International Journal of Advances in Scientific Research and Engineering (ijasre)

E-ISSN : 2454-8006

DOI: 10.31695/IJASRE.2022.8.11.8

Volume 8, Issue 11
November - 2022

Development of an Internet-Based Information/Data Resource Platform during the COVID-19 Pandemic Using Hierarchical Clustering

Achudume Nkechi Evelyn[1], Olumide Owolabi[2], Ebelogu Christopher Ubaka[*3], Amujo Oluyemi Enoch[4]

[1-4]Department of Computer Science

University of Abuja, FCT-Abuja

Nigeria.

ABSTRACT

There is heavy usage of Internet-based information resource platforms due to the COVID-19 pandemic. If it can present the search results as a list of groups of related terms and context, that would be ideal. An additional enhancement would be to display information in the search results in a classified manner. Therefore, the research is proposing an effective and efficient information retrieval system. The methodology adopted follows the conventional Information Retrieval (IR) system design except that Cosine similarity and hierarchical clustering were introduced for search result ranking. The Average Precision of the proposed system was compared with that of flat clustering such as K-means. The result of the evaluation shows that hierarchical clustering improved in precision, being10% better than flat clustering result retrieval result. This led to greater usability, user experience, and efficiency of the system.

Key Words: COVID-19, Hierarchical Clustering, Information Retrieval.

1. INTRODUCTION

The rapid and extensive development of information technologies has transformed contemporary industrialized societies into a network of societies (called the Global Village. Learning and Speech Technology [2] are no exception to this megatrend. ICT applications in the printing and publishing process have been meaningful additions to conventional procedures. [8]. [4] think that in this modern age of technology, it can be presumed that all students have access to electronic resources “due to the simple fact that most students can afford the internet”.

In the early 1800s, a shift began in the scientific community and by 1930 this trend had gained much popularity. Quantity overpowered quality. In the 1960s, the Science Citation Index was created that measured not just how many papers an author published, but how often those papers were cited by others [7].

Searching platforms have been accepted as a major tool used for information and educational purpose as they support the vision and mission of indexing and cataloging. There is a heavy demand on these platforms because the number of searchers keeps increasing in number and currently, COVID-19 has caused people to intensify their search for a cure, vaccines, and medical palliatives. Therefore, rendering the search results as a list of groups of related terms and contexts would be ideal. An additional enhancement would be to display information in the search results in a classified manner.

This work is aimed at proposing an effective and efficient way of organizing documents in an information retrieval system to facilitate rapid retrievals. The Coronavirus, code-named COVID-19, is a pandemic that has affected the whole world. Information searching and inquiry is on the high rise for solutions in terms of vaccines, cure, and palliatives. It is popularly agreed that the internet makes information more accessible to people than conventional libraries.

The study focuses on the hierarchical clustering of documents within an information-searching system. It emphasizes the fact that similar documents, in terms of context and content, may be grouped into clusters. A document can be in more than one cluster depending on the relevance of the contents. Document clustering provides two major benefits; efficiency and quality usability. If clustering is performed before searching, it provides efficiency and makes searching very easy while it provides quality usability when it is performed after searching, also as it is grouped similar results give ease of accessibility

2. RELATED WORKS

The first case of coronavirus was notified as cold in 1960 and by 2001, approximately 500patients were identified as having Flu-like systems. Corona was treated as a simpleton fatal virus till 2002. Several cases of severe acute respiratory syndrome caused by corona and more than 1000 patients were reported in 2003. This was the black year for microbiologists. For them to conclude and understand the pathogenesis of the disease after a deep exercise they discovered it was coronavirus but a total of 8,096 patients were confirmed as infected with the coronavirus. A “state of emergency” was declared in 2004 by the World health organization and centers for dis- ease control and prevention. A different discovery study report in Hong Kong confirmed 50 patients with severe acute respiratory syndrome while 30 of them were confirmed as coronavirus infected. COVID-19 was first identified and isolated from pneumonia patent in Wuhan, china [11].

