LUMEN ET VIRTUS
ISSN: 2177-2789

TEXT MINING: CLUSTERING APPLIED TO SCIENTIFIC ARTICLES IN CHEMISTRY, USING THE CASSIOPEIA MODEL

https://doi.org/10.56238/levv15n42-070

Submitted on: 26/10/2024 **Publication date:** 26/11/2024

Diego Sampaio Amariz[1] and Marcus Vinícius Carvalho Guelpeli[2]

ABSTRACT

Chemistry, by dedicating itself to understanding the submicroscopic nature of matter and its transformations, develops its own language and produces fundamental knowledge about nature. Its nature as basic knowledge led it, along with other natural sciences, to compose the knowledge of any citizen, whether to read and understand the natural world or transformed by the hand of man, or to continue studies at a higher or technical level in other areas or professions. However, assimilating and dealing with the large volume of information available, locating them quickly and accurately, has become a great challenge, within the diverse range of existing documents. With this, Text Mining Techniques can assist in this process, through the extraction of textual data. Thus, the objective of this research is to relate concepts of Chemistry by finding similar words in scientific articles in the area, which can demonstrate a connection between some concepts addressed in High School. Through the clustering technique with the use of the Cassiopeia model, in a corpus of academic texts related to Chemistry. The research was developed according to the following actions: bibliographic survey; construction of the corpus; collection of the corpus; statistical analysis of the corpus; text mining; clustering; and, finally, the analysis of the data from the generated clusters. The results obtained showed that the clustering carried out in the corpus provided the relationship between chemical concepts, finding similar words in the scientific articles that make up the corpus developed in this research, which demonstrate the connection of high school chemistry contents.

Keywords: Text Mining. Corpus. Chemistry. Clustering. Cassiopeia model.

[1] Master's student in the Graduate Program in Education – Federal University of the Jequitinhonha and Mucuri Valleys – UFVJM

[2] Professor of the Graduate Program in Education – Federal University of the Jequitinhonha and Mucuri Valleys - UFVJM

INTRODUCTION

Chemistry is a science that permeates all areas of knowledge and involves concepts, chemical reactions and transformation of matter, contributing considerably to the advancement and social and technological development of humanity (SCHNETZLER, 2003).

The discipline is an integral part of the curriculum of High School and some university courses, and seeks to develop scientific knowledge and understanding of chemical phenomena. The acquisition of knowledge has an educational purpose for the formation of scientific bases and goes further, preparing students to be more critical and reflective citizens (BRASIL, 2018).

The main difficulties detected by research in the area are related to the way in which the contents are transmitted, as well as the didactic methodology applied by the teachers and the structure of the environment, which is often inadequate for the consolidation of teaching with theoretical and practical activities (YAMAGUCHI; SILVA, 2019).

Research has shown that the teaching of Chemistry has been structured around activities that lead to the memorization of information, formulas and knowledge that limit students' learning and contribute to demotivation in learning and studying.

To broaden the conception of the teaching of Chemistry, it is necessary to seek improvement and detection of obstacles, so that teaching and learning can occur in a full way.

There are numerous discussions about the teaching of chemistry, its learning difficulties, the training of teachers and the didactic methodologies that can collaborate with a teaching that aims at a greater understanding of the contents, so that the teaching can be useful for the formation of the individual as a whole (SCHNETZLER, 2002).

The content of Chemistry can be divided into several branches, it is observed that some concepts are addressed separately in independent chapters. However, the Chemistry content has a linearization between chapters and concepts, which can be evidenced, for example, in the contents of chemical kinetics and chemical equilibrium. Where in chemical kinetics the processing speed of chemical reactions is studied and in the other the equilibrium between the substances involved in a chemical reaction, the speed of a reaction is an important variable for the equilibrium of a reaction to occur (ATKINS, 2007).

Another example that can be evidenced is the polarities of molecules and intermolecular interactions. The polarity of a molecule is directly related to the way in which electrons are distributed around atoms. If there is a symmetrical distribution, the molecule

will be nonpolar, however, if the distribution is asymmetric, and one of the parts of the molecule has a large electron density, then it will be a polar molecule (ATKINS, 2007).

An intermolecular interaction, on the other hand, occurs when two molecules come together, with an interaction of their magnetic fields, which causes a force to arise between them, which varies in intensity, depending on the type of molecule (polar or nonpolar). These concepts are evidenced in the contents of Inorganic Chemistry and Organic Chemistry.

Therefore, these concepts have a linearization of knowledge, evidencing their connectivity. This linearization will be the focus of study in this work with the use of the text mining technique, using the concepts of clustering through the *Cassiopeia model software (GUELPELI, 2012).*

In this sense, the concept of text mining is becoming increasingly popular as a method for information mining. Text mining (MT) is a set of methods used to navigate, organize, find, and discover information in textual databases. It can be seen as an extension of the *Data Mining* area, focused on text analysis. (ARANHA; PASSOS, 2006, p. 2).

Thus, TM is the process of extracting useful information (knowledge) from within a text document that is not structured. To do this, several other tools known as Natural Language Processing or Textual Based Knowledge Discovery (BARION; LAGO, 2018).

The text mining technique that will be used is clustering, which consists of a set of techniques used to gather a set of objects that have similar characteristics into distinct groups (ARORA; DEEPALI; VARSHNEY, 2015).

The basic idea of clustering is that elements belonging to the same group must present high similarity, however, they must be very distinct from objects from other *clusters*. According to (TAN *et al*., 2015), the greater the homogeneity within each *cluster* and the greater the heterogeneity between *clusters*, the better and more distinct the classification.

In this context, the general objective of this work is to relate the clustering applied in scientific articles of Chemistry, to find similar words in scientific articles of the area, which can demonstrate a connection between some concepts addressed in the EM. Through Text Mining, using the clustering technique with the use of the Cassiopeia model (GUELPELI, 2012).

METHODOLOGY

After reading some scientific articles referring to the journal Química Nova na Escola, http://qnesc.sbq.org.br/edicoes.php, which addresses articles related to education in

Chemistry, it is observed a similarity between some words among some contents of the Chemistry discipline, addressed in High School (AMARIZ; GUELPELI, 2023).

