Vol. 07, No. 01, January 2023, pp. 31-40

INTERNATIONAL JOURNAL OF INNOVATION IN ENTERPRISE SYSTEM

Journal homepage: https://ijies.sie.telkomuniversity.ac.id
e-ISSN: 2580-3050, DOI: https://doi.org/10.25124/ijies.v7i01.187

Forming Dataset of The Undergraduate Thesis using Simple Clustering Methods

Tio Dharmawan[1*], Chinta 'Aliyyah Candramaya[2], Vandha Pradwiyasma Widharta[3]

[1] University of Jember, Jember, Indonesia
tio.pssi@unej.ac.id

[2] University of Jember, Jember, Indonesia
chintaalya661@gmail.com

[3] Pukyong National University, Busan, Republic of Korea
vandhapw@pukyong.ac.kr

[*]tio.pssi@unej.ac.id

ARTICLE INFO

Article history:
Received 26 November 2022
Accepted 27 January 2023
Published 31 january 2023

ABSTRACT

Each university collects many undergraduate theses data but has yet to process it to make it easier for students to find references as desired. This study aims to classify and compare the grouping of documents using expert and simple clustering methods. Experts have done ground truth using OR Boolean Retrieval and keyword generation. The best cluster was discovered by the experiments using the K-Means, K-Medoids, and DBSCAN clustering methods and using Euclidean, Manhattan, City Block, and Cosine Similarity metrics. The cluster with the best Silhouette Score compared to the accuracy of the categorization of each document. The K-Means clustering method and the Cosine Similarity metric gave the best results with a Silhouette Score value of 0.105534. The comparison between ground truth and the best cluster results shows an accuracy of 33.42%. The result shows that the simple clustering method cannot handle data with Negative Skewness and Leptokurtic Kurtosis.

Keywords:
Document Clustering; Text Mining; Relevant Term; Information Retrieval; Topic Identification

1. INTRODUCTION

The students from the undergraduate program need to do research as a graduation condition. The research article has many topics based on the interest of students and supervisors. The University of Jember uses an information system to manage the data, so the data was collected well [1]. The data can be processed to give the insight to support the student doing their research. Currently, the data is categorized based on the student's faculty. The data categorization based on the topics is needed to make the student easier to find the related references.

Document classification can be a solution to categorize documents [2][3]. However, the data has not been labeled. Forming the dataset is needed for the classification model. It can be achieved by applying the information retrieval method to retrieve data. There are many information retrievals methods, such as Boolean Retrieval [4], [5], Jaccard [6], Sorensen Dice Index [7], and Cosine Similarity [8]. Information retrieval methods need any keyword to retrieve the data. The keywords can be generated with an expert's assistance or using the keyword generator algorithm [9].

Moreover, the clustering method can be used for [10][11] labeling the data. The simple clustering methods are K-Means [12], K-Medoids [13], and Density-Based Spatial Clustering of Applications with Noise (DBSCAN) [14]. The clustered data were then labeled to form the dataset. The clustering method depends on the distance between two documents. The distance metrics used [15][16] are Euclidean, Manhattan, City Block, and Cosine Similarity. The experiment needed to get the best result. K-Means and K-Medoids results will differ based on the initial centroid and the number of clusters. In contrast, the DBSCAN result is affected by the radius and the number of neighbors. This method is the primary method of clustering and has a unique approach to solving the clustering problem.

The performance of clustering is measured using a silhouette score as an evaluation to get the best clustering result. Silhouette score [17] calculates the difference between intra-cluster and inter-cluster distance. It shows how convergent the formed cluster is, but it cannot prove the validity of the cluster as needed. The challenge in forming a dataset using the clustering method is how to cluster the data considering the actual need. Therefore, validating the clustering result with expert reasoning is needed.

The study is conducted to make a dataset based on the clustering result. This study tries to solve the validity problem between the clustering result and the actual need. The simple clustering methods used are K-Means, K-Medoids, and DBSCAN. The expert's assistance is needed to evaluate the validity between the clustering result and the actual needs.