Among them are severe acute respiratory syndrome coronavirus (SARS-CoV) reported in November 2002 [19] and Middle East respiratory syndrome coronavirus (MERS-CoV) reported in September 2012, which emerged in human population from animal reservoirs and caused severe respiratory illness with high mortality rates [18]. Precautionary measures to contain the spread of this virus are being practiced throughout the globe; which include social distancing, isolation and quarantine, frequent washing of hands, national lockdowns, and travel restrictions.

On the 18th of April 2020 by 10:00 am CEST; Africa CDC reported, 19,895 confirmed cases, including 1,017 deaths and 4,642 recoveries, from 52 African countries, while two countries (Comoros and Lesotho) were still virus-free (CDC, 2020). Interestingly, most of the identified cases of COVID-19 in Africa have been imported from Europe and the United States, rather than from the original COVID-19 epic-center China [17]. When COVID-19 hit the continent of Africa, the economy was already struggling, which could further amplify the economic crisis. Africa is a unique continent so a unique COVID-19 response needs to be developed for Africa, where all these issues which make the continent more vulnerable and different from the rest of the world, will be taken into consideration. Although, past experiences show that there is a scope for suppressing transmission of COVID-19, provided governments and the public will change their behavior towards this virus as they did previously for Ebola, HIV, Polio, and other outbreaks. However, Africa needs global support because it cannot confront this alone to step ahead of this pandemic [17].

Most importantly, human coronaviruses targeting vaccines and antiviral drugs should be de- signed that could be used against the current as well as future epidemics. Many companies are working on the development of effective SARS-CoV-2 vaccines, such AsastraZenica which needs a third shot called booster [16], Modena Therapeutics, Inovio Pharmaceuticals, Novavax, Vir Biotechnology, Stermirna Therapeutics, Johnson & Johnson, VIDO-InterVac, GeoVax-BravoVax, Clover Biopharmaceuticals, CureVac, and Codagenix. Since these vaccines still require 3–10 months for commercialization there is a need for rapid human and animal-based trials. They should ban people from using wild animals and birds as a source of food. A strategy to rapidly diagnose SARS-CoV-2 in a suspected patient has also required aside the development of the most efficient drug. Therefore, as PCR-based testing is expensive and time-consuming, an accurate and rapid diagnostic kit or meter for the detection of SARS-CoV-2 in suspected patients is required. It is appreciable that the Chinese health workers have efficiently controlled the outbreak in china and limited the mortality rate to less than3% only. The therapeutic strategies used by Chinese healthcare authorities should also be followed by other countries [12].

According to [13] shows preparing the next generation of African Healthcare Workers and Scientists: In this work, we have developed a search ranked system based on hierarchical clustering and shown that the techniques can improve system usability and efficiency. Depending on the objective, clustering can be performed either before the search to cluster similar documents or after the search to cluster similar results.

Several researchers [1] have done the improvement of the efficiency of information retrieval systems using the clustering technique with a focus on document categorization within the corpus, before the search, based on their similarity. However, little has been done on improving the usability of the system by categorizing the results of the retrieved information after a search.

This research improved the usability of the information retrieval system using hierarchical clustering techniques. The result of information retrieval is categorized based on the similarity derived by the cosine similarity algorithm. The implication is that once the result is categorized using the proposed technique, they are displayed and ranked according to their relevance. The proposed work is validated by comparing with flat clustering and human-judged rank documents, and the result shows that the proposed solution outperformed flat clustering by 10%.

Table 1. Computation of Average Precision of hierarchical clustering and K-means clustering ranks of document retrieval results