This study aims to use text mining, through the clustering technique, using the Cassiopeia model, to demonstrate the existence of a connection between the words existing in these articles, being able to relate the contents related to chemistry (AMARIZ; GUELPELI, 2023).

The interest in learning chemical concepts is connected to the conception that the knowledge covered allows a more articulated and less fragmented view of a world. Contributing to the citizen seeing himself as a participant in a world in constant transformation.

The construction of a *linguistic corpus* presents factors that can certainly help researchers to obtain and organize information to create their own database of texts that help in the process of text mining.

A total of 120 scientific articles available in Portuguese were collected from the journal Química Nova na Escola. This journal was chosen because it deals exclusively with the content of Chemistry focused on education. The articles were collected randomly and the collection aimed to form a textual database, called "*Corpus*". These scientific articles refer to the years 1978 to 2021.

The pre-processing of the *corpus* was divided into two parts, first the conversion from the Portable document format (PDF) to the format that stores plain text (TXT), due to the Get Finecount program and the Cassiopeia model processing this format.

Statistical analysis of the *corpus* was performed, where calculations of word amplitude, word averages, standard deviation of words and coefficient of variation were performed.

Clustering consists of dividing them into groups, called *clusters*, so that they are more similar to other points in the same group than those of other groups, and can use statistical calculations for each group developed.

This work was carried out using the Cassiopeia model, which consists of grouping hierarchical texts to prove the connection between the chemical contents addressed in the MS.

According to Soares (2013), the definition for each stage of text mining is:

DATA COLLECTION

The premise of this work is the collection of texts, that is, the search for scientific articles related to Chemistry. The collection aims to form a textual database, called

"*Corpus*". It can be carried out in several ways, but all of them require great efforts, in order to obtain material of satisfactory quality and that serves as raw material for the continuity of the process and for the acquisition of knowledge.

PRE-PROCESSING

Its objective is to prepare the collected documents in order to obtain a form for better data processing. The entire system to be developed depends on a filtering of the texts, that is, a reduction in the number of words to obtain effective informativeness, which can provide a qualitative and quantitative gain for the processing of chemical articles.

INDEXING

It is responsible for establishing indexes in order to establish greater speed and agility for the retrieval of documents and their terms.

MINING

The processing uses hierarchical text grouping and an algorithm to join the texts with similarities. Clustering is performed by the Cassiopeia model, which identifies the characteristics of words in scientific articles, using relative frequency, which defines the importance of the terms, according to the periodicities in each text used. The Cassiopeia model provides the removal of *stopwords*, words that do not provide context to this search.

ANALYSIS

It consists of the evaluation of the data obtained. In this stage, the model groups scientific articles related to Chemistry by similarity, allowing a better evaluation of the data, with an efficient degree of informativeness.

Text mining, through clustering techniques, using the Cassiopeia model (Guelpeli, 2012) organizes scientific articles focused on Chemistry, according to the similarity between the words in each clustered article.

Thus, the set containing the one hundred and twenty chemistry articles were submitted to the Cassiopeia model. The model performed the grouping and regrouping process, with the set of one hundred and twenty articles thirty times, that is, the Cassiopeia model analyzed the *corpus* of this research repeatedly during these thirty interactions.

The results of these clusters were analyzed through qualitative analysis of the articles that constitute each cluster and quantitative, through the internal or unsupervised metric called Silhouette Coefficient, explained in the previous chapter.

The data analysis was carried out through the frequencies of similar words occurring in the articles, performing the calculations and a table of word frequencies, then the *clusters* generated during the clustering process were analyzed.

The *clusters* generated through clustering were analyzed and a selection of the best data generated was made, through quantitative results, which were the results of internal metrics, during the text mining process, and through qualitative results, which were the analysis of similar contents generated in each *cluster*.

RESULTS AND DISCUSSIONS

The *corpus* was converted from PDF format to TXT format for computational processing, due to the computer programs used in this research accepting the TXT format. In addition, in this conversion, images, graphs, tables, page numbers and all annotations that were not part of the body of the text were removed. The files were renamed following a sequential order, starting at 1 and ending at 120, and an analysis *software* called *Get Finecount* 2.6 was used to count the words.

The *Get Finecount* 2.6 software is a tool that provides analysis of a document. It analyzes a text and counts the number of words, characters, repetitions, spaces, redundant spaces, lines, sentences, and pages in an orthographic file. It was important to verify the number of words in each article that constitutes the *corpus* of this research.

In *the corpus*, the amplitude was calculated, through the number of words in the articles. Amplitude is a measure of dispersion that determines the degree of variation of numbers, this measure is determined by the difference between the maximum value found and the minimum value, according to Mathematical Equation 1.

$$R = X_{máximo} - X_{mínimo}$$

Where max X is the maximum word value found and min X is the minimum word value found.

The article that contains the largest number of words had 5,566 and the article that had the least number of words had 2,428 words. The amplitude found in the *chemistry corpus* was 3,138 words (AMARIZ; GUELPELI, 2023).

The average of the words in the articles of the corpus was also calculated. The Average is part of the concepts of Statistics. As far as the mean is concerned, it can be arithmetic (simple or weighted), geometric, harmonic, quadratic, cubic or biquadratic. Dealing specifically with the arithmetic mean, it is considered "the most basic concept of

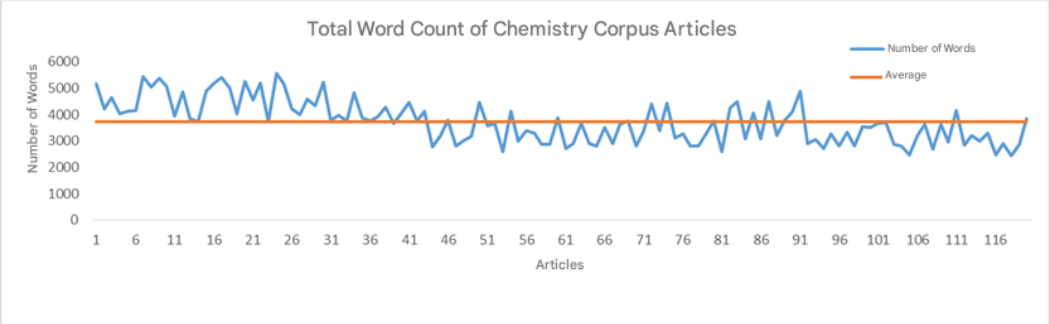
Statistics and experimental Science, it is also the most used in people's daily lives" (MAGINA; CAZORLA; GITIRANA; GUIMARÃES, 2010, p. 61-62), Equation 2 is used to calculate the mean.

$$\overline{X} = \frac{\sum X_i}{n}$$

Where, Xi is the total number of words in all articles and n the number of selected articles.

