iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** III **Month of publication:** March 2022

DOI: https://doi.org/10.22214/ijraset.2022.40758

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue III Mar 2022- Available at www.ijraset.com

Unsupervised Learning Algorithms for Keyword Extraction in an Undergraduate Thesis

Fred Torres-Cruz[1], Edelfré Flores-Velásquez[2], William E. Arcaya Coaquira[3], Irenio L. Chagua Aduviri[4], Marga I. Ingaluque-Arapa[5]
[1, 2, 3, 4, 5]*Maestría en Ingeniería de Sistemas, Universidad Nacional del Altiplano de Puno, P.O. Box 291, Puno-Perú*

Abstract: ***The amount of data managed in many academic institutions has increased in recent years, particularly in all the research work done by undergraduate students, who simply use empirical techniques for keyword selection, forgetting existing technical methods to assist their students in this process. Information and communication technologies, such as the platform for integrated research and academic work with responsibility (PILAR), which records information about research projects, such as titles, summaries, and keywords in their various modalities, have gained relevance and importance in the management of these. We proved algorithms with these records of research projects that have been analysed in this study, and predictions were made for each of the nine (09) models of unsupervised machine learning algorithms that were implemented for each of the 7430 records from the dataset. The most efficient way of extracting keywords for this dataset was the TF-IDF method, obtaining 72% accuracy and [0.4786, SD 0.0501] in average extraction time for each thesis file processed by this model.***
Keywords: Keyword extraction, Machine learning, thesis. university students, unsupervised learning.

I. INTRODUCTION

It is undeniable the increase in the volumes of information that are being generated through the implementation of knowledge information systems in organizations, with the aim of achieving efficient management and service, on the other hand, universities are implementing services with the help of technology of different types that allow transactions at different levels of institutional management storing information in all its processes [1], [2]. All the existing volume of information on the Internet is growing permanently and acquires different forms of representation, from simple text files on a personal computer or an electronic newspaper to digital libraries and much larger and complex spaces such as the web, this information in hordes increased even more with the use of digital media [3], [4] , the Universidad Nacional del Altiplano de Puno is no stranger to this affluent growth, so from the implementation of information systems such as the Platform for Integrated Research to Academic Work with Responsibility[5] (PILAR), as well as the other like research platform for the special fund for university teachers, through these applications, information is being collected as well as the management of knowledge, generated by students, graduates, but noting in their procedures no treatment at the time of choosing the keywords. Automatic keyword selection approaches are increasingly important to classify large volumes of documents, this process has become essential to make these documents more manageable and to obtain valuable information [6]. Automatic keyword extraction helps to filter and find better recommendation and retrieval of information based on the content of the text itself, thus has a representation based on the content being evaluated [6] .The goal of automatic keyword extraction is the application of the power and speed of current computational and computing capabilities to the problem of access and retrieval, with emphasis on information organization [7]. Likewise for information and knowledge extraction is the subject of considerable research interest in the fields of machine learning and data mining. text data mining and in particular text mining has become one of the most active sub fields of research in data mining [8].

Therefore, the present work addresses the use of the most representative techniques of automatic keyword extraction, using unsupervised machine learning models, being these techniques one of those that are rapidly implemented, as well as laying the groundwork for other similar studies and that this area of knowledge can continue to be studied.

II. LITERATURE REVIEW

A. *Definitions*

1) *Unsupervised Learning:* Unsupervised learning is the equivalent of grouping and this process is not subject to a contrast analysis, so this grouping is not previously identified, if we go to a specific section in the extraction of keywords with the unsupervised approach it is not necessary to need a manual or automatic labelling corpus. To perform these tasks there are different techniques in which the characteristics of the text are analysed to obtain a better extraction effect [9] .

©IJRASET: All Rights are Reserved

2) *Keyword Selection:* For the selection of keywords for research papers, articles and academic documents, the author must choose between 3 to 10 words [10] the same ones that should represent the idea that most of the time are presented in the abstract and title of the work repeatedly, for easy location and traceability[11]. After this main ide of keyword selection, we may talk about:
3) *Keyword Extraction:* Depending on the model, autonomous keyword extraction is the process of selecting words and phrases from a text document that, at best we may project the central idea of the document without any human intervention. [12].

