game	study	**pet**	learning	research	student
2	disease	covid	fire	disaster	disease	**stress**
3	**risk**	connective	tooth	plaque	bacteria	tissue

The researchers prepared three (3) word intrusion tasks to be presented to a subject. To construct the sets, the researchers selected at random a topic from the model. Then the researchers selected the five most probable words from that topic. In addition to these words, an intruder word, set as bold in the table above, is selected at random from a pool of words with low probability in the current topic. All six words were then shuffled and presented to the subject.

Similar to word intrusion, in topic intrusion subjects were asked to identify the “intruder” topic from groups of topics that made up documents. In this task, subjects were shown a title and a snippet from a document along with 4 topics. Three of the topics had a high probability of belonging to the document while the remaining topic had a low probability - the “intruder” topic. The success with which subjects could correctly choose the intruder topic helped to determine the level of coherence.

Table 10. Topic intrusion tasks prepared by the researchers.

SET	TOPICS							
1	Gamification in Industrial Shopfloor							
	Gamification is the art of making non-game context more engaging and fun, with the help of game mechanics. Production and logistics are good examples of non-game context, which could benefit a lot from increased engagement.							
	system	web	base	use	user	software	application	develop
	event	**decision**	**sentiment**	**human**	**analysis**	**topic**	**tweet**	**expert**
	use	game	develop	learning	student	research	test	study
	system	embed	energy	gaming	device	simulation	efficiency	power
2	Blue Card Management Information System							
	The study entitled “Blue Card Management Information System” falls under the E-government that utilizes the use of information technology to improve the services provided by the government to the public. The web-based system focused on managing member information and facilitating real-time updating of information in the offices involved.							
	Information	text	database	document	module	computer	use	extract
	system	web	base	use	user	software	application	develop
	intelligent	**image**	**system**	**treatment**	**model**	**plant**	**drug**	**diagnosis**
	information	text	database	documentn	module	computer	use	extract
3	World Wide Web							
	World Wide Web or commonly known as “web” is an information system where documents, media and other resources are identified by a Uniform Resource Locator or URL which may be interlinked with hypertext links readily accessible on the internet. Over the years, technology continually changes according to the needs of every user. A technology could be depleted or diminished when there is lesser demand of the system.							
	event	**decision**	**sentiment**	**human**	**analysis**	**topic**	**tweet**	**expert**
	system	web	base	use	user	software	application	develop
	system	embed	energy	gaming	device	simulation	efficiency	power
	information	text	database	documentn	module	computer	use	extract

The researchers prepared three (3) topic intrusion tasks. For the topic intrusion tasks, subjects were shown the title and a snippet from a document. Along with the document, they were presented with four topics (each topic is represented by the eight highest-probability words within that topic). Three of those topics were the highest probability topics assigned to that document. The remaining intruder topic was chosen randomly from the other low-probability topics in the model.

K-Means Clustering Models

	students	skills_descriptions	true_clusters
0	Faye Villanueva	I read a lot of article or study to gathered d...	1
1	Jian Carlo Carayap	I am good at programming language which is HTM...	0
2	John Christian Austria	I know a lot in C++ programming language. It h...	0
3	Limuel Leoncio	I am good at programming. I always used PHP, C...	0
4	Andrei Jimenez	I rely on what I read so my documentation is b...	1
5	Ellaine Salisi	I follow the instruction about my document so ...	1
6	Vincent Pestano	SQL programming is my favourite database query...	0
7	Allison Punzalan	I only know 2 programming language and it is P...	0
8	Carl Gutierrez	I love scripting language so read a lot about ...	0
9	Cedric James Landerito	I always need to communicate my work process t...	1

Figure 33. Dataset of classified students.

In order to evaluate the performance of the K-means clustering algorithm for student group member suggestions, ten (10) classified students, classified into their own clusters based on their skills, were prepared by the researchers and then compared with the prediction of the K-means clustering model for such task.

	technical_panels	titles	true_clusters
0	Technical Panel 1	Municipal Web-based Profiling System	1
1	Technical Panel 2	Information Management System for Profiling	0
2	Technical Panel 3	Municipal Energy Planning: Scope and Method De...	1
3	Technical Panel 4	Record Management System with SMS	0
4	Technical Panel 5	Municipal Energy Planning: Scope and Method De...	1
5	Technical Panel 6	Crime Prediction using Spatio Temporal	0
6	Technical Panel 7	Report Management System for Office of the Reg...	0
7	Technical Panel 8	Classroom Attendance Monitoring System for Sen...	0
8	Technical Panel 9	Android Game Development Integrated in the Web	0
9	Technical Panel 10	Information Management System for Dental Clinic	0

Figure 34. Dataset of classified technical panels.

Moreover, to test the performance of the K-means clustering algorithm for technical panel member suggestions, ten (10) classified technical panel members, classified into their own clusters based on their title works students, were prepared by the researchers and then compared with the prediction of the other K-means clustering model.

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}}$$

Figure 35. Accuracy formula.

To quantify the performance of the K-means clustering models, the total number of correct predictions was divided by the total number of predictions. As a result, its accuracy was calculated, with 1.00 being the highest perfect accuracy score.

Information Retrieval

The performance of the ranked information retrieval system of the web application assessed how well the system met the information needs of its users. It was examined based on recall, precision, average precision (AP), and mean average precision (MAP) on three (3) user search queries.

Recall

$$\text{Recall} = \frac{\text{No. of relevant retrieved documents}}{\text{No. of relevant documents}}$$

Figure 36. Recall formula.

It was the ratio of the number of relevant records retrieved to the total number of relevant records in the database. It is usually expressed

as a percentage. Recall was the fraction of the documents that were relevant to the query that were successfully retrieved.

Precision

$$\text{Precision} = \frac{\text{No. of relevant retrieved documents}}{\text{No. of retrieved documents}}$$

Figure 37. Precision formula.

It was the ratio of the number of relevant records retrieved to the total number of irrelevant and relevant records retrieved. It was usually expressed as a percentage. Precision was the fraction of the documents retrieved that were relevant to the user's information need.

Average Precision (AP)

$$\text{AveP} = \frac{\sum_{k=1}^{n} P(k) \times \text{rel}(k)}{\text{number of relevant documents}}$$

Figure 38. Average precision formula.

It was the sum of precisions divided by the number of relevant documents in the ranked list.

Mean Average Precision (MAP)

$$\text{MAP} = \frac{\sum_{q=1}^{Q} \text{AveP}(q)}{Q}$$

Figure 39. Mean average precision formula.

It was the mean of the average precision scores for each query where, Q was the number of queries in the set and *AveP(q)* was the average precision (AP) for a given query, *q*.

Table 11. Query sets prepared by the researchers.

SET	QUERY
Q1	“SMS notificaion”
Q2	“barangay profiling”
Q3	“municpality”

The researchers prepared three (3) user search query sets shown in the table above. The relevance of the documents retrieved to a query were based on human judgement. Recall and precision were calculated on each of the query sets, and their average precision (AP), and mean average precision (MAP) were calculated afterward.

CHAPTER IV

RESULTS AND DISCUSSION

This chapter would present the overview of the system and the methodologies used in carrying out the research questions of the study. The results of the specific objectives followed the order similar to how they were listed in Chapter I.

System Overview

The study entitled "Papernerds: A Research Digital Library and Undergraduate Thesis Portal with Text Mining Algorithms," developed a web application that stored research papers in a public digital library and dealt with some of the undergraduate thesis processes with research stakeholders for the College of Computer Studies (CCS) in Laguna State Polytechnic University SCC. Beyond being more than just a standard web application, the developed system was integrated with machine learning techniques, notably text mining algorithms. Three (3) text mining algorithms namely, topic modeling, clustering, and information retrieval were used for the web application.

With these being said, the system would be able to provide benefits for the stakeholders of the study. Topic modeling, which found the latent topics within texts, would be able to assign per-document topic distributions within a research paper. This enabled the system to be able to recommend more research papers for a user.

Information retrieval, which calculated the similarity between a user's search input and the abstracts of research papers, would give the user a search result with ranking based on their similarity.

Clustering, on the other hand, which groups unlabeled data, would allow the application to suggest a student member or technical panel member for the undergraduate student writing up their thesis.

1. To design and develop a web application that included a public digital library for research papers and a web portal for undergraduate theses for the College of Computer Studies (CCS) at Laguna State Polytechnic University SCC.

To design and develop Papernerds, the researchers conducted interviews with the Research Implementing Unit Head (RIUH) in order to collect necessary information, such as research papers, and understand the undergraduate thesis process.

PRODUCT BACKLOG						
ID	As a...	I want to be able to...	So that...	Priority	Sprint	Status
1	User	search reseach papers	I can find relevant documents to my query	High	1	Done
2	User	view related research papers	I can find more relevant research papers	High	1	Done
3	Student	access the undergraduate thesis portal	I can go through the undergraduate thesis process	High	1	Done
4	Student	be recommended group members	I can have an idea of the perfect group	High	1	Done
5	Student	be recommended technical panel members	I can have the most suitable technical panel members	High	2	Done
6	Guest	create an account	I can have full access to the digital library	Medium	2	Done
7	Guest	login	I can view full text research papers	Medium	2	Done
8	User	view full text reserch papers	I can be able to see the full context	Medium	2	Done
9	Student	upload deliverables	I can complete my requirements	Low	3	Done
10	RIUH/Dean	view all users	I can track the users registered	Low	3	Done
11	RIUH	monitor the groups' progress	I can see who is working and not working	Low	3	Done
12	RIUH	upload final manuscripts	I can store the graduate students' papers in the system	Low	3	Done
13	Technical Panel	check the group's progress	I can check their research papers	Low	3	Done

Figure 40. Final product backlog of the researchers.

Based on the interviews, the researchers also consolidated a list of features to be developed, listed in a product backlog, which was the first step in the SCRUM software methodology. It was essentially a “To-Do” list, wherein high-priority items were prioritized at the top of the list, while non-priority items were pushed to the bottom. Sprint planning then took place on each iteration in order to identify the researchers would be working on in a series of sprints. After each sprint, a usable software called an increment, would have been developed. Sprint review and sprint retrospective were then undertaken to review and reflect on the process respectively.

The developed research digital library and undergraduate thesis portal were shown in the following:

Research Digital Library

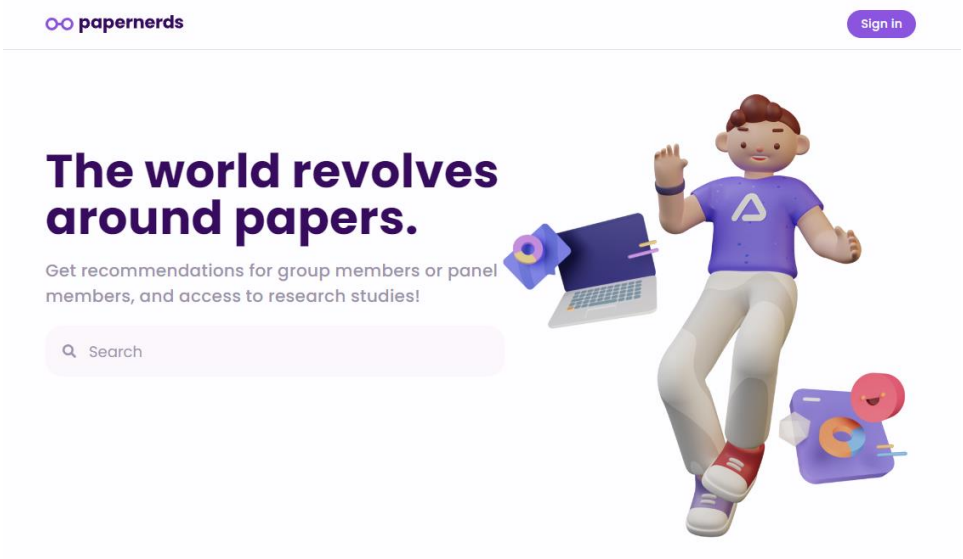


Figure 41. Landing page.

The search bar and the “Sign In” button were shown in the figure above. The search bar allowed anyone to access the public research digital library of the web application, while the “Sign in” button allowed registered users to log in to the web application.

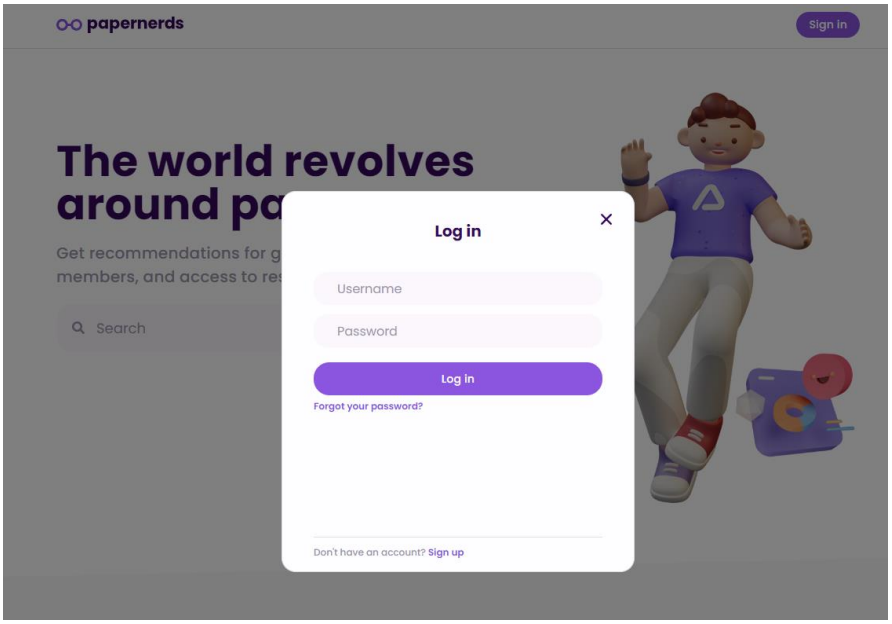


Figure 42. Login panel.

Upon clicking the “Sign in” button, the web application would show a login panel where users could log in with their username and password. Moreover, if a user forgot their password, they could click the text “Forgot your password?” Then, they would be redirected to another page to configure their forgotten password. Lastly, upon clicking the “Sign up” button, the user would be asked to enter their details to register.

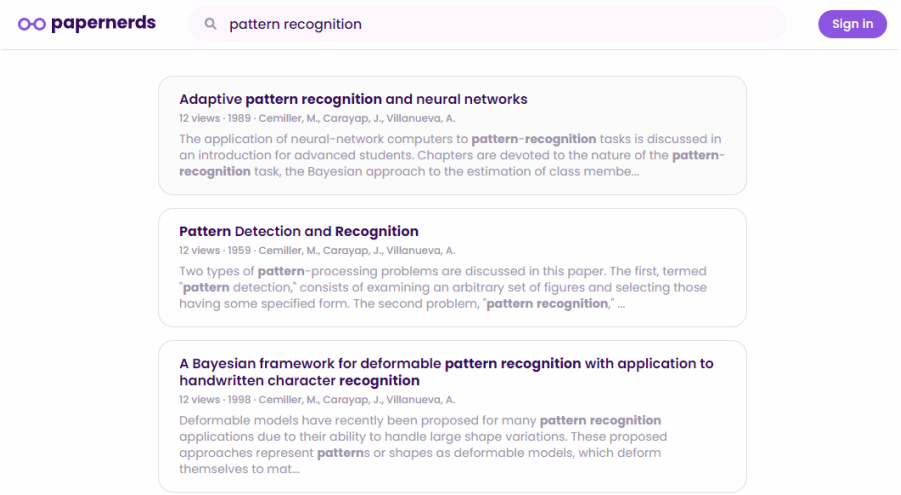


Figure 43. A ranked search results.

The search feature of the application allowed users to retrieve a ranked search results of their input from the stored research papers in the database of the web application. The users would be able to distinguish a non-relevant research paper with this.

