

# University of San Jose – Recoletos

Cebu City

## GRADUATE SCHOOL OF GSCICCT

Implementation of the Oral Defense Recommendations Form

(to be complied with prior to the refined copy submission)

Proponent : **CARLO D. PETALVER** Date of Oral Defense: **March 27, 2021**  
Degree Program : **MASTER IN INFORMATION TECHNOLOGY (MIT)**  
Title : **"BookShelf: A Document Categorization for Library using Text Mining"**

<b><u>Recommendations</u></b>	<b><u>Page of Implementation</u></b>	<b><u>Remarks</u></b>
1. Cite references to back up claims in the 2 <sup>nd</sup> and 3 <sup>rd</sup> paragraph of the introduction.	1 (Ms. Jovy Cuizon)	DDC and LCC discussions were removed since they have no relevant relationship with the project in terms of text classification.
2. Review of related literature. Cite references. Use IEEE citation. Organize thought process in RRL. Include in RRL other approaches used in text classification and justify why SVM was selected.	5 (Ms. Jovy Cuizon)	Included some IEEE citations in this project.
3. Theoretical Background: Discuss underlying concepts/approaches used in the paper a. OCR b. Machine Learning c. SVM d. Text Mining	2 (Ms. Jovy Cuizon)	Concepts already discussed.  OCR Machine Learning SVM Text Mining AI

4. Architectural diagram to show interaction of components, libraries/APIs.	40 (Ms. Jovy Cuizon)	An architectural diagram showing the interaction of components is depicted.
5. Conceptual diagram to show process of text mining.	37 (Ms. Jovy Cuizon)	The detailed process is discussed starting from the input, processes, and output of the system using text mining.
6. Use Case Diagram/Model – show features relevant to the study (from data capture (OCR) up to classification. Exclude CRUD functionalities.	8 (Ms. Jovy Cuizon)	Excluded the CRUD functionalities already.
7. Use case Narrative - provide user interaction on each process. One for each use case.  Format: Side heading, intro, figure, discussion	9 - 19 (Ms. Jovy Cuizon)	Heading, intro, figure, and discussion are already depicted in each use case.
8. Identify classification categories. Present run through example with snapshot of data at each process. a. Data capture b. Data processing c. Training d. Testing/Validation	56 - 70 (Ms. Jovy Cuizon)	Run through example of dataset and input data were captured from data cleansing, preprocessing, feature engineering, training, testing and up to the validation.

9. How are the document matrix stored? Training set description - no of documents, size of corpus -Test set description	67 - 68  (Ms. Jovy Cuizon)	The shape or size of the corpus is (50, 30179)  Split: The training set is 67% while the testing set is 33% (Standard split from sklearn) which is a good split. I have tried it against 60-30 split with this size of the corpus.
10. Discussion of SVM (concept) should be placed in theoretical background in chapter 3, discuss SVM as used in the project.	3 and 65-67  (Ms. Jovy Cuizon)	Already discussed in the following pages. It includes also the mathematics behind SVM and how it is used in the project.
11. Confusion matrix to present accuracy per category.	68  (Ms. Jovy Cuizon)	The confusion Matrix result is already depicted on this page showing the accuracy of each category during prediction.
12. Discussion of the integration to the web app, mobile app, OCR.	40  (Ms. Jovy Cuizon)	Integration of the web app, mobile app, and OCR is discussed together with its architecture.
13. Summary of findings - discuss insights or discoveries after conduction the experiment. No need for UI here.	77  (Ms. Jovy Cuizon)	UIs discussions were removed.
14. Conclusion – modify to show how the objectives were met. No need for another introduction.	78  (Ms. Jovy Cuizon)	It is already mentioned on this page.

15. Include Bibliography	80 – 81 <b>(Ms. Jovy Cuizon)</b>	Already included some recent references.
16. No need for definition of terms part.	<b>(Ms. Jovy Cuizon)</b>	Already excluded or removed.
17. SUMMARY OF FINDINGS is 4 pages?  The SUMMARY OF FINDINGS is another detailed description on the process. Should this be included in the BODY?  It is suggested that the focus on the illustration is the input/output presentation (similar to raw data/dashboard) and a brief description (few sentence) of the process applied.	77  <b>Sir Felipe Petralba</b>	Summary of Findings has been reduced to 1 page only. A brief discussion of the process applied is mentioned.
18. Discussion of RESTful web service	6, 40, 54, 78 <b>Sir Felipe Petralba</b>	The discussion of RESTful web service is mentioned on the following pages.
19. Cannot find in the Summary of Findings that you have designed and developed a web application.	77 <b>Sir Felipe Petralba</b>	Already mentioned on this page.

---

From the OBJECTIVES

"Save time and accuracy in classifying/ associating library materials to the course syllabus."

1

20. - Do you have gathered metrics on these. old system vs. your system. how long?

(Ms. Jovy Cuizon)

---

21. Why index pages and TOC only? Is this the standard way of classifying books? If it is not, then have we not considered other parts of a book? The assumption of the work is that the Index pages and the TOC are correctly done?