B. Background

The extraction of words is widely studied from the different aspects of computer science in particular from machine learning, for which we present the following works that support the development of this research work like Xu & Zhang who propose a new approach to extracting keywords from a text that combines characteristics such as word frequency and the association between them [9], in other field Kretschmann et al., develop a system for keyword assignment for scientific abstracts to be functional as recommendations of researchers[13]. YAKE is an unsupervised learning keyword extraction method that relies on statistical text characteristics extracted from individual documents [14]. As we can see many authors has been work in this field of study such as Guan et al. who show us in their research how improved TF-IDF for article keywords extraction specifically the accuracy of the algorithm[15], like them other representative study was made by Mahata et al., writing an article on keyword extraction called Key2Vec with unsupervised methods by taking advantage of the formation of multi-word phrase embeddings that are used for the thematic representation of scientific articles [16] , same as these examples that we show we may prove these methos an get an evidence of accuracy using a local dataset. The need to substantiate the effectiveness of unsupervised machine learning algorithms for the generation of keywords of research work at the National University of the Altiplano stems from the practicality required for the implementation of these in running work environments, these algorithms used in the different computational disciplines such as data mining with text mining and artificial intelligence with machine learning, both dedicated to the generation of keywords. However, despite the existence of these methods of knowledge generation, these algorithms are not currently used; therefore, it will be critical to provide an evaluation of the most effective methods. For this purpose, we will identify, implement and compare to verify the effectiveness extraction of keywords in undergraduate research projects at the Universidad Nacional del Altiplano de Puno.

III.METHODS

The methodology of this work, was based in test algorithms that could be used for the extraction of keywords that are found when analysing information stored through information systems that manage the research works, makes us have valid and reliable information with respect to the quality of it since they are regulated procedures within the Universidad Nacional del Altiplano de Puno. The research is non-experimental, cross-sectional since data is collected in a single moment, in a certain time. In order to describe the results based on the information collected from each of the tests performed in this study.

A. Population

The research works was made in the Universidad Nacional del Altiplano de Puno, all the research works carried out by undergraduate students, these projects that are registered in PILAR which is administered by the vice chancellor for research, from which the information was extracted in order to carry out this research, which are presented in greater detail PILAR. In this study we won't use sampling, we need to prove the algorithms with a large amount of data as possible the data was represented in (Table I).

TABLE I. Dataset Distribution

Knowledge Area	Status				
	Archived	Rejected	Project	Proposal	Completed
Biomedical	17	5	320	133	973
Business Economics	32	10	286	162	465
Engineering	60	24	847	368	1211
Social Sciences	34	20	891	291	1281
Partial	143	59	2344	954	3930
Total			7430		

Source: From database of PILAR - UNA PUNO 2016-2021. Exported on 01-09-2021

B. Data Source

For the development of this research, we obtain the export of each register of the research projects of students and graduates has been prepared, which are administered in the vice chancellor for Research of the Universidad Nacional del Altiplano de Puno, taking as an initial source the data structure (Table 2), with 7430 records detailed in Table 3, these records were exported and processed individually in text files individual in text files (.txt) due to the cost of storing in memory could cause an overflow of memory.

TABLE II
DATA STRUCTURE

N°	Number	Detail
01	Guy	*Type of research work*
02	Code	*Labor Code based* on Data
03	Title	*Title of the work registered by the* author.
04	Summary	*Summary registered by the* author.
05	Keywords	*Keywords registered by the* author. *[,]*

C. Data Processing

To develop this research work, the information detailed in (Table I), that was 7430 records of titles detailed in (Table I), whose structure is composed of titles, abstracts and keywords in the structure declared in (Table II), these data correspond to the research works of students and graduates who opted for the thesis modality to obtain their professional degree. This information that has been processed using text files, it began its treatment using regular expression methods in order to the text and exclude the special characters and stay only with the analysis of the text, to later perform the segmentation and procedures necessary to maintain the data an operable format. After that we applied the unsupervised methods to finally extract keywords.

