Welcome to JISST

Journal of Information Science, Systems and Technology (JISST) is the flagship bi-annual electronic academic journal of the Africa Regional Centre for Information Science, University of Ibadan, Nigeria.

Objectives and Scope

JISST welcomes manuscripts for publication on all topics of interest in the multidisciplinary field of information science and technology. The field of Information Science and Technology is defined broadly to include the academic and professional aspects of the following disciplines and subjects: Archives Studies/Management, Communication Arts/Science, Computer Science, Information Management, Information Science, Information Services, Information Systems, Information Technology, Innovation Management, Library Management, Library Science, Media Science, Media Studies, Records Management, Research and Development Management, Telecommunications, and other related fields, occupations and professions. These disciplines and subjects are concerned with different and overlapping perspectives, processes, antecedents and consequences of using data, information or knowledge in its various forms in society, as well as the methods, technologies and standards for creating, organizing, storing, transferring and using such data, information or knowledge. JISST provides a platform for the publication of research manuscripts in any of the above disciplines and subjects, and the exchange of academic and professional ideas through Short Communications, Book Reviews an Letters to the Editor on any issue in the field of Information Science and Technology.

Publication Frequency

JISST will publish two issues per volume/year, in May and November.

Access and Subscriptions

Access to issues of the journal will be free initially, until long term open access and subscription frameworks are finalized.

JISST Content

JISST Content

Journal of Information Science, Systems and Technology

Volume 5, Number 3

Table of Content

JISST
ISSN: 2545-5915

Journal of Information Science, Systems and Technology

Africa Regional Centre for Information Science, University of Ibadan, Nigeria
6 Benue Road, P. O. Box 22133, University of Ibadan, Nigeria 200001
Website: www.arcis.ui.edu.ng/jisst *Email:* jisst_arcis@ui.edu.ng ; jisst2015plus@gmail.com

Vol. 5, No. 3 (December 2021)

Application of Text Mining for Explicit Knowledge Retrieval at Kenya Coastal Development Project

Ednah Nyakerario Onkundi [1, 4]
eonkundi@gmail.com

Raymond Wafula Ongus [2]
raymondongus@gmail.com

Constantine Matoke Nyamboga [3]
constantinenyamboga@gmail.com

[1] *Kenya Marine and Fisheries Research Institute, P. O. Box 81651, Mombasa, Kenya*
[2] *School of Pure & Applied Sciences, Department of Computing & Informatics, Mount Kenya University, Thika Main Campus, P.O. Box 342-01000, Thika, Kenya*
[3] *Lukenya University, P.O Box 90-90128, Mtito Andei, Kenya*
[4] *Corresponding author*

Abstract

This study applied and evaluated a Text Mining model and tools in the retrieval of explicit knowledge at the Kenya Coastal Development Project (KCDP). The study identified text-mining techniques that could be used to develop and evaluate a text-mining model for the task, complemented with system analysis of the existing systems through surveys of 52 purposively sampled relevant staff of the KCDP using a questionnaire and focus group discussions (FGD) to collect data to establish the current situation at the KCDP in terms of records and knowledge management systems in place and how text mining may be used to retrieve knowledge from the existing systems. Text data were also collected from the project and related websites, using various Python programming language libraries including Python Request 2.22 and Beautiful Soup 3, and summarized using algorithms that included Luhnsummarizer, Lsansummarizer, Lexranksummarizer and Edmondsummarizer. Topic modelling was also performed with the text data using Latent Dirichlet Allocation (LDA) topic-modelling algorithm, and the model was evaluated to establish its performance. It was concluded that text mining and analysis could be used to analyze explicit knowledge from both structured and unstructured electronic data at the KCDP using the model. The study recommended that more research be done in the development and evaluation of proposed models and text analysis tools and code libraries should be developed to support other languages than English, such as Kiswahili.

Keywords: Data Mining, Explicit Knowledge, Retrieval Techniques, Text Mining, Text Analysis

Introduction

Increase in analysis of text that is available has gone high in recent years. This is because of growth in social media, blogging and use of bulletin board systems. In addition to these, many documents, feeds from news and articles are now stored in soft copy. Classification of text documents involves an important step of classifying categories and classes of known text documents in a corpus (Shah *et al.*, 2017). Text mining techniques have grown because of the presence and demand of data in massive volumes in the web in the form of text. According to Tan (1999), text data mining has experienced challenges, which are relevant and relate to the mining of explicit knowledge. To gain great competitive advantage in today's global economy, an organization has to make effective use of its knowledge and knowledge assets. The most important resource in an organization and one that can show the strategic direction an organization takes is its knowledge base, which determines its decision making process based on data and information available. The democratic access of knowledge by the entire organization provides for opportunities of innovation and hence giving it a competitive edge against its competitors (Jandhyala, & Phene, 2015).

Knowledge discovery provides an opportunity for an organization to assess knowledge assets that may contribute to the realization of its immediate project objectives in the short run and to idealize the knowledge requirements like in the coastal communities where the project was designed to intervene. Text Mining applies the principles of Data Mining (DM), which is growing and expanding at an alarming rate in relation to new technologies like big and open data. Through the application of data analysis techniques, artificial intelligence and machine learning to discover relevant trends and show relations found within data, data analysis techniques makes it easy to find patterns that are difficult to observe manually. Most of the knowledge in Kenya Coastal Development Project (KCDP) is not shared with staff, members and the business community at large (KMFRI, 2012). The extension of the project as part of the President's Big four agenda was unveiled on 26th June 2021 at Kenya's Coastal region. The President of the country, while launching the project assured Kenyans especially those at the coastal region that the project shall eventually contribute on completion to a sustainable exploitation of Kenya's marine fisheries. (President Kenyatta, 2021).

The aim of this research was to apply text mining techniques like text classification, clustering and summarization to support the Kenya Coastal Development Project (KCDP) activities for explicit Knowledge Discovery, retrieval and the knowledge management processes. KCDP is a World Bank project hosted by Kenya Marine and Fisheries Research Institute. The project has various types of information that comprise of fisheries information, which is collected through research on marine and aquatic life. (KCDP, 2012). The information currently resides mostly in different paper documents and files, and the KCDP seeks to provide systems for updating and retrieving the information to and from created databases, thereby, supporting improved decision making on environmental, economic and social management of the coastal areas of the country to positively impact the people and occupations in the coastal communities. The total amount of available data to be updated to databases is estimated to be over 20 terabytes to be stored in databases and made accessible through numerous portals (KCDP, 2012).

Research Problem

This study involved use of Text Mining for explicit Knowledge Discovery. Development of Big Data increased availability of intelligence mining tools is boosting exponential expansion of scientific knowledge discovery through evolving data and research professions and industries (Wilensky, 2015).

Among the core mandate of the Kenya Coastal Development Project (KCDP) is the dissemination of research findings and recommendation to various stakeholders and communities living at the coast of Kenya. Traditionally, such valuable information is mostly created and held by various individual scientists and institutions and not systematically collated

and shared through deployed central knowledge management system of the project. Identifying and collating such disparately held information into such system for subsequent retrieval or mining by or for various stakeholders would greatly facilitate meeting some of the key objectives of KCDP. The discovery and storage of large amounts of data, information and knowledge at KCDP, can be made more effective and efficient through applying best practices in collecting, storing, archiving and sharing knowledge. It may also be possible to collate and codify even some tacit embedded in KCDP staff experiences, feelings and personal abilities and preferences through various documented communications, meetings, seminars, conferences. Collating and storing both explicit and such documented tacit knowledge is a fundamental knowledge management activity that organizations need to invest in (Joia & Lemos, 2010). Knowledge management systems reduce the loss of intellectual capital from people leaving the institution, saves money that might have been wasted through reinventing the wheel for new upcoming projects, and provides faster problem solving approaches, which leads to timely decision-making processes.

The knowledge assets and information at KCDP are currently inadequately managed. Project related or relevant data and information at KCDP exist in both hard and soft copies - on paper and different digital media created with various software on various computers. The hard copies are retrieved manually from various files, while the soft copies are kept in different storage sites that can be accessed by authorized users (KMFRI, 2012).This study therefore explored the use of Text Mining techniques for retrieval of explicit knowledge at the Kenya Coastal Development Project in Mombasa, as an envisaged component of a formal automated Knowledge Management System to enable all project related or relevant knowledge from various internal and external sources to be centrally stored, shared to and accessed by project managers, policy makers, members of the research and local coastal communities, and other stakeholders.

