Cosine Similarity – A Computing Approach to Match Similarity Between Higher Education Programs and Job Market Demands Based on Maximum Number of Common Words

https://doi.org/10.3991/ijet.v17i12.30375

Ylber Januzaj[1(✉)], Artan Luma[2]
[1] Faculty of Economic, University "Isa Boletini", Mitrovica, Kosovo
[2] CST Faculty, South East European University, Tetovo, Northern Macedonia
ylber.januzaj@umib.net

Abstract—Comparing textual content is becoming more and more problematic due to the fact that nowadays data is very dynamic. The application of sophisticated methods enables us to compare how similar the documents are to each other. In our research we apply the Cosine Similarity method to compare the similarity of several documents with each other. We also apply the TF-IDF technique which enables us to normalize the results. Normalization of these results is necessary for the fact that there are some words that are repeated several times and from this repetition is determined their importance. Finally we can see a comparison between the similarity of documents with normalized and non-normalized results. As can be seen in the research, the normalization of results has a great value in comparing documents with textual content.

Keywords—machine learning, cosine similarity, tf-idf, document similarity

1 Introduction

Nowadays the volume of information which is found in electronic media is very dynamic, and increases from second to second. Such a large dynamic data, requires better management, where one of them is the comparison of similarities between documents. Documents that have a numerical content are not problematic to compare with each other, but when it comes to documents that have a textual content, then it becomes more difficult. There are different techniques of text similarity checker such as: Cosine similarity, Jaccard similarity and Euclidian distance, and each of them has its own characteristics [1-3].

Based on Njeru et al. [4] has argued that application of Cosine Similarity is required while analyzing and comparing textual content documents. Cosine similarity, from the contrast with other techniques that deals with the binary data of one set, addresses the match similarity between multiple textual documents [5-7].

In our research, first is present the way how Cosine similarity works by measuring the content of different textual documents. There are used four different documents with different words, in order to compare with each other and to measure the similarity.

The conversion of the words into a vector is more than necessary, in order to measure the distance between two vectors that are projected into the X and Y axes. Based on the shape of the cosine value, two vectors have the same orientation with cosine 1, and two vectors with completely opposite directions have cosine -1. While the vector having 90 degrees with each other, one in the X axis and the other in the Y axis at the cosine is 0. Cosine similarity as a measurement method of similarities between the documents varies between positive values of 0 to 1. The values that the vectors receive depend on the frequency. As long as these words appear in the document, the cosine similarity method allows us to find similarities between these vectors. Song et al. [8] presented that the number of vectors that can be located in the multidimensional space is not limited, as is the number of comparisons between these vectors.

2 Cosine similarity

The similarity between two vectors is measured using Cosine Similarity technique [9-10]. The way how it works, is by measuring the cosine of the angle between two documents which are expressed in vectors. Haskova et al. [11] stated that the angle between vectors, determines whether they are pointing in the same or different directions.

If vectors are pointing in the same directions, it means that documents are similar, the closer they are expressed on the axis, the more similar they are. Vice versa, the farther they are expressed on the axis the less similar they are.

Each document which is used for comparison, is presented in space as a plot, and cosine similarity captures the orientation of these plots. Based on recent studies [12,13], by using Cosine Similarity determination of document orientation is done, and in order to determine the quantity of any documents other techniques can be applied.

Below, documents are plotted in the multidimensional step, with identical identities, the smallest similarity, and the similarity of 0 among the words.

Figure 1. shows the case of two documents that are located in the multidimensional space at the same point. As mentioned above, the position of the vector in the multidimensional space between the points x and y depends on the expression of the words which are in their entirety. In this case, is a large number of documents, since there is no question as to how much distance the documents are in their own, but the relevant ones are between them in the range of 0. As a conclude, the smallest is the distance between documents, the bigger is the similarity between them, always approaching the value 1.