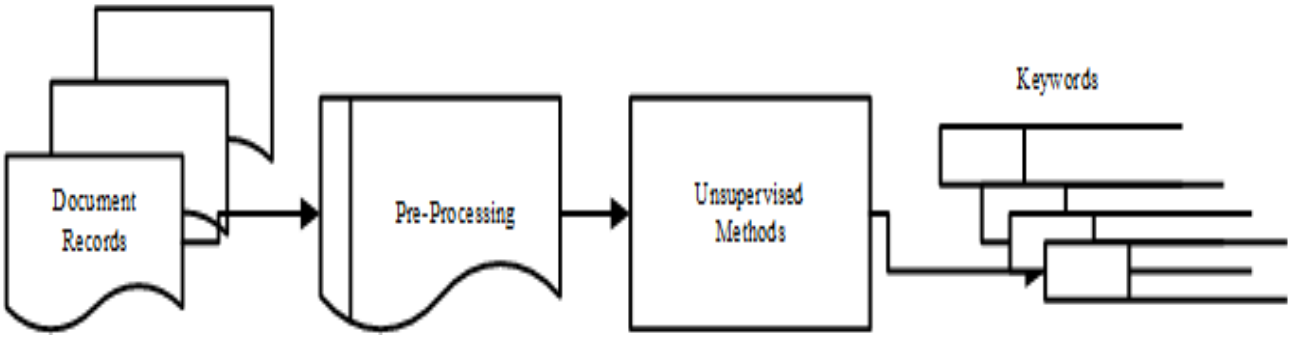


Fig. 1 Data processing representation

D. Model Implementation

For the present work the models of Python Keyword Extraction PKE [17] have been used, these models were encoded with the following structure of equivalence of file names for the present research work TF-IDF(**M1**), KPMiner[18](**M2**), YAKE[14](**M3**), TextRank[19](**M4**), SingleRank[20](**M5**), TopicRank[21](**M6**), TopicPageRank[22] (**M7**), PositionRank [23](**M8**), MultipartRank [24](**M9**), of these we present the evaluation and implementation in the result part. All of these models were implemented in same conditions using Python programming language,

E. Models Assessment

For the model assessment, we applied the following evaluation metrics (Table III) for which the sklearn library was used, as one of the most used in this type of evaluations, as well as its adaptability to the results obtained in the prediction of each group of keywords, after that we use R-Software to evaluate and get the results.

TABLE III
MODEL EVALUATION METRICS

Metric	Detail
Test F1 *F1-Score*	The F1 score can be interpreted as a harmonic mean of accuracy and recovery, where an F1 score reaches its best value at 1 and the worst score at 0.
Recall test *Recall Score*	Withdrawal is the ratio where the number of true positives and the number of false negative. The retreat is intuitively the ability of the classifier to find all positive samples.
Precision Test *Precision Score*	Accuracy is the reason where the number of true positives and the number of false positives. Accuracy is intuitively the classifier's ability not to label a sample that is negative as positive.
Scoring of Accuracy *Acuracy Score*	Calculates the accuracy of the subassembly: The label conjunction predicted for a sample must exactly match the corresponding label conjunction in the original label list
Average Loss *Hamming Score*	*Hamming* loss, is the number of tags that is incorrectly predicted.
Time *Time*	The time is detailed from the reading of the source resource to the final prediction.

Source: (Scikit learn, 2021) Metrics and scoring: quantification of the quality of predictions. Modules. https://scikit-learn.org/stable/modules/model_evaluation.html

IV. RESULTS

After implementing the methods to extract keywords, it was necessary to individually encode each one of nine models described in model implementation section in the previous part, then we have to made an analysis about time and accuracy evaluation was carried out according to the metrics detailed in (Table III) for each of the 7430 records, these prediction metrics allowed us to determine the efficiency of computed keyword extraction for each model. In the next figure we present the main difference between these models that we proved.