Literature Review

Text mining is the process of extracting important and exciting trends from textual databases. Text mining can also be defined as the discovery of new, unknown information of previous research by computers and the automatic extraction from different sources. The science of Text Mining has evolved over years in various fields. In the medical field, Text Mining has been applied in making sense of raw text. Text Mining and qualitative research are compatible epistemologically. Qualitative research approaches are applied in grounded theory of text mining since it supports open-mindedness and discourages preconceptions. It enables liberty and variety in manipulation of initial categories in an iterative fashion. Text mining applies extraction of common themes and threads using computer algorithms. Content analysis and text mining count words by extracting common themes (Yu, Jannasch-Pennell & DiGangi, 2011). The main and key element in text mining is the art of linking extracted information to form new facts and hypotheses to be discovered through experiments using conventional means (Hearst, 2003). Very early on, Italian researchers did discovery and integration of explicit knowledge and learning by example in recurrent networks by use of a unified approach, and the results from their studies led to the evolution of intelligent systems based on connection models (Bianchini *et al*., 1998*).*

Time mining research, models, algorithms and applications have since exploded. Among these are the following. In Tokyo University, multi-layered neural networks, which included the extraction of knowledge and its use, were applied to train explicit knowledge extraction in Information-Theoretic Supervised Multi-Layered (SOM). Connection weights were obtained at the knowledge extraction phase and were used to train. The method applied was the spam identification problem. The experiment results showed that the information theoretic Supervised Multi-Layered (SOM) improve learning and performance (Kamimura, 2014). Researchers in India also implemented a Document Management System (DMS) that captured explicit knowledge in organizational documents, and concluded that Document Management

System (DMS) contained useful information which was integrated with the existing online Human Resource Information System (HRIS) database (Khan *et al.,* 2015). An engineering department in Indonesia assessed the impact of tacit and explicit knowledge on small and medium enterprise, and the results indicated that tacit and explicit knowledge could be shared to create more knowledge (Rumanti, Samadhi & Wiratmadja, 2016). Researchers in South Korea applied the text mining methods of text pre-processing, clustering, categorization and visualization in the medical field to identify the relationship between gait-Parkinson's diseases (PD) from PD based research articles, and resolved that for assessing Parkinson's disease, gait related analysis was most important (Aich *et al.,* 2017). The knowledge had since been adopted by clinicians to improve on diagnosing and treating the disease. In Pakistan, researchers carried a study on technological verdicts using implicit and explicit knowledge mining of crowd sourced communities. The framework utilized text mining techniques that supported software development teams to get a wider spectrum of opinions in form of knowledge patterns discovered by the crowsourcing knowledge mining framework (Mushtaq *et al.,* 2018).

Explicit Knowledge Retrieval Issues

Most techniques were developed some years back with a combination ranging from natural language processing, information retrieval, information extraction, and Data Mining. Some of the techniques such as information retrieval involves using algorithms that make it possible to search data to satisfy user requirements. Continual improvement is key for any organization to move from one level to another. Adoption of and use of Text Mining may give breakthrough in promoting effective Knowledge Management and hence improve service deliveries and decision-making. In respect for continual improvement, organizations are collecting large volumes of data and storing them in various databases for future reference. Key issues identified and discussed within textual data and its classification would unearth and help the policy, knowledge workers and decision makers to better manage their activities (Ur-Rahman & Harding, 2012).

Retrieval of both tacit and explicit knowledge is not yet well researched on especially on the process of retrieval. Retrieval problems are dependent on the utility of a document based on their ranking and retrieval methods applied (Zhai *et al* 2015). In some instances, skills gained do not give any impact due to lack of knowledge sharing making projects fail, where their sustainability and continuity are hampered (Conger, 2015). Ethical challenges during data collection have been an issue when considering issues to do with privacy, storage and location of data, data interpretation, informed consent, identification, classification and management of data. Other ways taken to understand similar approaches depend on various ideologies and assumptions (Arnold & Pistilli, 2012). Key issues identified and discussed within textual data and its classification would unearth and help the policy, knowledge workers and decision makers to better manage their activities (Ur-Rahman & Harding, 2012). The major aim for clustering is to distribute cases for example people or objects into groups, so that the similarity degree is strong between objects in the same cluster and weak between objects in different clusters (Hajizadeh, *et al.* 2010). There is no pre-classified data in clustering and it does not distinct between independent and dependent variables. The most common learning models applied in the world today is classification and it is commonly used as a learning model in Data Mining techniques. Its main objective is to build a model that identifies the category an object belongs to *(*Ngai *et al.,* 2009*).* According to Dunham (2000), classification is viewed as a mapping from the database to the set of classes. Classification produces predefined, none overlapping and partitioned classes in an entire database. Summarization or text summarization reduces the length and detail of a document and at the same time avoids the distortion of data and information contained in a document. It applies models like LDA topic models that are probabilistic for analyzing text that are meaningful and useful (Jelodar *et al.,* 2019).

Research Objectives

The general objective of this study was to investigate the theoretical and practical prospects and challenges of applying a Text Mining model in the retrieval of explicit knowledge at the Kenya Coastal Development Project (KCDP). The specific objectives of the study were:

1. Design a Text Mining Model for retrieval of explicit knowledge at KCDP.
2. Explore Text Mining techniques and tools for retrieving explicit knowledge at KCDP.
3. Implement and Evaluate Use of Text Mining for Explicit Knowledge Retrieval at KCDP.

Methodology

1. Methodological framework

The methodological framework for the text mining implemented in the study is diagrammed in Figure 1.

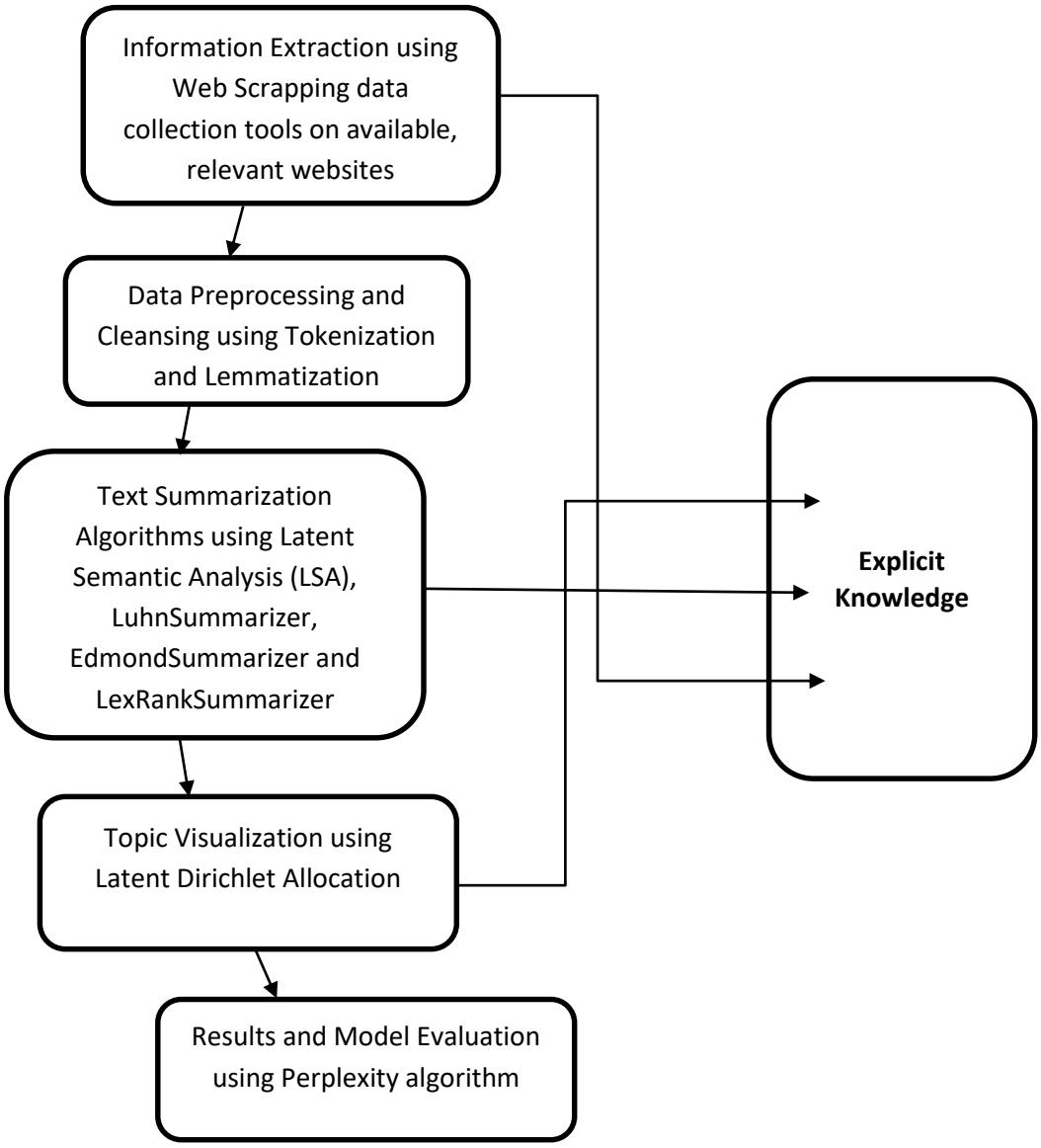

Figure 1: Text Mining Model for Retrieval of Explicit Knowledge at KCDP

Figure 1 illustrates the main stages for text mining implementation, which are: (i) Information Extraction using Web Scrapping data collection tools on available, relevant websites (ii) Data Preprocessing/Data Cleansing using Tokenization and Lemmatization; (iii) Text Summarization Algorithms using Latent Semantic Analysis (LSA), LuhnSummarizer,