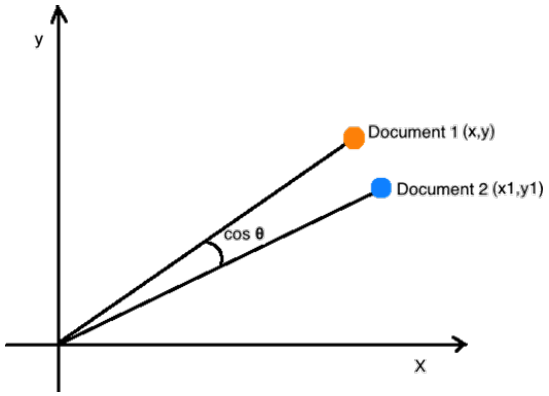

Fig. 1. Similar documents placed on multidimensional space at the same point

In Figure 2., there is a case when two documents that are not similar to each other, and the distance between them is bigger. As you can see, the distance between document 1 and document 2 is approximately 90 degrees. And the bigger the distance between the two documents the more, the smaller is similarity between the documents.

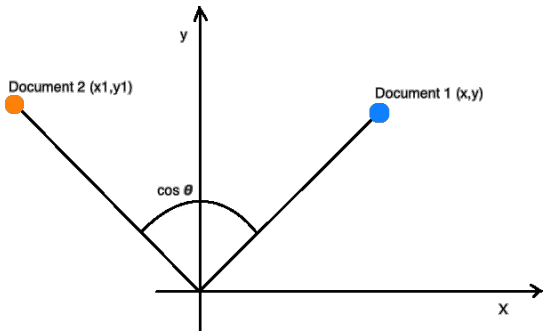

Fig. 2. Less similar documents placed on multidimensional space

Figure 3. shows two documents that are completely opposite to each other. As well as the first document is placed in the x and y axes at positive values, while the second document is placed at the negative values of the x and y constrains. Since the two documents are placed in diametrically opposite dimensions, the number of documents is larger, which can reach 180 degrees. With such a large angle, after calculating the cosine, its value will be negative. And when negative values occur during the calculation of cosine similarity, then the similarity between them is calculated as the value of the value 0.

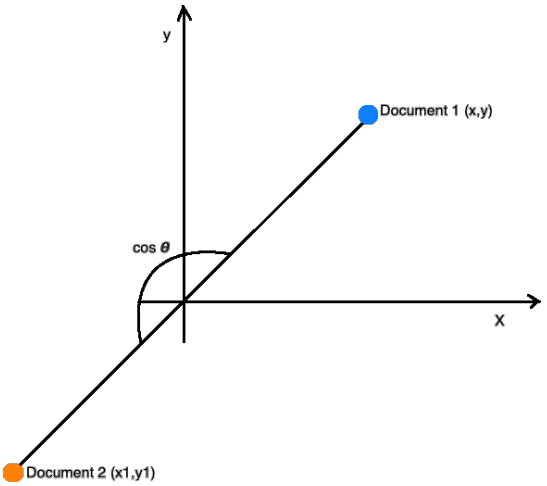

Fig. 3. Two completely different documents placed on multidimensional space

3 Methodology

In this is shown how Cosine Similarity works in order to calculate similarity between multiple different documents. The Cosine Similarity illustration will be used by several different documents that include different sentences within. The documents to be used are:

- **Document 1: Market demands are met by university programs.**
- **Document 2: University programs should be very appropriate.**
- **Document 3: Market demands are too high.**
- **Document 4: Students are interested in our university programs.**

Four different documents that will be anonymous compared with each other in order to calculate the similarity between them.

In order to continue with calculations of Cosine Similarity, ordering the words is necessary, ensuring that the words will not be repeated. The words that are used in the above-mentioned documents are: appropriate, are, be, by, demands, filled, for, high, interested, market, our, programs, should, students, too, university, very. Each of the listed words will be checked for how many times is mentioned in each document. Variously than each word has been presented, it will also be the value that it gains in placing it in the multidimensional space.