2. METHOD

The data used in this research is the title of the research article in the department of computer science, University of Jember. The title is in the Indonesian language. The data consist of 1.062 titles of the research article conducted between 2009 and 2022. The number of terms occurred after the documents were preprocessed using stopword removal and stemming using Sastrawi for the Indonesian language. The document is discretized using Term Frequency Inverse Document Frequency (TF-IDF) term weighting. The minimum and maximum terms in documents are 5 and 28, as shown in Figure 1. The mean terms in all documents are 14 terms.

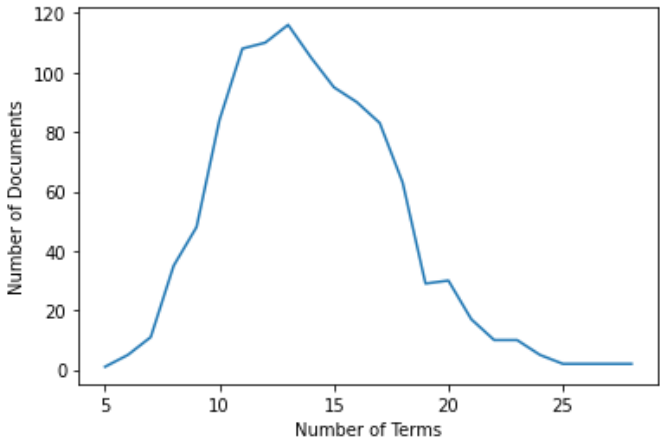

Figure 1 - Distribution of Documents by Terms

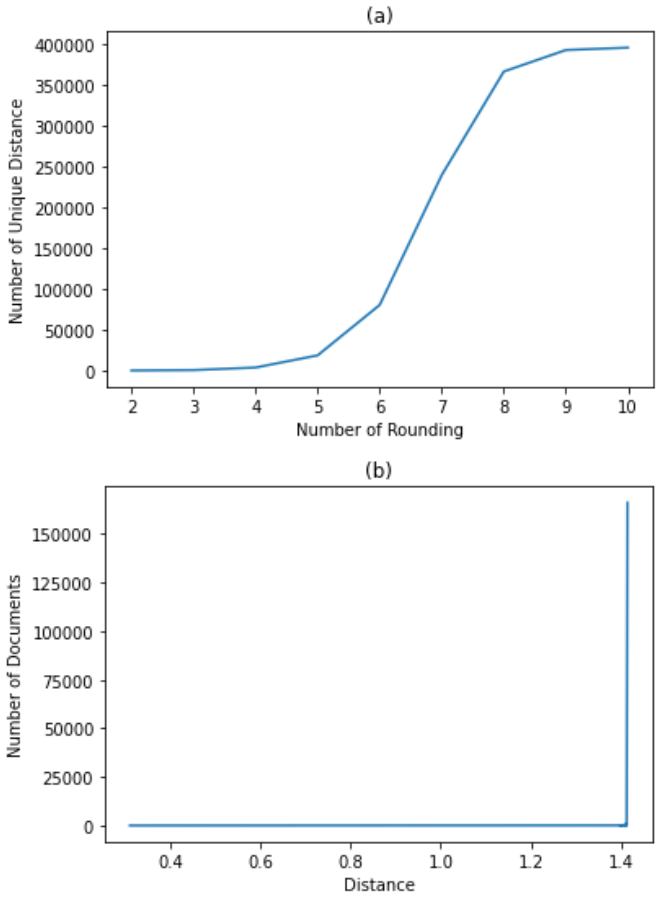

Figure 2 - (a) Distribution of Documents by Terms (b) Distribution of Documents by Distance

As an illustration, the distance of every document was calculated using Euclidean distance. The distances are rounded to make the distances general and then visualized by the number of rounding. Figure 2 (a) shows us that the number rounding to 5 gives the best set of distances because of the low level of variation. The best group was then plotted in Figure 2 (b) to present the data distribution based on the number of documents at each distance. The figure shows that the distance ranges from 0.31052 to 1.41421 Euclidean distance. It has 0.00107 of variance, -5.75313 skewness, and 54.71089 kurtoses. Based on that value, data distribution is negatively skewed and leptokurtic. The data with a distance of less than 1.4 is still used in this research, assuming every research article must have its category.

