Journal of Computer and Communications, 2021, 9, 24-37
https://www.scirp.org/journal/jcc
ISSN Online: 2327-5227
ISSN Print: 2327-5219

A Deep Look into Extractive Text Summarization

Jhonathan Quillo-Espino, Rosa María Romero-González, Ana-Marcela Herrera-Navarro

Facultad de Informática, Universidad Autónoma de Querétaro, Querétaro, México
Email: estudiantejquillo@gmail.com

How to cite this paper: Quillo-Espino, J., Romero-González, R.M. and Herrera-Navarro, A.-M. (2021) A Deep Look into Extractive Text Summarization. *Journal of Computer and Communications*, **9**, 24-37.
https://doi.org/10.4236/jcc.2021.96002

Received: May 6, 2021
Accepted: June 6, 2021
Published: June 9, 2021

Abstract

This investigation has presented an approach to Extractive Automatic Text Summarization (EATS). A framework focused on the summary of a single document has been developed, using the Tf-ldf method (Frequency Term, Inverse Document Frequency) as a reference, dividing the document into a subset of documents and generating value of each of the words contained in each document, those documents that show Tf-Idf equal or higher than the threshold are those that represent greater importance, therefore; can be weighted and generate a text summary according to the user's request. This document represents a derived model of text mining application in today's world. We demonstrate the way of performing the summarization. Random values were used to check its performance. The experimented results show a satisfactory and understandable summary and summaries were found to be able to run efficiently and quickly, showing which are the most important text sentences according to the threshold selected by the user.

Keywords

Text Mining, Preprocesses, Text Summarization, Extractive Text Sumarization

1. Introduction

The creation of internet, social networks, forums, information technologies spread in a revolving way, inducing the interaction of information increasingly difficult to understand, create, save, develop and store. The entire document still has to be read completely to decide if the information it contains is relevant or not, but it becomes a slow and overwhelming activity. But what if the information could be summarized in such a way as to obtain keywords that can help reduce time and effort in the decision-making process, therefore, automatic text summaries are the solution to this problem (ATS).

Before talking about text summaries, we must first clarify: What a text summary is (TS)? A topical TS is a text that contains the most important information of one or more texts in a simplified form. The common stages for a TS: Identification of the most relevant text; Interpretation of information and obtaining a summary with the interpreted information.

The objective of ATS is to reduce the amount of text while preserving the main idea of the original document, allowing the reader to interpret the information read in a faster way. The ATS has gained popularity due to the need for analysis of large amounts of textual information, for example: generate summaries of books, comics, reviews, news, scientific articles, internet, social networks, among others, any type of textual information can be summarized, the excessive growth of information has forced researchers to seek different ways to obtain summaries of text and even its scope is such that they have achieved and created tools that allow summarizing content with illustrated text. The ATS can be applied in a single document or in a multi-document, depending on the specifications required.

The research is divided by the description of the Automatic Text Summarization techniques, later the Extractive Automatic Text Summarization is described and analyzed, immediately, the different approaches to generate ATS are mentioned later the experiment to perform EATS is demonstrated and finally, the experimental results are clearly and satisfactorily shown.

2. ATS Techniques

There are two approaches, Abstractive Automatic Text Summarization (AATS), [1] determines that they aim to concisely paraphrase the information content in documents by creating new information. The automatic text summarization (ATS) [2] concludes that they are those that choose the most noticeable sentences in the documents for further concatenation, to form a summary. [3] mentions that the most successful systems use EATS approaches as they cut and join parts of the text to produce a reduced version, [4] determines, therefore that EATS results in summaries with information available from the original text without any changes.

2.1. Extractive Automatic Text Summarization (EATS)

[5] proposes that a typical EATS consists of 2 phases, the first are the pre-processes, [6] determines that its objective is to transform textual data into clear elements, eliminating inconsistencies for future interpretation. In addition, they can attach new sentences that are not contained in the original document and in the second phase, [7] indicates that it is the use of an approach and its objective is to reduce the length and detail of a document, preserving the sense and the most important points without changing its meaning. **Figure 1** shows the diagram of a typical EATS.

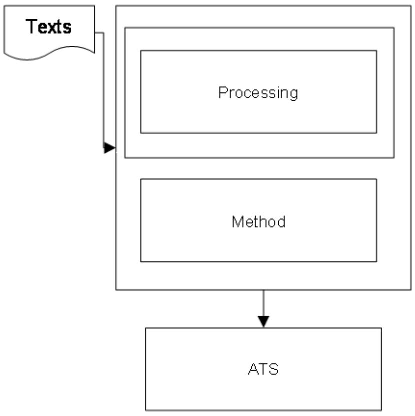


Figure 1. General EATS process [5].

2.2. Approaches to ATS

1) Graph-based [8] demonstrates that each statement is treated as a node, two sentences are connected with an edge if both sentences have some similarity, they are essentially a way of deciding the importance of a vertex within a graph, based on the information extracted from the graph structure.

2) Based on Machine Learning (ML), [9] states that the selection of important text is represented as a binary classification problem, dividing all sentences in the input into summary sentences, the probability that the sentence should be in the summary is the sentence score. Therefore, the classifier is the one that determines the score sentences by taking as input the sentence representation and as output the sentence score. The sentences with more punctuation are selected to form the summary.

3) Neural networks, [10] states that are those modified to discover the importance and unimportance to determine the value of the summary of each sentence in a document.

4) Grouping or Clustering [11] proposes the analysis of documents by grouping similar information for later comparison. The more the information is repeated, a summary can be constructed using a sequence of sentences related to the calculated clusters.