Once the table has been created and the numerical values obtained for each word, the mathematical calculations of cosine similarity will be made. The values that are obtained are presented as vector for each document. Therefore, each document will be a vector, and once the results are obtained, the calculation will be similar to each of the documents. At the end of the calculation, application of TF-IDF method contributes in results change because you are given a smaller value of the many words that are written. This helps a lot because it can be said that in the documents that are the calculations

there are words that are written and, as mentioned above, and the fact that the words are written does not mean that the words are the most relevant. Below a table is presented with words that are used in four documents.

In the Table 1, all the words used in the above-mentioned documents have been presented along with the number of times these words have been mentioned in all documents. There are different occasions when one word is presented in one of the four documents, and there are cases when a word is presented in three documents out of four. Since there are numerical values, parse of vector values for each document is done, because each document can have vector values. The resulting vector values will then be applied with the formula that computes the approximation of cosine similarity

Table 1. Ordered words of all documents

	D1	**D2**	**D3**	**D4**
Appropriate	0	0	0	1
Are	1	1	1	0
Be	0	0	0	1
By	1	0	0	1
Demands	1	0	1	0
Filled	1	0	0	0
For	0	1	0	0
High	0	0	1	0
Interested	0	1	0	0
Market	1	0	1	0
Our	0	1	0	0
Programs	1	1	0	1
Should	0	0	0	1
Students	0	1	0	0
Too	0	0	1	0
University	1	1	0	1
Very	0	0	0	1

Table 2. All words converted on vectors

Vector value	
D1	{0,1,0,1,1,1,0,0,0,1,0,1,0,0,0,1,0}
D2	{0,1,0,0,0,0,1,0,1,0,1,1,0,1,0,1,0}
D3	{0,1,0,0,1,0,0,1,0,1,0,0,0,0,1,0,0}
D4	{1,0,1,1,0,0,0,0,0,0,0,1,1,0,0,1,1}

The following table presents the value of each document that is presented as a vector based on the translation of the words in each document. The calculation of cosine similarity is done as follows:

$$cos\ similarity = \frac{D \cdot B}{||D||\ ||B||} \tag{1}$$

As can be seen in the equation (Eq. (1)), cosine similarity represents the division between the two product variables, in this case Di and Bi, and the multiplicity of the swatch vector variables obtained for both documents. Respectively, the formula above will be blurred and after the breakthrough it will take this form.

$$\frac{\sum_{i=1}^{n} Di\ Bi}{\sqrt{\sum_{i=1}^{n} Di2}\ \sqrt{\sum_{i=1}^{n} Bi2}} \tag{2}$$

Based on equation (Eq. (2)), which computed the values of the highest obtained vector, similarity between the most recent documents is presented. Below, the calculation of the first and second documents is done, then application of the procedure calculation for all the documents, in the end will appear next to the table that contains all the results between each document.

$$cos\ similarity = \frac{D1 \cdot D2}{||D1||\ ||D2||} \tag{3}$$

$$\frac{[0,1,0,1,1,1,0,0,0,1,0,1,0,0,0,1,0]*[0,1,0,0,0,0,0,1,0,1,0,1,1,0,1,0,1,0]}{Sqroot(1+1+1+1+1+1+1+1)*Sqroot(1+1+1+1+1+1+1+1)} \tag{4}$$

In equation (Eq. (3,4)) Calculation of dot product between the vector of the first and the second document vectors. This value is divided by the square root of each of these values. In the square root, the values are set to 1 because of the space in the document so that all the formulas are presented to us. After calculating:

$$\frac{1*1+1*1+1*1}{\sqrt{7}*\sqrt{7}} \tag{5}$$

In addition, more than just counts numbers greater than 0 because of the space you need for all the values. It is also the calculation of the values by which the value gained with the square root of the number 7 is multiplied by the same value. Since both of the two documents have a vector amount of 7, their calculation is then calculated and then their output. After the calculation:

$$\frac{3}{2.6*2.6} = \frac{3}{6.67} = 0.44 \tag{6}$$

As well as the value obtained after calculating the cosine similarity between the first and the second document is 0.44, which can be calculated as a similarity between the two documents of 44%. Next, calculation of all the documents with each other to present one of the results of the table in the form of one matrix.