EdmondSummarizer and LexRankSummarizer; (iv) Results and Model Evaluation using Perplexity algorithm; (v) Topic Visualization using Latent Dirichlet Allocation.
Information extraction and summarization are text-mining techniques widely used because of their flexibility and performance with different python text summarization algorithms like Latent Semantic Analysis (LSA), LuhnSummarizer, EdmondSummarizer and LexRankSummarizer. Topic modelling algorithms/libraries like Genism and Latent Dirichlet Allocation (LDA) models are then used to extract topics from the text. The text mining technologies used were information extraction, summarization and topic modelling/visualization. To determine the accuracy of information to be extracted both the precision and recall metrics were used, as follow:

$$\text{Precision metric} = \frac{\text{Number of relevant items retrieved}}{\text{Number of retrieved items}}$$

$$\text{Recall metric} = \frac{\text{Number of relevant items retrieved}}{\text{Total number of relevant items}}$$

2. System survey data collection

The study employed an exploratory research design to gather information about the existing data and information management systems in use at the KCDP. The exploratory survey design was considered best suited for the study because of the limited research that had been done in the area of text mining for retrieval of explicit knowledge. According to Oates (2005), exploratory research design is suitable for areas of study where little or no research had been done on an area of study.

A questionnaire was designed and used in the study to collect information about the existing methods and systems for information and knowledge management at the project. This assisted the study in the design and implementation of the text mining model in Figure 1. Questionnaires are a cheaper method of gathering data from potentially large numbers of respondents. It is a well-known technique to collect data and people's opinions (Preece *et al.,* 2002) Additionally, Focus Group Discussions was conducted to collect some qualitative data from the Knowledge Management and ICT departments of the Project. The collected information provided data management insights on the project that informed interpretation and discussion of the results of the study.

The study's population of interest comprised the 150 staff members of KCDP, as provided by the administrative officer of the project (I. Githaiga, Personal Communication, 11 January 2019). Nevertheless, purposive sampling technique was applied to ensure that the data sampled, collected and analysed could assist in achieving the objectives of the study, for the survey, which was to collect enough information perception data to assess the nature, prospects and challenges of the existing data, records and knowledge management systems being used at the KCDP. Purposive sampling also referred to as judgment or selective sampling, involved selecting a sample suited to the purpose of the study (Williamson & Johanson, 2017). The criteria used in purposive sampling involved ensuring sampling from the top management level, as well as research and administration, ICT and the knowledge management departments of the KCDP, on the purpose that the targeted staff would provide the required data and information from their available files, documents and experiences, and perceptions. Moreover, random sampling was used to select files and documents from these purposively sampled departments at the various other methodological stages of the research. A pilot study was conducted to pretest and assess the feasibility, cost, time, validity and reliability of the instruments of planned survey data collection. A pilot study helps to check out the face and content validity of questionnaire, s alongside opinions sought from professionals and experts

in the field of investigation (Mugenda & Mugenda, 2003). The pilot study was carried out in one of the Kenya Coastal Development Project agencies, the Kenya Marine and Fisheries Research Institute (KMFRI), and involved survey of 15 staff members, who were later excluded from the main survey of other agencies of the KCDP.

Information obtained from the knowledge management department provided great insights on how data and information acquired from previous and current research were managed and the challenges that needed to be addressed. The department also provided most of the information through focus group discussions and practical challenges that needed immediate solutions. This technique was chosen due to the unique subject matter in the research area to provide quick insights on explicit knowledge retrieval. The other sampled and surveyed departments sampled were those that contained much data and information about the KCDP project to enable meeting the objectives of the study. The total population of staff in the Kenya Coastal Development Project was one hundred and fifty (150), while the target sample size was fifty two (52). Various factors were considered in determining the sample size, such as objectives of the survey, population size, required sampling precision level, confidence level, etc. The precision level (sampling error) provided the range in percentage points where the true value of the population was estimated. The level of precision in this study was taken to be 11% due to the size of the population. The confidence level referred to the percentage of all possible samples that were distributed normally considering the true value. Since this was a normal distribution, a 95% level of confidence was selected for the research study. The degree of variability refers to the manner in which attributes of the population were distributed. A small sample size implied less variability in the population. A proportion of 50% was selected as it indicated the maximum variability in the population. The following formula was then used to calculate the adequate sample size (Israel, 1992).

$$n_0 = \frac{z^2 pq}{e^2}$$

Where n_0 is the sample size, z is the abscissa of the normal curve that cuts off an area $\propto$ at the tails, e is the precision level, p is the variability degree and $q = 1 - p$. Thus:

$$n_0 = \frac{(1.96)^2(0.5)(0.5)}{(0.11)^2} = 80$$

However, since the population is small, the sample size was slightly reduced and adjusted using:

$$n = \frac{n_0}{1 + \frac{(n_0 - 1)}{N}}$$

$$n = \frac{80}{1 + \frac{(80 - 1)}{150}} = 52$$

3. Text Data Collection, Cleaning, Analysis, Visualization

The text data collection methods used in this study involved the use of data collection tools like RapidMiner Studio 9.5 with its extensions Aylien Text Analysis, Rosette Text Analysis and web Application Programming Interfaces (API) like the Twitter Search API Open Source web scraping tools, which include Beautiful Soup 3 and Python Request 2.22, were also used. Text can be collected from websites and portals through web scrapping or using web application programming interfaces (Kobayashi *et. al*, 2018). Data collected from portals and websites that were involved in or related to the KCDP project were analyzed using text summarization algorithms like Luhnsummarizer, Lsansummarizer, Lexranksummarizer and Edmondsummarizer and visualized using a topic-modeling algorithm, the Latent Dirichlet Allocation (LDA). Data from the retrieved completed questionnaires were analysed using

Social Sciences (SPSS) version 25 statistical analysis software. The procedure involved validation of the questionnaire to check for clarity, legibility, relevance and appropriateness of the data. The questionnaires were then edited for completeness and consistency, coded using descriptive statistic and finally analysed in SPSS. Information collected and obtained from focus groups complimented the data and information collected and analyzed from both SPSS and the Text Mining tools.

Results and Discussion

From the data collected and analysed, it was established that there are several forms of databases at KCDP based on the department an employee works in. An Enterprise Resource Planning (ERP) Dynamics SL 2011 Software was used at KCDP to store data and information in form of a database for the respective departments. It was also established that some of the other forms of databases at KCDP included KOHA, an open source Integrated Library System, MYSQL, Microsoft Excel, Mat lab and the use of Geonetwork 2.0.3 software for data storage and manipulation. For some of the departments data was stored physically in files, cabinets and in hard copy form. Based on the study it was proposed to have data and information that is in hard copy form in its duplicate soft copy form for purposes of backup, data centralization, storage and archiving.

1. Survey findings

The results from the survey of the KCDP staff showed that almost all (96.2%) affirmed that they had databases at their places of work. But almost all (96%) of them indicated that they were not aware of automated explicit knowledge retrieval from the databases. Nevertheless, , responses to other questions revealed that the staff knew about or used before the following specific text mining processes: information extraction (32.7%), summarization and classification (15.38%), information extraction, summarization and classification (13.46%), information classification (11.54%) 7.69% used information extraction and summarization (7.69%), 5.77% used classification (5.77%), 1.92% applied summarization (1.92%), while the rest (11.54%) provided no responses. Majority (61.54%) of the respondents agreed that the project encouraged knowledge management, 21.2% agreed they understood very well the benefits of effective knowledge management, while 17.31% expressed they did not have any idea of knowledge management at KCDP. Also, 50% and 42.3% of the respondents strongly agreed or agreed that KCDP needed a model for retrieval of explicit knowledge, while only 7.7% did not know what the model could do and how it could benefit the project.

The focus group discussion was done to get an in-depth information to validate the data that was collected through the questionnaire. It was used as a qualitative approach to gain an in-depth understanding of social issues. The focus group discussion collated staff opinions on how data and information at KCDP could be accessed and used through explicit knowledge retrieval and mining to provide in-depth knowledge for decision makers at the project.. The focus group discussion took place at the KCDP offices on 19 November 2019. The main objective of the focus group discussion was to highlight where KCDP was in terms of data and information management and where it intended to be. Participants in the FGD showed understanding of what knowledge retrieval entailed through the various adequate descriptions of what knowledge retrieval entailed, including getting information from a specific storage, recovery and restoration of data from archives, obtaining data from an information management system and extraction of data and information using queries. They also understood knowledge management in terms of sharing of information for learning, access to expert information, availability of information when needed, and building of information assets within an organization. Despite these understandings, some of the participants still felt that more sensitization, training and subsequent audits of deployed Knowledge Management Systems needed to be carried out in all departments at KCDP. The majority of them understood

the whole concept of information and knowledge retrieval from the context of processes of collection, processing, storage and sharing of information.