Sir Felipe Petralba

---

22. Mentioned in INPUT-OUTPUT FORMATS. Related to above.

"A sample text extracted from the book's table of contents before and after sanitation"

The assumption of the paper is that the heading, subheading1, subheading2 ... have the same weight?

Sir Felipe Petralba

---

I cannot compare the old system vs. the new system in terms of metrics. The old system requires human conceptual ability by the professionals (e.g. librarians). While the new system can automatically predict by the use of machine learning techniques and results to be more precise in terms of categorization. It can improve the efficiency of human classifiers.

---

The study focused on the table of contents and index pages only since it carries the significant components of the document (topics, list of sections, and chapters) relevant for text categorization.

---

Yes. Text from table of contents and index pages have the same weight. They are stored in one table only as samples. A train test split of 67% and 33% has been performed.

---

### 23. SVM

"That the distance between the support vector and the hyperplane should be as far as possible or has the maximum distance to the support vectors of any class. The math behind SVM is very

simple, we take  $D^+$  which is the shortest distance to the closest positive point and  $D^-$  to the shortest

distance closest to negative point. The sum of  $D^+$  and  $D^-$  is called the distance margin. "

Can you give a simple example(numbers) and show the computation?

---

65 - 66

Sir Felipe Petralba

---

From the math intuition of SVM, giving this formula  $y = w^T x + b$ . Consider giving the value of  $(-4, 0)$ , and the slope represented by this formula  $w^T x + b = 0$ , giving us this matrix  $\begin{bmatrix} -1 \\ 0 \end{bmatrix}$  and we will compute it using the matrix multiplication will give us 4. Any value given is going to be always positive. Now, if there is a value that lies on the other side of the slope, for example giving the value of  $(4, 4)$ , again computing it with the given matrix  $\begin{bmatrix} -1 \\ 0 \end{bmatrix}$  as a slope using matrix multiplication, it will give us -4. Any value given is going to be negative. With this, we can consider all the positive values as one group or class, and all the negative values will be another group or class taking the  $D^+$  and  $D^-$  intuition.

---

24. Table 3. The books table with its contents  
Unless it is a USER's screen (for UI-software interaction) I suggest that you don't show a screenshot of an IDE tool. In this case this is a DB IDE showing the result of a select query. Rather a simple table will do.

---

71 - 72

Sir Felipe Petralba

---

Screenshots from the IDE tool (e.g. phpmyadmin table name: classifications, dataset, books) are replaced by a simple table.

---

25. Fig. 38 Class diagram for BookShelf

Can you give sample records of all the tables and which are interconnected?  
1 to 3 records will do

---

38, 73

Sir Felipe Petralba

---

A sample record from the books table is identified by the foreign key's *classification\_id* and *user\_id*.

---

26. Clearly define what is dataset. Is it books (table of contents), syllabus? How does it look - text corpus?

---

55

Sir Gregg Gabison

---

The dataset that was gathered came from the book's table of contents and index pages relevant to the specific course reference or syllabus. They are extracted from the image using OCR technology. Then, it is used in training the model.

---

27. How is Model Training/evaluation/generation done? How was SVM applied? How did you go through the process? Confusion Matrix? After testing, how did you evaluate the model?

---

55 - 70

Sir Gregg Gabison

---

After the creation of the dataset, it was split into training and testing. 67% went to training and 33% went to testing. Then, an SVM algorithm has been applied for the creation of the model using the OneVsRestClassifier with a linear kernel and a balanced class weight. The probability estimate of SVM has been used as well so that if the prediction threshold is 60% or greater, a class or category will be assigned to it, otherwise, it will be in 'no category'. The model accuracy has been evaluated using the k-fold cross-validation technique with our dataset and it surpasses and outperforms the other classifier algorithms.

---

28. Instead of Research Design, change it to Application Overview to include architectural and Conceptual diagram

37, 40

(Ms. Jovy Cuizon)

The conceptual diagram and Architectural diagram are already included as an Application Overview with an explanation.

---

29. Data preparation/Cleaning  
– Sample data can be presented for this.

55-64

Sir Gregg Gabison

---

Sample input text that is captured from a mobile app using OCR technology is presented on this page with pre-processing of text data (e.g. removal of numbers, punctuation, special characters, converting to lower case, tokenization, stopwords, and applying feature engineering techniques-TFIDF) are explained.

---

30. Deployment

Sir Gregg Gabison

---

Successfully deployed and tested in the native app (Android) and web app.

---

31. Summary of Findings associated with testing results, observation

Sir Gregg Gabison

---

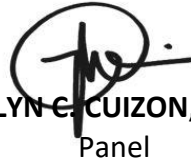
Already mentioned as per the advice of sir Gregg (that the SVM outperforms the other classifiers in terms of accuracy with the given dataset).



Panel Members:



**MARISA M. BUCTUANON, MCS**  
Panel

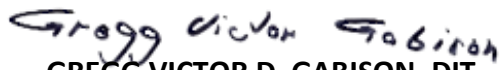


**JOVELYN C. CUIZON, DMHRM**  
Panel



**RODERICK A. BANDALAN, MSIT**  
Adviser

**FELIPE JR. PETRALBA, MS-Math, MSCS**  
Panel



**GREGG VICTOR D. GABISON, DIT**  
Chairman