5) Tf-Idf method [12] deduces that it is named after the document frequency (Tf) inverse document frequency (Idf), it is a statistical method that shows the importance of a token in a document. Where Tf (term) = number of times the term appears in the document, Idf = total logarithm of number of documents/number of documents containing. It is calculated with the formula:

$$\text{TfIdf} = \text{dn}\left(\log\left(1+\text{Tf}\right)/\log\left(\text{df}\right)\right)$$

where df = is the number of times the token appears in all documents, dn = is the number of documents.

3. Related Work

EATS over the years has gained interest from researchers, implementing multiple strategies to make EATS more efficient. Different approaches such as the following: [13], ratify that perform an analysis of advantages and disadvantages using logical Fuzzy algorithm in EATS, [14] [15] propose a semantic method for EATS multi-document using statistical methods based on machine learning which is based on graphs. [16] [17] expose a method based on genetic algorithms for obtaining EATS [1], develop neural network based on sequential model with the characteristic of offering visualization of predictions regarding the content information, [18] propose a model to extract individual sentences modeling the relationship between sentences, [19] determine the extraction of sub-sentences based on tree decisions, based on a neural model, [20] point out that it is oriented to RTAE of news through a hybrid algorithm between semantic analysis and random fields, coherent and detailed information.

4. Development

This research details the procedure for EATS development using the Tf-Idf method. **Figure 2** shows the proposed diagram for the EATS with the Tf-Idf method.

4.1. Preprocesses Application

Literature [6] shows that with the application of preprocesses it is ensured that the document information has been filtered, the idea is to standardize the text for subsequent analysis, and it is also a means to generate a future structuring in an efficient manner.

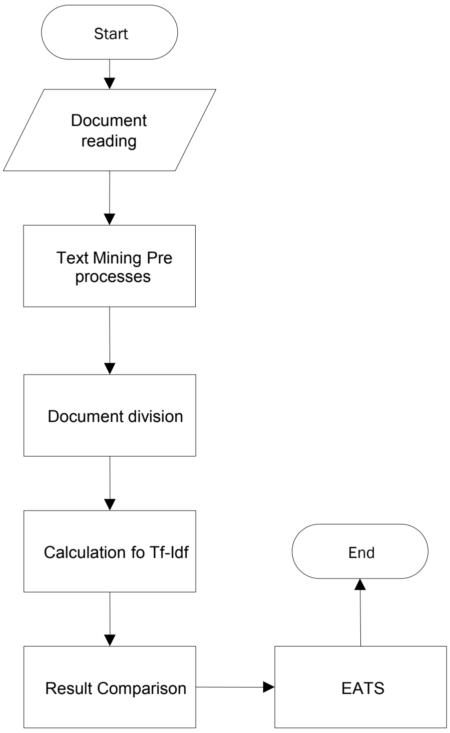


Figure 2. General process for EATS with Tf-Idf method.

4.2. Division of Documents

The division of the document into subsets of documents is done by [21], they determine that through the representation of the vector space, it is necessary to create a jagged array due to its characteristics of unequal rows and columns allows to generate the storage of each of the vectors coming from the documents, for each vector, the Tf-Idf method [22] concludes that it is a statistical method that reflects the importance of a token in a given document, is a combination of two values where [23] confirms that Tf is a term frequency used to measure the number of times a token appears in a document, and Idf shows how important the token is in the document [24], they point out that it is essentially a way of assigning weight to a word or token (term) with a document, therefore, the higher Tf the token will be more representative in the document. The calculation of Idf is done through the following formula:

$$\mathrm{Idf}(w) = 1 + \log\left(|\mathrm{d}|/\mathrm{df}(w)\right) \tag{1}$$

where df(w) = number of documents where the token appears, d = total number of documents. If Idf(w) is low and if the token appears in many documents it means that the token obtains low discrimination power, on the other hand, if Idf is high and the token appears in few documents it means that the token has a high discrimination level in the whole document.

4.3. Values of Tf-Idf

The value of Tf-Idf increases proportionally with the number of times a token appears in the document. Using the modified formula proposed by [25], they report that the calculation of Tf-Idf is performed as follows:

$$\mathrm{TfIdf}(w,s) = \mathrm{Tf}(w,s) * \mathrm{Idf}(w). \tag{2}$$

where Tf(w, s) = number of times the w token appears in the document. Idf(w) = is the number of documents in which the token appears. It is necessary to average Tf-Idf for each document with the equation:

$$\mathrm{Average}(\mathrm{tf}-\mathrm{idf}) = \sum_{i=1}^{w(s)} \mathrm{tf}-\mathrm{idf}(i,d)/w(s). \tag{3}$$

where w(s) number of total tokens in the document. Idf(i, d) = number of value obtained by calculating Idf for each token.

4.4. Threshold Value

The Threshold is the summary percentage level requested to the user, basically it is the amount of text to be summarized. The Threshold is used to calculate the maximum value of Tf-Idf. To calculate it, it is done through the following formula:

$$\mathrm{Threshold}(\mathrm{TfIdf}) = \%\ \mathrm{de\ threshold} * \max(\mathrm{TfIdf}). \tag{4}$$

The EATS is obtained from those documents whose average value (Tf-Idf) is equal to or higher than the Threshold (Tf-Idf).