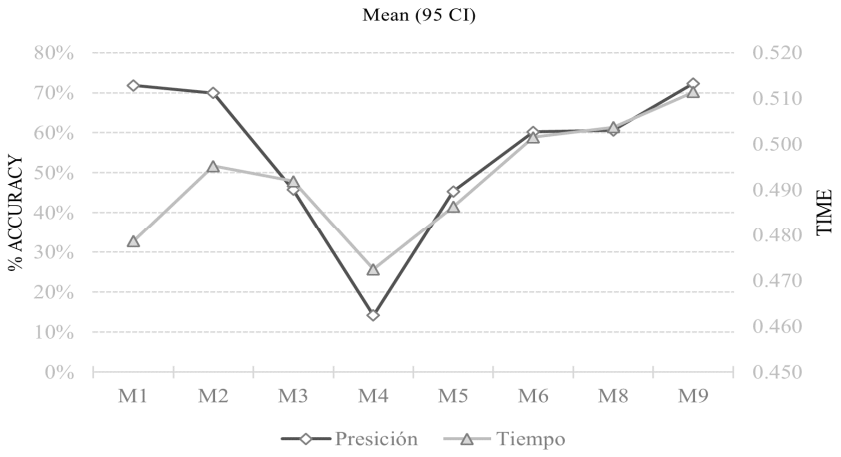


Fig. 2 Models assessment by time and accuracy

As we can see at the Fig.2 with a performing a visual analysis and observing the values time and accuracy, we only choose the model M1 implemented with the algorithm TF-IDF have result with 0.4786 in time evaluation and 72 % of accuracy optimizing key word extraction work in other hand the algorithm TextRank called in this research M4 have less time 0.4725 but the accuracy of 14 % provide us evidence to discard this model in future implementations with data same as we show in this study.

TABLE IV
KEYWORDS EXTRACTION MODELS ACURRACY

Model	*nE*	*nA*	Total	% E	% A
M1	2094	5330	7424	28 %	72 %
M2	2234	5196	7430	30 %	70 %
M3	4030	3400	7430	54 %	46 %
M4	6376	1054	7430	86 %	14 %
M5	4063	3367	7430	55 %	45 %
M6	2954	4476	7430	40 %	60 %
M7	1112	895	2007	55 %	45 %
M8	2927	4503	7430	39 %	61 %
M9	576	1503	2079	28 %	72 %

Legend: nE = number of errors, nA =number of successes, % E = error rate, % A = rate of success

The models based on TF-IDF, have been studied over many years always showing their efficiency, even with the improvements such as those developed in the work of Guan et al.[15], in which the recalculation of the prediction scores is carried out optimizing up to 20% of the evaluated indicators. The same happens with the TextRank model but in a smaller range in which in several studies and especially in the one that defines its implementation its precision is demonstrated by comparing its implementation with different groups of data [19], over time several studies have been demonstrating that the implementation of this Models that are theoretically simple demonstrate their great processing capabilities. There fore we demonstrate once that TF-IDF have better result than other models, that has been implemented in this research as we can see in the last table (Table IV).

An extra challenge in this research field is use natural language processing with autoregressive language models, also build a tree with all combinations of these models to up the accuracy an reduce the time in processing always trying to get the best result in this challenge of keyword extraction.

