EVALUATION OF MODEL THAT MAKES COMPARISON BETWEEN MARKET DEMANDS AND UNIVERSITY CURRICULA OFFER

Ylber Januzaj, Artan Luma, Besnik Selimi, Bujar Raufi

Faculty of Contemporary Sciences and Technologies
South East European University
e-mails: {yj16535, a.luma, b.selimi, b.raufi}@seeu.edu.mk
North Macedonia

Abstract: Meeting the demands of the labor market has a great importance nowadays, when these demands are on the rise. The methodology that is used for comparison between labor market demands and university curricula is very important. The division of labor market demand corpus into clusters helps us identify the fit between labor market demand and university curricula. The analysis used in our research helped us to derive optimal numbers of clusters that will be compared to university curricula. During our publication we present the analysis of clusters created with the syllabuses of the region's universities. Second, we present the list of words that have the most similarities and the smallest ones. And finally we compare the similarity between the created clusters.

Key words: Clusters, Data mining, Job market, University curricula, Web scraping.

1. INTRODUCTION

In order to make a more precise comparison between the demands of the labor market and university syllabuses, during our publication we present analyzes which make the result of their similarity more accurate [2,5,6]. Two textual content papers have been produced, one containing vacancies in the field of technology and the other one with the syllabus of each university. These two documents will be compared in the context of the textual content they possess.

In order to make a more accurate calculation of the textual similarity between the two documents, a clustering will first be done to know exactly which words the syllabuses are most similar to.

And to know the optimal number of clusters that will be created and used for analysis will be used the silhouette score technique. Through this technique we determine the optimal number of clusters that our research will contain, and we

graphically represent the objects of each cluster to know what distance they have to each other.

Once we have created the clusters, we begin with the analysis of the university syllabuses and clusters that have been created. Of course, through this analysis we will know exactly what the syllabus words are most used, and what the syllabus words are missing. In this way, we will be able to make recommendations regarding the words that should be included in the university syllabus.

In the following we present the mathematical calculations that are applied in order to perform silhouette score analysis, and then we present graphs and examples that present the comparative analysis between different documents.

2. DETERMINING THE NUMBER OF CLUSTERS THROUGH SILHOUETTE ANALYSIS

The method that shows how close an object is to its cluster compared to the other cluster is known as Silhouette analysis [1, 3, 4]. According to [9, 10], the average obtained by silhouette analysis shows exactly the optimal number of clusters. A higher average indicates that the number ***k*** of the clusters is optimal, and indicates the optimal number of clusters.

The values that can be obtained after applying Silhouette analysis are from ***-1*** to ***+1***. The higher the value, the closer the object is to its cluster, and vice versa, the smaller the value that is acquired, the farther away is the object with its cluster. After calculating these values, an average is obtained which shows the optimal number of clusters [2].

Whether or not an object is aligned with its cluster, it can be measured in several forms, but in the case of silhouette analysis this is done using the Euclidean distance method [9].

In the following we present the mathematical calculations that are used to perform the Silhouette analysis.

According to [2, 9], we present a case using clusters ***CM*** and ***CN***, and compare the distance ***a*** between O_i and other objects in ***CM***, and the distance ***b*** to between O_i and other objects in ***CN***.

$$a(O_i) = \frac{1}{|C_M|-1} \sum_{O_j \in C_M,\, O_j \neq O_i} d(O_i, O_j) \tag{1}$$

$$b(O_i) = min C_N \neq C_M \frac{1}{|C_B|} \sum_{O_j \in C_N} d(O_i, O_j) \tag{2}$$

$$silhouette(O_i) = \frac{b(O_i) - a(O_i)}{\max\{a(O_i), b(O_i)\}} \tag{3}$$

In the above equations we calculate the average of silhouette analysis. As we can see, three equations are presented which contain the calculation steps, starting from the first step presented in the first equation. According to 1, ***a*** (O_i) is equal to the division between ***1*** and the absolute value of the cluster ***CM*** minus ***1*** and the sum of the distance between O_i and O_j where both are objects of a cluster, but could not be equal to each other.