2.1 Cleansing

The data need to be preprocessed to clean the data of unnecessary characters. Regular expression is required to remove them. There are several patterns created to apply in the regular expression. Each pattern is used to find and remove unnecessary characters from the patterns presented in Table 1.

2.2 Case Folding

In a document, sometimes a word is written with different cases. It can happen because of a writing error. The terms need to be in a uniform case to make the computer recognize that the words have the same meaning. The case folding required making each word in lowercase. For example, "Analisis" and "aNalisis" have identical meanings. Thus, it must be changed to be the same form as "analisis".

Table 1 - Pattern for removing unnecessary character

Description	Pattern
HTML Tag	r'<[^<>]*>'
New Line	r' \"
Space Character	r' '
HTML Tag with Attribute	r'<.*?>'
Sentence Start with Dot	r'.+\s*{.*}'

Single Character	r'[^\w\s]'
Number	r'[\d]'
Start with Space	r'^\"
Double Space	r'\s\''

2.3 Stemming

Stemming is the process of making words to be basic word form. The purpose is to make any form of a word have the same meaning. It removes the particle of the term to be the primary word form. This process needs a corpus of many forms of the word and the basic word form. For the Indonesian language, Sastrawi is used. Stemming can increase the term frequency that involves the term weight.

2.4 Stopword Removal

Each sentence usually has many similar words that do not give meaning to the document, such as conjunction, prepositions, and other terms that often appear in many documents. Removing the stopwords will reduce the data dimensions and represent the document's meaning with the appropriate words. The fundamental corpus of stopwords is used from Sastrawi and expanded by observation of the terms in documents.

2.5 Term Weighting

Term weighting is the process of determining the term importance level. The less critical term will have a value near zero it shows at Equation 1. Term weighting is the discretization of the nominal form of a term to continue value. Thus, the data can be processed to another phase. The primary method is binary term weighting that gives a value of 1 if the term appears in the document and 0 if not.

$$w_i = 1 \; if \; d_i \in T \; else \; 0 \tag{1}$$

Term frequency is the improvement of the binary term weighting method (Equation 2). It emphasizes the contribution of terms by how many times the term appears in a document.

$$w_i = \sum t_i \in D \tag{2}$$

However, too frequently, a term appears in many documents, making the term not a unique identifier for the group of documents. To prevent that, the logarithmic document frequency of each term is applied in Term Frequency Inverse Document Frequency (TF-IDF), showed by Equation 3 and Equation 4. Inverse Document Frequency (IDF) is the weight of relevancy of the term in all documents. The IDF then multiplies by term frequency in each document to get the TF-IDF value.

$$w_i = TF * IDF \tag{3}$$

$$\text{idf} = \log\left(\frac{|N|}{f_{i,d}}\right) \tag{4}$$

2.6 Clustering Method

This study uses several simple clustering methods to discover the best-formed dataset suitable with expert reasoning. The clustering methods are K-Means, K-Medoids, and DBSCAN. The pairwise distance metrics used are Euclidean, Manhattan, and Cosine Similarity.

K-Means is a clustering method that uses the mean value of each cluster [18]. The means the value used to determine the new centroid. First, set the *k* value as the number of groups, then set the centroids randomly from the objects as the initialization. Calculate the pairwise distance of every object to the centroids. Afterward, compare the distance of an object to each of the centroids. Assign the object as a member of the centroid by the closer distance to centroids. Maintain the centroid using the mean value of the members and evaluate the clustering performance using Silhouette Score. Repeat the steps and compare the Silhouette Score to maintain the clustering process.

K-Medoids have similar steps to K-Means. The problem with K-Means is that the final centroids are not the point of cluster members. K-Medoids maintain the centroid using the medoids of cluster members [19]. Thus, the final centroids

are the member of the clusters. Furthermore, this method aims to minimize the sum of dissimilarities between points labeled in a group and the centroid. However, this method is not suitable for the clustering of non-spherical data.