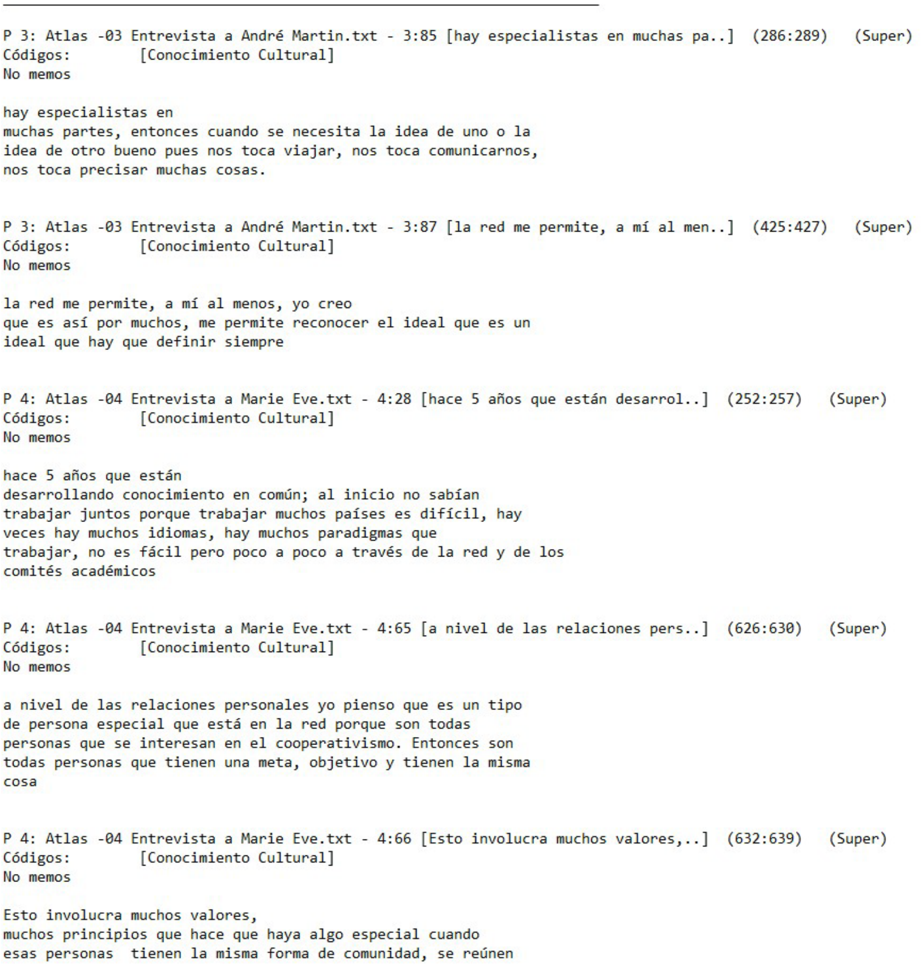
5. Results

This research was developed in Visual Studio Community 2019. A document with a total of 2431 tokens in Spanish language was used to evaluate the performance of the EATS. Figure 3 shows an example of the analyzed document.

Two threshold levels were randomly selected to verify the operation of the program. Two tests were performed with a Threshold level of 90% and 35%.

5.1. Execution Time

The total execution time with a Threshold of 90% and 30% respectively was 37.962 seconds for both, this is due to the fact that the calculation is performed on the same set of documents, therefore, there is no variation in both calculations. Figure 4 shows the execution time.



```
P 3: Atlas -03 Entrevista a André Martin.txt - 3:85 [hay especialistas en muchas pa..]  (286:289)   (Super)
Códigos:        [Conocimiento Cultural]
No memos

hay especialistas en
muchas partes, entonces cuando se necesita la idea de uno o la
idea de otro bueno pues nos toca viajar, nos toca comunicarnos,
nos toca precisar muchas cosas.


P 3: Atlas -03 Entrevista a André Martin.txt - 3:87 [la red me permite, a mí al men..]  (425:427)   (Super)
Códigos:        [Conocimiento Cultural]
No memos

la red me permite, a mí al menos, yo creo
que es así por muchos, me permite reconocer el ideal que es un
ideal que hay que definir siempre


P 4: Atlas -04 Entrevista a Marie Eve.txt - 4:28 [hace 5 años que están desarrol..]  (252:257)   (Super)
Códigos:        [Conocimiento Cultural]
No memos

hace 5 años que están
desarrollando conocimiento en común; al inicio no sabían
trabajar juntos porque trabajar muchos países es difícil, hay
veces hay muchos idiomas, hay muchos paradigmas que
trabajar, no es fácil pero poco a poco a través de la red y de los
comités académicos


P 4: Atlas -04 Entrevista a Marie Eve.txt - 4:65 [a nivel de las relaciones pers..]  (626:630)   (Super)
Códigos:        [Conocimiento Cultural]
No memos

a nivel de las relaciones personales yo pienso que es un tipo
de persona especial que está en la red porque son todas
personas que se interesan en el cooperativismo. Entonces son
todas personas que tienen una meta, objetivo y tienen la misma
cosa


P 4: Atlas -04 Entrevista a Marie Eve.txt - 4:66 [Esto involucra muchos valores,..]  (632:639)   (Super)
Códigos:        [Conocimiento Cultural]
No memos

Esto involucra muchos valores,
muchos principios que hace que haya algo especial cuando
esas personas  tienen la misma forma de comunidad, se reúnen
```

Figure 3. Analized document.