After this calculation, it was observed that the average number of words for the Chemistry articles in the *corpus* was 3,730 words. The number of words in relation to the mean is shown in Figure 6. This graph shows the dispersion of the number of words in relation to the average (AMARIZ; GUELPELI, 2023).

Figure 6: Dispersion of the number of words in relation to the Corpus mean.



Source: Author himself

With the mean value obtained, the standard deviation existing in this *corpus was calculated*. The standard deviation is the prototype of dispersion measurements by virtue of its mathematical properties and its use in sampling theory" (OLIVEIRA, 2017, p. 8). It is a measure of dispersion, which indicates how uniform the data set is. According to Mathematical Equation 3.

$$S = \sqrt{\frac{\sum \left(X - \overline{X}\right)^2}{n}}$$

Where, X is the individual value of each word in each article, $\overline{X}$ is the average of the data obtained and n total number of articles in the *corpus*.

The standard deviation for the data from the *Chemistry corpus* was 800 words. With the data obtained from the mean and standard deviation, the coefficient of variation of the data can be calculated.

Reliable experiments require the evaluation of the results by verifying their accuracy, which can be performed by the values of the coefficients of variation, CV, (NESI *et al*., 2010; STORCK *et al*., 2011). According to Steel *et al*. (1997), the CV allows the comparison of results of different experiments, involving the same variable or species, thus allowing the quantification of the accuracy of their research.

The VC is an important measure of the variability of experimental results, and can be useful in defining the number of repetitions of the assay, necessary to detect a difference between the means of treatments with a given probability (PIMENTEL-GOMES, 2009; NESI *et al*., 2010).

According to Storck *et al*. (2011), the distribution of VC makes it possible to establish ranges of values that guide researchers on the validity of their experiments. Thus, it can be said that the coefficient of variation is a way of expressing the variability of the data, excluding the influence of the order of magnitude of the variable. The coefficient of variation is equal to the standard deviation divided by the arithmetic mean, multiplied by 100% (LEVINE et al., 2014). Equation 4 shows how the calculation for the CV is performed.

$$CV = \frac{S}{\overline{X}} \times 100$$

Where, S is the standard deviation and $\overline{X}$ is the average of the data obtained

Since the coefficient of variation analyzes dispersion in relative terms, it will be given as a percentage. The lower the value of the coefficient of variation, the more homogeneous the data, i.e., the smaller the dispersion around the mean (LEVINE et al., 2014). In general, if the VC is less than or equal to 15%, the result presents a low dispersion of the data, homogeneous data. If the calculation of the data is between 15 and 30%, they present an average dispersion. And if it is greater than 30%, they have a high dispersion, heterogeneous data (LEVINE et al., 2014).

Therefore, when calculating the coefficient of variation of the data of the *Chemistry corpus* , 21.45% was obtained, which indicates an average homogeneity of the articles that make up this *corpus*. However, this value obtained is closer to 15% than 30%, which indicates a greater tendency towards homogeneity than a tendency towards heterogeneity.

All the calculations, amplitude, arithmetic mean, standard deviation and coefficient of variation made in this *corpus* are shown in Table 1.

Table 1: Calculations performed in the *Chemistry Corpus* .

Statistical Table	
Corpus	Amount
Arithmetic Mean	3,730 words
Standard Deviation	800 words
Amplitude	3,138 words
Coefficient of Variation	21.45%

Source: Author himself

When dealing with textual data, it is necessary to keep in mind that, in order to understand implicit knowledge in an agile and simplified way, it is necessary to find ways that can represent it to transmit some knowledge (SARGIANI et al., 2018). Some graphical tools, such as histograms and word clouds, can be used in the process to evaluate documents containing unstructured texts and to explore hidden knowledge in the texts (BRUNO, 2016).

The Cassiopeia model puts all letters in lowercase, in addition to other precautions, such as discarding all figures, tables, existing markings and the removal or not of *stopwords*. The function denoted for the *stopwords* can be configured by the user.

In addition, the Cassiopeia model identifies the characteristics of the words in the document, using the relative frequency, which defines the importance of a term, according to the frequency with which it is found in the document. The more a term appears in a document, the more important it is for that document (GUELPELI, 2012). Based on the weights of the words, obtained in relative frequency, the average is calculated over the total number of words in the document.

The text mining technique used in this research is clustering, through the Cassiopeia model, which separated the articles into *clusters*, sets of objects that have similarities to each other. Thus, the articles that have similarities between the words were in the same *cluster*.

The Cassiopeia model carried out the process of grouping and regrouping thirty times, that is, thirty interactions were carried out, in which each article was analyzed 30 times, thus, three thousand and six hundred analyses were totaled. With this, the model calculated the Silhouette Coefficient (CS) to perform the quantitative analysis of the *clusters*.

According to Guelpeli (2012), the Silhouette Coefficient is based on the idea of how similar an object is to the other members of its group, and how far this same object is from those of another group. Thus, this measure combines the cohesion and coupling measures. Equation 5 shows how the calculation for the Silhoette coefficient is performed.