Another algorithm is used in the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) method. This method can cluster the data into several groups based on the proximity of the data [20]. Thus, the number of clusters does not need to determine. DBSCAN needs the initialization of the minimal point to expand the group. The DBSCAN needs a minimal epsilon value to reach the next member of the cluster. Suppose the object is far from the minimal epsilon, so the object is not a group member. DBSCAN is not forcing every object to be a member of the cluster. Objects cannot be a cluster member; that is called noise.

Every clustering method needs a pairwise distance metric to describe the location of each data. The standard metric used in clustering is Euclidean distance. Euclidean calculated the distance using the Pythagorean Theorem (Equation 5). It tries to discover the diagonal distance between the two objects.

$$d(p,q) = \sqrt{\sum_{i=1}^{n}(p_i - q_1)^2} \tag{5}$$

Another distance metric that is commonly used is City Block (Equation 6). City Block has a different perspective on distance. It simulates that every data is a square object. Thus, side length is calculated to go to the destination.

$$d(p,q) = \sum_{i}^{n} |p_i - q_i| \tag{6}$$

The cosine Similarity concept is distinct from the two methods (Equation 7 and 8). This method measures the similarity degree instead of the range. It calculates the cosine degree between two objects. Cosine distance is obtained by subtracting one from the cosine similarity degree.

$$S_c(p,q) = \cos(\emptyset) = \frac{p.q}{||p||||q||} = \frac{\sum_{i=1}^{n} p_i.q_i}{\sqrt{\sum_{i=1}^{n} p_i^2}\sqrt{\sum_{i=1}^{n} q_i^2}} \tag{7}$$

$$cosine\ distance = 1 - \cos(\emptyset) \tag{8}$$

2.7 Ground Truth

This study aims to discover the best clustering result by comparing it with the ground truth. The ground truth is formed with the assistance of the expert in grouping documents into several clusters. The expert is a Thesis Committee member that experienced in grouping theses. Each group represents an undergraduate thesis topic between 2009 and 2022. The process of grouping utilizes the information retrieval method. The method used is OR Boolean Retrieval Method. This method needs keywords as the input to retrieve relevant documents. Consequently, the keywords for each topic are required with the assistance of experts. This study uses 22 groups of predefined undergraduate thesis topics of the Computer Science Faculty, there are Business Process Management, Computer Vision, Data Mining, Enterprise Management System, Game, IT Adoption, IT Audit, IT Evaluation, IT Infrastructure, IT Strategy, Information Security, IoT, Machine Learning, Music, Natural Language Processing, Network Management, Software Construction, Software Testing, Graph Theory, Text Mining, UI/UX, and E-Business.

A simple program is developed to make the grouping easier. The program's objective is to discover each document's keywords quickly and iterative, as shown in Figure 2. Firstly, the program displays the documents that have yet to be grouped. The expert then defined the keywords of the documents and mapped them to the groups. The keywords are then applied to the OR Boolean Retrieval method as the parameters to find the related documents after all documents are grouped, then the documents are labeled.

2.8 Evaluating the Cluster

The clusters obtained from ground truth and the clustering process were then evaluated using Silhouette Score (Equation 9). Silhouette Score is calculated in every document. The final score of the Silhouette Score is the mean of the Silhouette Score of each document. Silhouette Score needs to compute the mean distance of a document to other documents in the same cluster (a_i) (Equation 10) and compute the minimum value of the mean distance of a document in a group to other documents in another cluster (b_i) (Equation 11). The Silhouette Score was calculated by dividing the difference between a_i and b_i with the maximum value between a_i and b_i. The Silhouette Score value is between 1 and -1. If the score is one,

it means clusters are well apart from each other and distinguished. If the score is 0, it means the groups are indifferent. If the score is -1, it means the clusters are assigned incorrectly.