Once the first cluster is computed, we must define the second cluster that in our case we define with $b(O_i)$. This cluster is the minimum distance of one of the objects of the first cluster, but it must never be part of the first cluster. As we can see in 2, it is equal to the minimum of the first cluster that is different from the second cluster. Then this minimum value is multiplied by ***1*** partition for the absolute value of ***CN*** which in this case is the second cluster. And this value is also reduced by the sum of the distance between the objects O_i and O_j, where O_j is an element of the second ***CN*** cluster.

Once the second cluster is defined, then in the third equation we do the silhouette calculation.

According to 3, the silhouette equals the division between the subtraction of $b(O_i)$ and $a(O_i)$ and the maximum value between the first cluster a (O_i) and the second cluster $b(O_i)$.

If we want to calculate the classification quality of a single object, then we can extend the last silhouette equation [7,8]. Below we present the equation of calculating the quality of all objects that are part of a cluster, as well as the case of calculating the quality of clusters one by one.

$$Silhouette(C_i) = \frac{1}{|C_i|} \sum_{O_j \in C_i} silhouette(O_j) \quad (4)$$

This equation presents the case of calculating the quality of all objects that are part of a cluster. As can be seen in 4, the silhouette equals the division of the value ***1*** and C_i, and the multiplication of this value by the sum of the silhouette objects (O_j), where the object (O_j) must be an element of the first cluster. While calculating the quality value for each cluster according to the equation below.

$$Silhouette(C) = \overline{silhouette(m)} = \frac{1}{m} \sum_{i=1}^{m} silhouette\ (C_i) \quad (5)$$

In 5, we calculate the quality value for each cluster individually. As we can see, silhouette ***(C)*** is equal to silhouette ***(m)*** which is a vinculum, that is, the set of all cluster values. And this value is equal to the division between ***1*** and ***m*** cluster, and the output of this value is the sum of the silhouette ***(Ci)***, where ***i*** starts from ***1*** to ***m*** which is the number of clusters that are defined in the system.

So, as can be seen, the calculation of the quality of the cluster and the objects that are part of the cluster, through silhouette analysis, can be done in a very precise

way through the mathematical calculations as mentioned earlier presented by Kaufman and Rousseeuw.

Next in Figure 1 we show silhouette analysis in graphical form that shows the number of clusters and the average for each of them.

Fig.1. Silhouette score versus number of clusters

Figure 1 shows the number of clusters relative to the silhouette score that each cluster has. As we can see in this graph, the number of clusters is from ***2*** to ***10***. The smallest silhouette score reached by the cluster set is ***0.27*** where we have two clusters. Then the value of silhouette score for 3 clusters is ***0.36***, for 4 clusters we have the value of ***0.41***. The highest value is with 5 clusters, where the silhouette score reaches ***0.78***, which represents the optimal number of clusters that our research should contain. For 6 clusters we have a decrease in silhouette score, where its value reaches ***0.48***, and for 7 clusters we have ***0.49***. As for the last two groups with 9 clusters and 10 clusters we have values of ***0.44*** and ***0.34***.

Such an analysis helps us very much to determine the number of clusters that our research will contain. According to this analysis, our labor market demand corpus will contain 5 clusters, as it is the highest value of the silhouette score which is achieved by our model.

Since we have the optimal number of clusters, we use these 5 clusters to compare them with the university syllabuses that are part of the analysis. Below we present the graph of the full score silhouette, where based on it we will be able to see which

cluster is closer to each other, and which cluster objects are more aligned with other cluster objects. Next in Figure 2 we show silhouette plot score of our vacancy corpus.