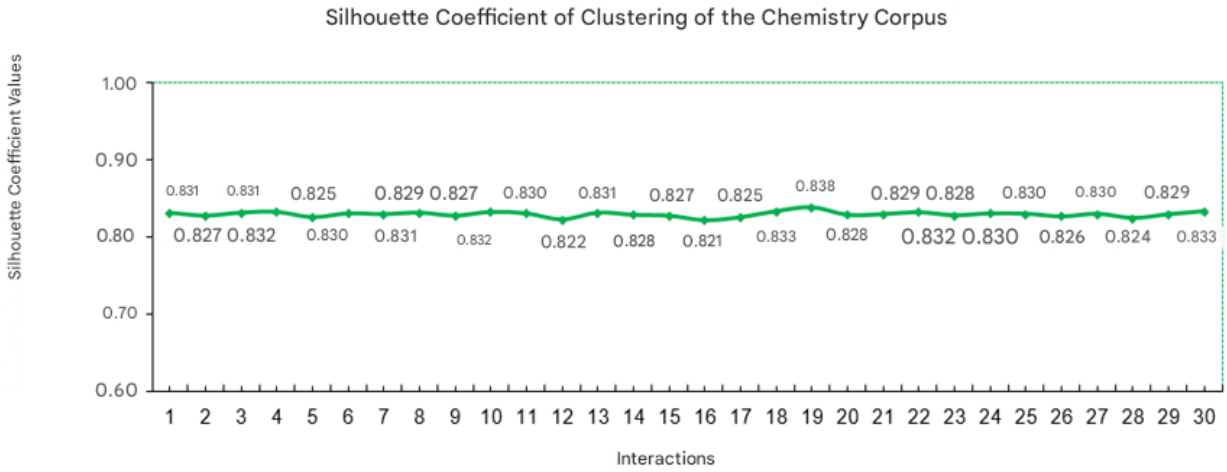
$$CS = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

Where a(i) is the average distance between the ith element of the group and the others in the same group. The b(i) is the minimum value of the distance between the ith element of the group and any other group, which does not contain the element, and max is the greatest distance between a(i) and b(i).

The Silhouette Coefficient of a group is the arithmetic mean of the coefficients calculated for each element belonging to the group, the CS value is in the range of 0 to 1 (GUELPELI, 2012).

For better organization of the results of the Silhoette Coefficients generated by clustering, by means of the Cassiopeia model, obtained over the 30 interactions. The results are presented in Figure 7.

Figure 7: Silhouette coefficient of the Chemistry *Corpus* Clustering .



Source: Author himself

The Silhouette coefficient or index (CS) is a value that measures how similar an object is to its own cluster (cohesion) compared to other clusters (coupling).

According to Guelpeli (2012), the coefficient varies between 0 and 1, in which values close to 1 indicate that the object is well related to its *cluster* and values close to 0 indicate that the object is not well related to its *cluster*.

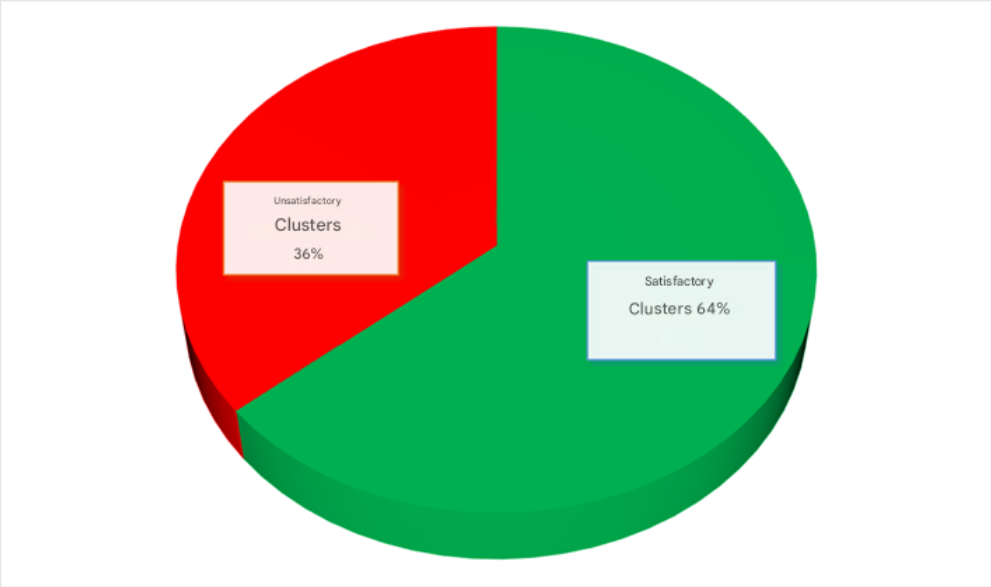
In all interactions, values for the Silhouette Coefficient were between 0.821, the minimum value found, and 0.838, the maximum value found in clustering, values that are very close to 1, which indicates that the scientific articles that make up the *corpus* of this research are well related within their respective *clusters*.

After the quantitative analysis of the *clusters,* qualitative analyses were performed and it was observed that clustering generated 36 *clusters*, that is, thirty-six sets of articles with similar words.

In this set, it is observed that some contained few Chemistry contents grouped together. While others had many contents grouped in the same *cluster* and clusters with the same grouped content also occurred, this was due to the similarity between the words.

In view of these sets of *generated clusters*, they were classified into *satisfactory clusters and* unsatisfactory clusters, as shown in Figure 8.

Figure 8: Cluster *Analysis*.



Source: Author Himself

According to Figure 8, the analysis generated 64% of *satisfactory clusters, i.e., 23* clusters, which grouped articles that contained two to three different chemistry contents addressed in the EM. These *clusters* presented a frequency of similar words in their articles, which is why they were classified as satisfactory.

Thus, 13 unsatisfactory *clusters* were generated , accounting for 36%, which did not present a considerable frequency of similar words in their articles and grouped several chemistry contents in the same *cluster*.

This analysis was due to the clustering performed by the Cassiopeia model, which indicates that it is satisfactory to analyze similar words in a *cluster* more frequently than to have different contents in the same *cluster*.