Results and Discussion

1. Specific Objectives

Specific Objective 1: Design a Text Mining Model for Retrieval of Explicit Knowledge at KCDP.

The general objective of this study was to investigate the prospects and challenges of applying a text mining model in the retrieval of explicit knowledge from digital documents at the Kenya Coastal Development Project (KCDP). To achieve this general objective, some specific objectives were also specified in terms of the expected processes and outputs from the use of various text mining tools for information extraction, categorization, classification, summarization and visualization, as illustrated in the framework in Figure 1.

Specific Objective 2: Explore Text Mining Techniques and Tools for Retrieving Explicit Knowledge at KCDP.

This involved evaluation of different text mining techniques like text data collection, cleaning, summarization, classification and categorization, and information extraction. These are some of the text mining techniques used in the development of text mining algorithms and models. RapidMiner Studio 9.5 tools and various available extensions, operators and Application Programming Interfaces (API's) were tested and eventually adopted and used to perform these techniques. The used RapidMiner Studio 9.5 extension for this study was the Aylien Text Analysis extension, which provided operators like language detection, summarization, and categorization and sentiment analysis. The use of Rosette Text Analysis extension involved the use of operators like entity extraction, entity linking, entity sentiments, name matching, name translation and morphology.

RapidMiner Studio 9.5 contains a twitter search application-programming interface. The Twitter application operators in RapidMiner Studio 9.5 was used to mine the top 100 most recent and most popular tweets that were related to KMFRI. Figure 2a and 2b show the results of using this twitter application operator was used to mine top 100 tweets of the Kenya Marine and Fisheries Research Institute (KMFRI), an institute that is participating in the Kenya Coastal Development Project. These results shows the importance of text mining techniques for mining and retrieval of explicit knowledge from unstructured data such as tweets.

Row No.	Created-At	Id	From-User	From-User-Id	To-User	To-User-Id	Language	..	Text
33	Jul 26, 2021 ...	1419732351...	Big Ship CBO	1397867851...	?	-1	en	.	RT @KmfriResearch: Earlier today, Senior Research Sci
34	Jul 26, 2021 ...	1419702842...	Caroline Njeri	1050373529...	?	-1	en	.	RT @Manyunyu_Corg: We celebrated the #WorldMangro
35	Jul 26, 2021 ...	1419702822...	Caroline Njeri	1050373529...	?	-1	en	.	RT @BigShip_CBO: PICTORIAL;
36	Jul 26, 2021 ...	1419701996...	Caroline Njeri	1050373529...	?	-1	en	.	Planting over 300 propagules,barefooted, sticking to the
37	Jul 26, 2021 ...	1419694636...	Big Ship CBO	1397867851...	?	-1	en	.	PICTORIAL;
38	Jul 26, 2021 ...	1419693274...	Big Ship CBO	1397867851...	?	-1	en	.	RT @Manyunyu_Corg: We celebrated the #WorldMangro
39	Jul 26, 2021 ...	1419689609...	George Maina	4254734482	?	-1	und	.	#WorldMangroveDay @Nature_Africa @KmfriResearch (
40	Jul 26, 2021 ...	1419686053...	Manyunyu Community Base...	9428012321...	?	-1	en	.	We celebrated the #WorldMangroveDay2021 in Mirironi J
41	Jul 26, 2021 ...	1419661837...	Kenya Projects	1306126684...	?	-1	en	.	RT @WeDontHaveTime: The shores of Lake Victoria are
42	Jul 26, 2021 ...	1419661466...	KMFRI	9021390783...	?	-1	en	.	Earlier today, Senior Research Scientist at KMFRI Dr Juc
43	Jul 26, 2021 ...	1419660924...	KMFRI	9021390783...	?	-1	en	.	RT @WeDontHaveTime: The shores of Lake Victoria are
44	Jul 26, 2021 ...	1419648506...	George Maina	4254734482	?	-1	und	.	#WorldMangrovesDay @NRT_Kenya @CarolineLumosi
45	Jul 26, 2021 ...	1419643931...	AgamirQueen Fashions	1253910014...	?	-1	en	.	RT @WeDontHaveTime: The shores of Lake Victoria are
46	Jul 26, 2021 ...	1419619109...	Joel Swagman	2807365255	?	-1	en	.	RT @WeDontHaveTime: The shores of Lake Victoria are

Figure 2a: Top Tweets table

(Source: KCDP text data)

Text
RT @KmfriResearch: Earlier today, Senior Research Scientist at KMFRI Dr Judith Okello following proceedings of the international day for th…
RT @Manyunyu_Corg: We celebrated the #WorldMangroveDay2021 in Mirironi Jomvu Sub County through the support of @GreengrantsFund togeth...
RT @BigShip_CBO: PICTORIAL;
Planting over 300 propagules,barefooted, sticking to the mud,is fun full, though tiring,but, knowing what it's actually doing to the environment,it's totall...
PICTORIAL;
RT @Manyunyu_Corg: We celebrated the #WorldMangroveDay2021 in Mirironi Jomvu Sub County through the support of @GreengrantsFund togeth...
#WorldMangroveDay @Nature_Africa @KmfriResearch @EmilyCLandis https://t.co/GIK8G4Te26
We celebrated the #WorldMangroveDay2021 in Mirironi Jomvu Sub County through the support of @GreengrantsFund together with the Chief Admini...
RT @WeDontHaveTime: The shores of Lake Victoria are clogged with water hyacinth, an invasive plant hurting Kenya's freshwater fishery, and…
Earlier today, Senior Research Scientist at KMFRI Dr Judith Okello following proceedings of the international day for the conservation of the mangrov...
RT @WeDontHaveTime: The shores of Lake Victoria are clogged with water hyacinth, an invasive plant hurting Kenya's freshwater fishery, and…
#WorldMangrovesDay @NRT_Kenya @CarolineLumosi @NRT_Kenya @KmfriResearch https://t.co/bRAQTy8L7W
RT @WeDontHaveTime: The shores of Lake Victoria are clogged with water hyacinth, an invasive plant hurting Kenya's freshwater fishery, and…
RT @WeDontHaveTime: The shores of Lake Victoria are clogged with water hyacinth, an invasive plant hurting Kenya's freshwater fishery, and…

Figure 2b: Top Tweets table (continuation of Text column)
(Source: KCDP text data)

Figure 3 shows visualizations of tweet counts from different users on different tweet topics, while Figure 4 shows the users tweeting the most on various topics. More information could be obtained on tweets like geo-location information, followers of a certain user, users a certain individual was following, date of joining twitter, website URL for individuals or institutions or companies.

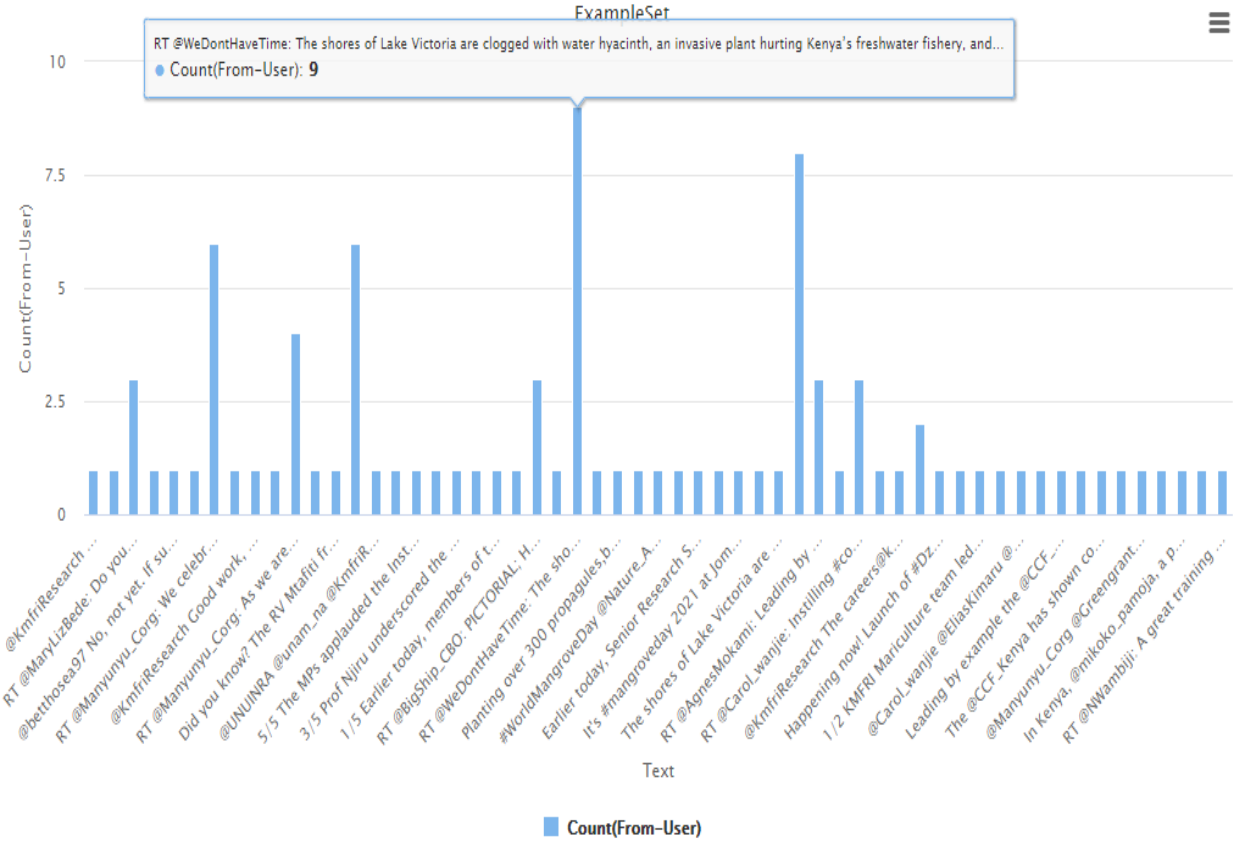

Figure 3: Top Tweets Counts Visualization

(Source: KCDP text data)