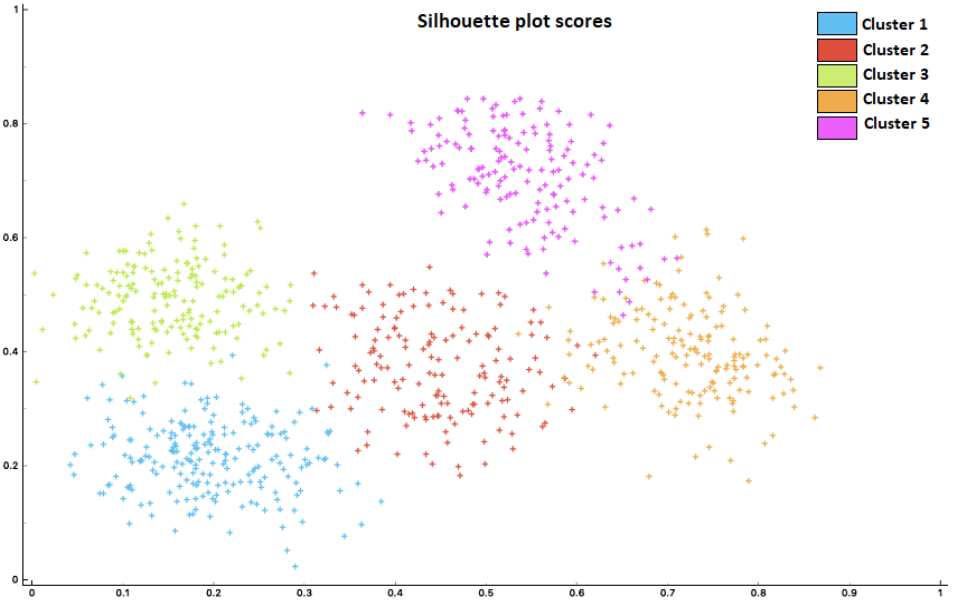

Fig.2. Silhouette plot score

Figure 2 shows the silhouette score after running the algorithm we built for such an analysis. As can be seen in Figure 2, only the optimal number of clusters defined by the system is shown here. So we have 5 clusters that are presented in the ***X*** and ***Y*** dimensions, where we can even guess which one is closest to the other. According to the graph we have the color split of all clusters, from ***1*** to ***5***, and we can also see which cluster is closest to the other. According to the graph we have a proximity between the objects of ***cluster 1*** and ***cluster 2***, since the distance between them is very small, but they are not equal to each other. There is also a small distance between ***cluster 1*** and ***cluster 3***, and between ***cluster 2*** and ***cluster 3***. We also have a small distance between ***cluster 2*** and ***cluster 5***, as well as ***cluster 3*** and ***cluster 5***, while there are large distances between ***cluster 5*** and ***cluster 1*** objects, as well as ***cluster 5*** and ***cluster 2***. According to the graph that made the representation of objects between clusters, there is no proximity between ***cluster 5*** objects and ***cluster 1***, and ***cluster 5*** and ***cluster 3***. Next we present the clusters created by the system to proceed with further analysis which validate our data accurately and efficiently.

2.1. Comparison of university syllabuses with created clusters

The inclusion of clustering in our research is very necessary as it helps us to identify the cluster of similarity words found in the syllabuses offered by universities in the field of technology.

Since in the previous chapter the analysis of the optimal number of clusters was done, then we have ready the number of clusters that will be used in our analysis. As shown in the graphs above, the number of clusters that will be part of our analysis will be 5, and each of them will be compared to the syllabus to find which words have the most similarity, and what are the words that have less similarity or are less mentioned in the university syllabus.

2.1. South East European University versus clusters

Figure 3 shows the relationship between clusters and similarity with the Computer Science syllabus of South East European University. As we can see, the textural similarity between the clusters and the syllabus ranges from ***0.036*** to ***0.078***. With the first cluster we have a similarity of ***0.036***, with the second cluster we have a similarity of ***0.038***. We have a greater similarity with the other two clusters since we have similarities of ***0.045*** and ***0.049***. And the highest similarity is achieved with the fifth cluster where we have a textual content similarity of ***0.078***. In Figure 4 we present the analysis that has been made between the University of Pristina and the clusters created by the system.