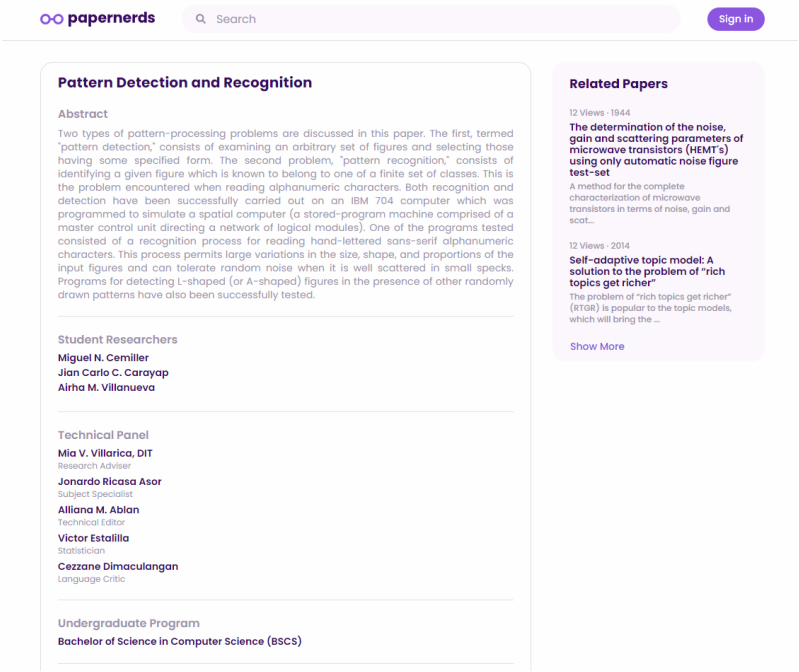


Figure 44. Research paper summary.

The page for the summary of a research paper was separated into two columns. The first column showed the abstract of the research paper, and the student researchers, technical panel, undergraduate program, year, number of views, and a button to view the full paper. The second column on the other hand simply showed the related research papers.

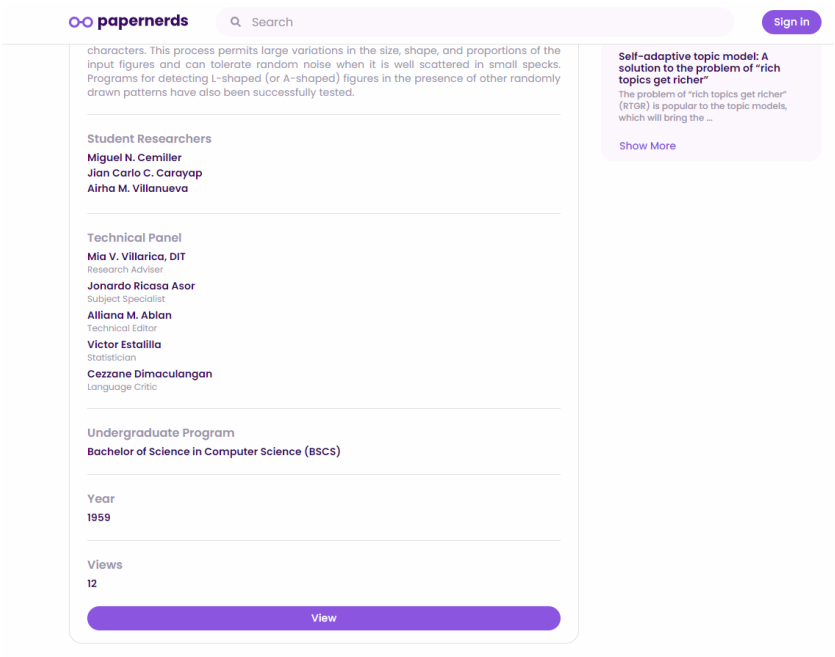


Figure 45. Research paper summary (scrolled).

Upon scrolling the page for the summary of the research paper, the button to view the full paper as a 10-pager in PDF format could be seen. However, if a user wasn't logged in, the user wouldn't be able to perform the action.

Undergraduate Thesis Portal

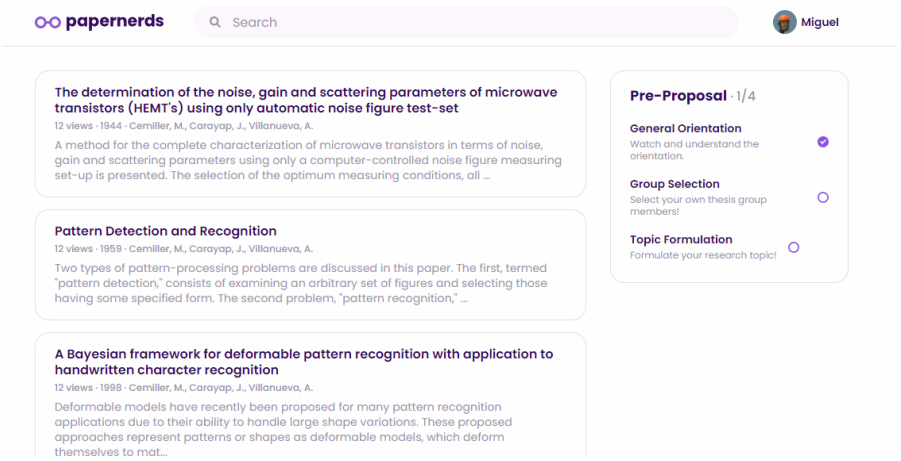


Figure 46. Student's homepage.

The undergraduate thesis process is divided into four (4) stages namely, pre-proposal, proposal, research colloquium, and final oral defense. The web application

works step by step in accomplishing all the steps for the undergraduate thesis process. An additional second column was added to represent the web portal of the web application.

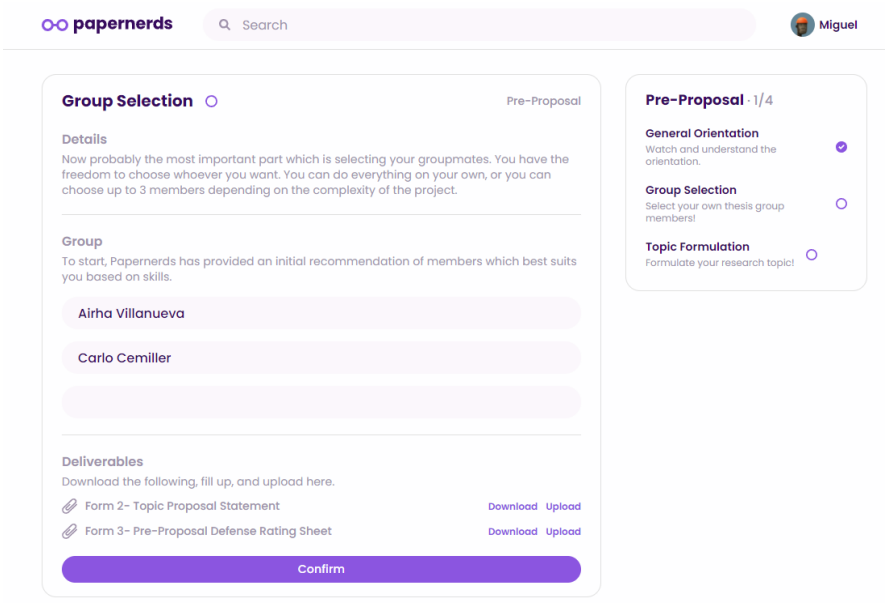


Figure 47. Group selection thesis process step.

For instance, the group selection step, under the pre-proposal stage, of the web application allowed users to get group member suggestions for their undergraduate thesis. It explained the details of each step under the "Details" section. Moreover, the deliverables for the step, where users could download and upload the documents, was indicated just right above the "Confirm" button.

2. To test LDA topic modeling based on coherence and perplexity and K-means clustering based on homogeneity, completeness, and V-measure score to identify which to integrate to the system for research papers recommendations and student group and technical panel member suggestions.

The researchers conducted a series of tests based on the datasets prepared. The datasets, abstracts in the form of paragraphs and skills in the form of keywords of not

more than five (5) words, were kept constant throughout the tests for fair comparison of LDA topic models and K-means clustering models.

LDA Topic Model Testing

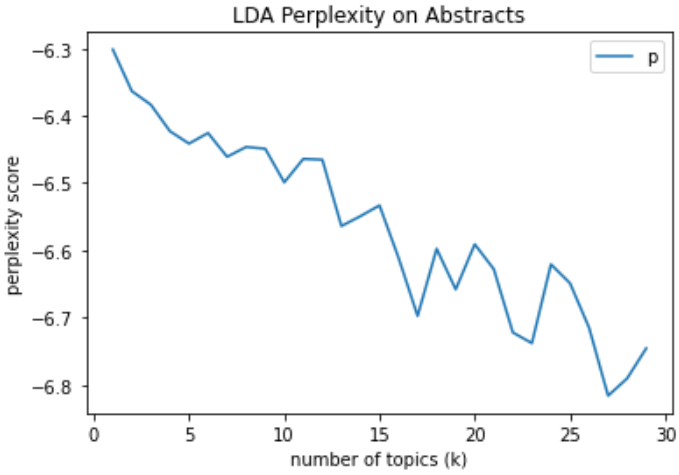


Figure 48. LDA perplexity graph on abstracts.

A graph with k number of topics, ranging from 0 to 30, was plotted in order to visualize the perplexity scores of the LDA topic model within the range. For instance, the topic model on *k* number of topics of 30 gave a perplexity score of -6.720. The score mentioned indicated that the topic model could very well predict a new sample, since a lower perplexity was better.

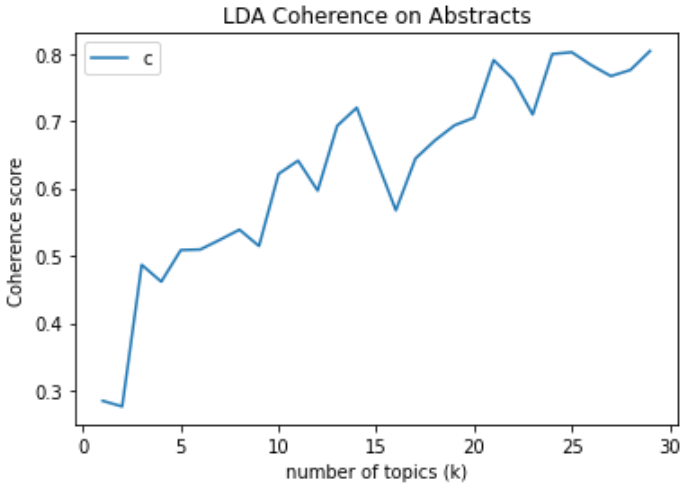


Figure 49. LDA coherence graph on abstracts.

Moreover, a graph with *k* number of topics, ranging from 0 to 30, was plotted once again to visualize the coherence scores of the LDA topic model within the range. For instance, the topic model on *k* number of topics of thirty (30) gave the highest coherence score of 0.756. The high coherence score implied that the topics that would be returned by the topic model would be highly interpretable to humans.

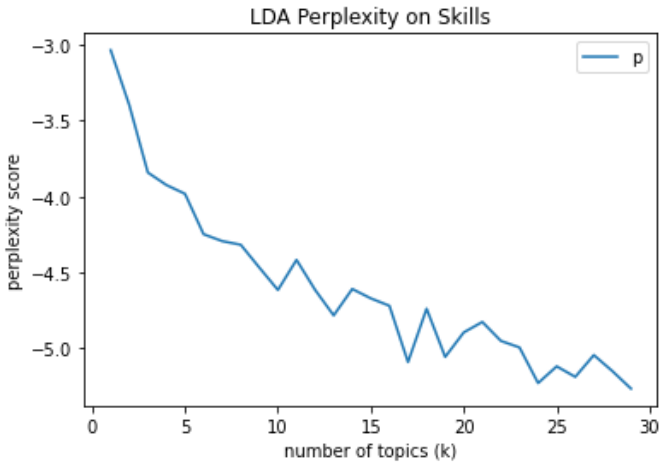


Figure 50. LDA perplexity graph on skills.

Finally, a graph with *k* number of topics, ranging from 0 to 30, was plotted in order to visualize the perplexity scores of the topics LDA topic model within the range. For instance, the topic model on thirty (30) *k* number of topics gave a perplexity score of -5.141. The very low perplexity score result indicated that the topic model could very well predict a new sample better than that of the previous test on the abstracts dataset.

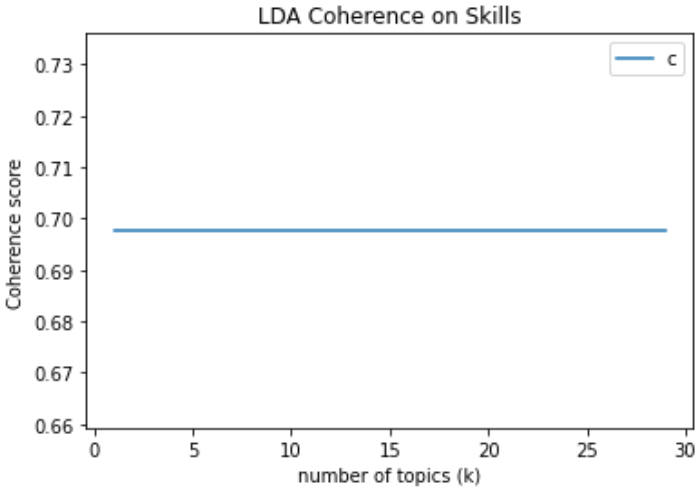


Figure 51. LDA coherence graph on skills.

In comparison to the abstracts dataset, another graph with *k* number of topics, ranging from 0 to 30, was plotted in order to visualize the coherence scores of the LDA topic model within the range. The topic model on all k number of topics gave a constant coherence score of 0.698. The result gave an interpretation that the topic model on all number of topics have the same level of interpretation which indicates that the topics would all just be similar to each other.

Therefore, the researchers determined that LDA topic models worked perfectly with longer texts in determining the topics for distributions since the topic model on the abstracts dataset was found to be more coherent than that of shorter texts with unreliable interpretations. Further, although the perplexity score on the shorter texts gave a better measure for predicting a new sample, it wasn't critical since its coherence score was unreliable.

K-Means Cluster Testing

Table 12. Results of K-means clustering on the skills dataset.

HOMOGENEITY	COMPLETENESS	V-MEASURE
1.000	1.000	1.000

For the testing of the K-means clustering on the skills dataset, it received a perfect result of 1.000 on all the metrics. It was implied that the K-means clustering model was perfectly homogenous and complete. The clustering was perfectly homogeneous since in each cluster the data points were of the same class label, and it was perfectly complete since all data points belonging to the same class were clustered into the same cluster. Thus, their harmonic mean, measured using a V-measure score, was also 1.00.

Table 13. Results of K-means clustering on the abstracts dataset.

HOMOGENEITY	COMPLETENESS	V-MEASURE
0.030	0.093	0.045

Lastly, for the testing of the K-means clustering on the abstracts dataset, a totally different result was received. The low homogeneity score of 0.030 indicated that the cluster contained several data points that were members of multiple classes. Moreover, the completeness score of 0.093 indicated that a few data points belonging to the same class were clustered into the same cluster. Thus, the clustering's V-measure score was 0.045, indicating a poor cluster.

The researchers, therefore, concluded that the K-means clustering algorithm worked well with the short texts from the skills dataset as opposed to longer texts from the abstracts dataset.

3. To examine the performance of the models developed integrated into the system through actual testing.

LDA Topic Model

The researchers conducted a series of tests in order to examine the performance of the topic model integrated into the system. In order to carry out the test, the researchers prepared three (3) word intrusion and three (3) topic intrusion tasks presented to one of their research members.

Table 14. Result of the word intrusion tasks.

SET	PREDICTED ANSWER	CORRECT ANSWER	RESULT
1	pet	pet	✓
2	stress	stress	✓
3	risk	risk	✓

The subject was tasked to identify the intruder word from the sets. It was shown in the table that the subject never got an incorrect prediction. Therefore, it was concluded that the LDA topic model integrated into the system was said to have an excellent coherence of words within a topic since the subject got all predicted answers right based on the three (3) sets of word intrusion tasks.

Table 15. Result of the topic intrusion tasks.