$$s(i) = \frac{b_i - a_i}{\max(b_i, a_i)} \quad (9)$$

$$a_i = \frac{1}{|C_i|-1} \sum_{j \in C_I, i \neq j} d(i,j) \quad (10)$$

$$b_i = \min_{J \neq I} \frac{1}{|C_J|} \sum_{j \in C_I, i \neq j} d(i,j) \quad (11)$$

2.9 Comparing the Clustering Results

The clusters that resulted from the clustering method were then examined to determine the topics. Figure 3 shows, expert assistance is needed to decide the topic of each group. The topics used are the same as those used in the ground truth-making process. Afterward, the clusters are compared with the ground truth to discover how good the clustering result is. The documents are compared then the accuracy of clustering is calculated. The accuracy was gained by correctly dividing the number of labeled documents by the number of all documents (Equation 12). The number of labeled precisely documents is the count of the document in the clustering result that has the same topic in the ground truth.

$$accuracy = \frac{labelled\ correctly}{all\ documents} \quad (12)$$

Figure 3 - Defining keywords assisted by expert

3. RESULT AND DISCUSSION

The experts use a simple program that implements OR Boolean Retrieval method that makes the document analysis easier to make a ground truth. The experts determine the keywords of each group iteratively. The keywords are then used as input for the OR Boolean Retrieval method. The implementation of the method uses the regular expression mechanism. The experts strung patterns from the keywords defined. The pattern is then used as input for the regular expression mechanism. The documents retrieved by the methods are labeled as the label keyword's group. The documents were then counted based on the labels. The number of documents in every group label is shown in Table 2.

Table 2 - Pattern of keyword of each topic

No	Label	Pattern	Count
1	BPM	r'\b(sop|business process|business proses|proses bisnis)\b'	9
2	Computer Vision	r'\b(lbp|gambar|citra|citra wajah|sift)\b'	9
3	Data Mining	r'\b(knearest neighbor|knn|Decision Tree|naive bayes|adaboost|pengelompokan|clustering|nearest|deskriptif|k-means|kmeans|c4.5|prediksi|jaringan syaraf|analisis indeks pretasi)\b'	55
4	Enterprise Management System	r'\b(Vogel Approximation Method|vam|lot sizing|poq|supply chain|supply|erp|enterprise resource planning)\b'	8

5	Game	r'\b(game|game design document|pengembangan game|GDD)\b'	4
6	IT Adoption	r'\b(technology acceptance|efektivitas implementasi|penerimaan|kesiapan pengguna|kesiapan penerapan|kesiapan|faktor penerimaan|adopsi|tram|tam|faktor-faktor efektivitas|performance expectancy|analisis kontribusi|Aydin|amp|Tasci|ovo)\b'	46
7	IT Audit	r'\b(kualitas layanan)\b'	19
8	IT Evaluation	r'\b(analisis pengaruh|analisis kualitas|hot fit|human organization technology|eservice|eloyalty|etrust|esatisfaction|studi kualitatif terhadap penggunaan|egovqual|webqual|eservqual|servqual|cobit|eucs|pls-sem|customer loyalty|utaut2|utaut|togaf|altman|kesuksesan|it bsc|balanced scorecard|kualitas pelayanan|Pemeringkatan EGovernment)\b'	40
9	IT Infrastructure	r'\b(domain driven|infrastruktur)\b'	3
10	IT Strategy	r'\b(investasi|pengembangan investas|activity based costing|sostac|strategi|search engine|information economics|brand image|risiko|resiko)\b'	21
11	Information Security	r'\b(pesan rahasia|steganografi|hashed message|enkripsi|Steganography|steganografi|encryption|Digital Forensic)\b'	8
12	IoT	r'\b(otomatitasi|otomatisasi|iot|wireless sensor network|fuzzy)\b'	17
13	Machine Learning	r'\b(goal programming|simple addictive|regresi|mdf|Martingale Theory|pemberian saran|Simple Multi Attribute Rating|neural network|gdss|hybrid filtering|single layer|forward chaining|backpropagation|profile matching|ant colony|sistem pendukung|case based|algoritma|penunjang keputusan|topsis|smart|weighted sum model|peramalan|holt-winters|exponential smoothing|single exponential|moving average|wma|proyeksi|ahp|term memory|weighted product|wp|forecasting|algoritma|machine learning|sistem pakar|simple additive|lstm)\b'	242
14	Music	r'\b(Media interaktif|musik elektronik)\b'	2
15	NLP	r'\b(chatbot)\b'	2
16	Network Management	r'\b(pcq|ospf|rip)\b'	2
17	Software Construction	r'\b(Knowledge Management System|v-model|flowthing|edwards|test-driven|testdriven|pxp|gamifikasi|pengembangan sistem|test driven development|perancangan dan implementasi|rancang bangun|extreme programming|sistem informasi|sistem monitoring|sistem pengacakan|sistem penghitung|aplikasi perhitungan|aplikasi penjualan|sistem penjualan|berbasis|e-learning|aplikasi mobile|perancangan dan pembuatan|scrum|vmodel)\b'	425
18	Software Testing	r'\b(blackbox)\b'	1
19	Graph Theory	r'\b(dominating|dominating set|graf)\b'	6
20	Text Mining	r'\b(information retrieval|tf-idf|twitter|media sosial|whatsapp|berita|ekstraksi web|kata kunci|sentimen|komentar|tfidf)\b'	23
21	UI/UX	r'\b(human centered design|user exprience|experience|ucd|uiux|ui|ux|user experience|user interface|antarmuka|usability|end user)\b'	107
22	e-Business	r'\b(e-marketing|emarketing|minat beli|smo|digital marketing|konten pemasaran|value model|endorsement|efektivitas iklan)\b'	13