Table 3. Calculated values without TF-IDF technique

	D1	**D2**	**D3**	**D4**
D1	1	0.44	0.51	0.59
D2	0.44	1	0.17	0.34
D3	0.51	0.17	1	0
D4	0.59	0.34	0	1

Based on the calculations that have been made for all the above-mentioned documents, the values of cosine similarity for all the documents have been presented. Just as it can be seen in the table above, the greater similarity is between the first and the second document. This percentage reaches a value of 0.59 of cosine similarity, or else it may appear as 59% of the similarity between them. A small percentage of adjustment is between the first and the third document, where a similarity of 51% is gained. On the other hand, two documents that did not appear to be similar to each other, based on the obtained results, are the third document and the second document since it has rated 0.

Below is shown the analysis of cosine similarity between these documents but with the application of the TF-IDF in order to make the normalization of the data. This normalization is even higher when there are some words that have been written many times in the text. Therefore, you will be given the smallest number of words you need when using the TF-IDF methods.

Initially, the calculation of TF or term frequency is calculated by calculating how many times each word appears in the report with all the words in the document. The number of words that are enclosed within a document is then used to calculate its logarithm. The logarithm of these values is known as IDF, and once both values have been found, the product is calculated between them in the order of the TF-IDF.

This value is considered as a normalized value and smaller compared to the direct values that were later calculated in the cosine similarity. The next step is to calculate the cosine similarity with the values that are normalized, and of course, the final values that will be gained in our calculations will be different and smaller compared to the previous values. A large difference can be found in documents with a large number of spelled words.

The following table presents the TF calculation for each word, which represents the number of words that is mentioned in relation to the total number of words in the document. In the first and second documents, there are seven words, while in the third document, there are have six words and in the fourth document, there are six words. The number of each word is divided by the total number of these words. Next, the calculation of IDF or inverse document frequency will be done, which in turn directly calculates the weight of the word swatch as much as it is relevant.

Calculating the inverse document frequency will be calculated by the sum of all values for each word. In this case, the calculation of the values of the first document will be calculated.

Table 4. TF calculation for all words of the documents

TF calculation				
	D1	**D2**	**D3**	**D4**
Appropriate	0/7	0/7	0/5	1/6
Are	1/7	1/7	1/5	0/6
Be	0/7	0/7	0/5	1/6
By	1/7	0/7	0/5	1/6
Demands	1/7	0/7	1/5	0/6
Filled	1/7	0/7	0/5	0/6
For	0/7	1/7	0/5	0/6

High	0/7	0/7	1/5	0/6
Interested	0/7	1/7	0/5	0/6
Market	1/7	0/7	1/5	0/6
Our	0/7	1/7	0/5	0/6
Programs	1/7	1/7	0/5	1/6
should	0/7	0/7	0/5	1/6
students	0/7	1/7	0/5	0/6
Too	0/7	0/7	1/5	0/6
university	1/7	1/7	0/5	1/6
Very	0/7	0/7	0/5	1/6

Table 5. IDF calculation for all words of the documents

	IDF calculation	**Log**	**Log**
appropriate	log(4/1)	log(4)	0.6
Are	log(4/3)	log(1.3)	0.1
Be	log(4/1)	log(4)	0.6
By	log(4/2)	log(2)	0.3
demands	log(4/2)	log(2)	0.3
Filled	log(4/1)	log(4)	0.6
For	log(4/1)	log(4)	0.6
high	log(4/1)	log(4)	0.6
interested	log(4/1)	log(4)	0.6
market	log(4/2)	log(2)	0.3
Our	log(4/1)	log(4)	0.6
programs	log(4/3)	log(1.3)	0.1
should	log(4/1)	log(4)	0.6
students	log(4/1)	log(4)	0.6
Too	log(4/1)	log(4)	0.6
university	log(4/3)	log(1.3)	0.1
very	log(4/1)	log(4)	0.6

In the next table can be used to calculate all the values that are now readily computed with the values earned by TF. The formula to be used for the calculation of TF-IDF is the above- mentioned formula.

$$TF - IDF = TF(SCORE) * IDF\ (SCORE) \tag{7}$$

Based on equation (Eq. (7)), the TF score for each word is gained. It will be multiplied by the IDF score, which was obtained by the IDF calculation. In this case, for the first "proper" keyword that has a TF value of 0/7 in the first document, and IDF of 0.6, the result is 0, since the output between the two values gives us the value of 0. For all other documents the values are 0 as with the calculation without TF-IDF, while the result of the fourth document is 0.1, since the calculation between is 1/6 * 0.6. When

comparing the value obtained without TF-IDF, which was 1, in this case the value is 0.1.