SET	PREDICTED ANSWER	CORRECT ANSWER	RESULT
1	event, decision, sentiment, human, analysis, topic, tweet, expert	event, decision, sentiment, human, analysis, topic, tweet, expert	✓
2	intelligent, image, system, treatment, model, plant, drug, diagnosis	intelligent, image, system, treatment, model, plant, drug, diagnosis	✓
3	information, text, database, document, module, computer, use, extract	event, decision, sentiment, human, analysis, topic, tweet, expert	✗

Afterwards, the subject was tasked to identify the intruder topic, represented by the eight highest-probability words within that topic, from the sets. The results show

that the subject got one (1) incorrect prediction. In conclusion, the LDA topic model integrated into the system is said to have a good coherence of topics, since based on the three (3) sets of topic intrusion tasks, the subject got two (2) correctly predicted answers out of 3.

K-Means Clustering Models

The web application developed by the researchers integrated two (2) K-means clustering models for suggestions. The performances of the models were tested on ten (10) datasets for each.

Table 16. Results for the performance evaluation of the K-means models.

K-MEANS MODELS	FALSE PREDICTIONS	TRUE PREDICTIONS	ACCURACY
Student group member	1	9	0.90
Technical panel member	2	8	0.80

The K-means model of the system for student group member suggestions got one (1) incorrect prediction out of ten (10), yielding an accuracy score of 0.90. On the other hand, the K-means model for technical panel member suggestions got two (2) incorrect predictions with 8 true predictions, yielding an accuracy score of 0.80.

The researchers, therefore, concluded that the K-means clustering models integrated into the web application were highly accurate since they had very few false predictions.

Information Retrieval

The researchers conducted tests on the information retrieval system of the web application using the three (3) query sets prepared in order to examine its performance. To quantify its performance, recall and precision were calculated on each of the query

sets, and their average precision (AP), and mean average precision (MAP) were calculated afterward.

Query Set 1: "SMS notification"

Table 17. Calculated recall and precision for query set 1.

RANK	RECALL	PRECISION
1	0.33	1.00
2	0.66	1.00
3	1.00	1.00

The ranked information retrieval system of the web application retrieved a total of 3 documents from the first query set, "SMS notification," ranked from the highest similarity to the lowest. The calculated recalls were 0.33, 0.66, and 1.00, while the calculated precisions were all 1.00.

Query Set 2: "barangay profiling"

Table 18. Calculated recall and precision for query set 2.

RANK	RECALL	PRECISION
1	1.00	1.00
2	1.00	0.50

The ranked information retrieval system of the web application retrieved a total of two (2) documents from the second query set, "barangay profiling," ranked from the highest similarity to the lowest. The calculated recalls were all 1.00, while the calculated precisions were 1.00 and 0.50.

Query Set 3: "municipality"

Table 19. Calculated recall and precision for query set 3.

RANK	RECALL	PRECISION
1	0.17	1.00
2	0.33	1.00
3	0.50	1.00

The ranked information retrieval system of the web application retrieved a total of three (3) documents from the second query set, "municipality," ranked from the highest similarity to the lowest. The calculated recalls were 0.17, 0.33, and 0.50, while the calculated precisions were all 1.00.

Table 20. Calculated AP and MAP for the query sets.

SET	QUERY	AP
Q1	"SMS notification"	1.00
Q2	"barangay profiling"	0.75
Q3	"municipality"	1.00
	MAP	**0.92**

The result of the performance of the ranked information retrieval system was calculated using the mean average precisions (MAP) from the calculated average precisions (AP) of the query sets prepared by the researchers. It showed that the mean average precision (MAP) is 0.92 or 92%.

CHAPTER V

SUMMARY, CONCLUSIONS, AND RECOMMENDATIONS

This chapter further discussed the results presented in the previous chapter. Based on these discussions, this chapter briefly summarized the study and provided the conclusions of this research. In the end, it described briefly the recommendations for future researchers.

Summary

The study aimed to design and develop a research digital library and an undergraduate thesis portal with text mining algorithms that stored research papers and dealt with some of the undergraduate thesis processes of the students. During the course of the study, the researchers were able to find that Latent Dirichlet Allocation (LDA) was still the best and most popular topic model to be used instead of LSI which was not an authentic topic model and pLSI which ran into overfitting. Further, the researchers found that the cosine similarity was a measure used for information retrieval systems while the Euclidean distance measure was mostly used for clustering. For clustering, the researchers discovered that the K-means clustering algorithm outperformed Hierarchical clustering when it comes to performance. Out of all the software development methodologies to be used, the SCRUM model was found to be the most suitable for this study since it was a combination of incremental and iterative development.

Conclusions

Based on the findings of the study from the objectives, the researchers have reached the following conclusions and generalizations:

1. In developing the web application, the researchers noticed that the development was fast and efficient using the Python programming language since it was used to create the machine learning algorithms and develop the other back-end functionalities, instead of using other programming languages for multiple tasks.
2. The researchers figured that topic models worked particularly well with longer texts as opposed to clustering algorithms which worked better with shorter texts.
3. After performing the performance evaluation of the models or algorithms, the researchers found the that performance testing conducted on the LDA topic model using word and topic intrusion tasks was effective.

Recommendations

Based on the conclusions, the researchers have reached the following recommendations:

1. The researchers recommended the use of the Python programming language when building a web application integrated with machine learning algorithms for faster development.
2. Based on the findings, the researchers recommended the use of K-means for clustering short texts and LDA topic modeling only for long texts for correct clustering and topic modeling respectively.
3. The researchers recommended the use of the word intrusion and topic intrusion tasks for evaluating the performance of the LDA topic models since they were effective in capturing the coherence of words and topics.

REFERENCES

Abbas, O. A. (2008). Comparisons Between Data Clustering Algorithms. *International Arab Journal of Information Technology*, 320-325.

Agrawal, A., & Gupta, U. (2014). Extraction based approach for text summarization using k-means clustering. *International Journal of Scientific and Research Publications*, *4*(11).

Alodadi, M., & Janeja, V. P. (2015). Similarity in Patient Support Forums Using TF-IDF and Cosine Similarity Metrics. *2015 International Conference on Healthcare Informatics*, 521-522. 10.1109/ICHI.2015.99

Anupriya, P., & Karpagavalli, S. (2015). LDA based topic modeling of journal abstract. *2015 International Conference on Advanced Computing and Communication Systems*, 1-5.

Arora, M., Kanjilal, U., & Varshney, D. (2016). Evaluation of information retrieval: precision and recall. *International Journal of Indian Culture and Business Management*, *12*(2), 224-236. 10.1504/IJICBM.2016.074482

Asenova, M., & Chrysoulas, C. (2019). ersonalized {Micro-Service} Recommendation System for Online News. *Procedia computer science*, *160*, 610-615. 10.1016/j.procs.2019.11.039

Ashraf, S., & Aftab, S. (2017). Latest Transformations in Scrum: A State of the Art Review. *International Journal of Modern Education and Computer Science*, *9*, 12-22.

Baeza-Yates, R., & Ribeiro, B. d. A. N. (1999). *Modern Information Retrieval*. ACM Press.

Baharudin, B., Lee, L. H., & Khan, K. (2010). A Review of Machine Learning Algorithms for Text-Documents Classification. *Journal of Advances in Information Technology*, *1*, 4-20.

Barysevich, A. (2019, October 25). *TF-IDF: Could It Really Help Your SEO?* Search Engine Journal. https://www.searchenginejournal.com/tf-idf-could-it-really-help-your-seo/331075/#close

Bearman, D. (2007). Digital Libraries. *Annual Review of Information Science and Technology*, *41*(1), 223-272. 10.1002/aris.2007.1440410112

Bhatt, S. (2018, March 19). *Reinforcement Learning 101. Learn the essentials of Reinforcement… | by Shweta Bhatt*. Towards Data Science. https://towardsdatascience.com/reinforcement-learning-101-e24b50e1d292

Bhuvaneswari, T., & Prabaharan, S. (2013). A Survey on Software Development Life Cycle Models. *International Journal of Computer Science and Mobile Computing*, *2*(5), 262-267.

Blei, D. M. (2012). Probabilistic Topic Models. 10.1145/2133806.2133826

Blei, D. M., Ng, A. Y., & Jordan, M. L. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*, 993-1022.

Borgman, C. L. (1999). What are digital libraries? Competing visions. *Information Processing and Management*, *35*(3), 227-243. 10.1016/S0306-4573(98)00059-4

Boyd-Graber, J., Mimno, D., & Newman, D. (2014). Care of Feeding of Topic Models: Problems, Diagnostics, and Improvements. In E. M. Airoldi, D. Blei, E. A. Erosheva, & S. E. Fienberg (Eds.), *Handbook of Mixed Membership Models and Their Applications*. CRC Press.

Brownlee, J. (2017, September 29). *How to Encode Text Data for Machine Learning with scikit-learn*. Machine Learning Mastery. https://machinelearningmastery.com/prepare-text-data-machine-learning-scikit-learn/

Cabrerizo, F. J., Morente-Molinera, J. A., Perez, I. J., Lopez-Gijon, J., & Herrera-Viedma, E. (2015). A decision support system to develop a quality management in academic digital libraries. *Information Sciences*, *323*, 48-8. 10.1016/j.ins.2015.06.022

Cai, D., Mei, Q., Han, J., & Zhai, C. (2008). Modeling Hidden Topics on Document Manifold. *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, 911-920. 10.1145/1458082.1458202

Celebi, M. E., & Aydin, K. (Eds.). (2016). *Unsupervised Learning Algorithms*. Springer International Publishing. 10.1007/978-3-319-24211-8

Chang, J., Boyd-Graber, J. L., Gerrish, S., Wang, C., & Blei, D. M. (2009). Reading Tea Leaves: How Humans Interpret Topic Models. *Advances in Neural Information Processing Systems*, *32*, 288-296.

Cheng, L. C., & Sharmayne, L. R. (2020). Analysing Digital Banking Reviews Using Text Mining. *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 914-918. 10.1109/ASONAM49781.2020.9381429

Cuaton, G. P., Caluza, L. J. B., & Neo, J. F. V. (2021). A topic modeling analysis on the early phase of COVID-19 response in the Philippines. *International Journal of Disaster Risk Reduction*, *61*. 10.1016/j.ijdrr.2021.102367

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the Americould Society for*

Information Science, *41*(6), 391-407. 10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9

Dehariya, V. K., Shrivastava, S. K., & Jain, R. C. (2010). Clustering of Image Data Set Using K-Means and Fuzzy K-Means Algorithms. *2010 International Conference on Computational Intelligence and Communication Networks*, 386-391. 10.1109/CICN.2010.80

del Rosario, E., del Rosario, J. D., Nieva, M., & Tan, T. (2016). CollaborateIT: A CCS IT Thesis Portal with Electronic Document Management System.

Deng, Z.-H., Luo, K.-H., & Yu, H.-L. (2014). A Study of Supervised Term Weighting Scheme for Sentiment Analysis. *Expert Systems with Application*, *41*(7), 3506-3513. 10.1016/j.eswa.2013.10.056

Dias, C. (2001). Corporate portals: a literature review of a new concept in Information Management. *International Journal of Information Management*, *21*(4), 269-287. 10.1016/S0268-4012(01)00021-4

Doll, T. (2018, June 24). *LDA Topic Modeling: An Explanation | by Tyler Doll*. Towards Data Science. https://towardsdatascience.com/lda-topic-modeling-an-explanation-e184c90aadcd

Ebrahim, M. (2021, August 7). *Remove Punctuation Using Python*. Like Geeks. https://likegeeks.com/python-remove-punctuation/

Eldrandaly, K. A., Shouman, M. M., & Abou-Ali, M. G. (2009). Ranking Web Portals Quality Criteria Based on Different Types of Personality. *Forth International Conference on Intelligent Computing and Information Systems*.

Forman, G. (2008). BNS Feature Scaling: An Improved Representation over Tf-Idf for Svm Text Classification. *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, 263-270. 10.1145/1458082.1458119

Fox, E. A., Akscyn, R. M., Furuta, R., & Leggett, J. J. (1995). Digital libraries. *Communications of the ACM*, *38*(4), 22-28. 10.1145/205323.205325

Garcia, P. G., Simonette, A., & Serrano, J. (2013). pen Knowledge Initiatives in the Philippines: The Vibal Foundation. *Open Educational Resources: An Asian Perspective*, 197-209.

Gharehchopogh, F. S., & Khalifelu, Z. A. (2011). Analysis and evaluation of unstructured data: text mining versus natural language processing. *2011 5th International Conference on Application of Information and Communication Technologies (AICT)*, 1-4. 10.1109/ICAICT.2011.6111017

Gupta, V., & Lehal, G. S. (2009). A Survey of Text Mining Techniques and Applications. *Journal of Emerging Technologies in Web Intelligence*, *1*(1), 60-76.

Himel, M. T., Uddin, M. N., Hossain, M. A., & Jang, Y. M. (2017). Weight based movie recommendation system using K-means algorithm. *2017 International Conference on Information and Communication Technology Convergence (ICTC)*, 1302-1306. 10.1109/ICTC.2017.8190928

Hoffmann, T. (2013). Probabilistic Latent Semantic Analysis. 289-296.

Huang, A.-L. (2008). Similarity Measures for Text Document Clustering.

Hull, D. A., & Grefenstette, G. (1996). Querying Across Languages: a Dictionarybased Approach to Multilingual Information Retrieval. *19th Ann Int ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'96)*.

Jacob, M., & John, A. (2016). Improved Clustering of Documents using K-means Algorithm. *International Journal of Engineering Research & Technology (IJERT)*, *5*(7), 378-382.

Jacobi, C., van Atteveldt, W., & Welbers, K. (2016). Quantitative analysis of large amounts of journalistic texts using topic modelling. *Digital Journalism*, *4*(1), 89-106. 10.1080/21670811.2015.1093271

Jiang, Y., Jia, A., Feng, Y., & Zhao, D. (2012). Recommending Academic Papers via Users' Reading Purposes. *Proceedings of the Sixth ACM Conference on Recommender Systems*, 241-244. 10.1145/2365952.2366004

Kameshwaran, K., & Malarvizhi, K. (2014). Survey on Clustering Techniques in Data Mining. *International Journal of Computer Science and Information Technologies*, *5*(2), 2272-2276.

Kaushik, M., & Mathur, B. (2014). Comparative Study of K-Means and Hierarchical Clustering Techniques. *International Journal of Software & Hardware Research in Engineering*, *2*(6), 93-98.

Kavvadias, S., Drosatos, G., & Kaldoudi, E. (2020). Supporting topic modeling and trends analysis in biomedical literature. *Journal of Biomedical Informatics*, *10*. 10.1016/j.jbi.2020.103574

Kayi, E. S., Yadav, K., & Choi, H. (2013). Topic Modeling Based Classification of Clinical Reports. In *ACL*.

Ko, Y. (2012). A Study of Term Weighting Schemes Using Class Information for Text Classification. *{Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1029-1030. 10.1145/2348283.2348453

Kotov, M. (2017). NLP Resources for a Rare Language Morphological Analyzer: Danish Case. *Proceedings of the 1st International Conference Computational Linguistics And Intelligent Systems*.

Kotsiantis, S. B. (2007). Supervised Machine Learning: A Review of Classification Techniques. *Informatica, 31,* 249-268.

Kramer, R. F. (1985). A Overview of Descriptive Research. *Journal of Pediatric Oncology Nursing, 2,* 41-45.

Kuang, Q., & Xu, X. (2010). Improvement and Application of TF•IDF Method Based on Text Classification. *2010 International Conference on Internet Technology and Applications*, 1-4. 10.1109/ITAPP.2010.5566113

Kučak, D., Juričić, V., & Đambić, G. (2018). Machine Learning in Education - A Survey of Current Research Trends (B. Katalinic, Ed.). *29th DAAAM International Symposium on Intelligent Manufacturing and Automation*, 406-410. 10.2507/29th.daaam.proceedings.059

Kumar, S. (2020, October 18). *Silhouette Method — Better than Elbow Method to find Optimal Clusters*. Towards Data Science. https://towardsdatascience.com/silhouette-method-better-than-elbow-method-to-find-optimal-clusters-378d62ff6891

Kumar, V. (2019, September 11). *Python Vs R: What's Best for Machine Learning*. Towards Data Science. https://towardsdatascience.com/python-vs-r-whats-best-for-machine-learning-93432084b480

Lan, M., Tan, C.-L., Low, H.-B., & Sung, S.-Y. (2005). A Comprehensive Comparative Study on Term Weighting Schemes for Text Categorization with Support Vector Machines. *Special Interest Tracks and Posters of the 14th International Conference on World Wide Web*, 1032-1033. 10.1145/1062745.1062854

Leau, Y. B., Loo, W. K., Tham, W. Y., & Tan, S. F. (2012). Software Development Life Cycle AGILE vs Traditional Approaches. *2012 International Conference on Information and Network Technology*.