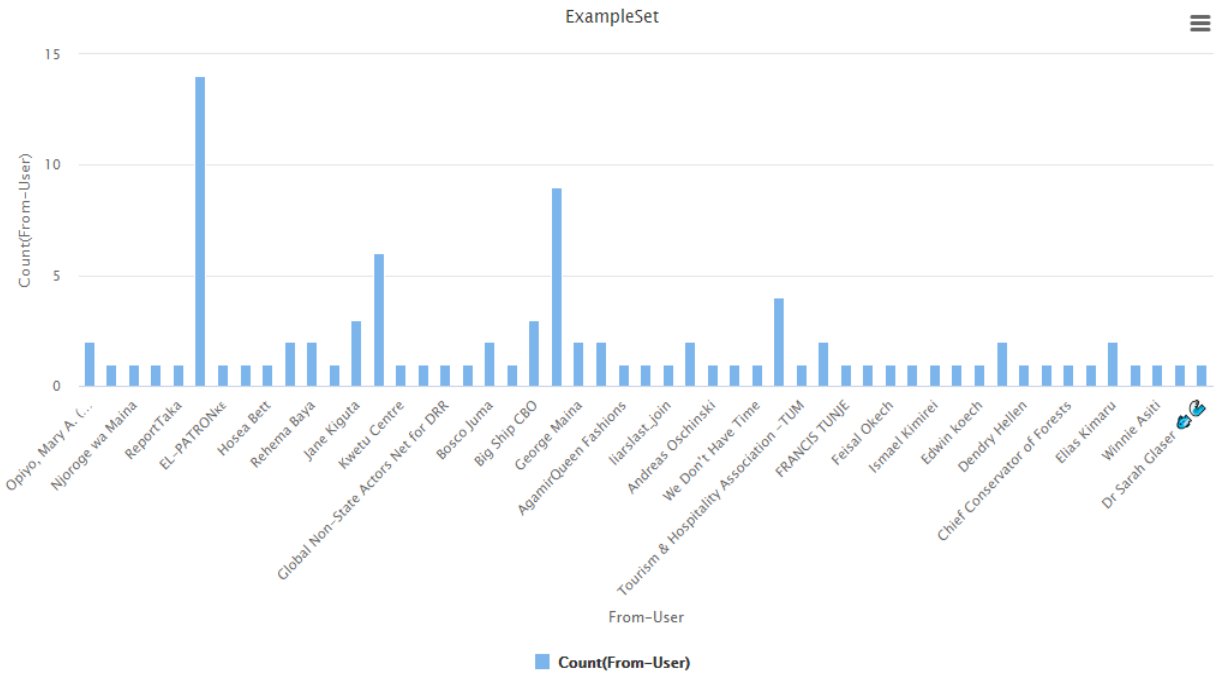

Figure 4: Top Users (Tweeters) Counts Visualization
(Source: KCDP text data)

The tweets provided valuable data and information on different topics that could be great sources of knowledge for decision-making, business intelligence, event prediction and analysis. RapidMiner Studio 9.5 as a tool was capable of reading, writing, analyzing and updating data in different formats like CSV, Excel, XML, MySQL, Access, Mails and Cloud sources like Amazon, Azure, Dropbox and Google Storage. Custom filters could be used to mine tweets examples could be retweets that are greater than a certain number, exporting tweets to a certain document like an excel document. The above twitter extension in rapid miner provided the researcher with insights in the development of the proposed model, especially in mining unstructured data.

Specific Objective 3: Implement Text Mining for Explicit Knowledge Retrieval at KCDP.
The development of the model was critically dependent on the availability of data. In the design of the text-mining model, the study narrowed down to the following text-mining techniques: information extraction; data cleansing; summarization; topic modelling. These techniques were selected because of their flexibility, compatibility and scalability in the use of text mining and topic modelling algorithms that could be used in implementing the model.

Information extraction and summarization are widely used text-mining techniques because of their flexibility and performance using different Python text summarization algorithms like Latent Semantic Analysis (LSA), LuhnSummarizer, EdmondSummarizer and LexRankSummarizer. Topic modelling algorithms/libraries like Genism and Latent Dirichlet Allocation (LDA) models were used to extract topics from text analyzed. Data was collected or mined using web scraping as a technique to get data in form of text from HTML and XML files.

The procedure involved the installation of web scraping python libraries that included Python Request 2.22 and Beautiful Soup 3. Text data was web scrapped from web pages, web links, research papers, reports, articles and policies on the publicly accessible websites and portals of the organizations that are involved in or connected with the Kenya Coastal Development Project, and saved in the *kcdp.txt* file for pre-processing and text mining in line

with developed model. Web scraping was the best method for information extraction from the web sites since it was user friendly, faster and did not require much of human intervention, only knowledge on how to install, configure and use Python web scrapping libraries for the required data extraction. Using RapidMiner studio 9.5 and its extensions, it was easier to design and operationalize the developed Text Mining Model in Figure 1, which could be used to retrieve explicit knowledge at KCDP.

Specific Objective 4: Validate the developed Text Mining model for retrieval of explicit knowledge at KCDP.

The model was tested and evaluated using the Google Colaboratory Platform ("Colab" as popularly known) which is a free Jupyter notebook environment rich in computing resources mainly RAM, GPU and storage that required no setup and run entirely in the cloud.. The figures below describe the process of evaluating the model. The process involved performing the various text mining operations in the model on the data web scrapped from the following websites:

https://www.kmfri.co.ke
http://kws.go.ke
http://www.kalro.org
https://planning.go.ke
https://www.kefri.org/ http://www.nema.go.ke/
https://cda.go.ke/
http://www.kenyaforestservice.org/
https://www.wiomsa.org/

Figure 5 shows the algorithms that was implemented to scap an dimport data from the websites.

```
[ ]  import nltk
     nltk.download('punkt')
     nltk.download('stopwords')
```

```
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Unzipping tokenizers/punkt.zip.
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Unzipping corpora/stopwords.zip.
True
```

```
from __future__ import absolute_import
from __future__ import division, print_function, unicode_literals

from sumy.parsers.html import HtmlParser
from sumy.parsers.plaintext import PlaintextParser
from sumy.nlp.tokenizers import Tokenizer
from sumy.summarizers.lsa import LsaSummarizer
from sumy.nlp.stemmers import Stemmer
from sumy.utils import get_stop_words
from sumy.summarizers.luhn import LuhnSummarizer
from sumy.summarizers.edmundson import EdmundsonSummarizer
from sumy.summarizers.lex_rank import LexRankSummarizer
```

Figure 7: Importing text for analysis
(Source: KCDP text data)

The scrapped text data were stored in two files, named *"kcdp.txt" or "speech.txt"*. The *speech.txt* file contained the 6775-word text of the speech by President Kenyatta of Kenya on State of the Nation Address in 2019 (speech-president-kenyatta-state-of-the-nation-address-2019/). The *kcdp.txt* file contained text data scrapped from above listed websites. Data

preprocessing process involved cleansing of the uploaded text where tokenization and lemmatization were performed. The preprocessing process involved the structured representation of original text to reduce dimensionality, inflectional endings and relationship of words by eliminating stop words like “a”, “and”, “the” …, and removal of punctuation marks. The elimination of the stop words and punctuation marks is what is referred to as tokenization. The other preprocessing process was lemmatization. This process involved the representation of lexically related works by the common prefix form.

Figure 5 shows the algorithm used for the data pre-processing process, which entailed cleansing of the uploaded text prior to subsequent tokenization and lemmatization. Data preprocessing involved the structured representation of original text to reduce dimensionality, inflectional endings and relationship of words by eliminating stop words like a, and the, and removal of punctuation marks. Tokenization involved the elimination of stop words and punctuation marks, while lemmatization involved the representation of words in their basic form, referred to as lemma. Thus, a different word tokens in the pre-processed text (e.g. words like “corruption”, “corrupted”, and “corrupting” being represented by their dictionary basic form or lemma, i.e. “corrupt”. Data preprocessing functionalities also provide indicators that determine the accuracy of the tokenization and lemmatization procedures.