Expected Rank	Hierarchical				K-Means			
	Accurately ranked?	C (i)	i-1	C(i) / i-1	Accurately ranked?	C (i)	i-1	C(i) / i-1
1	1	0	0	1.00000	1	0	0	1.00000
2	0	1	1	1.00000	0	1	1	1.00000
3	1	1	2	0.50000	1	1	2	0.50000
4	1	2	3	0.66667	1	2	3	0.66667
5	1	3	4	0.75000	0	3	4	0.75000
6	1	4	5	0.80000	1	3	5	0.60000
7	1	5	6	0.83333	1	4	6	0.66667
8	1	6	7	0.85714	1	5	7	0.71429
9	1	7	8	0.87500	1	6	8	0.75000
10	1	8	9	0.88889	1	7	9	0.77778
11	1	9	10	0.90000	1	8	10	0.80000
12	1	10	11	0.90909	0	9	11	0.81818
13	1	11	12	0.91667	1	9	12	0.75000
14	1	12	13	0.92308	1	10	13	0.76923
15	0	13	14	0.92857	0	11	14	0.78571
16	1	13	15	0.86667	1	11	15	0.73333
17	1	14	16	0.87500	1	12	16	0.75000
18	1	15	17	0.88235	1	13	17	0.76471
19	1	16	18	0.88889	1	14	18	0.77778
20	1	17	19	0.89474	1	15	19	0.78947
21	1	18	20	0.90000	1	16	20	0.80000
22	1	19	21	0.90476	1	17	21	0.80952
Summation (S)				17.96084				15.77334
p' = 0.047619 * S				0.85528				0.75111

3. METHODOLOGY

The cluster hypothesis offers an alternative to latent semantic indexing which is slow. Since there are many fewer clusters than documents, finding the closest cluster is fast; and since the documents thatching a query are all similar to each other, they tend to be in the same clusters. While this algorithm is not exact, the expected decrease in search quality is small. [10].

The approach follows conventional Information retrieval (IR) system design methodology except that we introduced a similarity matrix and hierarchical clustering at the result ranking stage which makes it differ from the conventional IR framework. [9].

The similarity matrix is a collection of all the inter-document similarity values for all the journal articles involved in the study. In our case, we shall have a 100 by 100 matrix.

To achieve this, we take the term-document matrix to document–document matrix approach, thus;

i. We apply the proximity matrix to all pairs of documents

ii. Creates the document-document matrix, which reports similarities/distances be- tween objects (documents)

iii. The diagonal is trivial (identity)

iv. As proximity measures are symmetric, the matrix is a triangle.

Thus;

Given: a set $of documents D = d_1 \ldots . d_n$of objects;

The goal is to compute the similarity function $sim: (D) \times (D)! R$

Hierarchical Clustering:

A complete hierarchical clustering algorithm by [14] shall be adopted to cluster the matrix. Assuming a set of N items to be clustered, and an N by N distance (or similarity) matrix, the basic process of hierarchical clustering goes thus:

i. Start by assigning each item to its cluster, so that if there exist N items, you now have N clusters, each containing just one item. Assume the distance or similarities of the clusters equal the distance or similarities between the items.

ii. Find the closest (most similar) pair of clusters and merge them into a single cluster, so that now you have one less cluster.

iii. Compute distances (similarities) between the new cluster and each of the old clusters.

iv. Repeat steps 2 and 3 until all items are clustered into a single cluster of size N.

Since we chose single-link clustering in this dissertation, we considered the distance between one cluster and another cluster to be equal to the shortest distance. If the data consist of similar clusters, we consider the similarity between one cluster and another cluster to be equal to the greatest similarity from any member of one cluster to any member of the other cluster. Step (iii) may be represented thus:

Single-link:

To obtain the initial partition:
(a) single-link:
For each couple of documents i and j such that d(i,j) < threshold :
- if i and j are not yet in a class, create a new one;
- if i and/or j are already allocated, merge all the documents of the class containing i (resp. j) with those of the class containing j (resp. i);