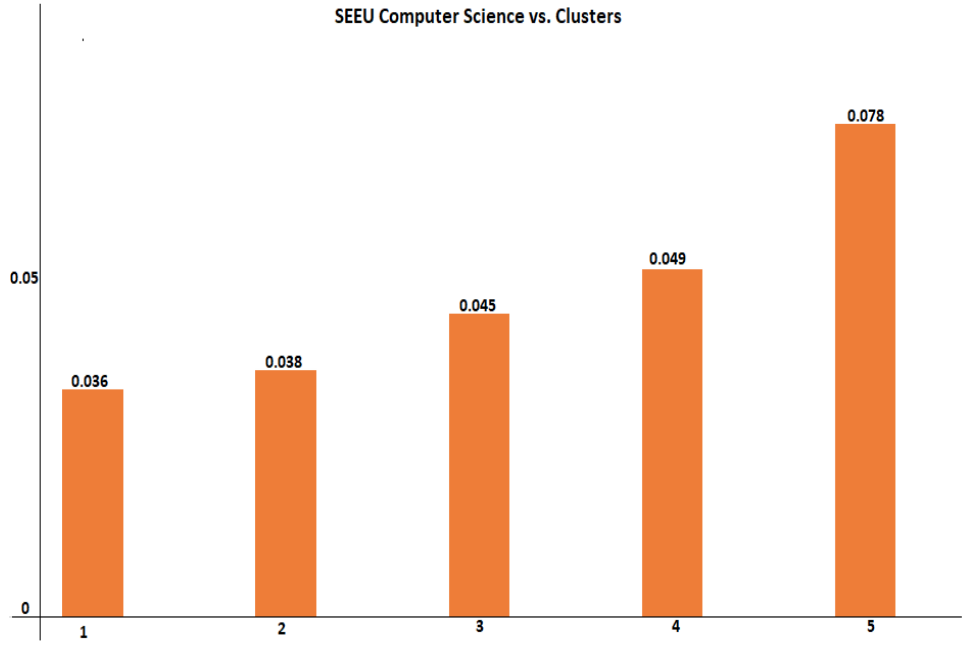

Fig.3. South East European University versus clusters

Figure 4 shows the analysis between the University of Pristina Computer Engineering syllabus and the clusters created by the system. In this analysis we have lower results than South East European University, since textural similarity ranges from ***0.026*** to ***0.060*** as the highest value. As we can see in figure 105, with the first

cluster we have a textural similarity of ***0.026***, also with the second cluster we have a similarity of ***0.026***. A greater similarity is achieved with the third and fourth clusters, since we have a similarity of ***0.035*** and ***0.038***. And the maximum value reached between the University of Pristina syllabus and clusters is ***0.060*** with the fifth cluster created by our system.