Table 6. Normalized results for the words after applying TF-IDF technique

		TF – IDF calculation		
	D1	***D2***	***D3***	***D4***
appropriate	0	0	0	0.1
Are	0.014	0.014	0.02	0
Be	0	0	0	0.1
By	0.04	0	0	0.05
Demands	0.04	0	0.06	0
filled	0.08	0	0	0
For	0	0.08	0	0
high	0	0	0.12	0
interested	0	0.08	0	0
market	0.04	0	0.06	0
Our	0	0.08	0	0
programs	0.01	0.01	0	0.01
should	0	0	0	0.1
students	0	0.08	0	0
Too	0	0	0.1	0
university	0.01	0.01	0	0.01
very	0	0	0	0.1

For each word in each document, there are values that are normalized and weighed less than the first case when the TF-IDF was not applied. Gained values are positive values from 0 to 0.1, compared to the first case when the values were 1 and 2. As mentioned above, these changes are even more emphasized when a number of twelve great for words that are inside a document. Since there are gained normalized values, calculation of the cosine similarity with the values so that the similarity between the documents is done.

$$\frac{[0.014,0.01,0.01]*[0.014,0.01,0.01]}{Sqroot(0.014+0.2+0.02)*Sqroot(0.014+0.32+0.02)} \tag{8}$$

After calculating the values:

$$\frac{0.014*0.014+0.01*0.01+0.01*0.01}{\sqrt{0.23}*\sqrt{0.35}} \tag{9}$$

After calculating the values:

$$\frac{0.000396}{0.27} = 0.014 * 100 = 0.14 \tag{10}$$

In Table 7, you can see the final results after the normalization of the data by TF-IDF techniques. As you can see, the bigger value that was 59% in the calculation after

TF-IDF, after normalization this value has rush to 0.7. This has happened because the words that have a great deal of significance have been lowered by their weight being accounted for as less significant. There are also documents that did not have the same D3 and D4 resemblance also resulted in 0%. While fewer documents are the D1 and D3 documents with 0.18, as well as D2 and D3 with 0.17.

Table 7. Cosine similarity with the normalized results after applying TF-IDF technique

	D1	**D2**	**D3**	**D4**
D1	1	0.14	0.18	0.7
D2	0.14	1	0.17	0.34
D3	0.18	0.17	1	0
D4	0.7	0.34	0	1

4 Conclusion

Extracting accurate results from the processing of documents with textual content is of great importance. The application of techniques in comparison to documents is usually applied by the type of analysis which must be done. As mentioned above, when dealing with comparability between two or more documents the cosine similarity technique is applied. In our research is shown how textual content is converted to vector. These vector values later are used to do mathematical calculations to highlight the similarities between multiple documents. Then application of data normalization in order to reduce the importance of the words that are used the most. And after normalization, there are derived satisfactory results in terms of similarity of documents. Therefore, as a conclusion, the application of cosine similarity techniques when dealing with document similarity is of great importance, and different algorithms are managed to implement it in different computer languages.

