Abstract

Background: As part of the transition of every higher education institution into an intelligent campus here in the Philippines, the Commission of Higher Education has launched a program for the development of smart campuses for state universities and colleges to improve operational efficiency in the country. With regards to the commitment of Camarines Sur Polytechnic Colleges in improving the accreditation operation and to resolve the evident problems in the accreditation process, the researchers propose this study as part of an Integrated Quality Assurance System that aims to develop an intelligent model that will be used in categorizing and automating tagging of archived documents used during accreditation.

Methods: As a guide in modeling the study, the researchers use an agile method as it promotes flexibility, speed, and, most importantly, continuous improvement in developing, testing, documenting, and even after delivery of the software. This method helped the researchers in designing the prototype with the implementation of the said model to aid the process in file searching and label tagging. Moreover, a computational analysis is also included to further understand the result from the devised model.

Results: As a result, from the processed sample corpus, the document labels are faculty, activities, library, research, and materials. The labels generated are based on the total relative frequencies which are 0.009884, 0.008825, 0.007413, 0.007413, 0.006354, respectively, that have been computed between the ratio on how many times the term was used in the document and the total word count of the whole document.

Conclusions: The devised model and prototype support the organization in file storing and categorization of accreditation documents. Through this, it is easier to retrieve and classify the data,
which is the main problem for the task group. Further, other patterns in clustering, modeling, and text classification can be integrated in the prototype.

Keywords
Latent Dirichlet allocation, document labels, natural language processing, accreditation, quality assurance, Intelligent model, CSPC

This article is included in the Artificial Intelligence and Machine Learning gateway.

This article is included in the IIARP Publications gateway.

Corresponding authors: Freddie Prianes (fprianes@cspc.edu.ph), Thelma Palaoag (tdpalaoag@uc-bcf.edu.ph)

Author roles: Prianes F: Conceptualization, Writing – Original Draft Preparation; Palaoag T: Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: The author(s) declared that no grants were involved in supporting this work.

Copyright: © 2023 Prianes F and Palaoag T. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Prianes F and Palaoag T. Modeling document labels using Latent Dirichlet allocation for archived documents in Integrated Quality Assurance System (IQAS) [version 1; peer review: 1 approved with reservations] F1000Research 2023, 12:105 https://doi.org/10.12688/f1000research.130245.1

First published: 27 Jan 2023, 12:105 https://doi.org/10.12688/f1000research.130245.1
Introduction

The creation of a smart campus is a step toward the creation of a smart city. Teaching and learning will be more difficult in the future as a result of the rapid advancements in information and communication technology (Kwok, 2015). With this rapid advancement, there is already a shift from the “smart” era to the “intelligent” era. A “smart phone,” “smart building” or “smart home” is one that is capable of adapting to changing conditions. The term “intelligent,” on the other hand, refers to more than just being smart; rather, it refers to having the ability to think, reason, and understand, as well as being able to adapt to changing conditions. If you apply this to a device example, “smart devices” can perform tricks, but “intelligent devices” can learn new tricks in response to their changing surroundings (Ng et al., 2010).

As part of the transition of every Higher Educational Institution (HEI) to being an intelligent campus, the Commission of Higher Education (CHED) has launched a program under CHED Memorandum Order No. 9 s. 2020 for the development of smart campuses for State Universities and Colleges (SUCs). In fact, CHED releases a budget to assist SUCs in the development of smart campuses in which HEIs use next-generation digital technologies woven seamlessly within a well-architected infrastructure in developing tools to enhance teaching and learning, research, and extension as well as to improve operational efficiency. On the other hand, as a requirement by CHED and maintaining the quality of education in HEIs, CHED gives the accountability and responsibility to the accrediting body, such as the Accrediting Agency of Chartered Colleges and Universities of the Philippines (AACCUP), Philippines Association of Colleges and Universities Commission on Accreditation (PACU-COA), Philippine Accrediting Association of Schools, Colleges, and Universities (PAASCU), and many others to assess and provide certifications of quality education in the accredited program/institution as stated in the CHED Memorandum Order No. 1 s. 200.