Data Preprocessing

```
# remove punctuations, numbers and special characters
clean_sentences = pd.Series(sentences).str.replace("[^a-zA-Z]", " ")

# change to lowercase
clean_sentences = [s.lower() for s in clean_sentences]
stop_words = stopwords.words('english')
# function to remove stopwords
def remove_stopwords(sen):
    sen_new = " ".join([i for i in sen if i not in stop_words])
    return sen_new
clean_text = [remove_stopwords(r.split()) for r in clean_sentences]
len(clean_sentences)
# clean_sentences
```

```
455
```

Figure 5: Data cleansing algorithm

The data cleansing process was followed by the installation of Sumy Library, a Python library used for extracting plain text and HTML pages. The installation was achieved with the algorithm shown in Figure 5. Sumy is a text analysis library that provided the study with the ability to use different text summarization algorithms like Luhn, Latent Semantic Analysis, Edmondson and LexRank. Figure 7 shows the process of importing text into the different text summarization algorithms namely LsaSummarizer, LuhnSummarizer, EdmondSummarizer and LexRankSummarizer to perform the text summarization processes.

```
!pip install sumy

Collecting sumy
  Downloading https://files.pythonhosted.org/packages/61/20/8abf92617ec80a2ebaec8dc1646a790fc9656a4a4377ddb9f0cc90bc9
     |████████████████████████████████| 92kB 2.5MB/s
Requirement already satisfied: docopt<0.7,>=0.6.1 in /usr/local/lib/python3.6/dist-packages (from sumy) (0.6.2)
Collecting pycountry>=18.2.23
  Downloading https://files.pythonhosted.org/packages/16/b6/154fe93072051d8ce7bf197690957b6d0ac9a21d51c9a1d05bd7c6fdb
     |████████████████████████████████| 10.0MB 8.7MB/s
Requirement already satisfied: nltk>=3.0.2 in /usr/local/lib/python3.6/dist-packages (from sumy) (3.2.5)
Requirement already satisfied: requests>=2.7.0 in /usr/local/lib/python3.6/dist-packages (from sumy) (2.21.0)
Collecting breadability>=0.1.20
  Downloading https://files.pythonhosted.org/packages/ad/2d/bb6c9b381e6b6a432aa2ffa8f4afdb2204f1ff97cfcc0766a5b7683fe
Requirement already satisfied: six in /usr/local/lib/python3.6/dist-packages (from nltk>=3.0.2->sumy) (1.12.0)
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.6/dist-packages (from requests>=2.7.0->su
Requirement already satisfied: urllib3<1.25,>=1.21.1 in /usr/local/lib/python3.6/dist-packages (from requests>=2.7.0-
Requirement already satisfied: idna<2.9,>=2.5 in /usr/local/lib/python3.6/dist-packages (from requests>=2.7.0->sumy)
Requirement already satisfied: chardet<3.1.0,>=3.0.2 in /usr/local/lib/python3.6/dist-packages (from requests>=2.7.0-
Requirement already satisfied: lxml>=2.0 in /usr/local/lib/python3.6/dist-packages (from breadability>=0.1.20->sumy)
Building wheels for collected packages: pycountry, breadability
  Building wheel for pycountry (setup.py) ... done
  Created wheel for pycountry: filename=pycountry-19.8.18-py2.py3-none-any.whl size=10627360 sha256=090f8578bdb106058
  Stored in directory: /root/.cache/pip/wheels/a2/98/bf/f0fa1c6bf8cf2cbdb750d583f84be51c2cd8272460b8b36bd3
  Building wheel for breadability (setup.py) ... done
  Created wheel for breadability: filename=breadability-0.1.20-py2.py3-none-any.whl size=21682 sha256=f6389ca29a81f30
  Stored in directory: /root/.cache/pip/wheels/5a/4d/a1/510b12c5e65e0b2b3ce539b2af66da0fc57571e528924f4a52
Successfully built pycountry breadability
Installing collected packages: pycountry, breadability, sumy
```

Figure 6: Sumy library Installation

```
[ ]
nt ("--LsaSummarizer--")
marizer = LsaSummarizer()
marizer = LsaSummarizer(Stemmer(LANGUAGE))
marizer.stop_words = get_stop_words(LANGUAGE)
 sentence in summarizer(parser.document, SENTENCES_COUNT):
 print(sentence,'\n')
```

```
--LsaSummarizer--
My first term laid the foundation for a better Kenya by building on the promise and aspirations of the new Constituti

Leading the string of innovators is Roy Allela who garnered global accolades for inventing smart gloves that convert

The consideration and approval by Parliament of various Protocols, Treaties and Sessional Papers continue to enhance

This Exercise, together with the National Integrated Identity Management System (NIIMS), will ensure that all persons

His Excellency Yoweri Kaguta Museveni, the President of the Republic of Uganda and a Great Statesman and Pan-Africani

As an island of peace in a conflict-prone and fragile region, Kenya nevertheless faces challenges from transnational

Key to this is continuing to strengthen our legal tools against these groups so that they are unable to take advantag

Corruption and Impunity create social distortions and divisions, fuel inequity and poverty, destroy the fabric of soc

This they did during the National Anti-Corruption Conference held in January this year, where they tasked me, the Spe

That is why we look to the Judiciary to do their part, to apply the law firmly and fairly; and for Parliament to uphc
```

Figure 7: Text Summarization using LsaSummarizer from Sumy Library
(Source: KCDP text data)

```
[ ] print ("--LuhnSummarizer--")
    summarizer = LuhnSummarizer()
    summarizer = LuhnSummarizer(Stemmer(LANGUAGE))
    summarizer.stop_words = ("I", "am", "the", "you", "are", "me", "is", "than", "that", "this",)
    for sentence in summarizer(parser.document, SENTENCES_COUNT):
        print(sentence, '\n')
```

```
--LuhnSummarizer--
In accordance with Article 132 of the Constitution, I am honoured to report to Parliament the measures taken and prog

My first term laid the foundation for a better Kenya by building on the promise and aspirations of the new Constituti

On behalf of a grateful Nation, I thank all of those Men and Women who serve the Republic in whatever capacity, who u

Devolution has received the full and firm support of my Administration, and, together with an enabling and supportive

The consideration and approval by Parliament of various Protocols, Treaties and Sessional Papers continue to enhance

My Administration has spearheaded the implementation of various environmental initiatives including: Interventions fo

We do so conscious of the fact that fidelity to international law and commitment to our international obligations is

Kenya’s election to the AU Peace and Security Council in 2019 and our strategic decision to vie for a non-permanent s

That is why we look to the Judiciary to do their part, to apply the law firmly and fairly; and for Parliament to uphc

In saying this, I do not presume to direct the Judiciary or Parliament, that is certainly not my constitutional place
```

Figure 8: Text Summarization using Luhnsummarizer from Sumy Library
(Source: KCDP text data)

```
[ ] words3 = ("another", "and", "some", "next")
    summarizer.null_words = words3
    for sentence in summarizer(parser.document, SENTENCES_COUNT):
        print(sentence, '\n')
```

```
--EdmundsonSummarizer--
By PSCU ,

No turning back on the war against corruption as it is a just war, a war to prevent misuse of public resources for se

We are not turning back because we are determined to gift our children a better Kenya than the one we inherited.

I look forward to continued positive engagement with Parliament in the quest for a better Kenya.

The State of our Economy is STRONG!!

Indeed, the Kenya Shilling held steady against major currencies, with an annual average exchange rate of Ksh.

In the ‘World Bank Ease-of-Doing-Business Index - 2019’, Kenya’s ranking improved 19 places to position 61 globally.

Overall, our economic outlook remains positive; underpinned by the implementation of our transformative development a

We remain true to our long-term strategy, the Kenya Vision 2030.

(c) Report on the State of Security of Kenya, 2018.
```

Figure 9: Text Summarization using EdmondsonSummariser from Sumy Library
(Source: KCDP text data)

The uploaded speech.txt file from the KCDP documentation contained 6,775 words. After preprocessing and cleansing the total text was reduced to 455. The different algorithms performed well where text were analyzed into a single page and information captured in each was meaningful. Luhnsummarizer text algorithm analyzed and reduced the text to 470, Lsansummarizer to 370 and EdmondSummarizer to 149. Overall EdmondSummarizer algorithm performed the best by summarizing the text in the best way possible and capturing

the important topics of discussion in form of analysed text. Though noteworthy is that the importance of information depends on each prospective user's point of view, and also bearing in mind that society contains different views, opinions and expectations.