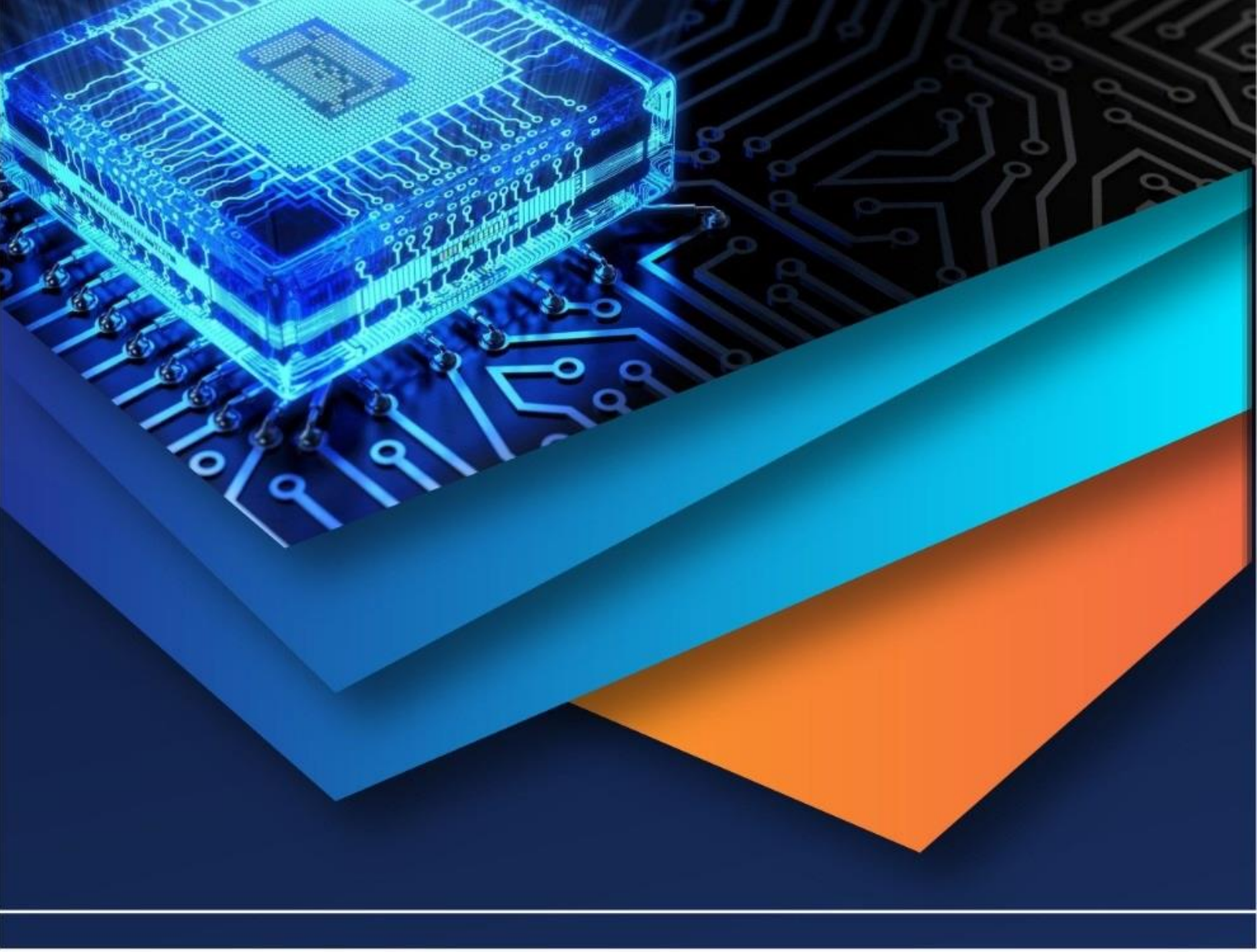
V. CONCLUSIONS

The most efficient way of extracting keywords for this dataset was the TF-IDF method, due to its relationship between time and the accuracy of the analysis, because it guarantees us a shorter processing time and high precision score in keywords extraction so we can consider it a simple and general model for this and other related works. In this research we prove nine (09) algorithms of unsupervised learning that could provide information to implementing in other universities form undergraduate thesis, also this research allowed us to exploit the computational resources in text processing, the challenge was the management and preprocessing, to finally evaluate more than fifty-six thousand records (> 56000). Finally, we show the difference between each model that we cand found for keyword extraction proposes.

REFERENCES

[1] E. Martín-Mora, S. Ellis, and L. M. Page, “Use of web-based species occurrence information systems by academics and government professionals,” PLoS One, vol. 15, no. 7 July, 2020, doi: 10.1371/journal.pone.0236556.
[2] S. Ranguelov, “Gestión de la Información y el Conocimiento en las Organizaciones,” Biblios, vol. 12, no. 1, pp. 1–7, 2012.
[3] G. H. Tolosa and F. R. a. Bordignon, Introducción a la Recuperación de Información Conceptos , modelos y algoritmos básicos, Pre-Edició. Laboratorio de Redes de Datos, 2008.
[4] V. Jayawardene, T. J. Huggins, R. Prasanna, and B. Fakhruddin, “The role of data and information quality during disaster response decision-making,” Prog. Disaster Sci., vol. 12, p. 100202, 2021, doi: 10.1016/j.pdisas.2021.100202.
[5] F. Torres-Cruz, “Plataforma web basada en cloud computing para el seguimiento de proyectos de tesis de pregrado UNA Puno 2016,” no. 051, 2016.
[6] A. Ahadh, G. V. Binish, and R. Srinivasan, “Text mining of accident reports using semi-supervised keyword extraction and topic modeling,” Process Saf. Environ. Prot., vol. 155, pp. 455–465, 2021, doi: 10.1016/j.psep.2021.09.022.
[7] B. Santosh Kumar, B. Korra Sathya, and J. Sanjay Kumar, “Automatic Keyword Extraction for Text Summarization: A Survey,” 2017, [Online]. Available: http://arxiv.org/abs/1704.03242.
[8] Z. Xuezhong, P. Yonghong, and L. Baoyan, “Text mining for traditional Chinese medical knowledge discovery: A survey,” J. Biomed. Inform., vol. 43, no. 4, pp. 650–660, 2010, doi: 10.1016/j.jbi.2010.01.002.
[9] Z. Xu and J. Zhang, “Extracting Keywords from Texts based on Word Frequency and Association Features,” Procedia Comput. Sci., vol. 187, pp. 77–82, 2021, doi: 10.1016/j.procs.2021.04.035.
[10] M. Gonzáles and S. Mattar, “Las claves de las palabras clave en los artículos científicos,” Rev. MVZ Cordoba, vol. 17, no. 2, pp. 2955–2956, 2012.
[11] C. Mack, “How to write a good scientific paper: title, abstract, and keywords,” J. Micro/Nanolithography, MEMS, MOEMS, vol. 11, no. 2, p. 020101, 2012, doi: 10.1117/1.jmm.11.2.020101.