Based on the ground truth, the infrequently topic is Software Testing with only 1 document. The frequent topic taken by students is software construction with 242 documents. The silhouette Score of the ground truth is 0.002429. That score is not significant, or the cluster is not significantly differentiated. It can happen because each document has a similar distance.
On the other side, the documents clustered using K-Means, K-Medoids, and DBSCAN. The K-Means and K-Medoids use 20 to 25 clusters. The number of predefined topics is used as a consideration of cluster numbers. For the DBSCAN, the radius is used by the pairwise distances between documents rounded to 2 decimal places. The minimal points used is 2 points. The experiments were conducted 230 times with the variation of hyperparameters in each clustering method. The results in Table 3 show that the most significant experiment using K-Means and Cosine Similarity as the distance metric with the number of clusters is 21. The score obtained is 0.105534 of the Silhouette Score. It gives a better score than the ground truth but does not reach the maximum score of the Silhouette Score. That means the result could be a better cluster.

The experts then label the best clustering results with the name of predefined topics. The labeling process does by considering the term frequency in a cluster. Afterward, the labeled documents were compared with the ground truth, as shown in Table 4. The clusters are grouped into ten groups because of the similarity of the topic in the clusters. Accuracy is calculated to evaluate the clustering result. The label between ground truth and the clustering result is compared to obtain accuracy. The accuracy of the clustering result is 43.78%. There are 465 documents that have the same label between ground truth and clustering result, and 596 documents have different labels.

Table 3 - Clustering Result

No	Method	Metric	Number of Clusters	Score
1	K-Means	Cosine Similarity	21	0.105534
2	K-Means	Manhattan	21	0.104197
3	K-Means	Cityblock	22	0.104036
4	K-Means	Cityblock	21	0.102866
5	K-Means	Cosine Similarity	24	0.101840

Table 4 - Number of each category in ground truth and clustering result

No	Category	Clustering Result	Ground Truth
1	Business Process Modelling	0	9
2	Computer Vision	17	9
3	Data Mining	0	55
4	Enterprise Management System	0	8
5	Game	0	4
6	IT Adoption	47	46
7	IT Audit	20	19
8	IT Evaluation	143	40
9	IT Infrastructure	0	3
10	IT Strategy	0	21
11	Information Security	0	8
12	IoT	34	17
13	Machine Learning	259	242
14	Music	0	2
15	Natural Language Processing	0	2
16	Network Management	0	2
17	Software Construction	436	425
18	Software Testing	0	1
19	Graph Theory	0	6
20	Text Mining	47	23
21	UI/UX	58	107

22	e-Business	0	13

Pivoting is needed to determine which of the miss labeled documents are categorized, as shown in Table 5. Mostly the documents are miss labeled in IT-Evaluation and Software Construction categories. It happens because the documents with that categories have many terms that appear in other categories.