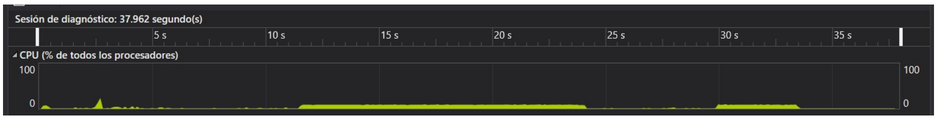

Sesión de diagnóstico: 37.962 segundo(s)

Figure 4. Execution time.

5.2. Number of Tokens per Document

By dividing the document into smaller documents, a total of 47 independent documents were obtained. A given number of tokens were obtained for each document. Figure 5 shows the result of tokens per document.

5.3. Tf per Token in the Document

Obtaining the Tf of each token inside of each document was essential, allowing the calculation of Tf-Idf, Figure 6 shows the example of the Tf calculation for document number 1.

5.4. Frequency of Token per Document

It was necessary to count the number of document in which a token appears, in order to perform the necessary operations for the Tf-Idf calculation. It is observed that the token much corresponding to token number 2 and number 17 of document number 1, but it is repeated a total of 19 times in the 47 documents, Figure 7 shows a clear example of the result of the Tf count per document in document number 1.

5.5. Idf per Document

Idf determines the tokens hierarchy within the document, the higher the Idf value the higher the relevance. It is observed that the highest values correspond to tokens number 11, 13, 15, 16 with a value of 3.87 respectively. Figure 8 shows the result of the Idf calculation for document number 1.

5.6. Tf-Idf Calculation Result

The results of Tf-Idf calculation for document number 1 shows the lowest value was Tf number 16 with a value of 2.77 and Tf number with highest value correspond to tokens with number 2 and 17 with a value of 99.81. Figure 9 shows the example of the results for the Tf-Idf calculation for document number 1.

Figures 4-7 show the obtained results from the calculations analysis of a total

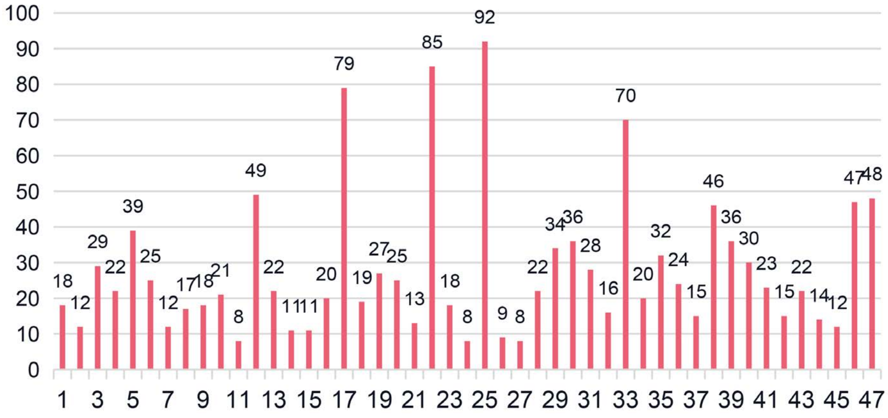


Figure 5. Number of tokens per document.

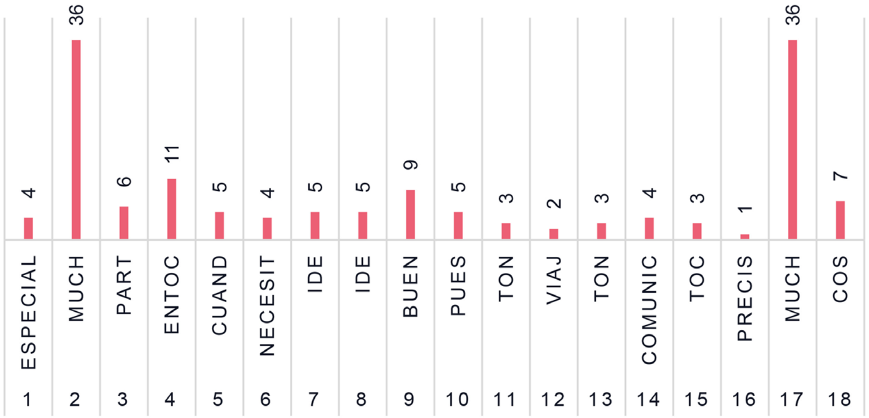


Figure 6. Tf for document number 1.

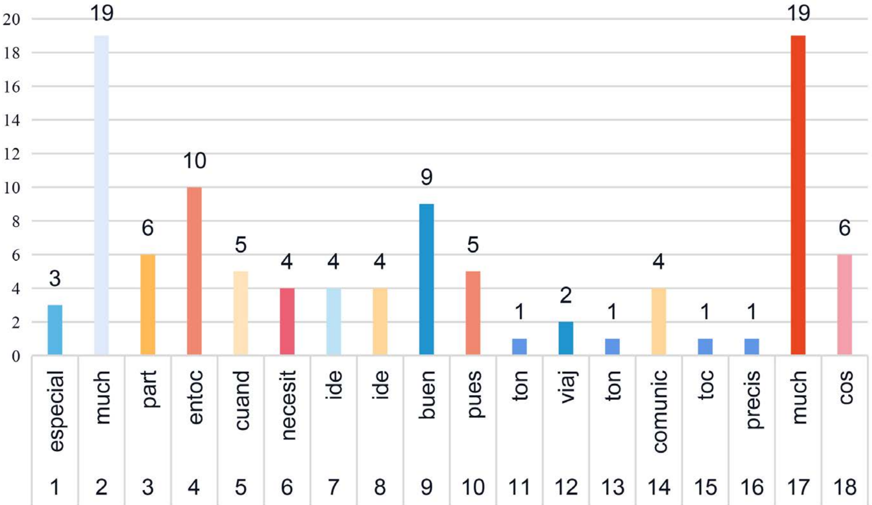


Figure 7. Number of document where each tokern appears.