Specific Objective 4: Extract and visualize representing topics from text data of the KCMP.
According to Jacobi, Attevedlt & Welbers (2016), topic modelling is a statistical modelling in machine learning that creates topics on the basis of patterns, frequencies and co-occurrences of words in analyzed text. Topic modeling is mostly used in text mining for discovery of hidden semantic structures in a body of text. The development of the text mining model for the retrieval of explicit knowledge in this study incorporated topic modelling functionalities and outputs. After analysis text are grouped into words and keywords, where topic modelling was performed based on the number of times a word reappeared. Tools and algorithms from the libraries pyLDAvis and Genism libraries of Python application development platform were used for the topic modelling. This involved the identification or creation of keywords to represent the topics from the text corpus to be visualized.

Figure 11 shows the algorithm from the pyLDAvis library that was used to extract and visualize topics from the KCMP text data, which could then be displayed visually in bar charts, line graphs, histograms, or pie charts. Topic modelling was performed on the text in the *kdcp.txt* file. The text were analyzed and modelled to display the chart shown in Figure 12. The distribution of the topics shows the most important topics at the top, to the least important topic toward the bottom. The visualized results show the relative frequencies and distances among the different topics, which can then be evaluated by different KCDP stakeholders. For instance, the modelled topics revealed that the Kenya Marine and Fisheries Research Institute (KMFRI) and the county governments had been having good communicative and working relationships towards achieving the objectives of the project, and that there was more concentration of the communications on fishery research than on other areas like mining and agriculture that could also improve the livelihood of the coastal people.

```
# Visualize the topics
pyLDAvis.enable_notebook()
vis = pyLDAvis.gensim.prepare(lda_model, corpus, id2word)
vis
```

Figure 10: Topic modelling algorithm

Specific Objective 5: Evaluate the reliability of the topics extracted from the KCMP text data
Topic modelling is however not an exact science. Topic models they are approximations based on statistical probabilies, which often vary as more input text data become available. Hence, there is always need to assess how well an estimated topic model of a body of text can be relied upon as basis for prediction and decision making. Perplexity is a statistical measure of how well a probability model based on test data would predict a topic from data. According Jurafsky (2012) and Muita (2020), the lower the perplexity value of a model for any given number of extracted topics, the better its performance in predicting occurrences of words in a body of text. The formula for the perplexity coefficient is:

$$Perplexity = N\sqrt{\frac{1}{p(w_1 w_2 ... w_N)}}$$

where N represented the number of words. Another useful metric for assessing the performance of a topic model over a corpus of text is the coherence score, which is used to measure how well the topics are extracted, and is calculated as:

$$Coherence\ Score = \sum_{i<j} Score(wi,wj),$$

where *wi*, *wj* are the top words of a topic. For one topic, the words i, j being scored in $\sum_{i<j}$Score(wi,wj) formula have the highest probability of occurring for that topic. So one needs to specify how many words in a topic for the overall score for each topic, and summatively for all the extracted topics.

The lower the perplexity value the better the performance of a topic model in predicting topics to represent words in a given text corpus. Figure 14 shows the program code used to compute the perplexity coefficient and coherence score for the visualized topics in Figure 13, which provided a perplexity score of -6.0455 and a coherence score of 0.3243.

: *a*

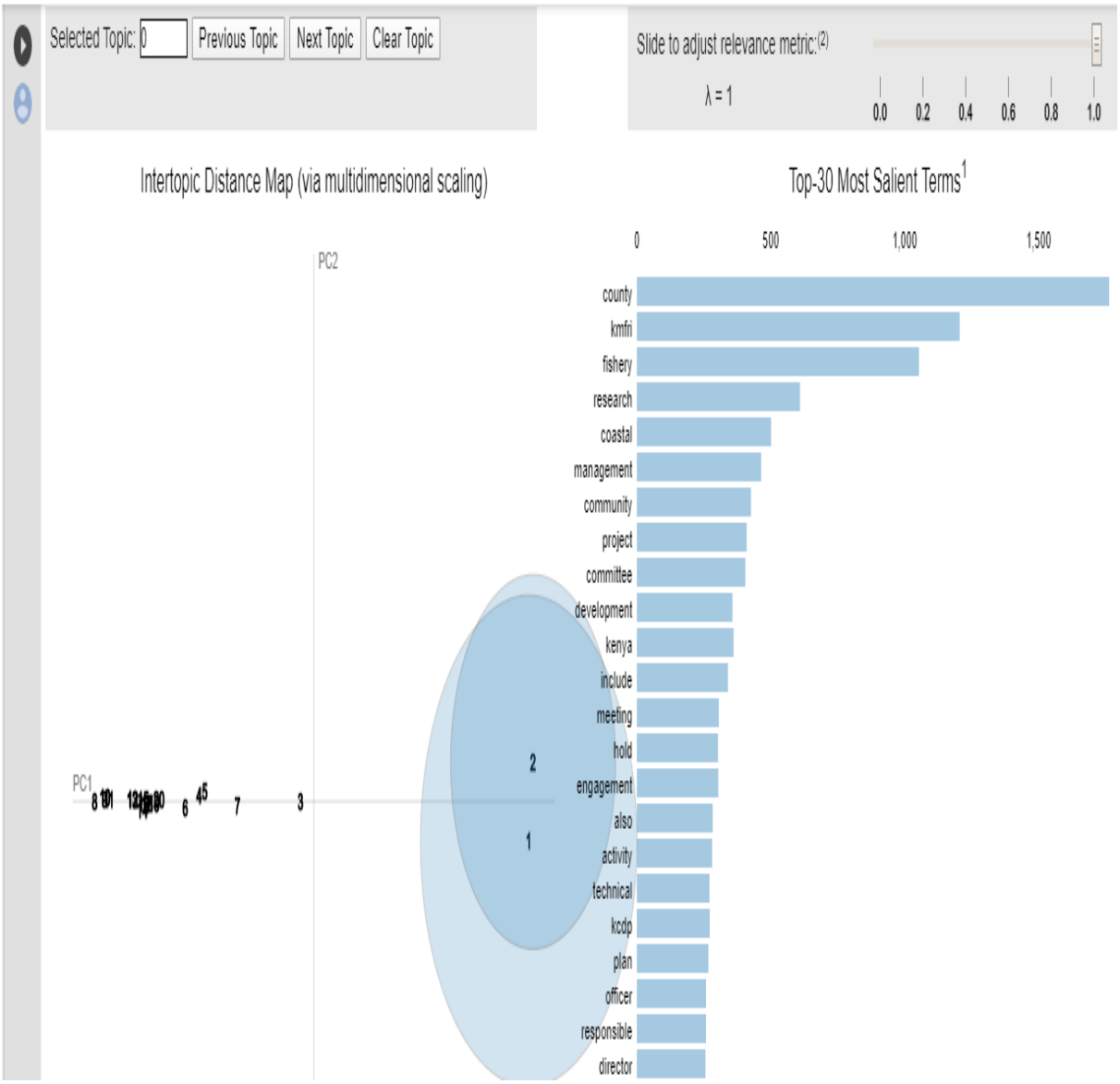

Figure 11: Statistical topic modelling and visualization
(Source: KCDP text data)

.

Conclusion

Data needs to be available and accessible to all intended users at all times, to facilitate knowledge transfer, sharing and better decision making across all parts of an organization. This study explored how Text Mining can be used in the retrieval of recorded explicit knowledge in the specific institutional context of the documented communications, and web content relevant to the activities and management of the Kenya Coastal Development Project.

A framework for how to do this was proposed and used to guide the identification, selection, testing and use of various available text techniques and algorithms in the retrieval of textual information being recorded and maintained by the project, to facilitate knowledge sharing, interpretation and use by various project stakeholders. However, data web scrapped from websites and portals of the different organizations participating in or associated with the KCDP was found out to be of variable formats and standards which was a major challenge for the text cleansing preparatory to analysis. The existing KCDP organizational policies that limited access to data from its database, e-mail, document and cloud storage platforms was also challenging, thereby forcing this study to consider resort to web scrapping using open source python tools as a data collection technique. However, because anonymous web scrapping of websites and portals of organizations is deemed illegal, the study could only test and evaluate the text mining model and text mining tools using only data available in HTML formats from the websites and portals of organizations connected with the KCDP. Despite these challenges and limitations the study demonstrated through the model that text mining could be used to retrieve and visualize explicit knowledge from both structured and unstructured text of the KCDP and similar organizations. Such text mining applications and functionalities for the retrieval of explicit knowledge should then serve as key components of required centralized Knowledge Management System for the KCDP.