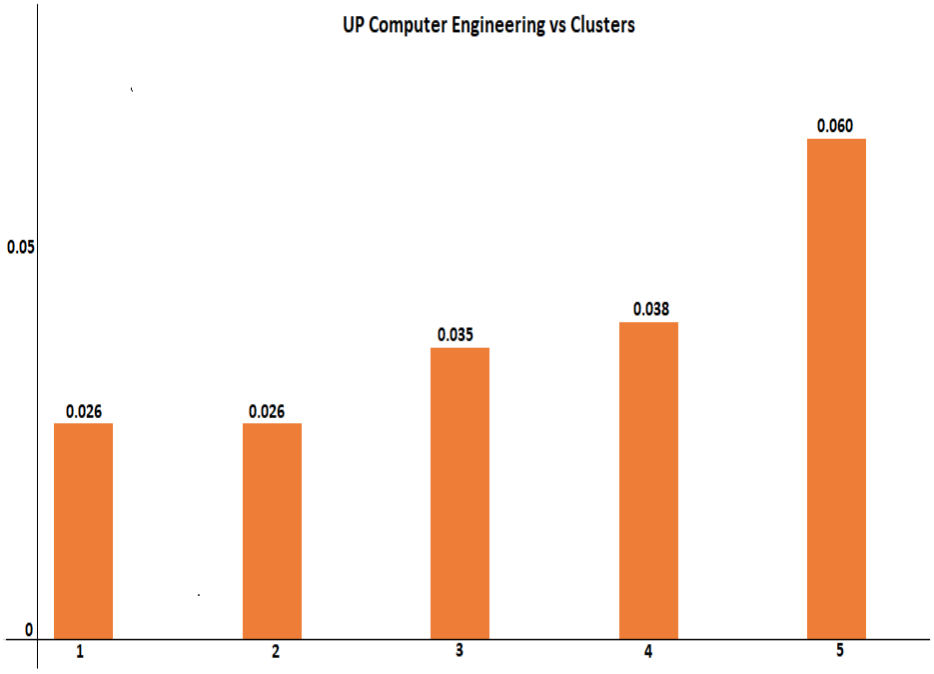

Fig.4. University of Pristina versus clusters

As mentioned at the outset, at the University of Pristina we have a smaller textual similarity between the Computer Engineering syllabus and the clusters created by the system. Certainly such an analysis is supported by the results of the above analysis which also make the University of Pristina the second university that has similarity of textual content to the demands of the labor market. In Figure 5 we present the analysis made between the University of Tirana and the clusters created by our system.

As for the two universities in the region, the same analysis was done for the University of Tirana, where the textual similarity between the Informatics syllabus and the clusters created by our system was compared. As can be seen in Figure 5, the similarity is very small compared to the two previous universities. The smallest value of textural similarity is ***0.016*** with the first cluster. We have a similarity of ***0.018*** with the second cluster, and we have a similarity of ***0.025*** with the third cluster. The maximum similarity is also with the fifth cluster, since the similarity value between the Informatics syllabus and the fifth cluster is ***0.029***.