Lesot, M.-J., Rifqi, M., & Benhadda, H. (2009). Similarity Measures for Binary and Numerical Data: A Survey. *International Journal Knowledge Engineering and Soft Data Paradigms*, *1*(1), 63-84. 10.1504/IJKESDP.2009.021985

Levandoski, A., & Lobo, J. (2018). *Document and Topic Models: pLSA and LDA* [Powerpoint slides]. https://people.cs.pitt.edu/~milos/courses/cs3750/lectures/class11.pdf

Liaw, S.-S., & Huang, H.-M. (2003). An investigation of user attitudes toward search engines as an information retrieval tool. *Computers in Human Behavior*, *19*(6), 751-765. 10.1016/S0747-5632(03)00009-8

Liu, T.-Y. (2009). Learning to Rank for Information Retrieval. *Foundations and Trends in Information Retrieval*, *3*(3), 225-331. 10.1561/1500000016

Macrohon, J. J. E., Bulaong, A. S., Bonsol, J. C. A., Gayeta, J. A. C., & Jeng, J.-H. (2021). Design and Development of a System for Managing Student Research: A Case Study for the School of Information Technology, Mapua University. *2021 IEEE International Conference on Educational Technology (ICET)*, 107-111. 10.1109/ICET52293.2021.9563107

Mahalakshmi, M., & Sundararajan, M. (2013). Traditional SDLC Vs Scrum Methodology – A Comparative Study. *International Journal of Emerging Technology and Advanced Engineering*, *3*(6), 192-196.

Mahmud, S. R. (2013). A Simple Information Retrieval Technique. *International Journal on Recent Trends in Engineering and Technology*, *8*(1), 68-69.

Mansourvar, M., & Yasin, N. M. (2010). Web portal As A Knowledge Management System In The Universities. *World Academy of Science, Engineering and Technology*, *70*.

Marchionini, G., & Hermann, M. (1995). The Roles of Digital Libraries in Teaching and Learning. *Communications of the ACM*, *38*(4), 67-75.

Mishra, A., & Dubey, D. (2013). Comparative Study of Different Software Development Life Cycle Models in Different Scenarios. *© 2013, IJARCSMS All Rights Reserved 64 | P a g e ISSN: 2321-7782 (Online) Volume 1, Issue 5, October 2013 International Journal of*, *1*(5), 64-69.

Motwani, M., & Tiwari, A. (2013). Comparative Study and Analysis of Supervised and Unsupervised Term Weighting Methods on Text Classification. *International Journal of Computer Applications*, *68*(10), 24-27.

Nagori, R., & Aghila, G. (2011). LDA based integrated document recommendation model for e-learning systems. *2011 International Conference on Emerging Trends in Networks and Computer Communications (ETNCC)*, 230-233. 10.1109/ETNCC.2011.6255892

Nam, J. (2014). Understanding the motivation to use web portals. *Computers in Human Behavior*, *36*, 263-273. 10.1016/j.chb.2014.03.074

Nasirun, N., Noor, S. M., Nor, Z. M., Ahmat, H., & Ahmad, Z. (2012). Perceived Web Service Quality for Students' Portal in Higher Learning Institution. *2nd International Conference on Sociality and Humanities*, *56*.

Niu, L.-Q., & Dai, X.-Y. (2015). Topic2Vec: Learning Distributed Representations of Topics.

Odhiambo, J., Onsongo, W., & Osman, S. (2020). An Analytical Comparison Between PythonVs R Programming Languages:Which one is the best for Machine Learning and Deep Learning?

Patil, L. H., & Atique, M. (2013). A novel approach for feature selection method TF-IDF in document clustering. *2013 3rd IEEE International Advance Computing Conference (IACC)*, 858-862.

Pham, D. T., Dimov, S. S., & Nguyen, C. D. (n.d.). Selection of K in K-means clustering. *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science*, *219*, 103-119.

Portugal, I., Alencar, P., & Cowan, D. (2018). The use of machine learning algorithms in recommender systems: A systematic review. *Expert Systems with Applications*, *97*, 205-227. 10.1016/j.eswa.2017.12.020.

Qaiser, S., & Ali, R. (2018). Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents. *International Journal of Computer Applications*.

Qimin, C., Qiao, G., Yangliang, W., & Xianghua, W. (2015). Text Clustering Using VSM with Feature Clusters. *Neural Computing and Applications*, *26*(4), 995-1003. 10.1007/s00521-014-1792-9

Quan, X., Wenyin, L., & Qiu, B. (2011). Term Weighting Schemes for Question Categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *33*(5), 1009-1021. 10.1109/TPAMI.2010.154

Rahutomo, F., Kitasuka, T., & Aritsugi, M. (2012). *Semantic Cosine Similarity*.

Ramos, J. (2003). Using TF-IDF to Determine Word Relevance in Document Queries.

Richey, R. C. (1994). Developmental Research: The Definition and Scope. *Proceedings of Selected Research and Development Presentations at the 1994 National Convention of the Association for Educational Communications and Technology Sponsored by the Research and Theory Division*.

Rosario, B. (2000). Latent Semantic Indexing: An overview.

Sagayam, R. (2012). A survey of text mining: Retrieval, extraction and indexing techniques. *International Journal of Computational Engineering Research*, *2*(5).

Sagayam, R., Srinivasan, S., & Roshni, S. (2012). A Survey of Text Mining: Retrieval, Extraction and Indexing Techniques. *International Journal of Computational Engineering Research*, *2*(5), 1443-1446.

Sah, S. (2020). Machine Learning: A Review of Learning Types. *Preprints*. 10.20944/preprints202007.0230.v1

Samant, S. S., Bhanu, M. N. L., & Malapati, A. (2019). Improving Term Weighting Schemes for Short Text Classification in Vector Space Model. *IEEE Access*, *7*. 10.1109/ACCESS.2019.2953918

Sathya Bama, S., Ifran Ahmed, M. S., & Saravanan, A. (2015). A Survey on Performance Evaluation Measures for Information Retrieval. *International Research Journal on Engineering and Technology (IRJET)*, *2*(2), 1015-1020.

Sawhney, P. (2021). *Introduction to Stemming and Lemmatization (NLP) | by Prateek Sawhney | Geek Culture*. Medium. https://medium.com/geekculture/introduction-to-stemming-and-lemmatization-nlp-3b7617d84e65

Schwaber, K., & Sutherland, J. (2017). *The Scrum Guide*.

Schwarz, C. (2018). ldagibbs: A command for Topic Modeling in Stata using Latent Dirichlet Allocation. *The Stata Journal*, 1-16.

Sendhilkumar, S., Srivani, M., & Mahalakshmi, G. S. (2017). Generation of Word Clouds Using Document Topic Models. *2017 Second International Conference on Recent Trends and Challenges in Computational Models (ICRTCCM)*, 306-308. 10.1109/ICRTCCM.2017.60

Sharma, S., & Hasteer, N. (2016). A comprehensive study on state of Scrum development. *2016 International Conference on Computing, Communication and Automation (ICCCA)*, 867-872. 10.1109/CCAA.2016.7813837

Singh, N., & Singh, D. (2012). Performance Evaluation of K-Means and Heirarichal Clustering in Terms of Accuracy and Running Time. *International Journal of Computer Science and Information Technologies*, *3*(3), 4119-4121.

Singhal, A. (2001). Modern information retrieval: a brief overview. *IEEE Computer Society Technical Committee on Data Engineering*, *24*(4), 35-43.

Sonagara, D., & Badheka, S. (2014). Comparison of Basic Clustering Algorithms. *International Journal of Computer Science and Mobile Computing*, *3*(10), 58-61.

Soucy, P., & Mineau, G. W. (2005). Beyond TFIDF Weighting for Text Categorization in the Vector Space Model. *Proceedings of the 19th International Joint Conference on Artificial Intelligence (IJCAI 2005)*, 1130-1135.

Srivastava, A., Bhardwaj, S., & Saraswat, S. (2017). SCRUM model for agile methodology. *2017 International Conference on Computing, Communication and Automation (ICCCA)*, 864-869. 10.1109/CCAA.2017.8229928

Stevens, K., Kegelmeyer, P., Andrzejewski, D., & Buttler, D. (2012). xploring Topic Coherence over Many Models and Many Topics. *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 952-961.

Stoica, M., Mircea, M., & Ghilic-Micu, B. (2013). Software Development: Agile vs. Traditional. *Informatica Economică*, *7*(4), 64-76.

Strzalkowski, T. (1995). Natural Language Information Retrieval. *Information Processing Management*, *31*, 397-417.

Swanson, D. R. (1988). Historical Note: Information Retrieval and the Future of an Illusion. *Journal of the American Society for Information Science*, *39*(2), 92-98.

Szalvay, V. (2004). An Introduction to Agile Software Development.

Tan, A.-H. (1999). Text mining: The state of the art and the challenges. *In Proceedings of the PAKDD 1999 Workshop on Knowledge Discovery from Advanced Databases*, *9*, 65-70.

Tatnall, A. (2005). Portals, Portals Everywhere. In *Web Portals: The New Gateways to Internet Information and Services* (p. 14). 10.4018/978-1-59140-438-5.ch001

Toutanova, K., & Johnson, M. (2007). A Bayesian LDA-based model for semi-supervised part-of-speech tagging.

Trotman, A. (2005). Learning to Rank. *Information Retrieval*, *8*, 359-381. 10.1007/s10791-005-6991-7

Tu, K., Ribeiro, B., Jiang, H., Wang, X., Jensen, D., Liu, B., & Towsley, D. (2014). Online Dating Recommendations: Matching Markets and Learning Preferences. *Proceedings of the 23rd International Conference on World Wide Web*, 787-792. 10.1145/2567948.2579240

Turpin, A., & Scholer, F. (2006). User Performance versus Precision Measures for Simple Search Tasks. *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 11-18. 10.1145/1148170.1148176

van Zaanen, M., & Kanters, P. H. M. (2010). Automatic mood classification using tf*idf based on lyrics (J. Downie & R. Veltkamp, Eds.). *11th International Society for Music Information Retrieval Conference (ISMIR 2010)*, 75-80.

Vijayarani, S., Ilamathi, J., & Nithya, S. (2015). Preprocessing Techniques for Text Mining-An Overview. *International Journal of Computer Science & Communication Networks, 5*(1), 7-16.

Voorhees, E. M. (1999). Natural Language Processing and Information Retrieval. *Information Extraction: Towards Scalable, Adaptable Systems*, 32-48.

Waegel, D. B. (2006). The Development of Text-Mining Tools and Algorithms.

Wagstaff, K., Cardie, C., Rogers, S., & Schrödl, S. (2001). Constrained K-Means Clustering with Background Knowledge. *Proceedings of the Eighteenth International Conference on Machine Learning*, 577-584.

Watstein, S. B., Calarco, P. V., & Ghaphery, J. S. (1999). Digital Library: keywords. *Reference Service Review*, *27*(4), 344-352.

Wei, X., & Croft, W. B. (2006). LDA-Based Document Models for Ad-Hoc Retrieval. *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 178-185. 10.1145/1148170.1148204

Wellers, D., Elliott, '., & Noga, M. (2017, May 31). *8 Ways Machine Learning Is Improving Companies' Work Processes*. Harvard Business Review. https://hbr.org/2017/05/8-ways-machine-learning-is-improving-companies-work-processes

Witten, I. H., Loots, M., Trujillo, M. F., & Bainbridge, D. (2002). The Promise of Digital Libraries in Developing Countries. *Communications of the ACM*, *44*(5), 82-85. 10.1145/374308.374357

Wynne, H. E., & Wint, Z. Z. (2019). Content Based Fake News Detection Using N-Gram Models. *Proceedings of the 21st International Conference on Information*

Integration and Web-Based Applications & Services, 669-673. 10.1145/3366030.3366116

Xie, H. I. (2006). Users' evaluation of digital libraries (DLs): Their uses, their criteria, and their assessment. *Information Processing & Management, 44*(3), 1346-1373. 10.1016/j.ipm.2007.10.003

Xu, J. (2018). Topic Modeling with LSA, PLSA, LDA & lda2Vec.

Xu, R., & Wunsch, D. (2005). Survey of clustering algorithms. *IEEE Transactions on Neural Networks*, *16*(3), 645-678. 10.1109/TNN.2005.845141

Xuan, N. P., Le Quang, H., Denoux, T., Tran, D., Le, A., & Pham, S. (2014). A New Improved Term Weighting Scheme for Text Categorization (V. Huynh, Ed.). *Knowledge and Systems Engineering. Advances in Intelligent Systems and Computing,, 244*. 10.1007/978-3-319-02741-8_23

Yao, M., Pi, D., & Cong, X. (2012). Chinese Text Clustering Algorithm Based k-means. *Physics Procedia*, *33*, 301-307.

Yi, B., Qiao, H., Yang, F., & Xu, C. (2010). An Improved Initialization Center Algorithm for K-Means Clustering. *2010 International Conference on Computational Intelligence and Software Engineering*, 1-4. 10.1109/CISE.2010.5676975

Yuan, C., & Yang, H. (2019). Research on K-Value Selection Method of K-Means Clustering Algorithm. *Multidisciplinary Scientific Journal*. 10.3390/J2020016

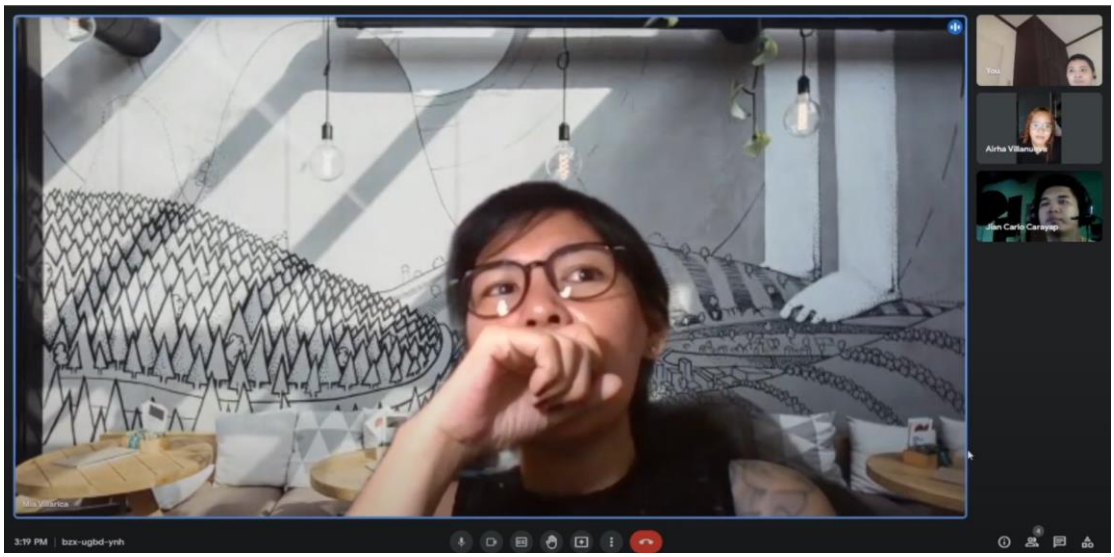
Zhao, W.-L., Deng, C.-H., & Ngo, C.-W. (2018). k-means: A revisit. *Neurocomputing*, *291*, 195-2006.