Thus, a *corpus* containing 120 articles was clustered, which were divided into two classes, namely, *satisfactory clusters and* unsatisfactory *clusters*. Analyses were performed to identify the number of articles that are present in the satisfactory clusters and in the unsatisfactory clusters. These divisions are evidenced in Figure 9.

Figure 9: Analysis of the divisions of the articles within the *Clusters*.

Source: Author Himself

The *corpus* presents 56% of its articles in *the satisfactory clusters, that is, 67 articles of the* corpus*, are grouped in* the *clusters* of interest for this research. These articles presented a frequency of similar words within the same *cluster.*

However, the *corpus* presents 44% of the articles within the clusters unsatisfactory, computing 53 articles, which presented a considerable frequency of similar words within their *cluster*, but grouped several chemistry contents in the same *cluster* or similar contents applied in the MS.

With the data acquired by clustering, an analysis of each cluster was carried out , relating the articles and their contents with the years in which they are taught during high school, as shown in Table 2.

Table 2: Relationship between the contents of each *Cluster* and the years of the MS.

Tabela relacionada aos conteúdos de cada *cluster* e os anos do Ensino Médio			
Clusters	**Ano referência do Ensino Médio**		
Cluster 1	1º Ano	2º Ano	
Cluster 2		2º Ano	3º Ano
Cluster 3	1º Ano	2º Ano	
Cluster 4		2º Ano	
Cluster 5	1º Ano	2º Ano	3º Ano
Cluster 6	1º Ano	2º Ano	3º Ano
Cluster 7	1º Ano	2º Ano	
Cluster 8		2º Ano	
Cluster 9		2º Ano	
Cluster 10		2º Ano	
Cluster 11	1º Ano	2º Ano	3º Ano
Cluster 12			3º Ano
Cluster 13			3º Ano
Cluster 14	1º Ano		
Cluster 15	1º Ano		3º Ano
Cluster 16	1º Ano	2º Ano	3º Ano
Cluster 17	1º Ano	2º Ano	3º Ano
Cluster 18		2º Ano	3º Ano
Cluster 19	1º Ano	2º Ano	
Cluster 20		2º Ano	3º Ano
Cluster 21	1º Ano	2º Ano	
Cluster 22	1º Ano	2º Ano	
Cluster 23	1º Ano		
Cluster 24	1º Ano		
Cluster 25	1º Ano	2º Ano	3º Ano
Cluster 26	1º Ano	2º Ano	3º Ano
Cluster 27	1º Ano		3º Ano
Cluster 28	1º Ano		3º Ano
Cluster 29	1º Ano	2º Ano	
Cluster 30	1º Ano		
Cluster 31	1º Ano	2º Ano	
Cluster 32	1º Ano	2º Ano	
Cluster 33	1º Ano	2º Ano	
Cluster 34	1º Ano	2º Ano	
Cluster 35		2º Ano	
Cluster 36	1º Ano		3º Ano

Source: Author himself

Thus, it is observed that, through clustering, the *corpus* of this research presents the contents addressed in the EM in 26 *clusters* containing contents of the 1st year and 2nd year of the EM and 16 *clusters* containing the contents taught in the 3rd year of the MS.

The 23 *satisfactory clusters generated in this research are shown in Table 3, which indicates the relationship between the contents of each satisfactory clusters and the years of the EM in which they are addressed.*

Table 3: Relationship between the contents of the *satisfactory Clusters and the years of MS.*

Tabela relacionada aos conteúdos de cada *cluster* satisfatórios e os anos do Ensino Médio			
Clusters	**Ano referência do Ensino Médio**		
Cluster 2		2º Ano	3º Ano
Cluster 3	1º Ano	2º Ano	
Cluster 4		2º Ano	
Cluster 8		2º Ano	
Cluster 9		2º Ano	
Cluster 10		2º Ano	
Cluster 12			3º Ano
Cluster 13			3º Ano
Cluster 14	1º Ano		
Cluster 17	1º Ano	2º Ano	3º Ano
Cluster 18		2º Ano	3º Ano
Cluster 20		2º Ano	3º Ano
Cluster 21	1º Ano	2º Ano	
Cluster 23	1º Ano		
Cluster 27	1º Ano		3º Ano
Cluster 28	1º Ano		3º Ano
Cluster 29	1º Ano	2º Ano	
Cluster 31	1º Ano	2º Ano	
Cluster 32	1º Ano	2º Ano	
Cluster 33	1º Ano	2º Ano	
Cluster 34	1º Ano	2º Ano	
Cluster 35		2º Ano	
Cluster 36	1º Ano		3º Ano

Source: Author himself

It is verified that in 12 *satisfactory clusters* there was the presence of contents that can be taught in the 1st year of MS, 15 *clusters* presented contents that can be taught in the 2nd year of MS and 9 *clusters* showed contents of the 3rd year of MS. With these results, five satisfactory *clusters* were selected to be addressed.

Cluster *3,* generated by the Cassiopeia model, presents similarity between the contents of electrochemistry and periodic table, contents taught in different periods of EM, because they presented words similar to these contents.

This occurred due to the analysis of the words found in the set of articles in this *cluster*. Words such as: electrolysis, solution, batteries, cathode, oxidation, reduction, electrons, electrode, metal and electronegative are present in all the articles of this textual set, with a certain relative frequency, as shown in Table 4.

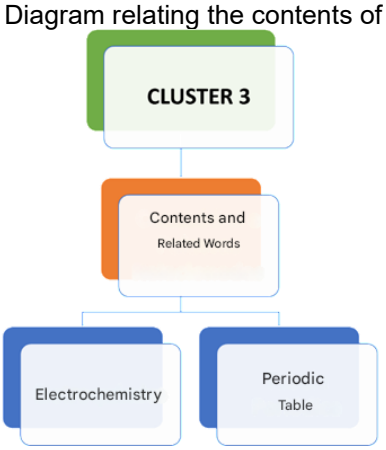
Table 4: Relative frequency of words in Cluster 3

Table of relative frequency of words in Cluster 3		
	Relative Frequency (%)	
Words	Article 44	Article 56
Electrolysis	70.6	29.4
Valence	48.9	51.1
Stacks	43.9	56.1
Cathode	60.9	39.1
Oxidation	46.7	53.3
Reduction	61.8	38.2
Electrons	45.0	55.0
Electrode	73.1	26.9
Metal	30.6	69.4
Electronegative	66.7	33,3

Source: Author himself

The relative frequency of the words that appeared most in the articles of this *cluster* shows the relationship between the contents of electrochemistry and the periodic table, as shown in Figure 10.

