MIT 3201 – Individual Project

Project Proposal Form - 2020

Candidate Details			
Name	BGSN Bambaranda		
E-Mail Address	Info.bambaranda@gmail.com	**Registration No.**	2016/MIT/006
Phone Number(s)	0717307831	**Index No.**	16550061
Project Details			
Proposed Project Title	IntelyDoC – Keyword based document classifier for unstructured documents		
Name of the Supervisor	Dr. DAS Athukorale		

Candidate's Project Attempt		**(✓)**
1st Attempt	**For students of 2016 intake**	
Re-Attempt	Submitting a new project proposal	
	Continuing the same project which was proposed in a previous year	✓
If Re-Attempt, identify stream	**MIT(General) Project : MIT3101**	✓
	MIT(Multimedia) Project : MIT3121	
	MIT(e-Learning) : MIT3111 Repeat Students, please use the e-Learning Template	

[1]. Problem

Over the last decade, the number of digital documents available for various purposes has grown enormously with the increasing availability of high capacity storage hardware and powerful computing platforms. The vivid increase of documents demands effectual organizing and retrieval methods mainly for large documents. Document classification is one of the key techniques in text mining to categorize the documents in a supervised manner.

Document classification is an age-old problem in information retrieval, and it plays an important role in a variety of applications for effectively manage large volumes of unstructured information in different industrial sectors of the country. The problem is to assign a document to one or more categories.

Document classification has two different methods: manual and automatic classification. In manual document classification, users interpret the meaning of text, identify the relationships between concepts and classify the documents. While this gives users more control over classification, manual classification is inconsistent, expensive, time consuming and lacks security. Automatic document classification can be defined as content-based assignment of one or more predefined categories to documents. This makes it easier to find the relevant information at the right time and for filtering and routing documents directly to users. Furthermore, Automatic document classification applies machine learning or other technologies to automatically classify documents; this results in faster, scalable, and more objective classification.

The problem domain selected for this project is inefficiency of manual document classification. Categorizing a large number of documents manually is time consuming, inefficient and error prone.

The proposed solution, will provide a smart solution to automatically classify large volume of scanned or computer-generated documents into classes

[2]. Project Objective(s)

- To develop a classifier which is able to classify the given data set accurately
- To develop a software which is capable of extracting keywords from the documents to classify a large volume of documents into predefined classes using machine learning techniques.
- To facilitate users with both automated document classification and keyword-based document classification.
- To allow different document types.*ei - .pdf,.txt,images*
- To identify duplicate documents before doing the classification