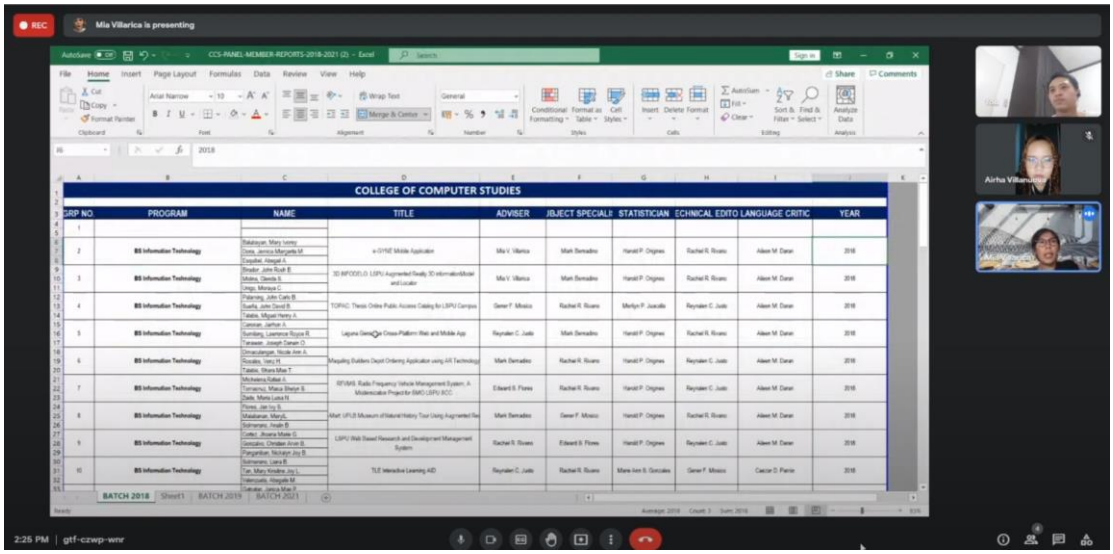
APPENDICES

APPENDIX A

Technical Background

Interview


You
Airha Villanueva
Jan Carlo Carayap
Mia Villarica
3:19 PM | bzx-ugbd-ynh


REC
Mia Villarica is presenting
Airha Villanueva
COLLEGE OF COMPUTER STUDIES
PROGRAM
NAME
TITLE
ADVISER
STATISTICIAN
YEAR
BS Information Technology
BATCH 2018
Sheet1
BATCH 2019
BATCH 2021
2:25 PM | gtf-czwp-wnr

APPENDIX B

Communication Letter & Forms

Request Letter

Republic of the Philippines
Laguna State Polytechnic University
Province of Laguna

COLLEGE OF COMPUTER STUDIES

Mrs. Reynalen C. Justo, MM-ITM
Dean/Associate Dean, CCS
College of Computer Studies, LSPU-SCC

August 11, 2021

Dear Ma'am,

Good day! We, students of College of Computer Studies currently taking Bachelor of Science in Computer Science from Laguna State Polytechnic University - Santa Cruz Campus, would like to request an interview regarding our system proposal as part of our requirements for Capstone Project 1. This will also be beneficial for your concern by giving you rights to use the system to possibly enhance your practices in handling research papers of undergraduate students. Furthermore, it would benefit the students of College of Computer Studies in gathering ideas and literatures related to their study.

To this, we ask for your permission to allow us conduct a preliminary investigation and ask a few questions and possibly recommendations for the betterment of the system.

We would highly appreciate whatever assistance you could extend to us that can be both beneficial to our study and yours personally.

Rest assured that we will abide/conform to the rules and regulations and follow strict confidentiality. No information will be divulged to unauthorized persons or organizations.

Thank you and we look forward for your kind consideration and support.

Respectfully yours,

Miguel N. Cemiller
Student/Researcher

Jian Carlo C. Carayap
Student/Researcher

Airha M. Villanueva
Student/Researcher

Noted:

Mia V. Villarica, DIT
Research Implementing Unit Head, CCS

APPROVED/DISAPPROVED:

Title Proposal Statement

Republic of the Philippines
Laguna State Polytechnic University
Province of Laguna

COLLEGE OF COMPUTER STUDIES

Title Proposal

Proponent/Researcher:	**Miguel N. Cemiller** Jian Carlo C. Carayap Airha M. Villanueva
Project Title: CCS Research Hub: A Research Digital Library Using Topic Modeling Algorithm with Chatbot	
Introduction	The Digital Revolution has affected nearly every aspect of library and museum services, from the automation of internal recordkeeping systems to the digitization of physical collections, and from the acquisition of new "born-digital" works of art or library publications to the use of technology to present collections and engage audiences. The digital library is the collection of services and the collection of information objects that support users in dealing with information objects available directly or indirectly via electronic/digital means (Leiner, 1998). However, digital libraries basically store materials in electronic format and manipulate large collections of those materials. Research into digital libraries is research into network information systems, concentrating on how to develop the necessary infrastructure to effectively mass-manipulate the information on the Net (NFS, 1999). The goal of the proposed study is to develop a digital library for capstone projects and undergraduate theses with Topic Modeling algorithm and chatbot for the 3rd year students of Laguna State Polytechnic University under the College of Computer Studies (CCS), research coordinator, and dean of the college. This will allow the research coordinator to see large themes in giant groups of documents as clusters of word and document by topic. Furthermore, a chatbot will also allow them to easily communicate with the students whenever the research coordinator isn't around. This will result in lesser tasks for the research coordinator. CCS Research Hub: A Research Digital Library Using Topic Modeling Algorithm with Chatbot will be a web application. The algorithm that will be used for Topic Modeling is Latent Dirichlet Allocation (LDA). It is a generative probabilistic model in which each document is assumed to be consisting of a different proportion of topics. Furthermore, the proposed system will consist of a chat feature for student's consultation and chatbot. The proponents will be using Rapid Application Development (RAD) and MySQL database in the development. Furthermore, system testing and evaluating of the algorithms will be conducted.
Statement of the Problem	Digital libraries basically store materials in electronic format and manipulate large collections of those materials. Thus, this can be hard to see into a large group of collections of research for the research coordinator. The proponents aspire to develop a digital library for capstone projects and undergraduate theses with Topic Modeling algorithm and chatbot for the 3rd year students of Laguna State Polytechnic University under the College of Computer Studies (CCS), research coordinator, and dean of the college. This will allow the research coordinator to see large themes in giant groups of documents as clusters of word and document by topic. Furthermore, a chatbot will also allow them to easily communicate with the students whenever the research coordinator isn't around. This will result in lesser tasks for the research coordinator.

	CCS Research Hub: A Research Digital Library Using Topic Modeling Algorithm with Chatbot will be a web application. The algorithm that will be used for Topic Modeling is Latent Dirichlet Allocation (LDA). It is a generative probabilistic model in which each document is assumed to be consisting of a different proportion of topics. Furthermore, the proposed system will consist of a chat feature for student's consultation and chatbot. The proponents will be using Rapid Application Development (RAD) and MySQL database in the development. Furthermore, system testing and evaluating of the algorithms will be conducted.
Objective of the Study	The study aims to develop, test and evaluate the algorithms of a digital library that would contribute to the productivity of the research coordinator and undergraduate students of CCS by having the functionality for data reporting, chat, and storage of data. Specifically, this study sought to find out the following: 1. To identify how to import the library of documents of students 2. To develop and design a research digital library a. Data reporting on users who download documents and recommendation on who their group members are based on interest and skills b. Chat for consultation and chatbot c. Online database 3. To test the major functionalities of the system to be developed a. Report generation b. Chat c. Storage of data 4. To evaluate the algorithms of the system to be developed
Scope and Limitation of the Study	The proposal is focused on development and design of the system, system testing, and algorithm testing. However, the proposal has its limitations in terms of methodology. The following are the scope of the study: ● The proposed system will focus on three main functionalities: Report generation, chat feature, and storage of data. ● The chatbot's answers will be based on the most frequently asked questions of the 3rd year undergraduate students of CCS. ● The study's primary respondents would be the 3rd year undergraduate students of CCS and their research coordinator, Ms. Mia Villarica. IT experts will also be involved as part of the system testing. ● The study will also cover the use of qualitative research methods for the proposed system's development and quantitative research methods for algorithm testing. The following are the limitations of the study: ● The study does not cover deployment or implementation.
Review of Literatures and Related Systems	The proponent collected information that is greatly related to the study. These related studies and literature are compiled in this section as a basis in the development of the system. Related studies and projects are listed below in thematic order:

Theme 1: Digital Libraries

Based on the study entitled "Digital Library Evaluation" written by Tefko Saracevic (2001), Digital libraries have a short yet turbulent and explosive history. A number of early visionaries it has a notion of libraries in the future being highly innovative and different in structure, processing, and access through heavy applications of technology. But, besides visionary and futuristic discussions and highly scattered research and developmental experimentation, nothing much happened in the next two decades. While the exciting history has yet to be written, Borgman's (1999) discussion of competing visions for digital libraries is a good beginning for understanding the forces and players involved. These competing visions and associated definitions come from several communities that are involved in digital library work. The work of two communities, research and practice, are reviewed below. While they work and proceed mostly independently of each other, they can be considered as two ends of a spectrum, which as yet have not met in the middle.

Theme 2: Topic Modeling of Digital Libraries

The previously stated literature shows how digital libraries are used in the society.

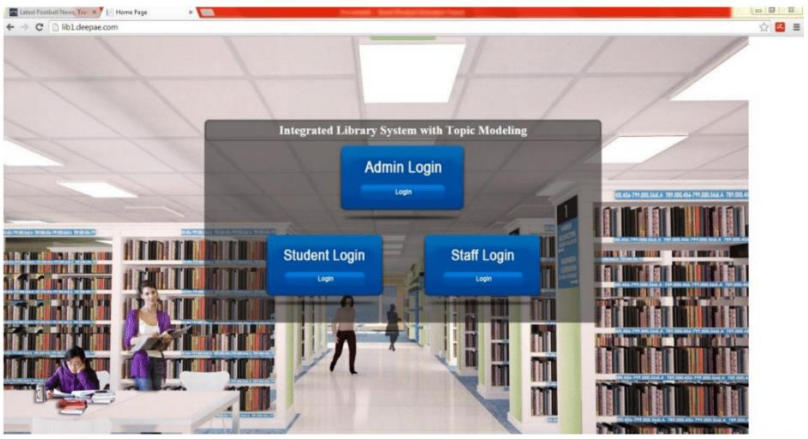
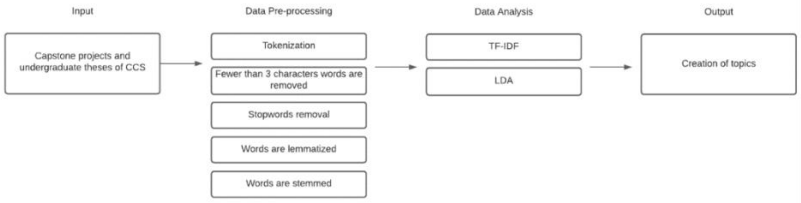
According to Newman, D., Noh, Y., Talley, E., Karimi, S., & Baldwin, T. (2010). Evaluating topic models for digital libraries, topic models could have a huge impact on improving the ways users find and discover content in digital libraries and search interfaces, through their ability to automatically learn and apply subject tags to each and every item in a collection, and their ability to dynamically create virtual collections on the fly. However, much remains to be done to tap this potential, and empirically evaluate the true value of a given topic model to humans. For all types of collections encompassing a wide range of genres and disciplines, the volume of online content accessible through digital libraries and search interfaces is always expanding. Hundreds of volumes are scanned every day in programs like the Google Books Library and the Internet Archive.

Theme 3: Chatbot in Digital Libraries

The previously stated literature shows the importance of using Topic Modeling in digital libraries.

Based on Chatbot for University Related FAQs(2017), chatbots are programs that mimic human conversation using Artificial Intelligence (AI). It is designed to be the ultimate virtual assistant, entertainment purpose, helping one to complete tasks ranging from answering questions, getting driving directions, turning up the thermostat in smart home, to playing one's favorite tunes etc. Chatbot has become more popular in business groups right now as they can reduce customer service cost and handle multiple users at a time. But yet to accomplish many tasks there is need to make chatbots as efficient as possible.

According to Anbang Xu, Zhe Liu, Yufan Guo, Vibha Sinha, Rama Akkiraju(2017). A New Chatbot for Customer Service on Social Media. Users are rapidly turning to social media to request and receive customer service; however, a majority of these requests were not addressed timely or even not addressed at all. To overcome the problem, they created a new conversational system to automatically generate responses for users requests on social media. The system is integrated with state-of-the-art deep learning techniques and is trained by nearly 1M Twitter conversations between users and agents from over 60 brands. The evaluation reveals that over 40% of the requests are emotional, and the system is about as good as human agents in showing empathy to help users cope with emotional situations. Results also show that their system outperforms information retrieval systems based on both human judgments and an automatic evaluation metric.

	In this section, the proponents will discuss the current state of technology and project related to the research topic. **Related Project:** The library system developed by Adebayo (2019) is a library management system with topic modeling to do topic extraction for well-informed decisions to save time flipping through already borrowed books to find relevant content the reader is looking for.  **Figure 1**. An Integrated Library System with Topic Modeling
Conceptual Framework	 **Figure 2**. Conceptual Framework of CCS Research Hub: A Research Digital Library Using Topic Modeling Algorithm with Chatbot

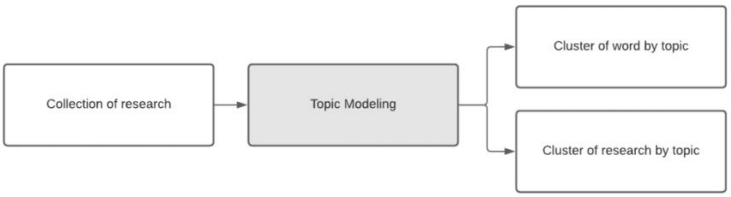


Figure 3. Latent Dirichlet Allocation (LDA) Algorithm

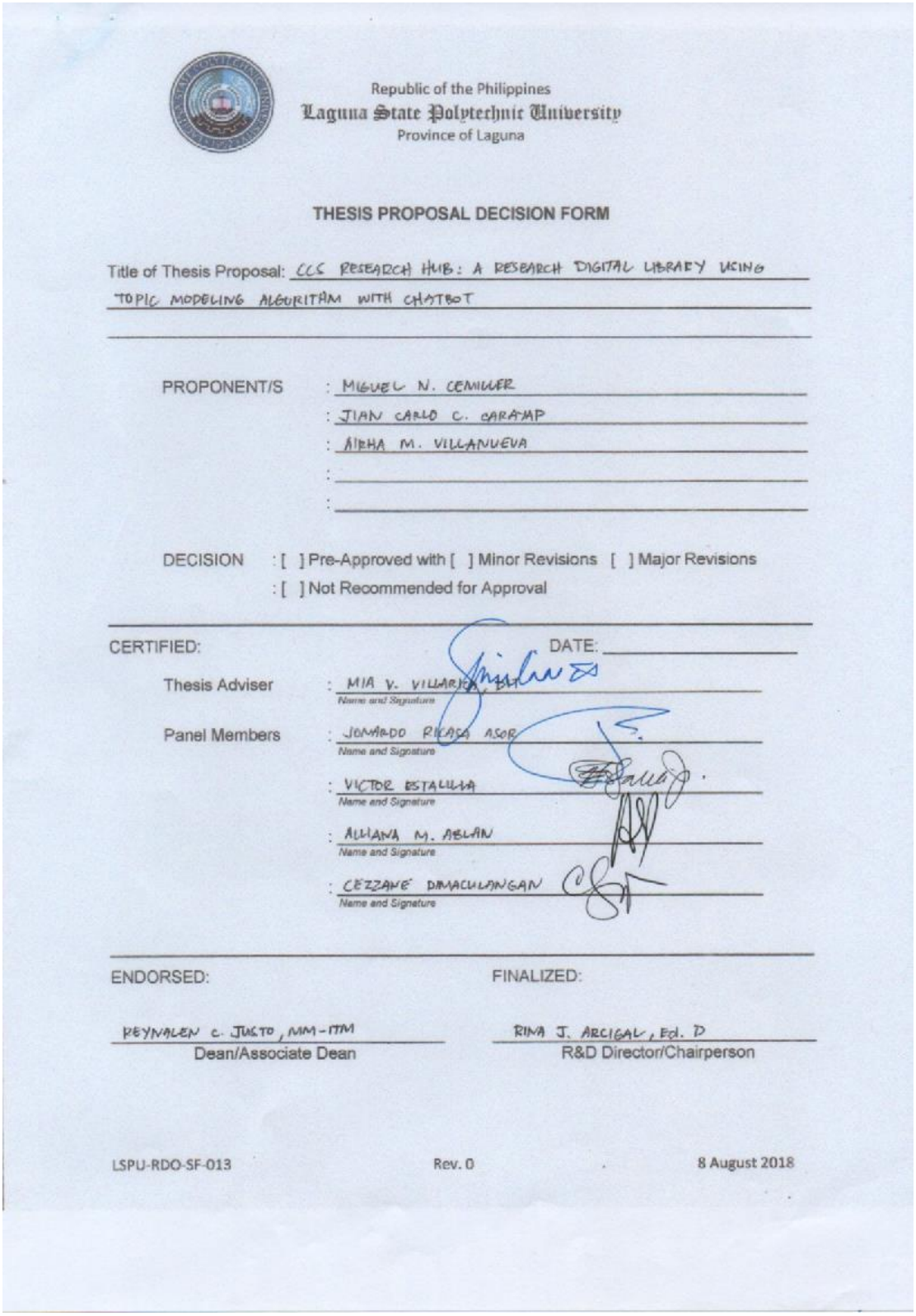
The conceptual framework of the study illustrates how the research will be conducted. The framework reflects the research objectives and activities to be conducted.

