Introduction

e learning algorithms . cluster ##ing 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 , re ##tri as the le ##mma . lower ##cas ##ing it refers to the conversion of all characters to lower ##case . 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 . prep ##ro ##ces ##sing 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 program ##mi ng 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 un ##st ##ructured text into a structured format in order to find new insights and significant patterns . t ##f - 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 import ##a nc ##e across all rent related literature and studies . digital libraries digital libraries had been around for a l on ##g time . in the study by bear ##man (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 t he pre - www context of 1991 - 1993 , the period of popular discovery of the internet , it was logical that digital libraries focused on fixing the tech ##n ic ##al issues that people creating multimedia material collections encountered . however , the focus of digital l ib ##rar ##y applications ha d shifted since the focus on discrete collections during that period . borg ##man (1999) , in her study entitled “ what are digital libraries ? competing visions , ” stated that the term “ digital library ” was a practical and well - known short ##h and for electronic collections , convey ##ing a feeling of deeper material and capabilities than terms like " database " or " information re ##tr ie ##val 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 ev ##oke d 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 computer ##ized . to others , who have studied library science , believe it calls for a new ap pro ##ach to library operations , one that includes new 13 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 inter ##link ##ed information , or a network ##ed multimedia information system . in this regard , wat ##stein et al . (199 9) , in their study entitled “ digital library : key ##words , ” stated that the word wide web (www) , an inter ##connected sy stem of public web ##page ##s or websites accessible via the internet , ha d been referred to as a “ digital library ” on occasion , with the emphasis on the inter ##connected system or collection of databases of websites or web ##page ##s , rather than the organization , intellectual accessibility , or service attributes . the arrival of the world wide web ha d enabled the proliferation of new sources of content and in formation . with the increasing popularity of the internet , digital libraries ha d become crucial tools for institutions . in the study by cab ##rer ##iz ##o et al . (2015) entitled “ a decision support system to develop a quality management in academic digital li ##bra ##rie s , ” 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 br ##id ##ging that gap . ti ##onal libraries . further , by making the broad range of information resources available to student s 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 wit ##ten et al . (2002) entitled “ the promise o f digital libraries in developing countries , ” it was stated that they could assist human development by providing a non ■ commercial mechanism for d ist ##ri ##bu ##ting humanitarian information on topics such as health , agriculture , nutrition , hygiene , sanitation , an d 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 vi ##bal foundation , ” fi ##lip ##in ian ##a . net serve d as a research portal and digital library on philippines studies wherein it collect ed , organize d , and di ##sse ##minate d fi ##l ar ##y . web portals ta ##t ##nall (20 05) , “ 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 access ##ing different resources on the web . in the study by el ##dran ##dal ##y et al . (2009) entitled “ ranking web portals quality criteria based on different types of personality , ” it was stated that a web portal takes information fr ##o m 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