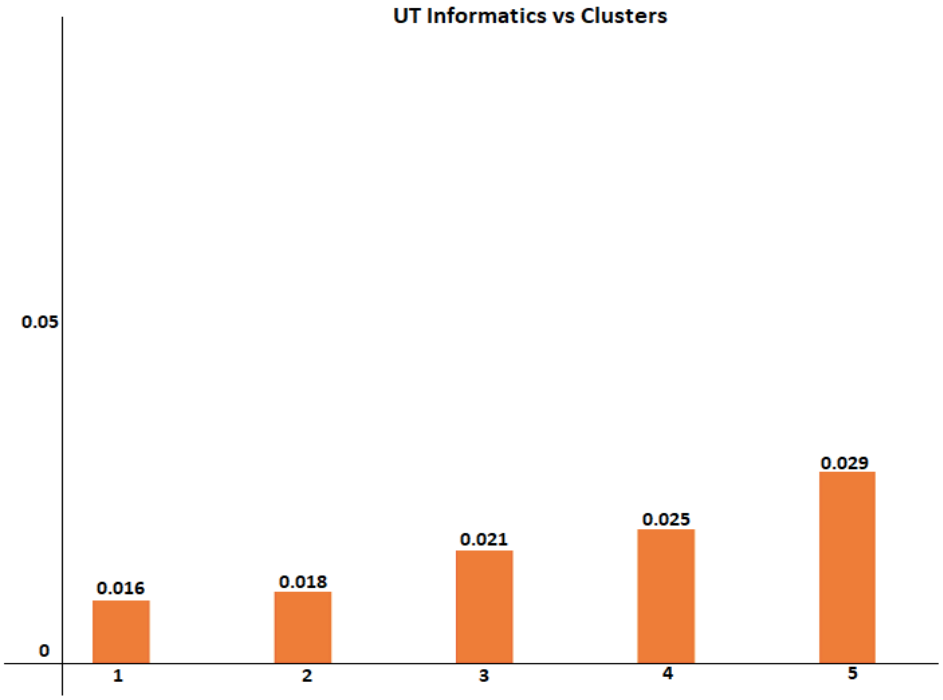

Fig.5. University of Tirana versus clusters

Compared to the other two universities, the similarity of the University of Tirana is very small, and it ranks as the third university for the similarity of textual content it has with the labor market requirements. Of course, even to this analysis we have a high support, as it supports all the analyses that have been done before, and that all have put the University of Tirana as the third university. Next we show three of analyses in Figure 6 that provides the maximum, average and minimum of similarities between clusters and syllabuses of universities.

Figure 6 shows a graph showing the maximum, average, and minimum values that universities have with the clusters created by the system.

Based on the graph, we can see that the maximum value for South East European University is ***0.078***, the average value is ***0.050*** and the minimum value for textural similarity is ***0.036***.

For the University of Pristina, the maximum value of textual similarity is ***0.060***, the average value is ***0.037*** and the minimum value of textual similarity is ***0.026***.

For the University of Tirana, the maximum value is ***0.029***, the average value is ***0.021***, and the minimum value for textual similarity is ***0.016***.

At all three universities, maximum similarity was achieved with ***cluster 5***, and minimum similarity was achieved with ***cluster 1***. Based on these values, we will make an analysis of these clusters about their textual content to analyze which words are the most mentioned, and which are less mentioned. In Figure 7 we present the

contents of ***cluster 5*** words to see which syllabus words are similar and which words are not.

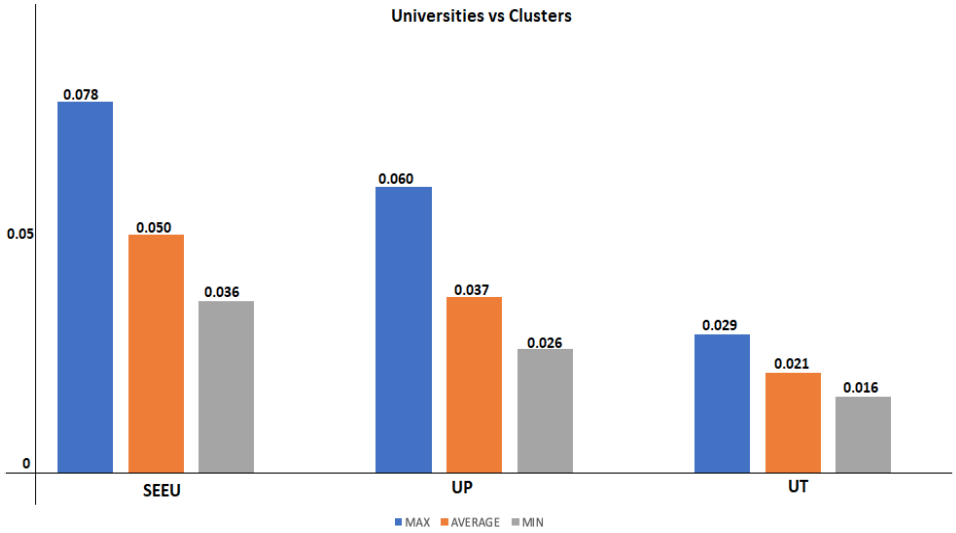

Fig.6. Universities versus clusters

Figure 7 shows the cluster of words belonging to ***cluster 5***, which is most similar to all syllabuses of the three universities.

Of course, each of the words that we have shown in Figure 7 have their importance and their weight based on the frequency that those words have in the labor market demand corpus. Next in Figure 8 we show the cluster with words that have less similarity with syllabuses of universities.