Figure 10: Diagram relating the contents of *Cluster 3*.



Source: Author Himself

Words such as valence, oxidation cells, electrons presented considerable relative frequencies in article 44 and article 56, which are part of these clusters. The words that appear most in the articles and characterize the electrochemistry content are electrolysis, batteries, cathode, oxidation, reduction, electrons and electrode. Likewise, the words electrons, metal, valence and electronegative are often addressed in the content of the periodic table.

Cluster 14, generated by the Cassiopeia model, presents similarity between the contents of the atomic models and chemical bonds, these contents are taught in the 1st

year of EM. This *cluster* is composed of two articles, article 7 and article 17 of the *corpus* that constitute this research.

This occurred due to the analysis of the words found in the set of articles that form this *cluster*. Words such as: Dalton, laws, atom, weights, bond, electrons, molecule, geometry, polarity and ionic are present in all the articles of this textual set, with a certain relative frequency, as shown in Table 5.

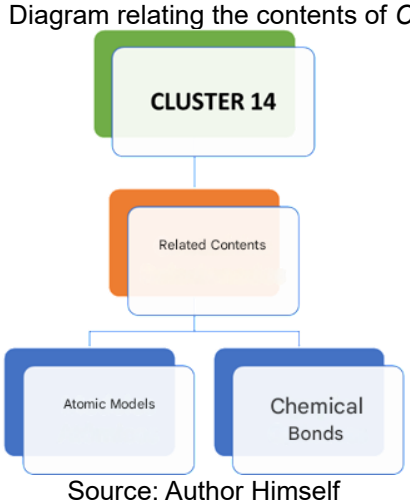
Table 5: Relative frequency of words in *Cluster 14.*

Relative frequency table of words in cluster 14		
	Relative Frequency (%)	
Words	Article 7	Article 17
Dalton	14.6	85.4
Laws	16.7	83.3
Atom	60.0	40.0
Ponder	33,3	66.7
Connection	69.4	30.6
Electrons	58.3	41.7
Molecule	61.5	38.5
Geometry	34.5	65.5
Polarity	58.8	41.2
Ionic	47.1	52.9

Source: Author himself

The relative frequency of the words that appear most in the articles that make up this *cluster* shows the relationship between the contents, atomic models and chemical bonds. Figure 11 shows a diagram relating the contents to *cluster* 14.

Figure 11: Diagram relating the contents of *Cluster 14*.



Source: Author Himself

Cluster 28, generated by the Cassiopeia model, presents similarity between the contents of atomic structure and polymers, contents taught in different periods of EM, because they presented words similar to these contents.

Thus, words such as plastic, polymer, electrons, shielding, atom, ionization, electronics, and orbitals are present in the set of articles in this *cluster*, with a certain relative frequency, as shown in Table 6.

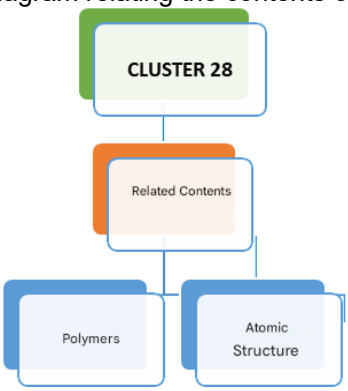
Table 6: Relative frequency of words in *Cluster 28*

Relative frequency table of words in cluster 28					
	Relative Frequency (%)				
Words	Article 14	Article 91	Article 108	Article 119	Article 120
Plastic	4.9	8.7	36.9	35.0	14.6
Polymer	11.9	20.3	28.8	25.4	13.6
Electrons	39.5	25.8	8.9	14.5	11.3
Shielding	66.7	33,3	0.0	0.0	0.0
Atom	31.4	14.7	17.6	15.7	20.6
Ionization	35.3	17.6	20.6	11.8	14.7
Electronics	75.0	25.0	0.0	0.0	0.0
Orbitals	25.0	39.6	6.3	18.8	10.4

Source: Author himself

The relative frequency of the words that appeared most in the articles of this *cluster* shows the relationship between the contents of atomic structure and polymers, as shown in Figure 12.

Figure 12: Diagram relating the contents of *Cluster 28*.



Source: Author Himself

Words such as polymers, electrons, atom and ionization presented considerable relative frequencies in the articles that make up this *cluster*. The words that appear most in the articles and characterize the content presented in Figure 12.

Cluster 32, generated by the Cassiopeia model, also presents similarity between some contents related to the teaching of chemistry to MS. The contents of environmental chemistry and acid-base concepts are evidenced in this *cluster*.

Therefore, words such as effect, greenhouse, infrared, absorption, gases, acid, base, ions, indicators, and reaction are present in the set of articles that constitute this *cluster*, as shown in Table 7.

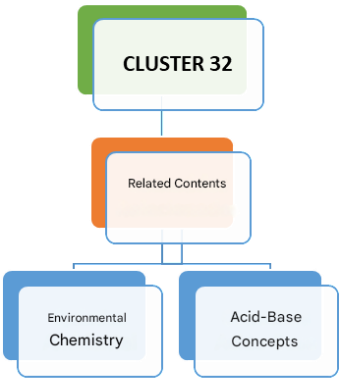
Table 7: Relative frequency of words in *Cluster 32.*

Relative frequency table of words in cluster 32		
	Relative Frequency (%)	
Words	Article 22	Article 23
Effect	61.2	38.8
Stove	61.2	38.8
Infra-red	61.5	38.5
Absorption	45.8	54.2
Gases	58.7	41.3
Acid	28.0	72.0
Base	36.8	63.2
Ions	40.9	59.1
Indicators	59.3	40.7
Reaction	55.0	45.0

Source: Author himself

The relative frequency of the words that appear most in the articles that constitute this *cluster* shows the relationship between the contents of environmental chemistry and acid-base concepts. Figure 13 shows a diagram relating the contents to *cluster* 32.