Achieving a smart/intelligent campus requires consideration of different areas by the institution. Based on the study of Ng et al., there are six main areas of intelligence, namely (1) iLearning, (2) iManagement, (3) iGovernance, (4) iSocial, (5) iHealth, and (6) iGreen. The accreditation process alone will fall under iManagement, however that entire aspect and purpose of accreditation falls in all the areas.

Camarines Sur Polytechnic Colleges (CSPC) as a state college will be one of the settings for the initial implementation of the system. As part of the goal of CSPC to be the center for development and center of excellence, the institution opted to go along with the launch of the CHED program to become one of the smart campuses in the region. In connection to this, the institution also undergoes continuous accreditation through AACCUP, as depicted in Figure 1, and ISO quality assurance to achieve the goal and gain the university status as the Polytechnic University of Bicol.

The accreditation process, as shown in Figure 1, passes through various phases or actions: (a) Application: An educational institution submits an application to AACCUP for accreditation. (b) Institutional self-survey: After the application has been approved, the applicant institution is expected to conduct an internal evaluation by its internal accreditors to evaluate whether the program is ready for an external review. (c) Preliminary survey visit: This is when external accreditors evaluate the program for the first time. The program is eligible to receive a Candidate status that is good for two years after passing the assessment. (d) The first formal survey visit reviews the program that has obtained Candidate status, and if it has met a higher standard of excellence, it is given a Level I Accredited status, which is valid for three years. (e) The second survey visit entails evaluating an accredited program, and if it has met the standards for a greater degree of quality than the survey visit that came before it, the program may be eligible to get a Level II Re-accreditation status that is valid for five years. (f) During the third survey visit, the accreditation level is completed by a program after five years of holding Level II Re-accreditation status. The program is reviewed and must perform exceptionally in four categories, namely instruction and extension, which are essential; and two other areas, which must be selected from among research, performance in licensure exams, faculty development, and links. (g) The fourth survey visit is a more difficult level that, if passed, may grant the organization institutional accreditation status.

Accompanied with the tedious accreditation process are many documents that needs to be produced. For most experiences in the current accreditation undertakings in CSPC, the majority of the tasks have been done manually. Though there are tools available for cloud storage and automation like Google Drive, Dropbox, etc., problems such as repetition of work, invalid instruments, inefficient resource utilization, and inefficient monitoring before, during, and after the accreditation are still experienced by the personnel. With this perceived problem, an integrated system dedicated to quality assurance processes is a must.

Upon the CSPC’s goal of becoming a university and becoming a smart/intelligent campus, the researchers propose a centralized system that will cater to the needs of the institution in the process of accreditation, which is part of quality assurance. Through this study, CSPC will benefit from being a smart/intelligent campus by means of utilization of the system in the iManagement area and, at the same time, it addresses the problems encountered during the accreditation processes.
Based on the problems identified and the commitment of the institution to be a smart/intelligent campus, the researchers propose this study as a component in the Integrated Quality Assurance System (IQAS) (RRID:SCR_023146). The study focuses on the documents archive needed for the accreditation process. The system will have a document repository of archived documents and these documents will be analyzed by the system through the use of intelligent modeling. Through the use of this, the documents will be categorized by means of the extracted labels.

In general, the study aims to create a model in support of the categorization and automated tagging of the archived documents used during accreditation.

**Related works**

Unstructured data make it more difficult and time-consuming to find a relevant document due to the exponential growth of electronic documents. Text document classification, which organizes unstructured documents into pre-defined classifications, is crucial to information processing and retrieval (Akhter et al., 2020). The text documents provide a number of difficult problems for data processing in order to retrieve the pertinent data. One of the well-liked methods for information retrieval based on themes from biomedical documents is topic modeling. Finding the correct subjects from the biological documents is a difficult task in topic modeling. Additionally, redundancy in biomedical text documents has a detrimental effect on text mining quality. As a result, the exponential rise of unstructured documents necessitates the
development of topic modeling machine learning approaches (Rashid et al., 2019). In the framework of document categorization, they have conducted a comparative analysis of three models for a feature representation of text documents. The most popular family of bag-of-words models, the recently suggested continuous space models Word2Vec and Doc2Vec, and the model based on the representation of text documents as language networks are all taken into consideration in detail (Martinčić-Ipšić et al., 2019).