[12] C. Zhang, H. Wang, Y. Liu, D. Wu, Y. Liao, and B. Wang, "Automatic keyword extraction from documents using conditional random fields," J. Comput. Inf. Syst., vol. 4, no. 3, pp. 1169–1180, 2008.
[13] M. Kretschmann, A. Fischer, and B. Elser, "Extracting Keywords from Publication Abstracts for an Automated Researcher Recommendation System," Digit. Welt, vol. 4, no. 1, pp. 20–25, 2020, doi: 10.1007/s42354-019-0227-2.
[14] R. Campos, V. Mangaravite, A. Pasquali, A. Jorge, C. Nunes, and A. Jatowt, "YAKE! Keyword extraction from single documents using multiple local features," Inf. Sci. (Ny)., vol. 509, pp. 257–289, 2020, doi: 10.1016/j.ins.2019.09.013.
[15] X. Guan, Y. Li, and H. Gong, "Improved TF-IDF for We Media Article Keywords Extraction," J. Phys. Conf. Ser., vol. 1302, no. 3, p. 032003, Aug. 2019, doi: 10.1088/1742-6596/1302/3/032003.
[16] D. Mahata, J. Kuriakose, R. R. Shah, and R. Zimmermann, "Key2Vec: Automatic Ranked Keyphrase Extraction from Scientific Articles using Phrase Embeddings," pp. 634–639, 2018, doi: 10.18653/v1/n18-2100.
[17] F. Boudin, "pke: an open source python-based keyphrase extraction toolkit," in Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations, 2016, pp. 69–73, [Online]. Available: http://aclweb.org/anthology/C16-2015.
[18] S. R. El-Beltagy and A. Rafea, "KP-miner: Participation in SemEval-2," ACL 2010 - SemEval 2010 - 5th Int. Work. Semant. Eval. Proc., no. July, pp. 190–193, 2010.
[19] R. Mihalcea and P. Tarau, "TextRank: Bringing order into texts," Proc. 2004 Conf. Empir. Methods Nat. Lang. Process. EMNLP 2004 - A Meet. SIGDAT, a Spec. Interes. Gr. ACL held conjunction with ACL 2004, vol. 85, pp. 404–411, 2004.
[20] X. Wan and J. Xiao, "CollabRank: Towards a collaborative approach to single-document keyphrase extraction," Coling 2008 - 22nd Int. Conf. Comput. Linguist. Proc. Conf., vol. 1, no. August, pp. 969–976, 2008.
[21] A. Bougouin, F. Boudin, and B. Daille, "TopicRank: Topic ranking for automatic keyphrase extraction," Rev. Trait. Autom. des Langues, vol. 55, no. 1, pp. 45–69, 2013.
[22] L. Sterckx, T. Demeester, J. Deleu, and C. Develder, "Topical word importance for fast keyphrase extraction," WWW 2015 Companion - Proc. 24th Int. Conf. World Wide Web, no. 2, pp. 121–122, 2015, doi: 10.1145/2740908.2742730.
[23] C. Florescu and C. Caragea, "PositionRank: An unsupervised approach to keyphrase extraction from scholarly documents," ACL 2017 - 55th Annu. Meet. Assoc. Comput. Linguist. Proc. Conf. (Long Pap., vol. 1, pp. 1105–1115, 2017, doi: 10.18653/v1/P17-1102.
[24] F. Boudin, "Unsupervised keyphrase extraction with multipartite graphs," NAACL HLT 2018 - 2018 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf., vol. 2, pp. 667–672, 2018, doi: 10.18653/v1/n18-2105.
[25] Scikit learn, "Métricas y puntuación: cuantificación de la calidad de las predicciones," Modules, 2021. https://scikit-learn.org/stable/modules/model_evaluation.html (accessed Sep. 29, 2021).



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)