[3]. Scope of the Project

Based on the objectives of the project, the scope is defined as follows

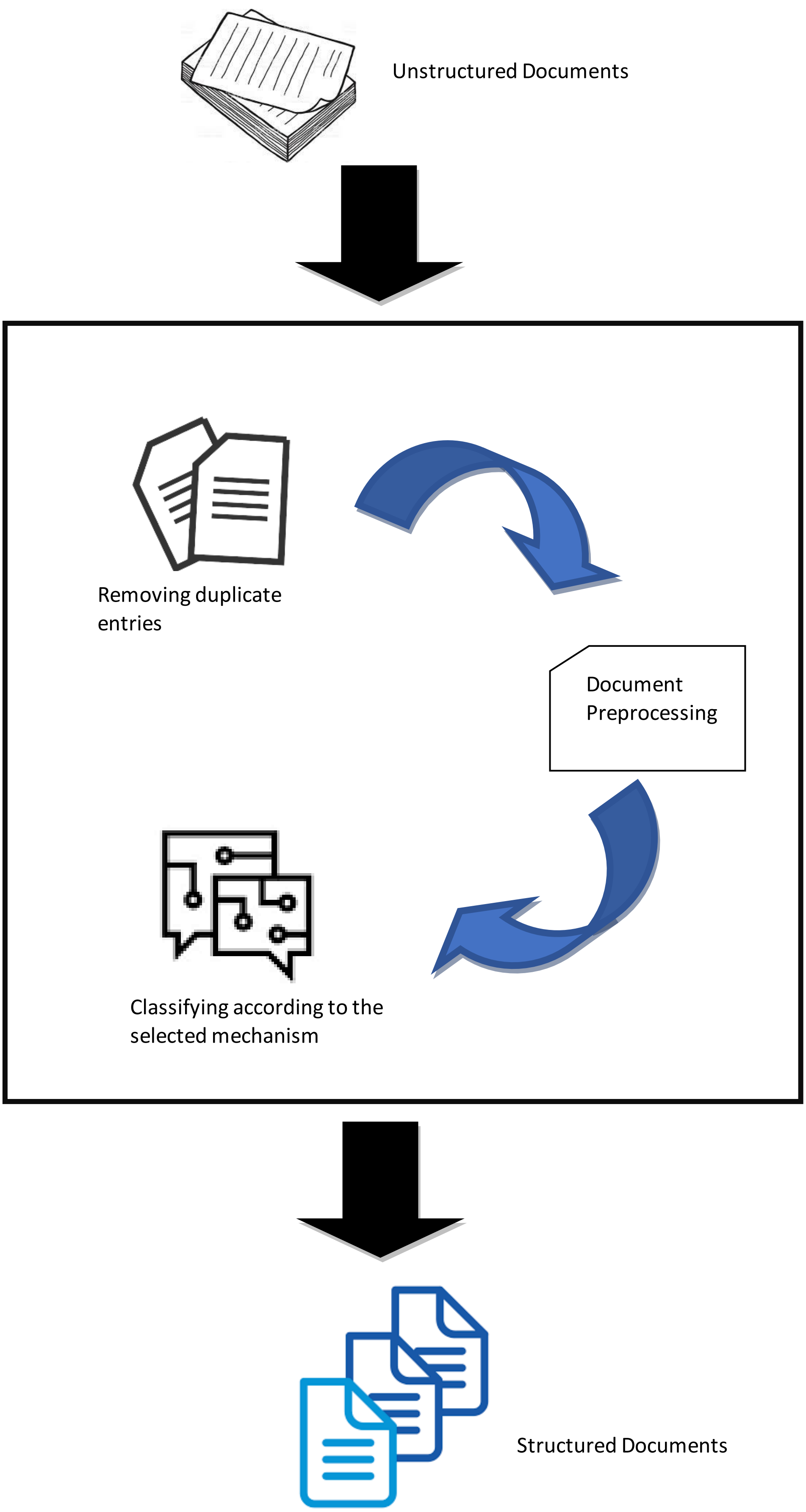
In Scope

- Automated document classification and keyword-based document classification are available as two options for classification
- Automated document classification will automatically classify the documents into classes or can provide a set of documents as examples for classes.
- Keyword-based document classification will ask for a set of keywords or metadata in order to classify.
- Several document types will be considered. *.ei - .pdf,.txt,images*
- Duplicates documents will be identified once the documents are uploaded.

Out Scope

- pictures and graphs are not considered when classifying.

Prototype feature Diagram


Unstructured Documents
Removing duplicate entries
Document Preprocessing
Classifying according to the selected mechanism
Structured Documents

[4]. Related Work/ Background Study

Text classification is the process of assigning tags or categories to text according to its content. It's one of the fundamental tasks in Natural Language Processing (NLP) with broad applications such as sentiment analysis, topic labeling, spam detection, and intent detection. Unstructured data in the form of text is everywhere: emails, chats, web pages, social media, support tickets, survey responses, and more.