In this study, word representation techniques were used to analyze how the similarity between English words is calculated. This work used the Word2Vec paradigm to express words as vectors. The 320,000 English Wikipedia articles included in this study’s model served as the corpus, and the similarity value was calculated using the cosine similarity calculation method (Jatnika et al., 2019). Real-world text categorization problems frequently involve a multitude of closely related categories arranged in a taxonomy or hierarchical structure. When processing huge sets of closely related categories, hierarchical multi-label text categorization has grown more difficult (Ma et al., 2021). A popular technique for clustering functional data is the functional k-means clustering algorithm. The derivative information is not further taken into account by this approach when determining how similar two functional samples are to one another. In actuality, the derivative information is crucial for spotting variances in trend characteristics among functional data. By including their derivative information, we establish a novel distance in this paper that is utilized to compare functional samples (Meng et al., 2018). Due to its capacity to analyze data from numerous sources or views, multi-view clustering has drawn a growing amount of interest in recent years. In the research, they presented a unique multi-view clustering method called Two-level Weighted Collaborative k-means (TW-Co-k-means) to simultaneously address the issues on consistency across different views and weighing the views for the improvement of cluster results. For multi-view clustering, a new objective function has been developed that leverages the unique information in each view while also cooperatively utilizing the complementarity and consistency between various views (Zhang et al., 2018). The various pattern matching algorithms are used to locate every instance of a constrained set of patterns inside an input text or input document in order to examine the content of the documents. This research utilized four string matching techniques that are now in use: the Brute Force approach, the Knuth–Morris–Pratt algorithm (KMP), the Boyer–Moore algorithm, and the Rabin–Karp algorithm (Bhagya Sri et al., 2018). All the literature listed has similarity in text clustering, modeling, and classification, and serves as a proof that the study is feasible, and the proposed intelligent model can be integrated to further assist in the accreditation process of CSPC.

Methods
As a guide in modeling the study, the researchers used the agile method (https://dx.doi.org/10.17504/protocols.io.n2bvj82mxygks5/v2) as it promotes flexibility, speed, and, most importantly, continuous improvement in developing, testing, documenting, and even after delivery of the software. Since the phases of this model are light, the teams are not bound by a rigid systematic-based process on pre-set constraints and restrictions as some other models, like the waterfall model, and can adjust changes whenever they are needed. This flexibility on every stage propagates creativity and freedom within processes. Furthermore, development teams can modify and re-prioritize the backlog, allowing for speedy implementation (Trivedi, 2021).

Following the agile methodology, the researchers adapted the stages, as presented in Figure 2. These are: (1) Plan: the researchers collected previous documents involved in the accreditation process, such as compliance reports under the
areas of student, faculty, facility, library, and administration. Also, understanding of the existing problems in tracking, tagging, and duplication of these documents during the accreditation process. (2) Design: the requirement specifications in this stage were identified in relation to the existing problem of the HEI in tracking, tagging, and duplication of the documents for accreditation and quality assurance. Along with this, the researchers also created the process of the intelligent model, which will be the basis of document labelling. (3) Develop: this stage is intended on the creation of the prototype, which involves the processing of the documents in order to identify the proper label for each document. (4) Deploy: the prototype undergoes a test run during this stage. (5) Review: the researchers conduct a checklist function review to check if each component is running properly. Lastly, (6) launch: wherein the prototype is embedded to the local system of the HEI.

Results and discussion

Intelligent model

The results from this intelligent model are used for visualization in the super word vector and histogram. The super word vector is presented in a cloud map word to visualize the frequency of the words in the corpus, and the histogram is used to present the relationship of the words per sentence in the form of line graphs. The extracted labels and generated word vector and histogram are tagged and linked to the uploaded document, as patterned in the process shown in Figure 3. This model is implemented in the IQAS to assist in categorization and searching in the file repository of accreditation documents.

Prototype

The design prototype presented in this section is focused on the label extraction feature for automatic tagging of the archive documents used in accreditation.