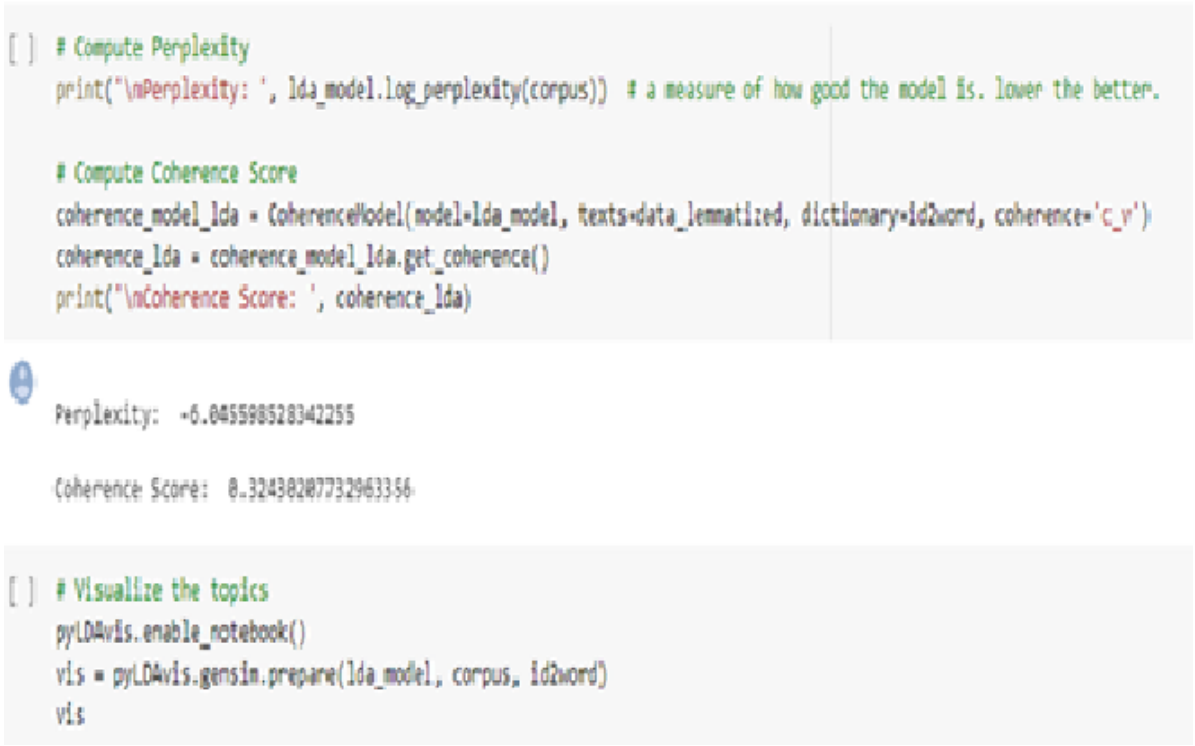

Figure 12: Computing the Perplexity and Coherence Scores

Recommendations

More research is required in evaluating text and topic mining models, algorithms and models beyond the objectives of this study. In particular, text mining models and algorithms should be proposed and tested for other languages than English, like Kiswahili and Arabic in the Kenyan and East and Southern Africa context to enable text mining for retrieval of explicit knowledge from text in these and other local languages, which might be valuable in some contexts.

References

Aich, S., Sain, M., Park, J., Choi, K. W., & Kim, H. C. (2017, November). A Text Mining approach to identify the relationship between gait-Parkinson's diseases (PD) from PD based research articles. Internation Conference on Inventive Computing and Informatics (ICICI), (pp. 481-485). New Jersey: IEEE.

Arnold, K. E. & Pistilli, M. D. (2012). Course signals at Purdue: Using learning analytics, articles. *Advances in Knowledge Discovery and Data Mining*, 12(1), 307-328.

Bianchini, M., Frasconi, P., Gori, M., & Maggini, M. (1998). Optimal learning in artificial neural networks: A theoretical view. *Neural Network Systems Techniques and Applications*, 1-51.

Conger, S. (2015). Knowledge management for information and communications technologies for development programs in South Africa. *Information Technology for Development*, 21(1), 113-134.

Hearst, M. (2003). What is text mining? SIMS, UC Berkeley. Published October 17, 2003. Retrieved June 20, 2020 from https://people.ischool.berkeley.edu/~hearst/text-mining.html

Dunham, H.M. (2003). *Data Mining: Introductory and Advanced Topics.* Prentice Hall/Pearson Education.

Israel, G.D. (1992) Determining Sample Size. University of Florida Cooperative Extension Service, Institute of Food and Agriculture Sciences, EDIS, Florida

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.

Jandhyala, S., & Phene, A. (2015). The role of intergovernmental organizations in cross-border knowledge transfer and innovation. *Administrative Science Quarterly*, 60(4), 712-743.

Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., & Zhao, L. (2019). Latent Dirichlet Allocation (LDA) and Topic modeling: models, applications, a survey. *Multimedia Tools and Applications*, *78*(11), 15169-15211.

Joia, L. A., & Lemos, B. (2010). Relevant factors for tacit knowledge transfer within organizations. *Journal of Knowledge Management*, 14(3), 410-427.

Jurafsky, D., & Manning, C. (2012). Natural language processing. *Instructor*, 212(998), 3482.

Kamimura, R. (2014). Explicit knowledge extraction in information-theoretic supervised multi-layered SOM. In *Foundations of Computational Intelligence (FOCI),* 2014 IEEE Symposium on (pp. 78-83).New York City IEEE.

Khan, S., Rani, U., Prasad, B. V. N., Srivastava, A. K., Selvi, S., & Gautam, D. K. (2015, March). Document management system: An explicit knowledge management system. In *Computing for Sustainable Global Development (INDIACom),* 2015 2nd International Conference on (pp. 402-405). New York, IEEE.

Kobayashi, V. B., Mol, S. T., Berkers, H. A., Kismihók, G., & Den Hartog, D. N. (2018). Text mining in organizational research. *Organizational Research Methods*, 21(3), 733-765.

Muita, S. (2020). *A Model for processing public participation feedback using topic Modeling* (Doctoral dissertation, University of Nairobi).

Ngai, E. W., Xiu, L., & Chau, D. C. (2009). Application of data mining techniques in customer relationship management: A literature review and classification. *Expert Systems with Applications*, 36(2), 2592-2602.

Oates, B. J. (2005). *Researching information systems and computing*. New York: Sage.

Preece, J., Rogers, Y., & Sharp, H. (2002) *Interaction design: Beyond Human-Computer Interaction.* New York: John Wiley & Sons, Inc.

Rumanti, A. A., Samadhi, T. A., & Wiratmadja, I. I. (2016, December). Impact of tacit and explicit knowledge on knowledge sharing at Indonesian Small and Medium Enterprise. In *Industrial Engineering and Engineering Management (IEEM),* 2016 IEEE International Conference (pp. 11-15). New York: IEEE.

Shah, M., Shinde, S., Sawant, R. S., & Wagh, P. (2017). Analysis of Text Review using Hybrid Classifier. *International Journal of Engineering Science*, Vol. 10914.

Tan, A. H. (1999, April). Text mining: The state of the art and the challenges. In *Proceedings of the PAKDD 1999 Workshop on Knowledge Discovery from Advanced Databases* (Vol. 8, (April), 65-70).

Ur-Rahman, N. & Harding, J. A. (2012). Textual data mining for industrial knowledge management and text classification: A business-oriented approach. *Expert Systems with Applications*, 39(5), 4729-4739.

Wilensky, H. L. (2015). *Organizational intelligence: Knowledge and policy in government and industry.* Quid Pro Books. 212p. ISBN: 978-1610272870

Williamson, K. & Johanson, G. (Eds.). (2017). *Research Methods: Information, Systems, and Contexts*. Sawston, Cambridge: Chandos Publishing.

Yu, C. H., Jannasch-Pennell, A. & DiGangi, S. (2011). Compatibility between text mining and qualitative research in the perspectives of grounded theory, content analysis, and reliability. *The Qualitative Report*, 16(3), 730-744.

Profiles of the Authors

Ednah Nyakerario ONKUNDI is an Information Technology professional specializing in databases and information security. She works at Kenya Marine and Fisheries Research Institute as a Senior ICT Officer. She holds a bachelor's degree in Business and Information Technology and currently completing a master's degree in Information Technology at Mount Kenya University, Kenya.

Raymond Wafula ONGUS holds doctorate degree and is Associate Professor at Mount Kenya University, Thika Main Campus Kenya, where he teaches in the School of Computing and Informatics. He had worked for four years at Mount Kenya University, Kigali Campus, Rwanda where he once served as Deputy Vice Chancellor-designate. He was also once Senior Assistant Librarian at Egerton University in Kenya, where he was in charge of the J. D. Rockefeller Research Library and its computer network. He has a PhD in Library and Information Science, M.Sc. In Information Science, and B.Ed. (Science) specialized in pure mathematics and statistics.

Constantine Matoke NYAMBOGA holds a PhD and is currently Vice Chancellor of Lukenya University, Mtito Andei Campus, Kenya. He was previously Associate Professor at Mount Kenya University, Kenya where he taught in the School of Computing and Informatics. He also worked for three years at Kisii University where he founded the Faculty of Information Science and Technology. He posseses Bachelor's, Master's and doctoral degrees in Library and Information Science, and had had lengthy experience in the field of Library and Information science at Egerton University in Kenya where he had supervised numerous postgraduate students.