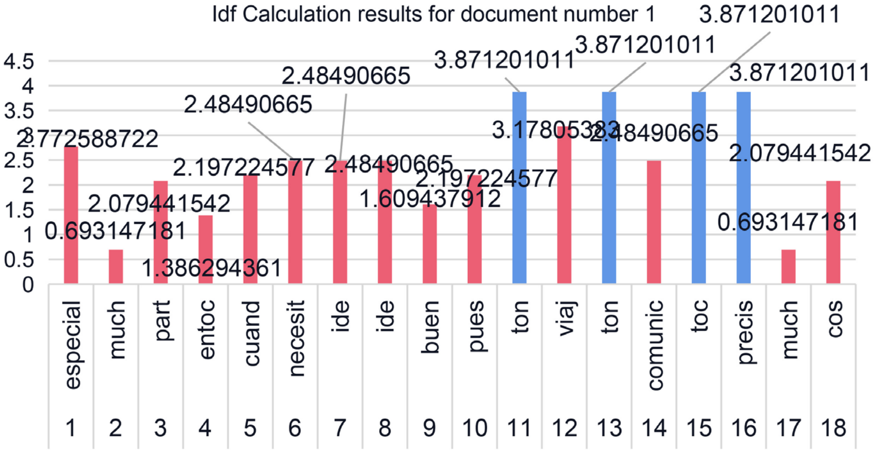


Figure 8. Idf for document number 1.

of 2431 tokens, corresponding to a total of 47 different documents. The results of document number 1 were the only ones exemplified.

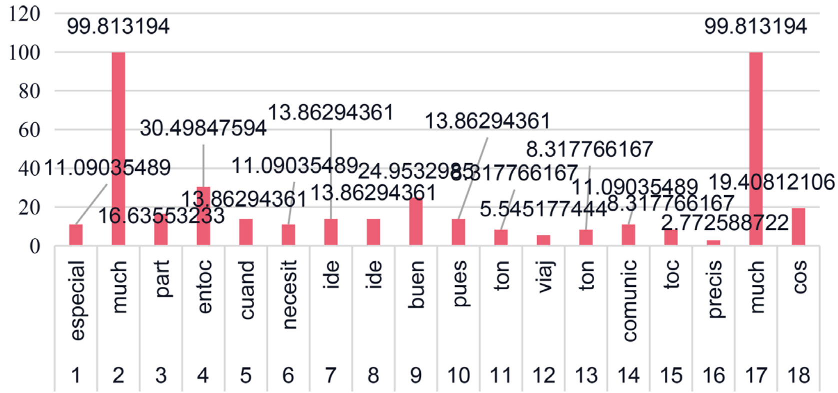


Figure 9. Tf-Idf calculation results for document number 1.

To check the performance of the EATS, a test was performed with a 90% Threshold value for the 47 documents, results are shown below.

5.7. Results of Calculation of Terms with a Threshold Value of 90%

The result of this type of Automatic Extractive Text Summarization calculation is due to the fact that it allows extracting a subset of the text with greater importance. Figure 10 shows the calculations results with a Threshold of 90%, it can be observed that the lowest value was document number 25 with a value of 4.62 and the maximum value was document number 12 with a value of 35.41.

5.8. Results of Values Equal to or Greater than a Threshold of 90%

When performing the calculation with a Threshold of 90%, a value of 31.87 was obtained. This value was compared with the values in Figure 10, document number 12 having a value of 35.41, therefore this is the most representative document. Figure 11 shows the result of the comparison with a threshold of 90%.

5.9. Final Result with a Value Equal to or Greater than a Threshold of 90%

Figure 12 shows the final result with a value equal to or greater a Threshold of 90%.

5.10. Final Results of Values Equal to or Greater than a Threshold of 35%

When the calculation was performed with a Threshold value of 35%, the value 12.39 was obtained. That value was compared with the values in Figure 10 being the quotes number 0, 1, 2, 3, 3, 4, 5, 5, 6, 7, 8, 8, 9, 10, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 22, 23, 24, 24, 27, 28, 28, 29, 30, 31, 32, 33, 34, 35, 37, 37, 38, 39, 40, 40, 41, 42, 44, 45. Quite a few citations were obtained because the Threshold value was small therefore, they fall within the range greater than or equal

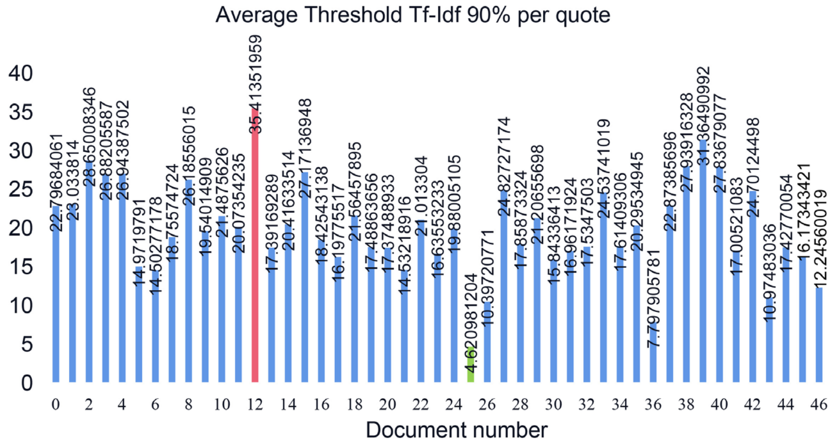


Figure 10. Average threshold of 90% per document.