Upload and clean

As shown in Figure 4, this phase allows the user to upload and clean the document through tokenization. Once uploaded, the user may set the configuration in cleaning the document. The options are removing numbers, symbols, and duplicates, adding and uploading additional stopwords, and showing and downloading the pre-processed data. There are other useful features particularly in managing the stopwords, such as showing the list of default stopwords and deleting the added and uploaded stopwords.

![Diagram of the intelligent model process (Figure 3)]
Phase II is intended for setting up the parameters for topic modeling, as presented in Figure 5. Right after uploading and cleaning the document, the user can set the topic modeling parameters that will be used in identifying and extracting the labels. The parameters included are the desired number of topics, frequency of iteration, the number of words per topic to
be generated, optimization interval, and the model’s name. These parameters are primarily the factors in modeling the topics and label identification for automatic tagging.

**Extract label**

This phase, as shown in Figure 6, provides the result of the processed corpus from the processing of the pre-processed document and the parameters that have been set up from the previous phase. This shows the number of documents uploaded, the total number of words in the document, the number of unique words, vocabulary density, readability index, average words per sentence, and most importantly the frequent words in the corpus. These frequently used words are extracted to be the label for automatic tagging later on. The user can also set the items to be shown in the most frequent word.

**Word cloud**

Along with the results of phase III, a word cloud is also generated. Phase IV, as depicted in Figure 7, is a super word vector view of the frequent words in the processed corpus. The most evident words in the word cloud are the frequently used words from the previous phase, which are *faculty, activities, library, research*, and *materials*. The font size of the word is based on how many times this word is used in the corpus.

**LDA visualization**

With the result generated during phase III, this phase provides the histogram presentation of the sample processed corpus with the support of the LDA visualization, as shown in Figure 8. The line graph provides the relative frequencies of each generated label per document segment.

**Auto-tagging to uploaded document**

After the five phases, automatic tagging of the generated labels takes place, as shown in Figure 9. The document is then stored in the file repository of the IQAS. The uploaded document will have corresponding metadata such as filename, file size, user, date created, tags, and the link of the processed model. The filename can also be updated, and adding and removing tags is also possible.
Figure 7. Phase IV—word cloud snapshot.

Figure 8. Phase V—LDA visualization snapshot.
For better understanding, this section provides the computational analysis of the actual result based on the processed document.

In reference to the results of phase III, there are four significant results evident in Figure 6. Vocabulary density is the ratio between the total number of words present in the corpus and the unique words. To obtain the vocabulary density, the total number of unique words is divided by the total number of words; for the sample computation see Equation 1.

\[
Vocabulary\ density\ (VD) = \frac{\text{Number of unique words (UW)}}{\text{Total word count (WC)}}
\]

\[
VD = \frac{720}{2,833}
\]

\[
VD = 0.254
\]

**Equation 1.** Vocabulary density computation.

The vocabulary density of the processed corpus is 0.254, which implies that the corpus contains complex text with many unique words. Moreover, the readability index and average word per sentence uses Java break iteration, which is a local sensitive class that has an imaginary cursor that points to the current boundary in a string of natural language text. This contains different kinds of boundaries such as for text character, words, sentence instance, and potential line breaks. These boundaries are the basis for the readability index and average words per sentence, which are 16.106 and 21.5, respectively. Frequently used words are identified based on the number counts of the word used in the processed corpus.

The LDA visualization is presented through the correlation of the relative frequency of the word per document segmentation, as shown in Figure 8. To identify the relative frequency, it is necessary to decide the number of document segmentations. For the purpose of this study, the researchers used 10 segments for the document. The grouping of words per segment is based on the total word count. The prototype now determines how many times a particular word is used per
segment. Upon determination, the identified number of counts is divided into the total word count. For the sample computation, see Equations 2 and 3.

\[
\text{Words per segment (WS)} = \frac{\text{Desired number of segments (DNS)}}{\text{Total word count (WC)}}
\]

\[
WS = \frac{10}{2,833}
\]

\[
WS = 283.3^*
\]

* First seven segments contain 283 words while the last three segments contain 284 words.