Method

i paper ##ner ##ds : 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 st ##a . cruz campus in partial fulfillment of the requirements for the degree bachelor of science in computer science by : miguel n . ce ##mill ##er jian carlo c . cara ##ya ##p air ##ha m . villa ##nu ##eva under the supervision of : mia v . villa ##rica , d . i . t . feb ##ru ##ar y 202 ##2 ii vision the laguna state polytechnic university is a center of sustainable development initiatives transforming lives and communities . mission l ##sp ##u provides quality education through responsive instruction , distinctive research , and sustainable extension and production services for improved quality of life . quality policy l ##sp ##u delivers quality education through responsive instruction , distinctive research , sustainable extension , and production services . thus , we are committed w it ##h continual improvement t o meet applicable requirements to provide quality , efficient and effective services to the university stakeholders ' highest level of satisfaction through an excellent management system im ##bu ##ed with utmost integrity , professional ##ism and innovation . college o f computer studies goal the college of computer studies graduates are expected to become globally competitive and innovative computing professionals im ##bu ##ed with utmost integrity , contributing to the country ’ s national development goals . program educational objective the bachelor of science in computer science (bsc ##s) 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 gr ##a du ##ate : 1 . apply knowledge of computing solutions from fundamental ##s to complex problems appropriate for the abstraction and conceptual ##ization of computing models . 2 . communicate effectively and recognize th ng ##age in life - long learning as foundation for professional development . iii approval sheet the thesis entitled “ paper ##ner ##ds : a research digital library and undergraduate thesis portal with text mining algorithm ” prepared and submitted by miguel n . ce ##mill ##er , jian carlo c . cara ##ya ##p , and air ##ha m . villa ##nu ##eva in partial fulfillment of the requirements for the degree of bachelor of science in computer science is here ##by recommended for approval and acceptance . approved by the committee on oral examination with a g ra ##de of _ _ _ _ _ _ _ _ _ . _ accepted and approved in partial fulfillment of the requirement for the degree of bachelor of science in computer science . date signed : _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ research contribution no . mia v . villa ##rica , di ##t thesis adviser jon ##ardo r . as ##or subject specialist all ##iana m . miranda - ab ##lan technical editor ma . ce ##zza ##ne d . dim ##ac ##ula ##nga ##n language critic victor a . est ##ali ##lla jr . stat ##istic ##ian rey ##nal ##en c . just ##o , mm - it ##m , lp ##t chairman rey ##nal ##en c . just ##o , mm - it ##m , lp ##t dean / associate dean dr . ri ##na j . arc ##iga ##l director / chairperson research and development iv acknowledge ##ments the researchers would like to express their heart ##felt gratitude to the following individuals who have contributed and supported them in the completion of the study : mrs . mia v . villa ##rica , d . i . t , their thesis adviser , for her t im ##e 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 . all ##iana m . miranda - ab ##lan , their technical editor , for correct ##ing the mistakes in the format and contents of the manuscript as well ; mr . jon ##ardo r . as ##or , their specialization expert , for sharing his much - needed knowledge and suggestions regarding the study ; mr . victor a . est ##ali ##lla jr . , their stat ##is ti ##cian , for sharing his inputs regarding this study ’ s data sampling design ; ms . ma . ce ##zza ##ne d . dim ##ac ##ula ##nga ##n , their language critic , for her assistance in criticizing the construction and grammar of the manuscript ; mrs . rey ##nal ##en c . just ##o , mm - it ##m , lp ##t , 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 . s discovered that topic models work particular l ##y well with longer texts as opposed to cluster ##ing algorithms which work better with shorter texts . ld ##a topic modeling on short texts returned a co ##her ##ence score of 0 . 69 ##8 while returning 0 . 75 ##6 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 . key ##words : digital library ; web portal ; undergraduate thesis ; text mining ; topic modeling ; cluster ##ing ; inform ##a ev ##ing , and often di ##sse ##minating recorded data , especially via the use of a computer ##ized system . k - means cluster ##ing it refers to an un ##su ##per ##vis ##ed learning algorithm , which groups the un ##lab ##ele ##d data ##set into different clusters . late ##nt dir ##ich ##let allocation (ld ##a) it refers to a topic model that is used to cat ##ego ##rize text in a document to a particular topic . xiv le ##mma ##ti ##zation it refers to the removal of in ##fle ##ction ##al endings only and to return the base or dictionary form of a word , which is known 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 & gr ##efe ##nst ##ette (1996) , was implemented . 2 next , the integration of topic modeling , a technique for finding late ##nt or hidden topics within document collections according to an ##up ##ri ##ya & ka ##rp ##aga ##val ##li (2015) , recommended re ##lat leg ##e of computer studies (cc