References

1. Seadle, M., & Greifeneder, E. (2007, March 18). *Defining a digital library*. https://edoc.hu-berlin.de/bitstream/handle/18452/9987/5.pdf?sequence=1.
2. *Saracevic, T. (2001). Digital Library Evaluation: Toward an Evolution of Concepts. https://www.ideals.illinois.edu/bitstream/handle/2142/8343/librarytrendsv49i2i_opt.pdf.*
3. *Newman, D., Noh, Y., Talley, E., Karimi, S., & Baldwin, T. (2010). Evaluating topic models for digital libraries. Proceedings of the 10th Annual Joint Conference on Digital Libraries - JCDL '10. doi:10.1145/1816123.1816156*
4. *Singh, S. (2017, September). (PDF) chatbot for university related FAQs. ResearchGate. https://www.researchgate.net/publication/321507021_Chatbot_for_university_related_FAQs.*
5. *Leiner, B. (1988, December). The ncstrl approach to open architecture for the confederated digital library. https://dlib.org/dlib/december98/leiner/12leiner.html.*

Prepared by: **Miguel N. Cemiller**

ISO Forms



Republic of the Philippines
Laguna State Polytechnic University
Province of Laguna

THESIS PROPOSAL DECISION FORM

Title of Thesis Proposal: CCS RESEARCH HUB: A RESEARCH DIGITAL LIBRARY USING TOPIC MODELING ALGORITHM WITH CHATBOT

PROPONENT/S : MIGUEL N. CEMILLER
: JIAN CARLO C. CARAMP
: AIRHA M. VILLANUEVA
:
:

DECISION : [ ] Pre-Approved with [ ] Minor Revisions [ ] Major Revisions
: [ ] Not Recommended for Approval

CERTIFIED: DATE:

Thesis Adviser : MIA V. VILLARIN
Name and Signature

Panel Members : JONARDO RICASA ASOR
Name and Signature
: VICTOR ESTALILIA
Name and Signature
: ALLIANA M. ABLAN
Name and Signature
: CEZZANE DIMACULANGAN
Name and Signature

ENDORSED: FINALIZED:

REYNALEN C. JUSTO, MM-ITM
Dean/Associate Dean

RINA J. ARCIGAL, Ed. D
R&D Director/Chairperson

LSPU-RDO-SF-013 Rev. 0 8 August 2018

Republic of the Philippines
Laguna State Polytechnic University
Province of Laguna

THESIS ADVISER NOMINATION

Student/s Name and Signature : MIGUEL N. CEMICLER

: JIAN CARLO C. CARAYAP

: AINHA MONTOYA VILLANUEVA

:

Degree Sought : BACHELOR OF SCIENCE IN COMPUTER SCIENCE

Major Field (if applicable) :

I/we, hereby nominate as Thesis Adviser

Name of Faculty : MIA V. VILLARICA, DIT

College : COLLEGE OF COMPUTER STUDIES

Field of Specialization :

I, hereby indicate my willingness to serve as the Thesis Adviser.

Signature of Faculty : _____ 2/14/22
Date Signed

Endorsed By:

MIA V. VILLARICA, DIT
Research Coordinator

Recommending Approval:

REYNALEN C. JUSTO, MM - ITM
Dean/Associate Dean

APPROVED/DISAPPROVED:

DINA J. ARCIGAL, Ed. D
R&D Director/Chairperson

Republic of the Philippines
Laguna State Polytechnic University
Province of Laguna

THESIS PANEL MEMBER NOMINATION

Student/s Name and Signature : MIGUEL M. CENILLER

: JUAN CARLO C. CARIÑO

: AINHOA MONTOYA VILLANUEVA

:

Degree Sought : BACHELOR OF SCIENCE IN COMPUTER SCIENCE

Major Field (if applicable) :

I/we, hereby nominate as Panel Member (TECHNICAL EDITOR)

Name of Faculty : ALLIANA M. ABLAN

College : COLLEGE OF COMPUTER STUDIES

Field of Specialization :

I, hereby indicate my willingness to serve as a Panel Member.

Signature of Faculty : _____ Date Signed

Endorsed By:

MIA D. VILLARICA, DIT
Research Coordinator

Recommending Approval:

REYNALEN C. JUSTO, MM-ITM
Dean/Associate Dean

APPROVED/DISAPPROVED:

RINA J. ARLIGAL, Ed. D
R&D Director/Chairperson

Republic of the Philippines
Laguna State Polytechnic University
Province of Laguna

THESIS PANEL MEMBER NOMINATION

Student/s Name and Signature : MIGUEL N. CEMILLER

: JUAN CARLO C. CARAYAP

: ALYSSA MARIONA VILLANUEVA

:

Degree Sought : BACHELOR OF SCIENCE IN COMPUTER SCIENCE

Major Field (if applicable) :

I/we, hereby nominate as Panel Member (SUBJECT SPECIALIST)

Name of Faculty : JONARDO RICASA ASOR

College : COLLEGE OF COMPUTER STUDIES - LOS BAÑOS

Field of Specialization :

I, hereby indicate my willingness to serve as a Panel Member.

Signature of Faculty : 2/16/22
Date Signed

Endorsed By:

MA V. VILLARICA, DIT
Research Coordinator

Recommending Approval:

REYNALEN C. JUSTO, MM-ITM
Dean/Associate Dean

APPROVED/DISAPPROVED:

RINA J. ARCIGAL, Ed. D
R&D Director/Chairperson

Republic of the Philippines
Laguna State Polytechnic University
Province of Laguna

THESIS PANEL MEMBER NOMINATION

Student/s Name and Signature : MIGUEL M. CEMILLER

: JUAN CARLO C. CARAYAP

: AIRA MONTOYA VILLANUEVA

:

Degree Sought : BACHELOR OF SCIENCE IN COMPUTER SCIENCE

Major Field (if applicable) :

I/we, hereby nominate as Panel Member (STATISTICIAN)

Name of Faculty : VICTOR ESTALILLA

College : COLLEGE OF ARTS AND SCIENCES

Field of Specialization :

I, hereby indicate my willingness to serve as a Panel Member.

Signature of Faculty : _____ _____
Date Signed

Endorsed By:

MARY V. VILLARICA, DIT
Research Coordinator

Recommending Approval:

REYNALDO C. JUSTO, MM-ITM
Dean/Associate Dean

APPROVED/DISAPPROVED:

RINA J. ARCIBAL, Ed. D
R&D Director/Chairperson

Republic of the Philippines
Laguna State Polytechnic University
Province of Laguna

THESIS PANEL MEMBER NOMINATION

Student/s Name and Signature : MIGUEL N. CAMILLER

: JUAN CARLO C. CARAYAP

: ALYSSA MONTOYA VILLANUEVA

:

Degree Sought : BACHELOR OF SCIENCE IN COMPUTER SCIENCE

Major Field (if applicable) :

I/we, hereby nominate as Panel Member (LANGUAGE CRITIC)

Name of Faculty : CEZZANE DIMACULANGAN

College : COLLEGE OF ARTS AND SCIENCES

Field of Specialization :

I, hereby indicate my willingness to serve as a Panel Member.

Signature of Faculty : _____ _____
Date Signed

Endorsed By:

MIA V. VILLARICA, DIT
Research Coordinator

Recommending Approval:

REYNALEN C. JUSTO, MM-ITM
Dean/Associate Dean

APPROVED/DISAPPROVED:

RINA J. ARCIGAL, Ed. D
R&D Director/Chairperson

Republic of the Philippines
Laguna State Polytechnic University
Province of Laguna

CONSENT FOR CO-AUTHORSHIP FORM

AUG. 11, 2021
Date

I/we, MIGUEL N. CEMILLER,
JIAN CARLO C. CARIAGA,
AIRHA M. VILLANUEVA,
_____, the primary author/s of the thesis entitled "CCS RESEARCH HUB: A RESEARCH DIGITAL LIBRARY USING TOPIC MODELING ALGORITHM AND CHATBOT _____," hereby acknowledge the significant intellectual contributions of the following faculty members in the completion of the study.

[ ] Thesis Adviser	MIA V. VILLARICA, DIT
[ ] Panel Member/s	JONARDO RICASA ASOR
	VICTOR ESTALILLA
	ALLIANA M. ABLAN
	CEZZANE DIMACULANGAN

I/we, in full knowledge, extend ALL RIGHTS AND PRIVILEGES as CO-AUTHOR/S to the above-mentioned name/s in any form of written publications and/or presentations.

CONFORME:

MIGUEL N. CEMILLER Student's Name and Signature	JIAN CARLO C. CARIAGA Student's Name and Signature
AIRHA MONTOYA VILLANUEVA Student's Name and Signature	_____ Student's Name and Signature

NOTED:

REYNALDO C. JUSTO, MM-ITM Dean/Associate Dean	RINA J. ARCILAL, Ed. D Director/Chairperson, R&D

NOTARIZED BY:

Defense Rating Sheets

Republic of the Philippines
Laguna State Polytechnic University
Province of Laguna

COLLEGE OF COMPUTER STUDIES

Topic Proposal Rating Sheet

Adviser: Mia V. Villarica Specialization Expert: Jonardo R. Asor
Date: _____ Group No. CS3A-05 Time: _____
Name of Proponents: Cemiller, Miguel N.
Caravap, Jian Carlo C.
Villanueva, Airha M.
Research Title: Papernerds: A Research Digital Library and Undergraduate Thesis Portal with Text Mining Algorithms

MANUSCRIPT (20%)		
CRITERIA	WEIGHT (%)	EVALUATED SCORE
1. Format (documentation, chapter division, style including neatness and organization of details.)	5	
2. Research Problems and Objectives (discuss the problems encountered by the client and answered with appropriate and adequate solutions)	5	
3. Related Literature and Studies (includes 10 for literature and 10 studies and summarize the discussion on synthesis)	5	
4. Research Methodology (appropriateness of methods of study, statistical treatment, analysis and interpretations)	5	
SUB – TOTAL	20	
1. Presentation (content and creativity of visual aid and/ or graphics and mastery of study evidenced by logical presentation of the conclusion)	10	
2. Defense (Ability to answer reasoning capability and ability to justify interpretation and conclusion)	10	
SUB – TOTAL	20	
1. Innovation	20	
2. Application, Relevance and Impact	20	
3. Research Thrust Impact	20	
SUB – TOTAL	60	
OVER ALL TOTAL	100	

PROPOSAL VERDICTS

[ ] **APPROVED**. Minor revisions are necessary but they do not have to be presented in front of and checked by all panelists. 86 – 100
[ ] **APPROVED WITH REVISIONS**. Major revisions shall be incorporated in the final copy of the revised Project Proposal summary. These must be checked by the panelists. 75 – 85
[ ] **DISAPPROVED**. The Proponents/Researchers failed to propose a researchable or scholarly Thesis / Capstone Project. Below 75

THESIS ADVISER
Role on the Defense

MIA V. VILLARICA
Signature over Printed Name of Evaluator

Form 4
CCS-SR-

Republic of the Philippines
Laguna State Polytechnic University
Province of Laguna

COLLEGE OF COMPUTER STUDIES

Topic Proposal Rating Sheet

Adviser: Mia V. Villarica Specialization Expert: Jonardo R. Asor
Date: _____ Group No. CS3A-05 Time: _____
Name of Proponents: Cemiller, Miguel N.
Caravap, Jian Carlo C.
Villanueva, Airha M.
Research Title: Papernerds: A Research Digital Library and Undergraduate Thesis Portal with Text Mining Algorithms

MANUSCRIPT (20%)		
CRITERIA	WEIGHT (%)	EVALUATED SCORE
1. Format (documentation, chapter division, style including neatness and organization of details.)	5	
2. Research Problems and Objectives (discuss the problems encountered by the client and answered with appropriate and adequate solutions)	5	
3. Related Literature and Studies (includes 10 for literature and 10 studies and summarize the discussion on synthesis)	5	
4. Research Methodology (appropriateness of methods of study, statistical treatment, analysis and interpretations)	5	
SUB – TOTAL	20	
1. Presentation (content and creativity of visual aid and/ or graphics and mastery of study evidenced by logical presentation of the conclusion)	10	
2. Defense (Ability to answer reasoning capability and ability to justify interpretation and conclusion)	10	
SUB – TOTAL	20	
1. Innovation	20	
2. Application, Relevance and Impact	20	
3. Research Thrust Impact	20	
SUB – TOTAL	60	
OVER ALL TOTAL	100	

PROPOSAL VERDICTS

[ ] **APPROVED**. Minor revisions are necessary but they do not have to be presented in front of and checked by all panelists. 86 – 100
[ ] **APPROVED WITH REVISIONS**. Major revisions shall be incorporated in the final copy of the revised Project Proposal summary. These must be checked by the panelists. 75 – 85
[ ] **DISAPPROVED**. The Proponents/Researchers failed to propose a researchable or scholarly Thesis / Capstone Project. Below 75

TECHNICAL EDITOR
Role on the Defense

ALLIANA M. TABLAN
Signature over Printed Name of Evaluator

Form 4
CCS-SR-

Republic of the Philippines
Laguna State Polytechnic University
Province of Laguna

COLLEGE OF COMPUTER STUDIES

Topic Proposal Rating Sheet

Adviser: Mia V. Villarica Specialization Expert: Jonardo R. Asor
Date: _____ Group No. CS3A-05 Time: _____
Name of Proponents: Cemiller, Miguel N.
Carayap, Jian Carlo C.
Villanueva, Airha M.
Research Title: Papernerds: A Research Digital Library and Undergraduate Thesis Portal with Text Mining Algorithms

MANUSCRIPT (20%)		
CRITERIA	WEIGHT (%)	EVALUATED SCORE
1. Format (documentation, chapter division, style including neatness and organization of details.)	5	
2. Research Problems and Objectives (discuss the problems encountered by the client and answered with appropriate and adequate solutions)	5	
3. Related Literature and Studies (includes 10 for literature and 10 studies and summarize the discussion on synthesis)	5	
4. Research Methodology (appropriateness of methods of study, statistical treatment, analysis and interpretations)	5	
SUB – TOTAL	20	
1. Presentation (content and creativity of visual aid and/ or graphics and mastery of study evidenced by logical presentation of the conclusion)	10	
2. Defense (Ability to answer reasoning capability and ability to justify interpretation and conclusion)	10	
SUB – TOTAL	20	
1. Innovation	20	
2. Application, Relevance and Impact	20	
3. Research Thrust Impact	20	
SUB – TOTAL	60	
OVER ALL TOTAL	100	

PROPOSAL VERDICTS

[ ] **APPROVED**. Minor revisions are necessary but they do not have to be presented in front of and checked by all panelists. 86 – 100
[ ] **APPROVED WITH REVISIONS**. Major revisions shall be incorporated in the final copy of the revised Project Proposal summary. These must be checked by the panelists. 75 – 85
[ ] **DISAPPROVED**. The Proponents/Researchers failed to propose a researchable or scholarly Thesis / Capstone Project. Below 75