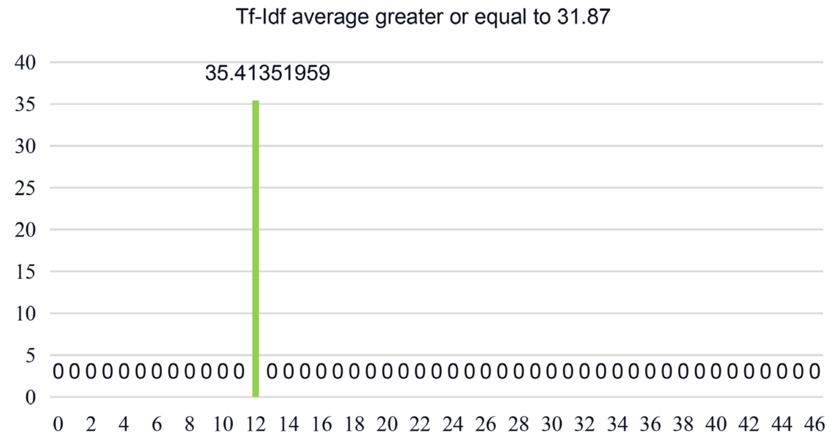


Figure 11. Result of the comparison with a threshold of 90%.

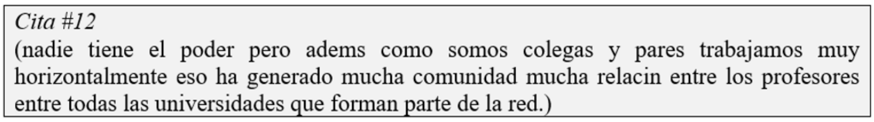

Cita #12
(nadie tiene el poder pero adems como somos colegas y pares trabajamos muy horizontalmente eso ha generado mucha comunidad mucha relacin entre los profesores entre todas las universidades que forman parte de la red.)

Figure 12. Final result with a value equal to or regreater a threshold of 90%.

to the Threshold. Figure 13 shows the result of the comparison with a Threshold value of 35%.

5.11. Final Results of Values Equal to or Greater than a Threshold of 35%

Figure 14 shows an example of the final result with a value equal to or greater than a Threshold of 35%.

6. Software and Hardware Specifications

6.1. Software Specifications

Windows 10, 20H2 version (OS Build 19042.630). Visual Microsoft Visual Studio Community 2019, version 16.7.7.

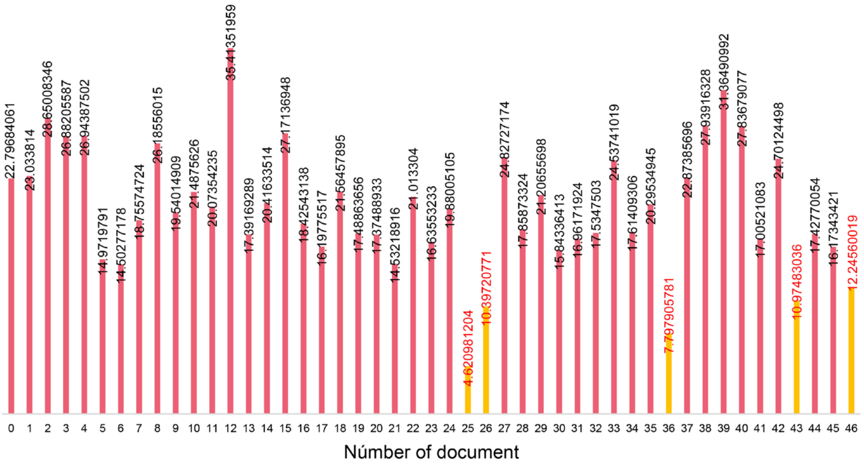
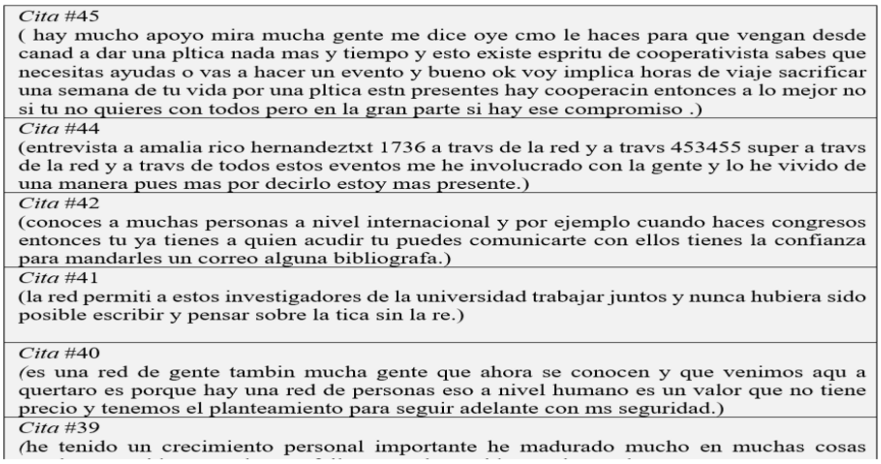


Figure 13. Tf-Idf value higher or greater than 35% of threshold or 12.39.