4. CONCLUSION

The experiment result shows that the simple clustering method with distance metrics Euclidean, Manhattan, City Block, and Cosine Similarity does not give a good result. It cannot cluster well with the distance distribution is negative skewness and leptokurtic. DBSCAN fails at all experiments because it clusters to only one cluster or clusters into the amount of data. The best combination is the K-Means method and Cosine Similarity, which gives a 0.105534 Silhouette Score and a more excellent score than the ground truth. To discover how good the clustering result is compared to the ground truth. It shows that the accuracy of clustering is 43.78% which is not a fair result to form the data as a training set. Many terms of IT-Evaluation and Software Construction categories appear in other categories. It may make the clustering not go well. Adding the information of documents like abstracts or all the content is needed to improve the clustering result.

Table 5 - List of miss-labeled categories

No	Category	Miss Labeled
1	BPM	IT-Evaluation, Software Construction
2	Computer Vision	Software Construction
3	Data Mining	IT-Evaluation, Machine Learning, Software Construction, Text Mining
4	Enterprise Management System	IT-Evaluation, Software Construction
5	Game	IT-Evaluation, Software Construction
6	IT Adoption	IT-Adoption, IT-Evaluation, Software Construction
7	IT Audit	IT-Evaluation, e-Governance
8	IT Evaluation	IT-Adoption, IT-Audit, IT-Evaluation, Software Construction, e-Governance
9	IT Infrastructure	IT-Evaluation, Software Construction
10	IT Strategy	IT-Evaluation, Software Construction
11	Information Security	IT-Evaluation, Software Construction
12	IoT	Machine Learning, Software Construction, e-Governance
13	Machine Learning	Computer Vision, IT-Adoption, IT-Evaluation, Machine Learning, Software Construction, Text Mining
14	Music	Software Construction
15	NLP	Software Construction
16	Network Management	Software Construction
17	Software Construction	Computer Vision, IT-Adoption, IT-Audit, IT-Evaluation, IoT, Machine Learning, Text Mining, UI/UX, e-Governance
18	Software Testing	Software Construction
19	Graph Theory	Machine Learning, Software Construction
20	Text Mining	IT-Evaluation, Software Construction
21	UI/UX	IT-Adoption, IT-Audit, IT-Evaluation, Software Construction
22	e-Business	IT-Evaluation, Software Construction

Disclaimer

The authors whose names are written certify that they have no conflict of interest

REFERENCES

[1] R. P. Soesanto, A. F. Rizana, and L. Andrawina, "Design of Reporting, Evaluation, and Monitoring Application for Student Organization in University," *International Journal of Innovation in Enterprise System*, vol. 3, no. 01, pp. 53–57, Jan. 2019, doi: 10.25124/ijies.v3i01.34.

[2] S. Yang, R. Wei, J. Guo, and H. Tan, "Chinese semantic document classification based on strategies of semantic similarity computation and correlation analysis," *Journal of Web Semantics*, vol. 63, p. 100578, Aug. 2020, doi: 10.1016/j.websem.2020.100578.