Text can be an extremely rich source of information but extracting insights from it can be hard and time-consuming due to its unstructured nature. Businesses are turning to text classification for structuring text in a fast and cost-efficient way to enhance decision-making and automate processes.Professor Airi Salminen, Professor Pasi Tyrväinen,Anne Honkaranta, Ph.D Lecturer, Mikko Jäkälä, Lecturer, Jussi Koskinen, Assistant Professor Virpi Lyytikäinen, Ph.D., Researcher in Department of Computer Science and Information Systems, University of Jyväskylä has done a Document Management Research. The research program was developing methods and techniques for the management of digital documents in the networked multimedia environments of enterprises. Document management was regarded as a means for information management in organizations. The research concerned documents as part of information technology solutions on the one hand, and as a means of communication in the activities of social communities, on the other hand. Especial focus was on structured documents like SGML and XML documents, EDI documents, and software documents. The major approach was constructive; the researchers acted for developing useful innovations, for finding out problems in the current document management, and for developing better solutions .

Mohd Haizam Mohd Saudi of Southern Cross University has done a research about the effects of the performance management system and the organizational culture on the employees' attitude in Malaysian government statutory bodies. This research focuses on the effects of the attitude on the employees in one of the biggest statutory bodies in Malaysia, Majlis Amanah Rakyat known as MARA .

Hesham S. Ahmad*, Issa M. Bazlamit, Maha D. Ayoush of Al-Zaytoonah University of Jordan, P.O. Box: 130 Amman, 11733 Jordan, Investigation of Document Management Systems in Small Size Construction Companies in Jordan .

The processing of text classification involves two main problems are the extraction of feature terms that become effective keywords in the training phase and then the actual classification of the document using these feature terms in the test phase. The most popular document classification systems are advanced AI-based machine learning algorithms that automatically learn how to classify documents based on samples and user feedback. Text classification can be used for document filtering and routing to topic-specific processing mechanisms such as information extraction and machine translation. Various methods are used for document classification such as Naive Bayes, Support Vector Machine, K-Nearest Neighbor, Fuzzy C-means, Neural Networks, Decision trees and Rule based learning algorithms

[5]. Methodology and Implementation Considerations

In the proposed work, the keywords are extracted from documents using TF-IDF. There are limited number of words are selected from each document. Based on the extracted keywords, documents are classified using machine learning techniques.

The problem can have two approaches: unsupervised and supervised learning. Supervised learning makes use of data that has been labeled with the correct classes or topics, while the unsupervised algorithms use input data that has not been hand-annotated with the correct class or topic. Generally, the unsupervised learning is more complex and yields less accurate results than the supervised learning. Nevertheless, the volume of the data that has not been labeled is much greater than the ones that have the correct classes assigned to it, and in some situations an unsupervised algorithm is the only option.

Incremental waterfall methodology will be used to carry out this project. Advantages of using this methodology is some working functionality can be developed quickly and early in the life cycle. In this whole requirement is divided into various builds. During each iteration, the development module goes through the requirements, design, implementation, and testing phases. Each subsequent release of the module adds function to the previous release. The process continues till the complete system is ready as per the requirement.

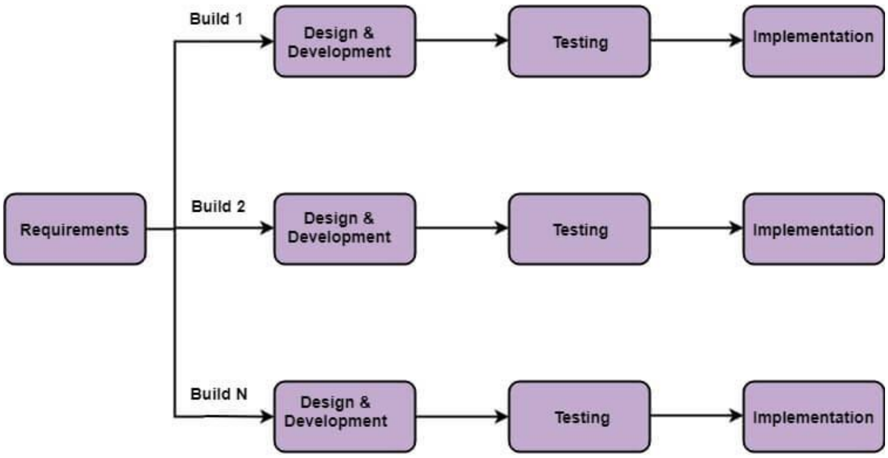


Fig: Incremental Model

Requirement Analysis phase

In the first phase of the incremental model, the product analysis expertise identifies the requirements. And the system functional requirements are understood by the requirement analysis team. To develop the software under the incremental model, this phase performs a crucial role.