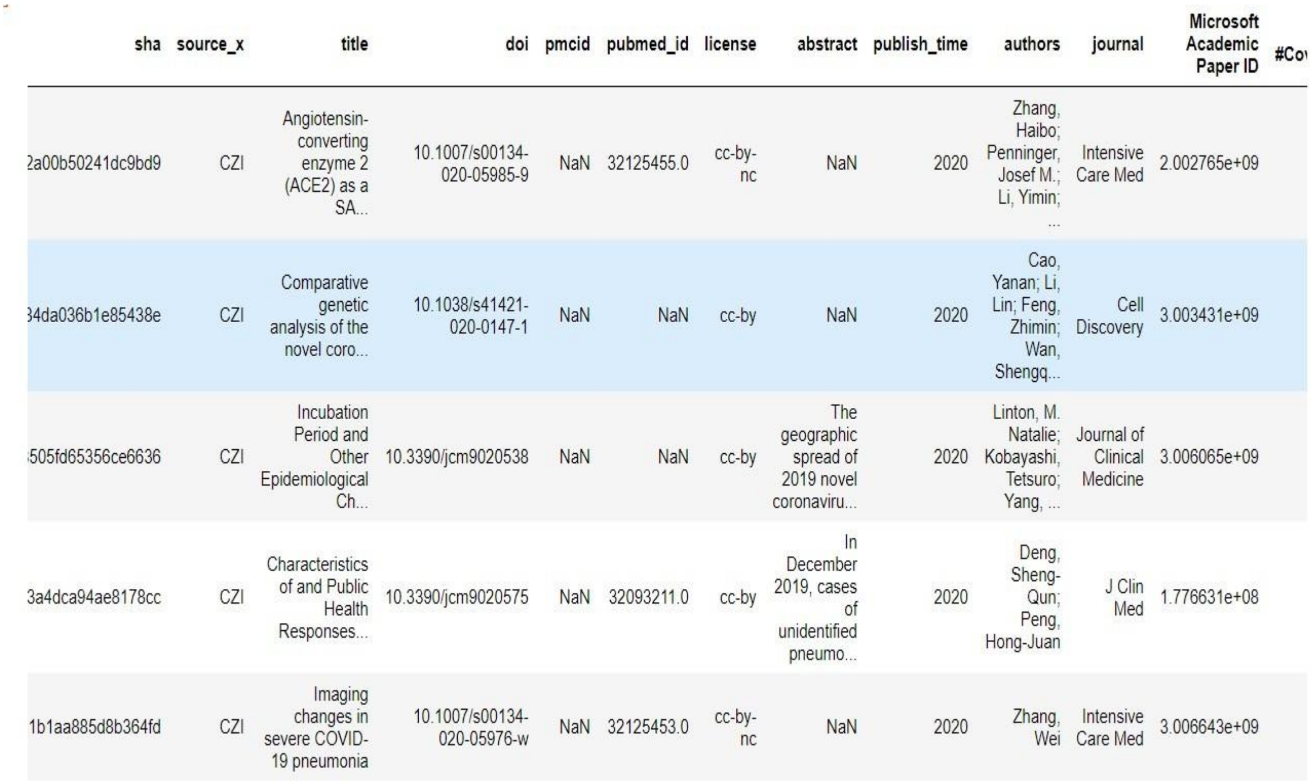
(b) partial hierarchical classification:
After this step, the number of classes may be greater than the number of clusters wanted.

So, as long as the number of classes is greater than the predefined one, we can:
- compute class representatives;
- compute distances between every pair of classes (triangular matrix);
- merge the two closest classes. [14]

The threshold value is chosen so that the number of documents assigned at the end of the step is greater than half the total number of documents.

4. RESULTS AND DISCUSSION

This involves the practical demonstration of the solution of document hierarchical clustering. The system development and implementation are done using Python Programming. Python is chosen because of its versatility in a scripting language such as machine learning. In addition, Jupyter Lab - a Python coding environment was used within the Anaconda framework:



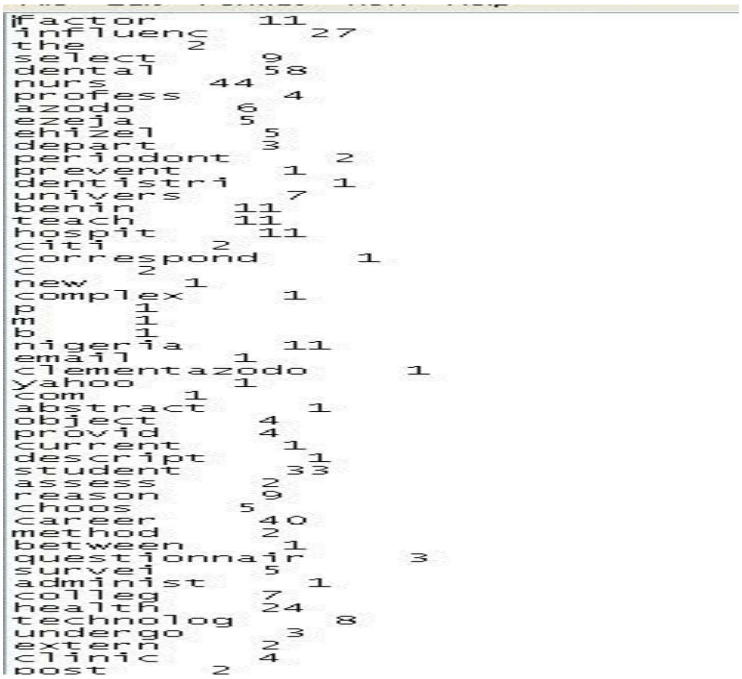
sha	source_x	title	doi	pmcid	pubmed_id	license	abstract	publish_time	authors	journal	Microsoft Academic Paper ID	#Co
2a00b50241dc9bd9	CZI	Angiotensin-converting enzyme 2 (ACE2) as a SA...	10.1007/s00134-020-05985-9	NaN	32125455.0	cc-by-nc	NaN	2020	Zhang, Haibo; Penninger, Josef M.; Li, Yimin; ...	Intensive Care Med	2.002765e+09	
34da036b1e85438e	CZI	Comparative genetic analysis of the novel coro...	10.1038/s41421-020-0147-1	NaN	NaN	cc-by	NaN	2020	Cao, Yanan; Li, Lin; Feng, Zhimin; Wan, Shengq...	Cell Discovery	3.003431e+09	
505fd65356ce6636	CZI	Incubation Period and Other Epidemiological Ch...	10.3390/jcm9020538	NaN	NaN	cc-by	The geographic spread of 2019 novel coronaviru...	2020	Linton, M. Natalie; Kobayashi, Tetsuro; Yang, ...	Journal of Clinical Medicine	3.006065e+09	
3a4dca94ae8178cc	CZI	Characteristics of and Public Health Responses...	10.3390/jcm9020575	NaN	32093211.0	cc-by	In December 2019, cases of unidentified pneumo...	2020	Deng, Sheng-Qun; Peng, Hong-Juan	J Clin Med	1.776631e+08	
1b1aa885d8b364fd	CZI	Imaging changes in severe COVID-19 pneumonia	10.1007/s00134-020-05976-w	NaN	32125453.0	cc-by-nc	NaN	2020	Zhang, Wei	Intensive Care Med	3.006643e+09	

Figure 1:1 Data-frame presentation of the dataset

Table 1. Comparison of 10 recorders of expected, hierarchical clustering and K-means clustering ranks of document retrieval results

Rank	Expected Ranked Result (L0)	Hierarchical-based Ranked Result (L1)	K-Means Ranked Result (L2)
1	The Human Sodium Iodide Symporter as a Reporter Gene for Studying Middle East Respiratory Syndrome Coronavirus Pathogenesis	The Human Sodium Iodide Symporter as a Report	The Human Sodium Iodide Symporter as a Report...
2	Pathogenic characteristics of persistent feline enteric coronavirus infection in cats	Programs might representa cost-efficient oppo...	Programs might represent a cost-efficient oppo...
3	Access to AIDS medicines stumbles on trade rules.	Access to	Access to
4	Rapid Genome Sequencing of RNA Viruses	Rapid Genome Sequencing of RNA Viruses	Rapid Genome Sequencing of RNA Viruses
5	Structural basis for viral 5â€²-PPP-RNA recognition by human IFIT proteins	Structural basis for viral5′-PPP-RNA recognition.	Value of Pharmacy-Based Influenza Surveil- lance â€” Ontario, Canada, 2009
6	Stability of Middle East Respiratory Syndrome Coronavirus in Milk	Stability of Middle East Respiratory Syndrome ...	Stability of Middle East Respiratory Syndrome ...
7	Structural basis for human coronavirus attachment to sialic acid receptors	Structural basis for human coronavirus attachment	Structural basis for human coronavirus attachment
8	Common and unique features of viral RNA-dependent polymerases	Common and unique features of viral RNA-depend...	Common and unique features of viral RNA-depend...
9	Innate immune responses and neuroepithelial degeneration and regeneration in the mouse olfactory mucosa induced by intranasal ad-ministration of Poly(I:C)	Innate immune responses and neuroepithelial de...	Innate immune responses and neuroepithelial de...
10	The Axl receptor tyrosine kinaseis a discriminator of macrophage function in the inflamed lung	The Axl receptor tyrosine kinase is a discrimi...	The Axl receptor tyrosine kinase is a discrimi...