SUBJECT SPECIALIST
Role on the Defense

JONARDO R. ASOR
Signature over Printed Name of Evaluator

Form 4
CCS-SR-

Republic of the Philippines
Laguna State Polytechnic University
Province of Laguna

COLLEGE OF COMPUTER STUDIES

Topic Proposal Rating Sheet

Adviser: Mia V. Villarica Specialization Expert: Jonardo R. Asor
Date: _____ Group No. CS3A-05 Time: _____
Name of Proponents: Cemiller, Miguel N.
Carayap, Jian Carlo C.
Villanueva, Airha M.
Research Title: Papernerds: A Research Digital Library and Undergraduate Thesis Portal with Text Mining Algorithms

MANUSCRIPT (20%)		
CRITERIA	WEIGHT (%)	EVALUATED SCORE
1. Format (documentation, chapter division, style including neatness and organization of details.)	5	
2. Research Problems and Objectives (discuss the problems encountered by the client and answered with appropriate and adequate solutions)	5	
3. Related Literature and Studies (includes 10 for literature and 10 studies and summarize the discussion on synthesis)	5	
4. Research Methodology (appropriateness of methods of study, statistical treatment, analysis and interpretations)	5	
SUB – TOTAL	20	
1. Presentation (content and creativity of visual aid and/ or graphics and mastery of study evidenced by logical presentation of the conclusion)	10	
2. Defense (Ability to answer reasoning capability and ability to justify interpretation and conclusion)	10	
SUB – TOTAL	20	
1. Innovation	20	
2. Application, Relevance and Impact	20	
3. Research Thrust Impact	20	
SUB – TOTAL	60	
OVER ALL TOTAL	100	

PROPOSAL VERDICTS

[ ] **APPROVED**. Minor revisions are necessary but they do not have to be presented in front of and checked by all panelists. 86 – 100
[ ] **APPROVED WITH REVISIONS**. Major revisions shall be incorporated in the final copy of the revised Project Proposal summary. These must be checked by the panelists. 75 – 85
[ ] **DISAPPROVED**. The Proponents/Researchers failed to propose a researchable or scholarly Thesis / Capstone Project. Below 75

STATISTICIAN
Role on the Defense

VICTOR A. ECTALIMA JR.
Signature over Printed Name of Evaluator

Form 4
CCS-SR-

Republic of the Philippines
Laguna State Polytechnic University
Province of Laguna

COLLEGE OF COMPUTER STUDIES

Topic Proposal Rating Sheet

Adviser: Mia V. Villarica Specialization Expert: Jonardo R. Asor
Date: _____ Group No. CS3A-05 Time: _____
Name of Proponents: Cemiller, Miguel N.
Carayap, Jian Carlo C.
Villanueva, Airha M.
Research Title: Papemerds: A Research Digital Library and Undergraduate Thesis Portal with Text Mining Algorithms

MANUSCRIPT (20%)		
CRITERIA	WEIGHT (%)	EVALUATED SCORE
1. Format (documentation, chapter division, style including neatness and organization of details.)	5	
2. Research Problems and Objectives (discuss the problems encountered by the client and answered with appropriate and adequate solutions)	5	
3. Related Literature and Studies (includes 10 for literature and 10 studies and summarize the discussion on synthesis)	5	
4. Research Methodology (appropriateness of methods of study, statistical treatment, analysis and interpretations)	5	
SUB – TOTAL	20	
1. Presentation (content and creativity of visual aid and/ or graphics and mastery of study evidenced by logical presentation of the conclusion)	10	
2. Defense (Ability to answer reasoning capability and ability to justify interpretation and conclusion)	10	
SUB – TOTAL	20	
1. Innovation	20	
2. Application, Relevance and Impact	20	
3. Research Thrust Impact	20	
SUB – TOTAL	60	
OVER ALL TOTAL	100	

PROPOSAL VERDICTS

[ ] **APPROVED**. Minor revisions are necessary but they do not have to be presented in front of and checked by all panelists. 86 – 100
[ ] **APPROVED WITH REVISIONS**. Major revisions shall be incorporated in the final copy of the revised Project Proposal summary. These must be checked by the panelists. 75 – 85
[ ] **DISAPPROVED**. The Proponents/Researchers failed to propose a researchable or scholarly Thesis / Capstone Project. Below 75

LANGUAGE CRITIC
Role on the Defense

MA. CEZZARE M. DIMAUNAHAN
Signature over Printed Name of Evaluator

Form 4
CCS-SR-

Republic of the Philippines
Laguna State Polytechnic University
Province of Laguna

COLLEGE OF COMPUTER STUDIES

Research Colloquium Rating Sheet

Adviser: Mia V. Villarica Specialization Expert: Jonardo R. Asor
Date: December 16, 2021 Group No. CS3A-05 Time: 10:00-10:30AM
Name of Proponents: Cemiller, Miguel N.
Carayap, Jian Carlo C.
Villanueva, Airha M.
Research Title: Papernerds: A Research Digital Library and Undergraduate Thesis Portal with Text Mining Algorithms

MANUSCRIPT (20%)		
CRITERIA	WEIGHT (%)	EVALUATED SCORE
1. Format (documentation, chapter division, style including neatness and organization of details.)	5	
2. Research Problems and Objectives (discuss the problems encountered by the client and answered with appropriate and adequate solutions)	5	
3. Related Literature and Studies (includes 10 for literature and 10 studies and summarize the discussion on synthesis)	5	
4. Research Methodology (appropriateness of methods of study, statistical treatment, analysis and interpretations)	5	
SUB – TOTAL	20	
1. Presentation (content and creativity of visual aid and/ or graphics and mastery of study evidenced by logical presentation of the conclusion)	10	
2. Defense (Ability to answer reasoning capability and ability to justify interpretation and conclusion)	10	
SUB – TOTAL	20	
1. Innovation	20	
2. Application, Relevance and Impact	20	
3. Research Thrust Impact	20	
SUB – TOTAL	60	
OVER ALL TOTAL	100	

PROPOSAL VERDICTS

[ ] **APPROVED**. Minor revisions are necessary but they do not have to be presented in front of and checked by all panelists. 86 – 100
[ ] **APPROVED WITH REVISIONS**. Major revisions shall be incorporated in the final copy of the revised Project Proposal summary. These must be checked by the panelists. 75 – 85
[ ] **DISAPPROVED**. The Proponents/Researchers failed to propose a researchable or scholarly Thesis / Capstone Project. Below 75

THESIS ADVISER
Role on the Defense

MIA V. VILLARICA
Signature over Printed Name of Evaluator

Form 4
CCS-SR-

Republic of the Philippines
Laguna State Polytechnic University
Province of Laguna

COLLEGE OF COMPUTER STUDIES

Research Colloquium Rating Sheet

Adviser: Mia V. Villarica Specialization Expert: Jonardo R. Asor
Date: December 16, 2021 Group No. CS3A-05 Time: 10:00-10:30AM
Name of Proponents: Cemiller, Miguel N.
Carayap, Jian Carlo C.
Villanueva, Airha M.
Research Title: Papernerds: A Research Digital Library and Undergraduate Thesis Portal with Text Mining Algorithms

MANUSCRIPT (20%)		
CRITERIA	WEIGHT (%)	EVALUATED SCORE
1. Format (documentation, chapter division, style including neatness and organization of details.)	5	
2. Research Problems and Objectives (discuss the problems encountered by the client and answered with appropriate and adequate solutions)	5	
3. Related Literature and Studies (includes 10 for literature and 10 studies and summarize the discussion on synthesis)	5	
4. Research Methodology (appropriateness of methods of study, statistical treatment, analysis and interpretations)	5	
SUB – TOTAL	20	
1. Presentation (content and creativity of visual aid and/ or graphics and mastery of study evidenced by logical presentation of the conclusion)	10	
2. Defense (Ability to answer reasoning capability and ability to justify interpretation and conclusion)	10	
SUB – TOTAL	20	
1. Innovation	20	
2. Application, Relevance and Impact	20	
3. Research Thrust Impact	20	
SUB – TOTAL	60	
OVER ALL TOTAL	100	

PROPOSAL VERDICTS

[ ] **APPROVED**. Minor revisions are necessary but they do not have to be presented in front of and checked by all panelists. 86 – 100
[ ] **APPROVED WITH REVISIONS**. Major revisions shall be incorporated in the final copy of the revised Project Proposal summary. These must be checked by the panelists. 75 – 85
[ ] **DISAPPROVED**. The Proponents/Researchers failed to propose a researchable or scholarly Thesis / Capstone Project. Below 75

TECHNICAL EDITOR
Role on the Defense

ALLIANA M. ABLAN
Signature over Printed Name of Evaluator

Form 4
CCS-SR-

Republic of the Philippines
Laguna State Polytechnic University
Province of Laguna

COLLEGE OF COMPUTER STUDIES

Research Colloquium Rating Sheet

Adviser: Mia V. Villarica Specialization Expert: Jonardo R. Asor
Date: December 16, 2021 Group No. CS3A-05 Time: 10:00-10:30AM
Name of Proponents: Cemiller, Miguel N.
Carayap, Jian Carlo C.
Villanueva, Airha M.
Research Title: Papernerds: A Research Digital Library and Undergraduate Thesis Portal with Text Mining Algorithms

MANUSCRIPT (20%)		
CRITERIA	WEIGHT (%)	EVALUATED SCORE
1. Format (documentation, chapter division, style including neatness and organization of details.)	5	5
2. Research Problems and Objectives (discuss the problems encountered by the client and answered with appropriate and adequate solutions)	5	4
3. Related Literature and Studies (includes 10 for literature and 10 studies and summarize the discussion on synthesis)	5	5
4. Research Methodology (appropriateness of methods of study, statistical treatment, analysis and interpretations)	5	5
SUB – TOTAL	20	19
ORAL DEFENSE (20%)		
1. Presentation (content and creativity of visual aid and/ or graphics and mastery of study evidenced by logical presentation of the conclusion)	10	9
2. Defense (Ability to answer reasoning capability and ability to justify interpretation and conclusion)	10	10
SUB – TOTAL	20	19
CAPSTONE/THESIS PROJECT (60%)		
1. Innovation	20	18
2. Application, Relevance and Impact	20	18
3. Research Thrust Impact	20	18
SUB – TOTAL	60	54
OVER ALL TOTAL	100	92

PROPOSAL VERDICTS

[ ] **APPROVED**. Minor revisions are necessary but they do not have to be presented in front of and checked by all panelists. 86 – 100
[ ] **APPROVED WITH REVISIONS**. Major revisions shall be incorporated in the final copy of the revised Project Proposal summary. These must be checked by the panelists. 75 – 85
[ ] **DISAPPROVED**. The Proponents/Researchers failed to propose a researchable or scholarly Thesis / Capstone Project. Below 75

SUBJECT SPECIALIST
Role on the Defense

JONARDO R. ASOR
Signature over Printed Name of Evaluator

Form 4
CCS-SR-

blic of the Philippines

Laguna State Polytechnic University

Province of Laguna

COLLEGE OF COMPUTER STUDIES

Research Colloquium Rating Sheet

Adviser: Mia V. Villarica Specialization Expert: Jonardo R. Asor
Date: December 16, 2021 Group No. CS3A-05 Time: 10:00-10:30AM
Name of Proponents: Cemiller, Miguel N.
Caravap, Jian Carlo C.
Villanueva, Airha M.
Research Title: Papernerds: A Research Digital Library and Undergraduate Thesis Portal with Text Mining Algorithms

MANUSCRIPT (20%)		
CRITERIA	WEIGHT (%)	EVALUATED SCORE
1. Format (documentation, chapter division, style including neatness and organization of details.)	5	5
2. Research Problems and Objectives (discuss the problems encountered by the client and answered with appropriate and adequate solutions)	5	4
3. Related Literature and Studies (includes 10 for literature and 10 studies and summarize the discussion on synthesis)	5	5
4. Research Methodology (appropriateness of methods of study, statistical treatment, analysis and interpretations)	5	5
SUB – TOTAL	20	19
ORAL DEFENSE (20%)		
1. Presentation (content and creativity of visual aid and/ or graphics and mastery of study evidenced by logical presentation of the conclusion)	10	9
2. Defense (Ability to answer reasoning capability and ability to justify interpretation and conclusion)	10	10
SUB – TOTAL	20	19
CAPSTONE/THESIS PROJECT (60%)		
1. Innovation	20	18
2. Application, Relevance and Impact	20	19
3. Research Thrust Impact	20	19
SUB – TOTAL	60	54
OVER ALL TOTAL	100	94

PROPOSAL VERDICTS

[ ] **APPROVED**. Minor revisions are necessary but they do not have to be presented in front of and checked by all panelists. 86 – 100

[ ] **APPROVED WITH REVISIONS**. Major revisions shall be incorporated in the final copy of the revised Project Proposal summary. These must be checked by the panelists. 75 – 85

[ ] **DISAPPROVED**. The Proponents/Researchers failed to propose a researchable or scholarly Thesis / Capstone Project. Below 75

Statistician
Role on the Defense

Victor A. Estalilla Jr.
Signature over Printed Name of Evaluator

Form 4
CCS-SR-

Republic of the Philippines
Laguna State Polytechnic University
Province of Laguna

COLLEGE OF COMPUTER STUDIES

Research Colloquium Rating Sheet

Adviser: Mia V. Villarica Specialization Expert: Jonardo R. Asor
Date: December 16, 2021 Group No. CS3A-05 Time: 10:00-10:30AM
Name of Proponents: Cemiller, Miguel N.
Carayap, Jian Carlo C.
Villanueva, Airha M.
Research Title: Papernerds: A Research Digital Library and Undergraduate Thesis Portal with Text Mining Algorithms

MANUSCRIPT (20%)		
CRITERIA	WEIGHT (%)	EVALUATED SCORE
1. Format (documentation, chapter division, style including neatness and organization of details.)	5	
2. Research Problems and Objectives (discuss the problems encountered by the client and answered with appropriate and adequate solutions)	5	
3. Related Literature and Studies (includes 10 for literature and 10 studies and summarize the discussion on synthesis)	5	
4. Research Methodology (appropriateness of methods of study, statistical treatment, analysis and interpretations)	5	
SUB – TOTAL	20	
1. Presentation (content and creativity of visual aid and/ or graphics and mastery of study evidenced by logical presentation of the conclusion)	10	
2. Defense (Ability to answer reasoning capability and ability to justify interpretation and conclusion)	10	
SUB – TOTAL	20	
1. Innovation	20	
2. Application, Relevance and Impact	20	
3. Research Thrust Impact	20	
SUB – TOTAL	60	
OVER ALL TOTAL	100	

PROPOSAL VERDICTS

[ ] **APPROVED**. Minor revisions are necessary but they do not have to be presented in front of and checked by all panelists. 86 – 100
[ ] **APPROVED WITH REVISIONS**. Major revisions shall be incorporated in the final copy of the revised Project Proposal summary. These must be checked by the panelists. 75 – 85
[ ] **DISAPPROVED**. The Proponents/Researchers failed to propose a researchable or scholarly Thesis / Capstone Project. Below 75

LANGUAGE CRITIC
Role on the Defense

MA. CEZANNE D. DIMACULANGAN
Signature over Printed Name of Evaluator

Form 4
CCS-SR-

Republic of the Philippines
Laguna State Polytechnic University
Province of Laguna

COLLEGE OF COMPUTER STUDIES

Final Oral Defense Rating Sheet

Date: February 5, 2021 Group No. CS3A-05 Time: 10:00-10:30am

Name of Proponents: Cemiller, Miguel N.
Carayap, Jian Carlo C.
Villanueva, Airha M.