Figure 13: Diagram relating the contents of *Cluster* 32.



Source: Author Himself

Cluster 32 presents the content acid-base concepts, which is currently taught in the first year of EM and the concept of environmental chemistry is taught in the middle of EM, in the second year.

Cluster 36, generated by the Cassiopeia model, presents similarity between the contents of intermolecular interactions and organic compounds. This *cluster* is composed of two articles, article 81 and article 83 of the *corpus* that constitutes this research.

This occurred due to the analysis of the words found in the set of articles in this *cluster*. Words such as polar, nonpolar, bond, dipole, hydrogen, molecule, detergent and organic are present in all articles in this textual set, with a certain relative frequency, as shown in Table 8.

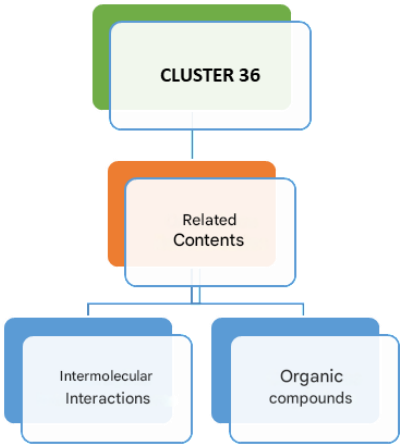
Table 8: Relative frequency of words in *Cluster 36.*

Relative frequency table of words in cluster 36		
	Relative Frequency (%)	
Words	Article 81	Article 83
Polar	58.8	41.2
Nonpolar	46.7	53.3
Connection	61.1	38.9
Dipole	47.5	52.5
Hydrogen	61.1	38.9
Molecule	60.9	39.1
Detergent	100.0	0.0
Organic	41.7	58.3

Source: Author himself

The relative frequency of the words that appeared most in the articles of this *cluster* shows the relationship between the contents of electrochemistry and the periodic table, as shown in Figure 14.

Figure 14: Diagram relating the contents of *Cluster 36*.



Source: Author Himself

Cluster 36 presents the content intermolecular interaction, which is currently taught in the first year of EM and the concept related to organic compounds is taught in the final year of EM, in the third year.

The 13 unsatisfactory clusters generated in this research are shown in Table 9, indicating the relationship between the contents of each unsatisfactory cluster and the years of the EM in which they are addressed.

Table 9: Relationship between the contents of the unsatisfactory Clusters and the years of the MS.

Tabela relacionada aos conteúdos de cada *cluster* insatisfatório e os anos do Ensino Médio			
Clusters	**Ano referência do Ensino Médio**		
Cluster 1	1º Ano	2º Ano	
Cluster 5	1º Ano	2º Ano	3º Ano
Cluster 6	1º Ano	2º Ano	3º Ano
Cluster 7	1º Ano	2º Ano	
Cluster 11	1º Ano	2º Ano	3º Ano
Cluster 15	1º Ano		3º Ano
Cluster 16	1º Ano	2º Ano	3º Ano
Cluster 19	1º Ano	2º Ano	
Cluster 22	1º Ano	2º Ano	
Cluster 24	1º Ano		
Cluster 25	1º Ano	2º Ano	3º Ano
Cluster 26	1º Ano	2º Ano	3º Ano
Cluster 30	1º Ano		

Source: Author himself

These *clusters* were classified as unsatisfactory because they did not have similar words among the articles that compose them. In addition, they demonstrate a large number of articles in the same *cluster*, providing a similarity between several chemical contents at the same time. This similarity between several articles, which have varied contents, can be considered as an imprecision in clustering.

CONCLUSION

Due to the large amount of texts available, especially in the area of education, there is a need for text mining, which can facilitate the search and retrieval of textual information. Thus, this research used clustering, one of the techniques used in text mining, as a way to contribute to the solution of some of the problems evidenced in this work, such as the isolated treatment of some concepts addressed in the discipline of Chemistry.

The *corpus* of this research consisted of 120 scientific articles, word count was performed in each article that constitutes the *corpus* and a statistical analysis was performed. The statistical analysis determined an average of the words in *the corpus*, which provides a small variation between the number of words between the articles (AMARIZ; GUELPELI, 2023).

With the mean value, the standard deviation of the words of the articles was obtained, it is observed that this value is small in relation to the number of words in the texts of each article in the *corpus*. This indicates that the words are condensed close to the mean, indicating a tendency towards homogeneity of the *corpus* (AMARIZ; GUELPELI, 2023).

This homogeneity can be evidenced by calculating the coefficient of variation, which proves a more homogeneous trend with the result made for the *corpus* at 21.45%. Thus, obtaining a more homogeneous than heterogeneous trend for the chemical *corpus* developed in this research.

After statistical analysis, text mining was performed using the Cassiopeia model. The Cassiopeia model removed all the undesirable items from the text, leaving only the important words for text mining to be used.

The clustering technique is used in this research to perform text mining. Thus, returning to the central question of this research, it is possible to conclude that the Cassiopeia model provided the relationship between chemical concepts, finding similar words in the scientific articles that make up the *corpus* developed in this research. This can be measured quantitatively by means of the Silhouette coefficient and qualitatively by the analyses of the constituent articles of each cluster.

Thus, the main objective of this research was to relate concepts of Chemistry by finding similar words in scientific articles in the area, which can demonstrate a link between some concepts addressed in High School. Through Text Mining, using the clustering technique with the use of a *software* called Cassiopeia, in a *corpus* of academic articles. From the evaluations of the articles that make up the *corpus*, statistical analyses, which proved the homogeneity of this *corpus* and the clustering, it was found that the research hypothesis was confirmed.

The creation of a *corpus* related to the content of Chemistry is highlighted as a contribution of this research, which can be used by researchers in future works. In addition, the relationship between words in several articles of the *corpus is* highlighted, which demonstrate the connection of Chemistry contents addressed in the EM (AMARIZ; GUELPELI, 2023).