[3] A. Y. Muaad *et al.*, "An effective approach for Arabic document classification using machine learning," *Global Transitions Proceedings*, vol. 3, no. 1, pp. 267–271, Jun. 2022, doi: 10.1016/j.gltp.2022.03.003.
[4] O. Karnalim, "IR-based technique for linearizing abstract method invocation in plagiarism-suspected source code pair," *Journal of King Saud University - Computer and Information Sciences*, vol. 31, no. 3, pp. 327–334, Jul. 2019, doi: 10.1016/j.jksuci.2018.01.012.
[5] N. K. Seong, J. H. Lee, J. B. Lee, and P. H. Seong, "Retrieval methodology for similar NPP LCO cases based on domain specific NLP," *Nuclear Engineering and Technology*, Oct. 2022, doi: 10.1016/j.net.2022.09.028.
[6] S. Bag, S. K. Kumar, and M. K. Tiwari, "An efficient recommendation generation using relevant Jaccard similarity," *Inf Sci (N Y)*, vol. 483, pp. 53–64, May 2019, doi: 10.1016/j.ins.2019.01.023.
[7] A. Gragera and V. Suppakitpaisarn, "Relaxed triangle inequality ratio of the Sørensen–Dice and Tversky indexes," *Theor Comput Sci*, vol. 718, pp. 37–45, Mar. 2018, doi: 10.1016/j.tcs.2017.01.004.
[8] M. Hanifi, H. Chibane, R. Houssin, and D. Cavallucci, "Problem formulation in inventive design using Doc2vec and Cosine Similarity as Artificial Intelligence methods and Scientific Papers," *Eng Appl Artif Intell*, vol. 109, p. 104661, Mar. 2022, doi: 10.1016/j.engappai.2022.104661.
[9] J. Pascual Espada, J. Solís Martínez, I. Cid Rico, and L. Emilio Velasco Sánchez, "Extracting keywords of educational texts using a novel mechanism based on linguistic approaches and evolutive graphs," *Expert Syst Appl*, vol. 213, p. 118842, Mar. 2023, doi: 10.1016/j.eswa.2022.118842.
[10] S. Behpour, M. Mohammadi, M. v. Albert, Z. S. Alam, L. Wang, and T. Xiao, "Automatic trend detection: Time-biased document clustering," *Knowl Based Syst*, vol. 220, p. 106907, May 2021, doi: 10.1016/j.knosys.2021.106907.
[11] K. Thirumoorthy and K. Muneeswaran, "A hybrid approach for text document clustering using Jaya optimization algorithm," *Expert Syst Appl*, vol. 178, p. 115040, Sep. 2021, doi: 10.1016/j.eswa.2021.115040.
[12] E. Mohamed and T. Celik, "Early detection of failures from vehicle equipment data using K-means clustering design," *Computers and Electrical Engineering*, vol. 103, p. 108351, Oct. 2022, doi: 10.1016/j.compeleceng.2022.108351.
[13] S. Harikumar and S. PV, "K-Medoid Clustering for Heterogeneous DataSets," *Procedia Comput Sci*, vol. 70, pp. 226–237, 2015, doi: 10.1016/j.procs.2015.10.077.
[14] Q. Zhu, X. Tang, and A. Elahi, "Application of the novel harmony search optimization algorithm for DBSCAN clustering," *Expert Syst Appl*, vol. 178, p. 115054, Sep. 2021, doi: 10.1016/j.eswa.2021.115054.
[15] S. I. Rizo Rodríguez and F. de A. Tenório de Carvalho, "Clustering interval-valued data with adaptive Euclidean and City-Block distances," *Expert Syst Appl*, vol. 198, p. 116774, Jul. 2022, doi: 10.1016/j.eswa.2022.116774.
[16] J. Yin and S. Sun, "Incomplete multi-view clustering with cosine similarity," *Pattern Recognit*, vol. 123, p. 108371, Mar. 2022, doi: 10.1016/j.patcog.2021.108371.
[17] E. Aytaç, "Unsupervised learning approach in defining the similarity of catchments: Hydrological response unit based k-means clustering, a demonstration on Western Black Sea Region of Turkey," *International Soil and Water Conservation Research*, vol. 8, no. 3, pp. 321–331, Sep. 2020, doi: 10.1016/j.iswcr.2020.05.002.
[18] J. A. Hartigan and M. A. Wong, "Algorithm AS 136: A K-Means Clustering Algorithm," *Appl Stat*, vol. 28, no. 1, p. 100, 1979, doi: 10.2307/2346830.
[19] L. Kaufman and Rousseuw Peter J., "Partitioning Around Medoids (Program PAM)," pp. 68–125. doi: 10.1002/9780470316801.ch2.
[20] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise," in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, 1996, pp. 226–231.