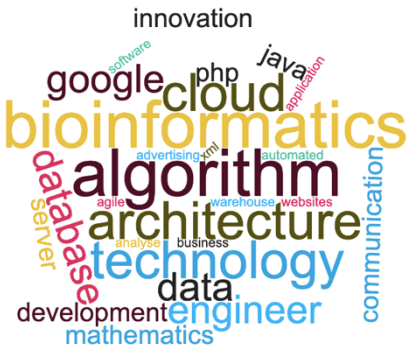

Fig.7. Words of most similar cluster

Fig.8. Words of less similar cluster

Figure 8 shows the list of ***cluster 1*** words with which the three universities have less similarities. As we can see in Figure 8, we have words sorted by weight, and

these are exactly the words that make syllabus resemblance smaller. Next in Figure 9 we compare syllabuses with each other for similarity.

Figure 9 presents an analysis of the textual similarity between the clusters created by our system. As can be seen in the graph above, the similarity between ***cluster 1*** and ***cluster 2*** is very low since we have a similarity of ***0.05***. Whereas for the similarity between ***cluster 1*** and ***cluster 3*** we have a textural similarity of ***0.07***. The greatest similarity begins with other clusters, as the textual similarity between ***cluster 1*** and ***cluster 4*** reaches ***0.16***. While the highest textural similarity is between ***cluster 1*** and ***cluster 5*** as it reaches a level of ***0.64***. Also the similarity between ***cluster 2*** and ***cluster 3*** is high as it reaches a level of ***0.59***. We have a smaller similarity between ***cluster 2*** and ***cluster 4***, since we have a similarity of ***0.15***. A very high similarity is observed between ***cluster 2*** and ***cluster 5***, since the value reaches ***0.59***. Finally we have the similarity between ***cluster 3*** and ***cluster 4*** where we have a value of ***0.14***, while ***cluster 3*** and ***cluster 5*** have a similarity of ***0.58***. As well as the last comparison is between ***cluster 4*** and ***cluster 5***, and the value obtained is lower than the other values because we have a value of ***0.16***.

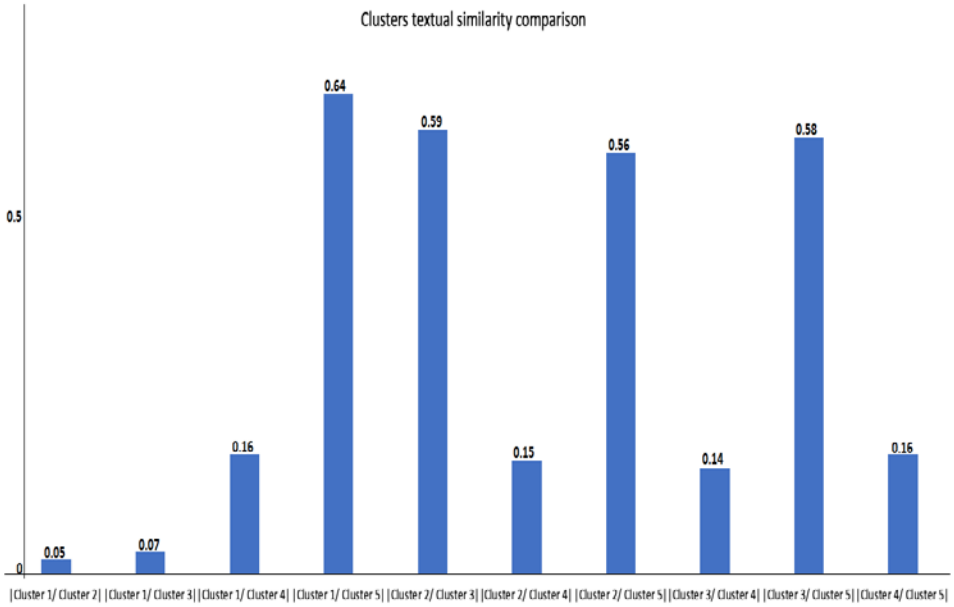

Fig.9. Cluster textual similarity comparison

3.CONCLUSION

The application of methods that increase the accuracy of comparison between different documents has a great importance. During our work we presented the form of identifying the optimal number of clusters that our analysis contains. After

identifying the optimal number, we created these clusters, and used them to compare them with the textual content of the university syllabuses that were included in this research.