REFERENCES

1. Amariz, D. S., & Guelpeli, M. V. C. (2023). Linguística de corpus aplicada a artigos científicos de química. International Journal of Development Research, 13(2), 61813–61815.

2. Araujo, A. C. F., de Oliveira Félix, M. E., & da Silva, G. N. (n.d.). Relato das dificuldades em aprender química de alunos da educação básica de uma escola pública de Campina Grande.

3. Aranganayagil, S., & Thangavel, K. (2007). Clustering categorical data using silhouette coefficient as a relocating measure. In International conference on computational Intelligence and multimedia applications, ICCIMA 2007 (pp. 13–17). Los Alamitos: IEEE.

4. Atkins, P., Jones, L., & Laverman, L. (2007). Princípios de química: Questionando a vida moderna e o meio ambiente. Bookman Editora.

5. Bizerra, A. M. C., et al. (2020). Dificuldades e motivações no ensino de química: Uma análise da perspectiva docente. VI CONEDU (Vol. 1, pp. 1406–1420). Campina Grande: Realize Editora. Disponível em: https://editorarealize.com.br/artigo/visualizar/65351. Acesso em: 26 nov. 2022.

6. Cavalcante, D. B. (2020). Análise do desempenho de parques eólicos por meio de clusterização de aerogeradores.

7. Cavalcante, M., & da Costa, J. G. (2021). Considerações sobre planejamento experimental e adequabilidade do uso de testes estatísticos em Ciências Agrárias. Diversitas Journal, 6(4), 3706–3723.

8. Cruz, L. A. (2019). Modelo para recuperação de informação em repositórios institucionais utilizando a técnica de sumarização a partir da seleção de atributos do Cassiopeia.

9. de Aguiar, L. H. G., et al. (2017). Uma coleção de artigos científicos de português compondo um no domínio educacional. Plurais Revista Multidisciplinar, 2(2), 107–119.

10. de Lima Yamaguchi, K. K. (2021). Ensino de química inorgânica mediada pelo uso das tecnologias digitais no período de ensino remoto. Revista Prática Docente, 6(2), e041–e041.

11. Giuliani, R., et al. (2022). Clusterização de trajetórias multiaspecto usando árvores de decisão.

12. Guelpeli, M. V. C. (2012). Cassiopeia: Um modelo de agrupamento de textos baseado em sumarização (Tese de doutorado, Universidade Federal Fluminense).

13. Guelpeli, M. V. C., Branco, A. H., & Garcia, A. C. B. (2009). Cassiopeia: A model based on summarization and clusterization used for knowledge discovery in textual bases. In 2009 International Conference on Natural Language Processing and Knowledge Engineering (pp. 1–8). IEEE.

14. Kunz, T., & Black, J. P. (1995). Using automatic process clustering for design recovery and distributed debugging. IEEE Transactions on Software Engineering, 515–527.

15. Levin, J., & Fox, J. A. (2004). Estatística para ciências humanas (pp. xv, 497).

16. Luhn, H. P. (1958). The automatic creation of literature abstracts. IBM Journal of Research and Development, 2, 159–165.

17. Magina, S., & Fonseca, S. (2016). A aprendizagem da média aritmética simples a partir de materiais didáticos distintos: Uma comparação entre duas propostas de ensino em teia. Revista de Educação Matemática e Tecnológica Iberoamericana, 7(1).

18. Maximiano, F. A. (2018). Princípios para o currículo de um curso de química. Estudos Avançados, 32, 225–245.

19. McHugh, M. L. (2003). Descriptive statistics, part II: Most commonly used descriptive statistics. Journal of Special Pediatric Nursing, 8(3), 111–116.

20. Oliveira, C. G., et al. (2019). Desvio padrão e imprecisão de leitura: Paquímetro. Caderno de Graduação-Ciências Exatas e Tecnológicas-UNIT-SERGIPE, 5(3), 27.

21. Oliveira, H. B. (2019). Framework Oráculo: Camada de coleta e mineração de textos para o Twitter.

22. Pinto, I. de J. P. (2021). Corpus para o domínio acadêmico: Modelos e aplicações (Tese de doutorado, PUC-Rio).

23. Reis, L. A., et al. (2021). Comportamento ingestivo de ovinos em pastos heterogêneos de capim-marandu com mesma altura média.

24. Santiago, J. C., et al. (2015). Compostos orgânicos versus inorgânicos: Um estudo sobre as propriedades físico-químicas entre essas duas classes de compostos. Enciclopedia Biosfera, 11(21).

25. Schimldt, E. R., et al. (2017). Coeficiente de variação como medida da precisão em experimentos de alface. Revista Agro@ mbiente On-line, 11(4), 290–295.

26. Silva, R. de A., et al. (2021). Uma metodologia para criação de um corpus textual adequada ao reconhecimento de entidades nomeadas em português.

27. Tan, P. N., Steinbach, M., & Kumar, V. (2006). Introduction to Data Mining. Addison-Wesley.

28. Twycross, A., & Shields, L. (2004). Statistics made simple. Part 1. Mean, medians and modes. Paediatric Nursing, 16(4), 32.

29. Rodrigues, C. F. de S., Lima, F. J. C. de, & Barbosa, F. T. (2017). Importância do uso adequado da estatística básica nas pesquisas clínicas. Revista Brasileira de Anestesiologia, 67, 619–625.

30. Veiga, M. S. M., Quenenhenn, A., & Cargnin, C. (2012). O ensino de química: Algumas reflexões. I Jornada de Didática-O Ensino como FOCO-I Fórum de professores de Didática do Estado Do Paraná, UTFPR.

31. Vieira, H. M. S. (2020). O uso da mineração de textos para o incremento da segurança dentro de sistemas de recuperação da informação da área financeira (Tese de doutorado, Mestrado em Sistemas de Informação e Gestão do Conhecimento).

32. Zoubi, M. B., & Rawi, M. (2008). An efficient approach for computing silhouette coefficients. Journal of Computer Science, 4, 252–255.