Title: Papernerds: A Research Digital Library and Undergraduate Thesis Portal with Text Mining Algorithms

MANUSCRIPT (20%)		
CRITERIA	WEIGHT (%)	EVALUATED SCORE
1. Format (documents, chapter division, style including neatness and organization of details)	5	
2. Research problems and objectives (discuss the problems encountered by the client and answered with appropriate and adequate solutions)	5	
3. Related Literature and Studies (includes 10 for literature and 10 studies and summarize the discussion of synthesis)	5	
4. Research Methodology (appropriateness of methods of study statistical treatment, analysis and interpretation)	5	
SUB - TOTAL	20	
1. Presentation (content and creativity of visual aid and / or graphics and mastery of study evidenced by logical presentation of the conclusion)	10	
2. Defense (Ability to answer reasoning capability and ability to justify interpretation and conclusion)	10	
SUB - TOTAL	20	
1. Innovation	20	
2. Application Relevance and Impact	20	
3. Research Thrust Impact	20	
SUB - TOTAL	60	
OVER ALL TOTAL	100	

VERDICTS

[ ] **APPROVED**. Minor revisions are necessary but they do not have to be presented in front of and checked by all panelists. 86 – 100

[ ] **APPROVED WITH REVISIONS**. Major revisions shall be incorporated in the final copy of the revised Project Proposal summary. These must be checked by the panelists. 70 – 85

[ ] **DISAPPROVED**. The Proponents/Researchers failed to propose a researchable or scholarly Thesis / Capstone Project. Below 70

THESIS ADVISER
Role on the Defense

MIA V. VILLARICA
Signature over Printed Name of Evaluator

Form 4
CCS-SR

Republic of the Philippines
Laguna State Polytechnic University
Province of Laguna

COLLEGE OF COMPUTER STUDIES

Final Oral Defense Rating Sheet

Date: February 5, 2021 Group No. CS3A-05 Time: 10:00-10:30am

Name of Proponents: Cemiller, Miguel N.
Carayap, Jian Carlo C.
Villanueva, Airha M.

Title: Papernerds: A Research Digital Library and Undergraduate Thesis Portal with Text Mining Algorithms

MANUSCRIPT (20%)		
CRITERIA	WEIGHT (%)	EVALUATED SCORE
1. Format (documents, chapter division, style including neatness and organization of details)	5	
2. Research problems and objectives (discuss the problems encountered by the client and answered with appropriate and adequate solutions)	5	
3. Related Literature and Studies (includes 10 for literature and 10 studies and summarize the discussion of synthesis)	5	
4. Research Methodology (appropriateness of methods of study statistical treatment, analysis and interpretation)	5	
SUB - TOTAL	20	
1. Presentation (content and creativity of visual aid and / or graphics and mastery of study evidenced by logical presentation of the conclusion)	10	
2. Defense (Ability to answer reasoning capability and ability to justify interpretation and conclusion)	10	
SUB - TOTAL	20	
1. Innovation	20	
2. Application Relevance and Impact	20	
3. Research Thrust Impact	20	
SUB - TOTAL	60	
OVER ALL TOTAL	100	

VERDICTS

[ ] **APPROVED**. Minor revisions are necessary but they do not have to be presented in front of and checked by all panelists. 86 – 100

[ ] **APPROVED WITH REVISIONS**. Major revisions shall be incorporated in the final copy of the revised Project Proposal summary. These must be checked by the panelists. 70 – 85

[ ] **DISAPPROVED**. The Proponents/Researchers failed to propose a researchable or scholarly Thesis / Capstone Project. Below 70

TECHNICAL EDITOR
Role on the Defense

ALLIANA M. ABLAN
Signature over Printed Name of Evaluator

Form 4
CCS-SR

Republic of the Philippines
Laguna State Polytechnic University
Province of Laguna

COLLEGE OF COMPUTER STUDIES

Final Oral Defense Rating Sheet

Date: February 5, 2021 Group No. CS3A-05 Time: 10:00-10:30am

Name of Proponents: Cemiller, Miguel N.
Carayap, Jian Carlo C.
Villanueva, Airha M.

Title: Papernerds: A Research Digital Library and Undergraduate Thesis Portal with Text Mining Algorithms

MANUSCRIPT (20%)		
CRITERIA	WEIGHT (%)	EVALUATED SCORE
1. Format (documents, chapter division, style including neatness and organization of details)	5	
2. Research problems and objectives (discuss the problems encountered by the client and answered with appropriate and adequate solutions)	5	
3. Related Literature and Studies (includes 10 for literature and 10 studies and summarize the discussion of synthesis)	5	
4. Research Methodology (appropriateness of methods of study statistical treatment, analysis and interpretation)	5	
SUB - TOTAL	20	
1. Presentation (content and creativity of visual aid and / or graphics and mastery of study evidenced by logical presentation of the conclusion)	10	
2. Defense (Ability to answer reasoning capability and ability to justify interpretation and conclusion)	10	
SUB - TOTAL	20	
1. Innovation	20	
2. Application Relevance and Impact	20	
3. Research Thrust Impact	20	
SUB - TOTAL	60	
OVER ALL TOTAL	100	

VERDICTS

[ ] **APPROVED**. Minor revisions are necessary but they do not have to be presented in front of and checked by all panelists. 86 – 100

[ ] **APPROVED WITH REVISIONS**. Major revisions shall be incorporated in the final copy of the revised Project Proposal summary. These must be checked by the panelists. 70 – 85

[ ] **DISAPPROVED**. The Proponents/Researchers failed to propose a researchable or scholarly Thesis / Capstone Project. Below 70

SUBJECT SPECIALIST
Role on the Defense

JENARD R. ASOR
Signature over Printed Name of Evaluator

Form 4
CCS-SR

Republic of the Philippines
Laguna State Polytechnic University
Province of Laguna

COLLEGE OF COMPUTER STUDIES

Final Oral Defense Rating Sheet

Date: February 5, 2021 Group No. CS3A-05 Time: 10:00-10:30am

Name of Proponents: Cemiller, Miguel N.
Carayap, Jian Carlo C.
Villanueva, Airha M.

Title: Papernerds: A Research Digital Library and Undergraduate Thesis Portal with Text Mining Algorithms

MANUSCRIPT (20%)		
CRITERIA	WEIGHT (%)	EVALUATED SCORE
1. Format (documents, chapter division, style including neatness and organization of details)	5	4
2. Research problems and objectives (discuss the problems encountered by the client and answered with appropriate and adequate solutions)	5	5
3. Related Literature and Studies (includes 10 for literature and 10 studies and summarize the discussion of synthesis)	5	5
4. Research Methodology (appropriateness of methods of study statistical treatment, analysis and interpretation)	5	4
SUB - TOTAL	20	16
ORAL DEFENSE (20%)		
1. Presentation (content and creativity of visual aid and / or graphics and mastery of study evidenced by logical presentation of the conclusion)	10	9
2. Defense (Ability to answer reasoning capability and ability to justify interpretation and conclusion)	10	9
SUB - TOTAL	20	18
CAPSTONE/THESIS PROJECT (60%)		
1. Innovation	20	17
2. Application Relevance and Impact	20	17
3. Research Thrust Impact	20	18
SUB - TOTAL	60	52
OVER ALL TOTAL	100	86

VERDICTS

[ ] **APPROVED**. Minor revisions are necessary but they do not have to be presented in front of and checked by all panelists. 86 – 100

[ ] **APPROVED WITH REVISIONS**. Major revisions shall be incorporated in the final copy of the revised Project Proposal summary. These must be checked by the panelists. 70 – 85

[ ] **DISAPPROVED**. The Proponents/Researchers failed to propose a researchable or scholarly Thesis / Capstone Project. Below 70

STATISTICIAN
Role on the Defense

VICTOR A. ESTACILLA JR.
Signature over Printed Name of Evaluator

Form 4
CCS-SR

Republic of the Philippines
Laguna State Polytechnic University
Province of Laguna

COLLEGE OF COMPUTER STUDIES

Final Oral Defense Rating Sheet

Date: February 5, 2021 Group No. CS3A-05 Time: 10:00-10:30am

Name of Proponents: Cemiller, Miguel N.
Carayap, Jian Carlo C.
Villanueva, Airha M.

Title: Papernerds: A Research Digital Library and Undergraduate Thesis Portal with Text Mining Algorithms

MANUSCRIPT (20%)		
CRITERIA	WEIGHT (%)	EVALUATED SCORE
1. Format (documents, chapter division, style including neatness and organization of details)	5	
2. Research problems and objectives (discuss the problems encountered by the client and answered with appropriate and adequate solutions)	5	
3. Related Literature and Studies (includes 10 for literature and 10 studies and summarize the discussion of synthesis)	5	
4. Research Methodology (appropriateness of methods of study statistical treatment, analysis and interpretation)	5	
SUB - TOTAL	20	
1. Presentation (content and creativity of visual aid and / or graphics and mastery of study evidenced by logical presentation of the conclusion)	10	
2. Defense (Ability to answer reasoning capability and ability to justify interpretation and conclusion)	10	
SUB - TOTAL	20	
1. Innovation	20	
2. Application Relevance and Impact	20	
3. Research Thrust Impact	20	
SUB - TOTAL	60	
OVER ALL TOTAL	100	

VERDICTS

[ ] **APPROVED**. Minor revisions are necessary but they do not have to be presented in front of and checked by all panelists. 86 – 100

[ ] **APPROVED WITH REVISIONS**. Major revisions shall be incorporated in the final copy of the revised Project Proposal summary. These must be checked by the panelists. 70 – 85

[ ] **DISAPPROVED**. The Proponents/Researchers failed to propose a researchable or scholarly Thesis / Capstone Project. Below 70

LANGUAGE CRITIC
Role on the Defense

MA. LEZIANE D. DIMALANGAN
Signature over Printed Name of Evaluator

Form 4
CCS-SR

Summary of Recommendations

SUMMARY OF RECOMMENDATIONS
Research Colloquium

Working Title:

Papernerds: A Research Digital Library and Undergraduate Thesis Portal with Text Mining Algorithms

Group Code :	CS3A-05		
Students :	Cemiller, Miguel N.		
	Carayap, Jian Carlo C.		
	Villanueva, Airha M.		
Section :	BSCS 4A		
Schedule :	Thursday	December 16, 2021	10:00-10:30
	Day	**Date**	**Time**

Location : Online via Google Meet

Panel Member	Comments Recommendations / Suggestions	Status
Mr. Jonardo Asor	• Don't just concentrate on accuracy when testing an experiment. Always check on other factors in the algorithm; FNR/ FPR. • Use dendrogram for testing K-means. • Add division of testing, training, validation test on Chapter 3. • Give at least 3 solutions on the Designing the Application. • Challenge: Determine if unsupervised/supervised/hybrid to be used for the tasks.	Done
Mrs. Alliana Ablan	• Revise the introduction by discussing the importance of the study. • Arrange the significance of the study. • Instead of "Beneficiaries", replace with "Population" • APA format of table and figures. • Remove initials of citations in Chapter 2.	Done

Checked by:

_____
Panel Member

_____
Panel Member

_____
Panel Member

_____
Panel Member

Noted by:

_____
Dr. Mia Villarica
Adviser

_____
Reynalen C. Justo
Associate Dean, CCS

SUMMARY OF RECOMMENDATIONS
Final Oral Defense

Working Title:

Papernerds: A Research Digital Library and Undergraduate Thesis Portal with Text Mining Algorithms

Group Code :	CS3A-05		
Students :	Cemiller, Miguel N.		
	Carayap, Jian Carlo C.		
	Villanueva, Airha M.		
Section :	BSCS 4A		
Schedule :	Saturday	February 05, 2022	10:00-10:30
	Day	**Date**	**Time**

Location : Online via Google Meet

Panel Member	Comments Recommendations / Suggestions	Status
Mr. Jonardo Asor	• Conduct experiments on the algorithms used, particularly with LDA and K-means algorithms for topic modeling and clustering. • Remove findings on supervised and unsupervised algorithms in synthesis of Chapter 2.	Done
Mrs. Alliana Ablan	• Conclusions must be based on the objectives. • Recommendations must be based on your conclusion. • Explain figures. • Explain advantages and disadvantages in a paragraph. • Indicate year of research papers the researchers collected. • Check figure and table format. • Make texts readable in conceptual framework. • Summarize specific objective no. 2 and remove the bullet points. • Present operational and technical terms in two columns. • Improve the abstract.	Done

Checked by:

_____
Panel Member

_____
Panel Member

_____
Panel Member

_____
Panel Member

Noted by:

_____
Dr. Mia Villarica
Adviser

_____
Reynalen C. Justo
Associate Dean, CCS

APPENDIX C

Curriculum Vitae

MIGUEL CEMILLER

Address Santa Cruz, Laguna, 4009
Email miguelcemiller@gmail.com
Phone 0920-236-9782
LinkedIn linkedin.com/in/miguelcemiller

Self-motivated student seeking for an internship position in an organization where my full potential can be discovered while working for the company dedicatedly.

Education

August 2018 – present

Bachelor of Science in Computer Science

Laguna State Polytechnic University – Santa Cruz (Main Campus)
Thesis: "Papernerds: A Research Digital Library and Undergraduate Thesis Portal with Text Mining Algorithms"
Coursework: Algorithms, Data Structures, Computer Architecture, Machine Learning, Programming, Operating System
Expected Graduation: August 2022

Affiliations and Achievements

- Civil Service Exam Passer – Professional (2018)
- Computer Science Organization – President (2018-2019)

Personal Projects

- Sent AJAX request from the back-end and returned JSON data in the front-end
- Trained a machine learning model to classify texts with Natural Language Processing
- Trained a neural network for image recognition to predict a dog's breed
- Scraped the web to extract data from a website

Technical Skills

Languages: Python, HTML, CSS, JavaScript, SQL, C, C#, Java
Frameworks: Django
Tools: VS Code, Git, NetBeans, Google Colab, Jupyter Notebook
Database Management: MySQL, SQLite

Soft Skills

Communication
Teamwork
Confidence
Problem Solving

JIAN CARAYAP

Address Pagsanjan, Laguna, 4008

Email jiancarlocarayap2212@gmail.com

Phone 0939-968-6368

I am a hardworking individual looking for internship work experience where I can apply and enhance both my technical and personal skills. I am a good communicator who can work well with a team but is also capable of working independently.

Education

August 2018 – present

Bachelor of Science in Computer Science

Laguna State Polytechnic University – Santa Cruz (Main Campus)

Thesis: "Papernerds: A Research Digital Library and Undergraduate Thesis Portal with Text Mining Algorithms"

Coursework: Algorithms, Data Structures, Computer Architecture, Machine Learning, Programming, Operating System

Expected Graduation: August 2022

Affiliations and Achievements

- STI College Sta. Cruz Click Club - President (2017-2018)
- STI College Sta. Cruz Journalism Club –Head Photographer (2016-2017)

Personal Projects

- Developed a personal website with basic HTML and CSS
- Made a simple quiz app with C#
- Edited videos for relatives and friends for their special occasions

Technical Skills

Languages: HTML, CSS, JavaScript, SQL, C#, Java

Tools: VS Code, Visual Studio, NetBeans, Jupyter Notebook

Database Management: MySQL

Other Tools: Adobe Premiere, MS Word, MS Excel, MS PowerPoint

Soft Skills

Communication ●●●●●

Teamwork ●●●●○

Confidence ●●●●○

Problem Solving ●●●●○

AIRHA VILLANUEVA

Address Santa Cruz, Laguna, 4009 **Email** airhamay542@gmail.com

Phone 0906-912-7075

I am a hardworking person with years of experience in dealing with paper-related tasks, seeking to apply my skills to grow the business.

Education

August 2018 – present

Bachelor of Science in Computer Science

Laguna State Polytechnic University – Santa Cruz (Main Campus)

Thesis: "Papernerds: A Research Digital Library and Undergraduate Thesis Portal with Text Mining Algorithms"

Coursework: Algorithms, Data Structures, Computer Architecture, Machine Learning, Programming, Operating System

Expected Graduation: August 2022

Affiliations and Achievements

- Certified Computer System Servicing NCII Holder (2018)
- Computer Science Organization – President (2019-2020)

Personal Projects

- Built a simple quiz game with C# using Visual Studio
- Made an animation of a stick man walking
- Developed a simple personal web page with HTML and CSS

Technical Skills

Languages: HTML, CSS, SQL, C#, Java

Tools: VS Code, Visual Studio, NetBeans, Jupyter Notebook

Database Management: MySQL

Other Tools: MS Word, MS Excel, MS PowerPoint

Soft Skills

Communication ●●●●●

Teamwork ●●●●○

Confidence ●●●●●

Problem Solving ●●●●○