Design phase

In which a software solution to meet the requirements is designed. This may be a new design, or an extension of an earlier design.

Implementation phase

Implementation phase enables the coding phase of the development system. It involves the final coding that design in the designing and development phase and tests the functionality in the testing phase. After completion of this phase, the number of the product working is enhanced and upgraded up to the final system product

Testing phases

In the incremental model, the testing phase checks the performance of each existing function as well as additional functionality. In the testing phase, the various methods are used to test the behavior of each task.

[6]. Evaluation

Measure the following points in evaluation planning,

- Achievements against planned objectives.
- Cost-effectiveness.
- Outcomes and impacts against desired benefits. Learning and Practicing.
- This product must be certified by a doctor, that the assumptions made, and the resulting outcomes are correct.

[7]. Is this Project Work Related?

No, it is not.

[8]. List of Deliverables

- SaaS to intelligently classify documents uploaded.
- User manual for the system.

[9]. List of References

[1]. Pay, Tayfun. (2016). Totally automated keyword extraction. 3859-3863. 10.1109/BigData.2016.7841059.

[2]. Sebastiani F., "Machine Learning in Automated Text Categorization", ACM Computing Surveys, vol. 34 (1),2002, pp. 1-47.

[3]. Han X., Zu G., Ohyama W., Wakabayashi T., Kimura F., Accuracy Improvement of Automatic Text Classification Based on Feature Transformation and Multi-classifier Combination, LNCS, Volume 3309, Jan 2004, pp. 463-468

[4]. Kehagias A., Petridis V., Kaburlasos V., Fragkou P., "A Comparison of Word- and Sense-Based Text Categorization Using Several Classification Algorithms", JIIS, Volume 21, Issue 3, 2003, pp. 227-247.

[5]. Salminen, A. (2003). Document analysis methods. In C.L. Bernie (Ed.), *Encyclopedia of Library and Information Science, Second Edition, Revised and Expanded* (pp. 916-927). New York: Marcel Dekker.