After comparing clusters with university syllabuses, we presented the cluster which has less similarity, and the cluster which has the most similarity with the university syllabuses. While presenting these two clusters, we identified the words that are part of these clusters.

After comparing clusters with university syllabuses, we also compared clusters with each other, to identify which clusters are closely resemble each other.

Finally, we can conclude that such a methodology of comparing textual content between labor market demands and university curricula will directly contribute to improving the curricula offered by universities.

REFERENCES

[1] Chi-Hoon, L., Osmar, R., Z., Ho-Hyun, P., Jiayuan, H., R., Clustering high dimensional data: a graph-based relaxed optimization approach, In *Information Sciences,* 2008.

[2] Jun, Y., Cosine similarity measures for intuitionistic fuzzy sets and their applications, In *Mathematical and Computer Modelling, ELSEVIER,* vol. 53, 2011, pp. 91-97.

[3] Grigori, S., Alexander, G., Helena, G., David, P., Soft Similarity and Soft Cosine Measure: Similarity of Features in Vector Space Model, In *Computacion y Sistemas,* vol. 18, no. 3, 2014, pp. 491-504.

[4] Jun, Y., Improved cosine similarity measures of simplified neutrosophic sets for medical diagnoses, In *Artificial Intelligence in Medicine, ELSEVIER,* 2015, pp. 171-179.

[5] Yunjae, J., Haesun, P., Ding-Zhu, D., Barry, L., D., A Decision Criterion for the Optimal Number of Clusters in Hierarchical Clustering, In *ACM Digital Library,* vol. 25, 2003, pp. 91-111.

[6] Mohamed, B., Shengrui, W., An objective approach to cluster validation, Pattern Recognition Letters. In *ACM Digital Library,* October 2006.

[7] Tichy, L., Chytry, M., Hajek, M., Talbot, S., Dukat, Z., OptimClass: Using species-to-cluster fidelity to determine the optimal partition in classification of ecological communities. In *Journal of Vegetation Science,* 2010.

[8] Dae-Won, K., Kwang, L., Doheon, L., On cluster validity index for estimation of the optimal number of fuzzy clusters. In *Elsevier Inc,* April 2004.

[9] Clustering algorithms and evaluations, In *University of Stuttgart.* https://www.ims.uni-stuttgart.de/institut/mitarbeiter/schulte/theses/phd/algorithm.pdf

[10] Caiming, Zh., Duoqian, M., Pasi, F., Minimum spanning tree based split-and-merge: a hierarchical clustering method. In *Information Sciences*, vol. 181, Issue 15, 2011, pp. 3397-3410.

Information about the authors:

Ylber Januzaj - has graduated the Faculty of Contemporary Sciences and Technologies in South East European University. He is PhD Candidate in South East European University, in e-Technologies program, specialization in Data Mining.

Artan Luma – has graduated in Faculty of Contemporary Sciences and Technologies in South East European University in Tetovo. He holds a PhD diploma in Computer Sciences from 2010.

Besnik Selimi – has graduated in Joseph Fourier University (Grenoble 1), Grenoble, France. He holds a PhD diploma in Computer Sciences from 2009.

Bujar Raufi - has graduated in Faculty: French Faculty of Electrical Engineering Technical University of Sofia, Sofia, Bulgaria. He holds a PhD diploma in Computer Sciences from 2011.

Manuscript received on 14 October 2019