DOI: 10.31695/IJASRE.2022.8.11.8



```
factor          11
influenc                27
the     2
select          9
dental          58
nurs    44
profess         4
azodo   6
ezeja   5
ehizel          5
depart          3
periodont               2
prevent         1
dentistri               1
univers         7
benin   11
teach   11
hospit          11
citi    2
correspond              1
c       2
new     1
complex         1
p       1
m       1
b       1
nigeria         11
email   1
clementazodo            1
yahoo   1
com     1
abstract                1
object          4
provid          4
current         1
descript                1
student         33
assess          2
reason          9
choos   5
career          40
method          2
between         1
questionnair            3
survei          5
administ                1
colleg          7
health          24
technolog               8
undergo         3
extern          2
clinic          4
post    2
```

Figure 1.2: Result of Porter stemming algorithm [5]

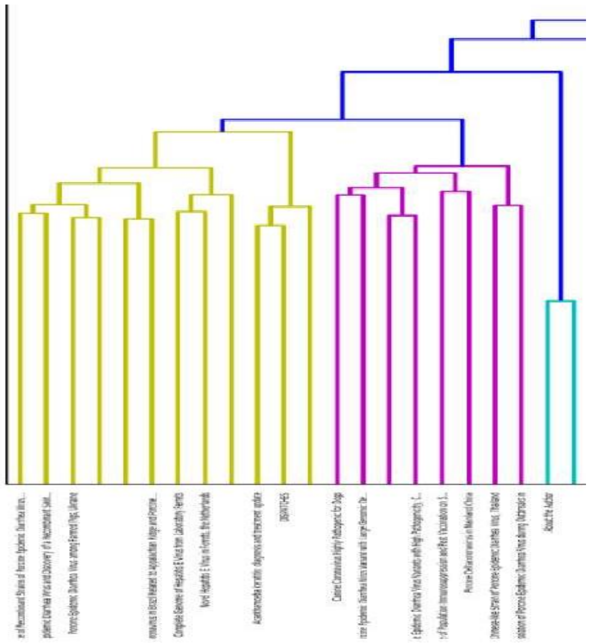


Figure 1.3: Showing top clusters of the 200-articleDendrogram

The idea is that since documents that are retrieved at the top of a list are more important than the documents towards the bottom, average precision assigns more weight to the errors made toward the top of a ranking than the errors towards the bottom. Therefore, the structure of hierarchical plays an advantageous role in such a way that there is a high probability that the closest document is ranked next to a document on the retrieval result list.

5. CONCLUSION

This study sought to propagate a semantic web – web of data by presenting a dendrogram (taxonomy) for 200 articles on COVID-19. On design and implementation of the study, 1,426 articles were downloaded but 200 were serialized, and stemmed, an average of fifteen terms were selected from each journal article. Based on our objectives, an information retrieval system was developed which defers the traditional IR by applying the cosine similarity function to determine the similarity between documents as well as adopting hierarchical clustering for retrieved result ranking. The bigger picture is that people tend to be more concerned about the

DOI: 10.31695/IJASRE.2022.8.11.8

retrieved documents and the relationship between successive documents on the result list.