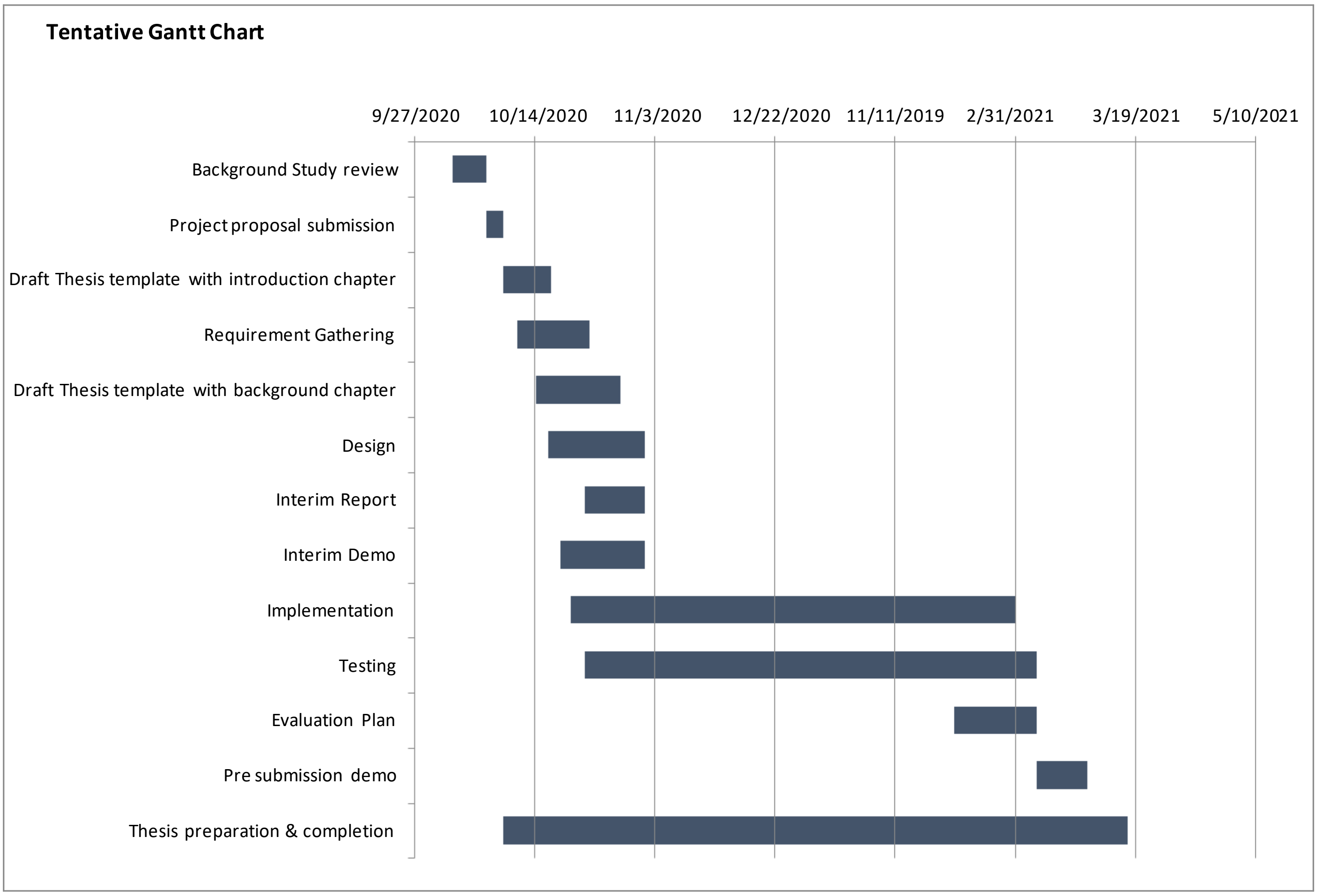
[6]. V. (2003), Analysing requirements for content management. In O. Camp, J. Filipe, S. Hammoudi, & M. Piattini Eds.), *Proceedings of the 5th International Conference on Enterprise Information Systems*,Vol. 3 (pp. 104-111). Portugal: Escola Superior de Technologia do Instituto Politécnico de Setúbal.

[7]. Document management overview. A guide to the benefits, technology and implementation essentials of digital document management solutions. 10th ed. USA: Compulink management center, Inc.; 2007.

[8]. Backblom M, Ruohtula A, Bjork B. Use of document management systems – A case study of the Finnish construction industry. ITcon 2003;8:367–380.

[9]. Björk BC. Electronic Document Management in Construction – Research Issues and Results. ITcon 2003;8:105–117.

[10]. Work Breakdown/ Project Plan



[11]. Additional Information

16/10/2020

....

Signature of the Candidate

Date

[12].Details of Project Supervisor(s):

Supervisor *			
Supervisor's Comments			
Name	Dr. Ajantha Atukorale, PhD (UQ), BSc Hons (CMB), MCS(SL), MIEEE		
Designation	Deputy Director/UCSC and Head / Department of Computation and Intelligent Systems	**Signature**	
		Date	

***- Supervisor must be a UCSC academic staff member**

Advisor +	
Advisor's Comments	

Name			
Designation		**Signature**	
		Date	

+ - Advisor can be selected as candidate wish