**Equation 2.** Words per segment computation.

Relative frequency (RF) = \[
\frac{\text{Word count per segment (WCS)}}{\text{Total word count (WC)}}
\]

\[
RF = \frac{2}{2,833}
\]

\[
RF = 0.0007060
\]

**Equation 3.** Sample computation for relative frequency (Word: research|2nd Segment).

For the overall results of the histogram, Tables 1 and 2 present the tabular representation of the relative frequency per label and per segment.

**Table 1. Word count of labels per document segment.**

<table>
<thead>
<tr>
<th>Labels</th>
<th>Word count per document segment</th>
<th>Total count</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Faculty</td>
<td>1</td>
<td>13</td>
</tr>
<tr>
<td>Activities</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>Library</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Research</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Materials</td>
<td>3</td>
<td>6</td>
</tr>
</tbody>
</table>

**Table 2. Relative frequency of labels per document segment.**

<table>
<thead>
<tr>
<th>Labels</th>
<th>Relative frequency per document segment</th>
<th>Total count</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Faculty</td>
<td>0.000353</td>
<td>0.004589</td>
</tr>
<tr>
<td>Activities</td>
<td>0</td>
<td>0.0001059</td>
</tr>
<tr>
<td>Library</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Research</td>
<td>0</td>
<td>0.000706</td>
</tr>
<tr>
<td>Materials</td>
<td>0.001059</td>
<td>0.002118</td>
</tr>
</tbody>
</table>
Conclusions
CSPC is in an exploratory phase when it comes to solving this particular problem involving accreditation. It is evident that there are problems encountered by the organization pertaining to the accreditation process. Therefore, the researchers devised a model that supports the organization for accreditation. In addition, the researchers also designed a prototype with the implementation of the model to help the organization through the process. As a result, it is easier to retrieve and classify the data, which is the main problem of the task group. Furthermore, other text classification patterns may also be integrated into the system and the results compared with given parameters.

Software availability
Software available from: https://github.com/CraigList056/iqas/tree/v1.0.0-alpha
Source code available from: https://github.com/CraigList056/iqas
Archived source code at time of publication: https://www.doi.org/10.5281/zenodo.7507492
License: MIT License

Acknowledgements
We would like to express our great appreciation to our colleagues and friends for their undeniable support and for uplifting our spirits to make this research paper possible. We would like to also extend our appreciation to our respective institutions (Camarines Sur Polytechnic Colleges and University of the Cordilleras), which have been our second home and witness of our efforts during the research process.

References
Open Peer Review

Current Peer Review Status:  

Version 1

Reviewer Report 29 March 2023

https://doi.org/10.5256/f1000research.142987.r165836

© 2023 Naseem S. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Shahid Naseem
Department of Information Sciences, Division of Science & Technology, University of Education, Lahore, Pakistan

By looking at paper overall structure, presentation and above all the provided contents, I would say the authors of the paper requires minor changes to accept it for indexing.

1. In this study, the indexing of the tagging/titles, sub-titles is missing.

2. Number of sentences and grammar mistakes in different sections of the paper.

3. In result section, number of students studied in the batch to be accredited, financial statement, and infrastructure must also be included because these documents are also required in accreditation process.

4. In related work, there should be structured or labelled data instead of predefined data items.

5. In second paragraph of related work, the authors defined four types of machine learning techniques, but didn’t explain for what purpose, these four techniques were used in this study.

6. In figure 2, there should be one more step i.e. maintenance included.

7. All the equations used in this study must be numbering.

8. In equation 1, explain the procedure to calculate VD, from where we get the used valued to calculate VD.

9. Numbers of references used to validate this study are too short. There must be some more literature review to authenticate this study be used in this research.

Is the work clearly and accurately presented and does it cite the current literature?
Yes

**Is the study design appropriate and is the work technically sound?**
Yes

**Are sufficient details of methods and analysis provided to allow replication by others?**
Yes

**If applicable, is the statistical analysis and its interpretation appropriate?**
Yes

**Are all the source data underlying the results available to ensure full reproducibility?**
Partly

**Are the conclusions drawn adequately supported by the results?**
Yes

**Competing Interests:** No competing interests were disclosed.

**Reviewer Expertise:** Artificial Intelligence, Machine Learning, and Deep learning for analyzing healthcare data

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

---

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com