REFERENCES

1. Amer A. A.,&Abdalla I. A. (2020). A set theory-based similarity measure for text clustering and classification. Journal of Big Data.Volume 7:74. https://doi.org/10.1186/s40537- 020-00344-3
2. Neustein, 2021 International Journal of Speech Technology https://www.springer.com/journal/10772
3. Anna O., Sandro G., David S., &Carlo S. (2020). The first 10 000 COVID-19 papers in perspective: Are we publishing what we should be publishing? European Journal of Public Health, Volume 30, Issue 5, Pages 849–850, https://doi.org/10.1093/eurpub/ckaa170
4. Ariffin and Bakar (2013). Electronic resources used by distance learners at the University of Namibia By MeamenoUtunaNampaHamutumwa.https://researchspace.ukzn.ac.za/xmlui/bitstream/handle/10413/12148/Hamutumwa_Me ameno_Utuna_Nampa_2014.pdf?sequence=1&isAllowed=y
5. Christopher D. M., Prabhakar R. & Hinrich S. (2009). An Introduction to Information Retrieval. Cambridge University Press Cambridge, England
6. Enago Academy (2020). A Brief History of the Catalogue of Scientific Papers. https://www.enago.com/academy/a-brief-history-of-the-catalogue-of-scientific-papers/
7. Holly E. (2020). How a torrent of COVID science changed research publishing — in seven charts. Nature 588, 553 (2020). DOI: https://doi.org/10.1038/d41586-020-03564-y
8. Ihebuzor, L. (2013). Book Publishing in Nigeria: Theories and Issues. Ibadan: College Press and Publishers Limited.
9. Jennifer. P, A. Muthukumaravel (2019) International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8 Issue-1S2, May 2019 https://www.singaporeanjbem.com/pdfs/SG_VOL_4_(12)/3.pdf https://www.ijrte.org/wp-content/uploads/papers/v8i1S2/A00650581S219.pdf
10. Kehinde Agbele (2018), Information Retrieval & Storage https://scholar.google.com/citations?view_op=search_authors&hl=en&mauthors=label:information_retrieval_%26_storage
11. Kumar, D., Malviya, R., Kumar., Sharma, P. (2020). Corona Virus: A Review of COVID-19. Eurasian Journal of Medicine and Oncology (EJMO) 2020;4(1):8–25.
12. Muhammad, A.S., Suliman, Khan, A.K., Nadia, B., Rabeea, S., (2020). COVID-19 infection: Emergence, transmission, and characteristics of human coronaviruses. Journal of Advanced Research. Volume 24. Pp. 91-98. https://doi.org/10.1016/j.jare.2020.03.005.
13. Oraebosi M.I., Chia T., Oyeniran O.I. (2020). https://www.sciencedirect.com/science/article/abs/pii/S2352552520300736
14. Patrice Bellot and Marc El-Bèze (2000). Clustering using Unsupervised Decision Trees or Hierarchical and K-means-like Algorithm: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.40.7762&rep=rep1&type=pdf
15. Porter M.F. (2006). An algorithm for suffix stripping. Program: Electronic Library and Information Systems
16. Rebecca Robbins (2021). A third dose of the AstraZeneca vaccine is found to boost immune response. https://www.nytimes.com/2021/06/28/world/astrazeneca-vaccine-booster-shot.html
17. Ruth, M. (2020). Africa braces for coronavirus, but slowly. The New York Times: Available from: https://www.nytimes.com/2020/03/17/world/africa/coronavirus-africa-burkina- faso.html
18. Shabir, A.L. &Aijaz, A. (2020). COVID-19 pandemic – an African perspective. Emerging Microbes & Infections. Volume 9, 2020 - Issue 1. pp.1300-1308. DOI:https://doi.org/10.1080/22221751.2020.1775132
19. Zhong, N.S., Zheng, B.J., Li, Y.M., (2003). Epidemiology and cause of severe acute respiratory syndrome (SARS) in Guangdong, People's Republic of China, in February. Lancet. 2003;362:1353–1358. doi:10.1016/s0140-6736(03)14630-2

DOI: 10.31695/IJASRE.2022.8.11.8