Cita #45
(hay mucho apoyo mira mucha gente me dice oye cmo le haces para que vengan desde canad a dar una pltica nada mas y tiempo y esto existe espritu de cooperativista sabes que necesitas ayudas o vas a hacer un evento y bueno ok voy implica horas de viaje sacrificar una semana de tu vida por una pltica estn presentes hay cooperacin entonces a lo mejor no si tu no quieres con todos pero en la gran parte si hay ese compromiso .)

Cita #44
(entrevista a amalia rico hernandeztxt 1736 a travs de la red y a travs 453455 super a travs de la red y a travs de todos estos eventos me he involucrado con la gente y lo he vivido de una manera pues mas por decirlo estoy mas presente.)

Cita #42
(conoces a muchas personas a nivel internacional y por ejemplo cuando haces congresos entonces tu ya tienes a quien acudir tu puedes comunicarte con ellos tienes la confianza para mandarles un correo alguna bibliografa.)

Cita #41
(la red permiti a estos investigadores de la universidad trabajar juntos y nunca hubiera sido posible escribir y pensar sobre la tica sin la re.)

Cita #40
*(*es una red de gente tambin mucha gente que ahora se conocen y que venimos aqu a quertaro es porque hay una red de personas eso a nivel humano es un valor que no tiene precio y tenemos el planteamiento para seguir adelante con ms seguridad.)

Cita #39
*(*he tenido un crecimiento personal importante he madurado mucho en muchas cosas

Figure 14. Example of the final result with a value equal to or greater than a threshold of 35%.

6.2. Hardware Specifications

Laptop Intel (R) Core^{TM} i7-8750 H CPU A 2.20Ghz. NVIDIA GeForce GTX 1650 graphics card. 32 Gb Ram memory.

7. Conclusions

The application and execution of text mining (TM) preprocesses are fundamental for the optimal functioning of EATS, the application of a spell checker promotes a decrease in the execution time of the general process of TM preprocesses.

The applications of the pre-process provide the necessary elements for text structuring, in addition, they facilitate the counting of those similar elements for obtaining Tf of the tokens.

With the application of the Tf-Idf method, mathematical operations of statistical type are generated which results in request to be stored, product of the division of the analyzed document to a subset of independent documents. For such reason it is recommended the creation and use of jagged arrays since with their characteristic of having unequal rows and columns facilitate the organization and storage, favoring the speed in the manipulation, search, route, insertion of the results of the mathematical operations.

The Tf-Idf method allows weighting the vectors, the used approach shows that it is an effective and functional method, the characteristics of the text depend on the language used, in this research it was adapted for Spanish language, working correctly. The value of Idf demonstrates the importance of the token within the documents.

The EATS process is a laborious process, but is performed in a careful way the results are satisfactory and clear and it can also be applied to a single document or several documents.

The values obtained through the Tf-Idf method allow to compare results using the comparing value will depend on the Threshold level requested by the user.

EATS aims to show the most relevant text according to the Threshold measure requested by the reader thus with this investigation it is verified that it works correctly way, with the experimental results presented, it is verified that texts can be summarized through the Tf-Idf method efficiently.

The main idea of EATS is to help reducing the total reading time, the final results are statistical in nature but the results are presented showing the most important text. The performance of the execution of this model is not affected on computers with low hardware resources, it does not consume large memory resources so it can be applied on any computer.

All the EATS examples carried out in this research have been applied for Spanish language and the results are satisfactory, therefore the method works correctly.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Nallapati, R., Zhai, F. and Zhou, B. (2016) SummaRuNNer: A Recurrent Neural Network Based Sequence Model for Extractive Summarization of Documents. *Proceedings of the* 31*st AAAI Conference on Artificial Intelligence* (*AAAI*-17). https://arxiv.org/abs/1611.04230

[2] El-Kassas, W.S., Salama, C.R., Rafea, A.A. and Monhamed, H.K. (2020) Automatic Text Summarization: A Comprehensive Survey. *Expert Systems with Applications*, **165**, 113679. https://doi.org/10.1016/j.eswa.2020.113679

[3] Rush, M.A., Chopra, S. and Weston, J. (2015) A Neural Attention Model for Sentence Summarization. *Proceeding of the* 2015 *Conference on Empirical Methods in Natural Language Processing*, Lisbon, September 2015, 379-389. https://doi.org/10.18653/v1/D15-1044

[4] Castañeda, N.H., Hernández, R.A.G., Ledeneva, Y. and Castañeda, A.H. (2020) Evolutionary Automatic Text Summarization Using Cluster Validation Indexes.

Hernández Castañeda, **24**, 583-595. https://doi.org/10.13053/cys-24-2-3392

[5] Quillo-Espino, J. and Romero-Gonzalez, R.-M. (2021) Where Are the Automatic Text Summaries Located in the 2021? A Review. *International Journal of Advanced Research in Computer and Communication Engineering*, **10**, 11-16. https://doi.org/10.17148/IJARCCE.2021.10402

[6] Quillo-Espino, J., Romero-Gonzalez, R.M. and Lara-Guevara, A. (2018) Advantages of Using a Spell Checker in Text Mining Preprocesses. *Journal of Computer and Communications*, **6**, 43-54. https://doi.org/10.4236/jcc.2018.611004

[7] Gaikwad, S.V., Chaugule, A. and Patil, P. (2013) Text Mining Methods and Techniques. *International Journal of Computer Applications*, **85**, 42-45. https://doi.org/10.5120/14937-3507

[8] alZahir, S., Fatima, Q. and Cenek, M. (2015) New Graph-Based Text Summarization Method. 2015 *IEEE Pacific Rim Conference on Communications*, *Computers and Signal Processing* (*PACRIM*), Victoria, 24-26 August 2015, 396-401. https://doi.org/10.1109/PACRIM.2015.7334869

[9] Nenkova, A. (2012) A Survey of Text Summarization Techniques. In: Aggarwal, C. and Zhai, C., Eds., *Mining Text Data*, Springer, Boston, 43-76. https://doi.org/10.1007/978-1-4614-3223-4_3

[10] Kaikhah, H. (2004) Automatic Text Summarization with Neural Networks. 2004 2*nd International IEEE Conference on* "*Intelligent Systems*", Varna, 22-24 June 2004, 40-44.