5 References

[1] Agaoglu, M., (2016). "Predicting Instructor Performance Using Data Mining Techniques in Higher Education". *IEEE Access*. https://doi.org/10.1109/ACCESS.2016.2568756

[2] Member, B., G., Sheng, V., S., Tay, K., Y., Romano, W., Li, Sh., (2015). "Incremental Support Vector Learning for Ordinal Regression". *IEEE Transactions on Neural Networks and Learning Systems,* Vol. 26, No. 7. https://doi.org/10.1109/TNNLS.2014.2342533

[3] Xie, T., Zheng, Q., Zhang, W., Qu, H., (2017). "Modeling and Predicting the Active Video – Viewing Time in a Large – Scale E – Learning System". *IEEE Access*. https://doi.org/10.1109/ACCESS.2017.2717858

[4] Njeru, A., M., Omar, M., S., Yi, S., (2017). "IoTs for Capturing and Mastering Massive Data Online Learning Courses". *IEEE Computer Society*, ICIS, Wuhan, China. https://doi.org/10.1109/ICIS.2017.7959975

[5] Heartfield, R., Loukas, G., Gan, D., (2016). "You are probably not the weakest link: Towards Practical Prediction of Susceptibility to Semantic Social Engineering Attacks". *IEEE Access.* https://doi.org/10.1109/ACCESS.2016.2616285

[6] Fortuny, E., J., Martens, D., (2015). "Active Learning – Based Pedagogical Rule Extraction". *IEEE Transaction on Neural Network and Learning Systems*, Vol. 26, No. 11. https://doi.org/10.1109/tnnls.2015.2389037

[7] Mukhopadhyay, A., Bandyopadhyay, S., (2014). "A Survey of Multiobjective Evolutionary Algorithms for Data Mining: Part I". *IEEE Transaction on Evolutionary Computation*, Vol. 18, No. 1. https://doi.org/10.1109/TEVC.2013.2290086

[8] Song, Zh., Kusiak, A., (2009). "Optimization of Temporal Processess: A Model Predictive Control Approach". *IEEE Transaction on Evolutionary Computation*, Vol. 13, No. 1. https://doi.org/10.1109/TEVC.2008.920680

[9] Malgaonkar, S., Soral, S., Sumeet, Sh., Parekhji, T., (2016). "Study on Big Data Analytics Research Domain*". International Conference on Reliability, Infocom Technologies and Optimization ICRITO,Noida*, India. https://doi.org/10.1109/ICRITO.2016.7784952

[10] Anicic, K., P., Divjak, B., Arbanas, K., (2017). "Preparint ICT Graduates for Real – World Challenges: Results of a Meta – Analysis". *IEEE Transactions on Education,* Vol 60, No. 3. https://doi.org/10.1109/TE.2016.2633959

[11] Genc, Z., Babieva, N. S., Zarembo, G. V., Lobanova, E. V., & Malakhova, V. Y. (2021). The Views of Special Education Department Students on the Use of Assistive Technologies in Special Education. *International Journal of Emerging Technologies in Learning (iJET)*, *16*(19), pp. 69–80. https://doi.org/10.3991/ijet.v16i19.26025

[12] Jayakodi, K., Bandara, M., Perera, I., & Meedeniya, D. (2016). WordNet and Cosine Similarity based Classifier of Exam Questions using Bloom's Taxonomy. *International Journal of Emerging Technologies in Learning (iJET),* 11(04), pp. 142–149. https://doi.org/10.3991/ijet.v11i04.5654

[13] Karajeh, W., Hamtini, T. M., & Hamdi, M. (2016). Designing and Implementing an Effective Courseware for the Enhancement of e-Learning. *International Journal of Emerging Technologies in Learning (iJET),* 11(04), pp. 70–76. https://doi.org/10.3991/ijet.v11i04.5384

6 Authors

Ylber Januzaj is an Assistant Professor in University "Isa Boletini", Mitrovica, Faculty of Economic, Kosovo. He is Professor in the area of Informatics. He holds a PhD diploma on E-Technologies. His research interests are: Machine Learning, Computer Networks, and Database (email: ylber.januzaj@umib.net).

Artan Luma is a Full Professor in South East European University, CST Faculty, Tetovo, Northern Macedonia. He holds a PhD diploma on Computer Sciences. His research interests are: Computer security, Networking (email: a.luma@seeu.edu.mk).

Article submitted 2022-02-23. Resubmitted 2022-04-13. Final acceptance 2022-04-14. Final version published as submitted by the authors.