Result

sm ##ates , 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 . ce ##mill ##er , miguel n . cara ##ya ##p , jian carlo c . villa ##nu ##eva , air ##ha m . vi abstract topic modeling is one of the most powerful approaches in text mining for data mining , late ##nt data discovery , and finding relationships among data and text documents . on the other hand , cluster ##ing in text mining has als documents . token ##ization it refers to the process of exchanging sensitive data for non - sensitive data called " token ##s " 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 . xv un ##su ##per ##vis ##ed 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 tr ##ai ve ##r the internet . i t was used to imply the output system of the study . chapter i introduction and its background data gr e ##w exponential ##ly every day . almost all types of institutions and organizations store d data digitally . according to saga ##yam (2012) , this generally c am ##e as textual data in digital libraries and other rep ##osi ##tori ##es . in the college of computer studies (cc ##s) of laguna state polytechnic university santa cruz campus , the research implementing unit head (ri ##uh) or the research co ##or and technical panel members through text mining algorithms for the undergraduate thesis process of students in th e college of computer studies (cc ##s) at laguna state polytechnic university sc ##c . research problem according to pat ##il & at ##ique (2013) , massive amounts of data were generated every day , resulting in the digit ##ization of the majority of information in govern ##men t , industry , business , and institutions . as stated in an interview conducted with the research implementing unit head (ri ##uh) of the college , they also utilize d the use of online rep ##osi ##tori ##es and public libraries for the storage of research papers . the res ##e arch papers were kept private through a cloud - based file storage service , google drive . however , this prevent ed the general public , particularly college student researchers , from access ##ing the research papers , especially in the midst of the co ##vid - 19 pan ##de ##m ic , when access to public libraries was limited . moreover , it was figured that searching relevant resources for a 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 retrieve ##s 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 per ##s 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 soft ware and applications which were re ##lat ms were prior ##iti ##zed at the top of the list , while non - priority items were pushed to the bottom . once the product back ##log was built , it was regularly maintained by the researchers to make sure prior ##iti ##zation is correct and new items were added or removed . 74 sprint pl ##a n ##ning in sprint planning , the researchers then addressed which product back ##log 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 c ri ##nt , such as “ search research papers ” , “ view related research papers ” , etc . sprint following the creation of the sprint back ##log , the researchers then began the sprint . when ever a sprint begins , the researchers or the developer start their work . the developer bu ##il t the features identified with the help of the sprint back ##log with a limit to a set period of time during which the work ha d to be completed . during the sprint , the res ##e archers held a daily sc ##rum , or a short daily stand ##up meeting in l 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 (ri ##uh) , a key stake ##holder , reviewed the outcome of a sprint in an event called sprint review . in sprint review , the researchers and the ri ##uh me t to review the completed work and address ed any concerns . sprint retrospective after the sprint review , the researchers convened for another meeting in less than an hour to wr ap 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 commit ted to improving in the next sprint . for the next sprint , the processes in the sc ##rum work ##flow , starting with sprint planning , were repeated . 76 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 en ##ume ##rating or ga ##mi ##fication 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 soft ##war e application develop event decision sent ##ime ##n t human anal ##ys ##i s topic t ##wee ##t expert use game develop learning student res ##ear ##c h test study system em ##bed energy gaming device sim ##ula ##t ion efficiency power 2 blue card mana ##g g en ##t image system treat ##me nt model plant drug dia ##gno ##s is inform ##a ti ##on text database doc ##ume nt ##n module com ##put er use extract 3

Discussion

e 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 , algorithm ##ic pri ##nc ##ip les 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 e v dedication the researchers whole ##hearted ##ly de ##dicate 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 de ##dicate their study to the following groups of individuals : family of the researchers , who have been their source of inspiration and gave them the mo ##t iv ##ation to keep pushing forward . the study would not have been possible without their constant love and support ; and lastly , their cl ##as o been proven to be a very effective m et ##ho ##d 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 ld ##a topic modeling and k - means cluster ##ing on different sets of textual data . based on the findings , it wa ti ##on retrieval ; late ##nt dir ##ich ##let allocation ; k - means ; co ##sin ##e similarity ; t ##f - idf ; vii table of contents