[11] Meena, Y.K., Jain, A. and Gopalani, D. (2014) Survey on Graph and Cluster Based Approaches in Multi-Document Text Summarization. *International Conference on Recent Advances and Innovations in Engineering* (*ICRAIE*-2014), Jaipur, 9-11 May 2014, 1-5. https://doi.org/10.1109/ICRAIE.2014.6909126

[12] Manjari, K.U., Rousha, S., Sumanth, D. and Devi, J.S. (2020) Extractive Text Summarization from Web Pages Using Selenium and TF-IDF Algorithm. 2020 4*th International Conference on Trends in Electronics and Informatics* (*ICOEI*), Tirunelveli, 15-17 June 2020, 648-652. https://doi.org/10.1109/ICOEI48184.2020.9142938

[13] Kyoomarsi, F., Hhosravi, H., Eslami, E., Dehkordy, K.P. and Tajoddin, A. (2008) Optimizing Text Summarization Based on Fussy Logic. 7*th IEEE*/*ACIS International Conference on Computer and Information Science*, Portland, 347-352.

[14] Bidoki, M., Monsavi, M.R. and Fakhramahgmad, M. (2020) A Semantic Approach to Extractive Multi-Document Summarization: Applying Sentence Expansion for Tuning of Conceptual Densities. *Information Processing & Management*, **57**, 102341. https://doi.org/10.1016/j.ipm.2020.102341

[15] Verma, P. and Om, H. (2016) Extraction Based Text Summarization Methods on User Review Data: A Comparative Study. In: Unal, A., Nayak, M., Mishra, D.K., Singh, D. and Joshi, A., Eds., *Smart Trends in Information Technology and Computer Communications*, Springer, Singapore, 346-354. https://doi.org/10.1007/978-981-10-3433-6_42

[16] Rojas, S.J., Ledeneva, Y. and Garcia-Hernandez, R.A. (2018) Calculating the Significance of Automatic Extractive Text Summarization Using a Genetic Algorithm. *Journal of Intelligent & Fuzzy Systems*, **35**, 293-304. https://doi.org/10.3233/JIFS-169588

[17] Abuobieda, A., Salim, N., Albaham, A.T., Osman, A.H. and Kumar, Y.J. (2012) Text Summarization Features Selection Method Using Pseudo Genetic-Based Model. International Conference on Information Retrieval & Knowledge Management, Kuala Lumpur, 13-15 March 2012, 193-197.

https://doi.org/10.1109/InfRKM.2012.6204980

[18] Zhong, M., Liu, P., Chen, Y., Wang, D., Qiu, X. and Huang, X. (2020) Extractive Summarization as Text Matching. *Proceedings of the* 58*th Annual Meeting of the Association for Computational Linguistics*, July 2020, 6197-6208. https://doi.org/10.18653/v1/2020.acl-main.552

[19] Zhou, Q., Wei, F. and Zhou, M. (2020) At Which Level Should We Extract? And Empirical Analysis on Extractive Document Summarization. *Proceedings of the* 28*th International Conference on Computational Linguistics*, Barcelona, December 2020, 5617-5628. https://doi.org/10.18653/v1/2020.coling-main.492

[20] Muneera, M.N. and Sriramya, P. (2020) Extractive Text Summarization for Social News Using Hybrid Techniques in Opinion Mining. *International Journal of Engineering and Advanced Technology*, **9**, 2109-2115. https://doi.org/10.35940/ijeat.B3356.029320

[21] Salton, G. and Buckley, C. (1988) Term-Weighting Approaches in Automatic Text Retrieval. *Information Processing & Management*, **24**, 513-523. https://doi.org/10.1016/0306-4573(88)90021-0

[22] Liu, C.-Z., Sheng, Y.-X., Wei, Z.-Q. and Yang, Y.-Q. (2018) Research of Text Classification Based on Improved TF-IDF Algorithm. 2018 *IEEE International Conference of Intelligent Robotic and Control Engineering* (*IRCE*), Lanzhou, 24-27 August 2018, 218-222. https://doi.org/10.1109/IRCE.2018.8492945

[23] Qaiser, S. and Ali, R. (2018) Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents. *International Journal of Computer Applications*, **181**, 25-29. https://doi.org/10.5120/ijca2018917395

[24] Khusna, N.A. and Agustina, I. (2018) Implementation of Information Retrieval Using Tf-Idf Weighting Method on Detik.Com's Website. 12*th International Conference on Telecommunication Systems*, *Services*, *and Applications* (*TSSA*), Yogyakarta, 4-5 October 2018, 1-4. https://doi.org/10.1109/TSSA.2018.8708744

[25] Larocca, N.J. and Santos, D.A. (2000) Document Clustering and Text Summarization. https://orcid.org/0000-0001-9825-4700