LUMEN ET VIRTUS
ISSN: 2177-2789

MINERAÇÃO DE TEXTO: A CLUSTERIZAÇÃO APLICADA EM ARTIGOS CIENTÍFICOS DE QUÍMICA, POR MEIO DO MODELO CASSIOPEIA

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

Data de submissão: 26/10/2024 **Data de publicação:** 26/11/2024

Diego Sampaio Amariz
Mestrando do Programa de Pós-Graduação em Educação – Universidade Federal dos Vales do Jequitinhonha e Mucuri – UFVJM

Marcus Vinícius Carvalho Guelpeli
Professor do Programa de Pós-Graduação em Educação – Universidade Federal dos Vales do Jequitinhonha e Mucuri - UFVJM

RESUMO

A química ao se dedicar à compreensão da natureza submicroscópica da matéria e suas transformações, desenvolve uma linguagem própria e produz um conhecimento fundamental sobre a natureza. Sua natureza como conhecimento básico a levou, junto com outras ciências da natureza, a compor o conhecimento de qualquer cidadão, quer seja para ler e compreender o mundo natural ou transformado pela mão do homem, quer seja para se continuar os estudos em nível superior ou técnico em outras áreas ou profissões. No entanto, assimilar e lidar com o grande volume de informação disponível, localizando-as de forma rápida e precisa, tornou-se um grande desafio, dentro da diversa gama documental existente. Com isso, Técnicas de Mineração de Texto podem auxiliar nesse processo, por meio da extração de dados textuais. Dessa forma, o objetivo dessa pesquisa é relacionar conceitos de Química encontrando palavras similares em artigos científicos de área, que possam demonstrar uma ligação entre alguns conceitos abordados no Ensino Médio. Por meio da técnica de clusterização com a utilização do modelo Cassiopeia, em um corpus de textos acadêmicos relacionados a Química. A pesquisa foi desenvolvida segundo as seguintes ações: levantamento bibliográfico; construção do corpus; coleta do corpus; análise estatística do corpus; mineração de texto; a clusterização; e, por fim, a análise dos dados a partir dos clusters gerados. Os resultados obtidos mostraram que a clusterização, efetuada no corpus proporcionou a relação entre conceitos químicos, encontrando palavras similares nos artigos científicos que compõem o corpus desenvolvido nessa pesquisa, que demonstram a ligação de conteúdos de Química do Ensino Medio.

Palavras-chave: Mineração de Texto. Corpus. Química. Clusterização. Modelo Cassiopeia.

1 INTRODUÇÃO

A química é uma ciência que permeia todas as áreas do conhecimento e envolve conceitos, reações químicas e transformação da matéria, contribuindo de forma considerável para o avanço e desenvolvimento social e tecnológico da humanidade (SCHNETZLER, 2003).

A disciplina é parte integrante da grade curricular do Ensino Médio e de alguns cursos universitários, e busca de forma geral desenvolver o conhecimento científico e a compreensão dos fenômenos químicos. A aquisição dos conhecimentos possui finalidade educativa para a formação de bases científicas e vai além, preparando os discentes para serem cidadãos mais críticos e reflexivos (BRASIL, 2018).

As principais dificuldades detectadas pelas pesquisas na área relacionam-se com a forma de como os conteúdos são repassados, bem como a metodologia didática aplicada pelos professores e a estrutura do ambiente que muitas vezes é inadequada para a consolidação do ensino com atividades teóricas e práticas (YAMAGUCHI; SILVA, 2019).

Pesquisas têm demonstrado que o ensino de Química vem sendo estruturado em torno de atividades que levam à memorização de informações, fórmulas e conhecimentos que limitam o aprendizado dos alunos e contribuem para a desmotivação em aprender e estudar.

Para ampliar a concepção sobre o ensino de Química, faz-se necessária a busca pelo aprimoramento e detecção dos entraves, para que ocorra o ensino e aprendizagem de forma plena.

Existem inúmeras discussões sobre o ensino da química, suas dificuldades de aprendizagem, a formação dos professores e as metodologias didáticas que podem colaborar com um ensino que visa a maior compreensão dos conteúdos, de forma que o ensino possa ser útil para a formação do indivíduo como um todo (SCHNETZLER, 2002).

O conteúdo de Química pode ser dividido em vários ramos, observa-se que alguns conceitos são abordados separados em capítulos independentes. Contudo, o conteúdo de Química possui uma linearização entre os capítulos e conceitos, que pode ser evidenciado, por exemplo, nos conteúdos de cinética química e no de equilíbrio químico. Onde na cinética química estuda-se a velocidade de processamento das reações químicas e no outro o equilíbrio entre as substâncias envolvidas em uma reação química, sendo que a velocidade de uma reação é uma variável importante para que ocorra o equilíbrio de uma reação (ATKINS, 2007).

Outro exemplo que se pode evidenciar são as polaridades das moléculas e as interações intermoleculares. A polaridade de uma molécula está diretamente relacionada à forma na qual os elétrons são distribuídos ao redor dos átomos. Se houver uma distribuição simétrica, a molécula será apolar, porém, se a distribuição for assimétrica, e uma das partes da molécula possuir uma grande densidade eletrônica, então será uma molécula polar (ATKINS, 2007).

Já uma interação intermolecular ocorre quando duas moléculas se aproximam, havendo uma interação de seus campos magnéticos, o que faz surgir uma força entre elas, que variam de intensidade, dependendo do tipo da molécula (polar ou apolar). Esses conceitos são evidenciados nos conteúdos de Química Inorgânica e Química Orgânica.

Logo, esses conceitos possuem uma linearização do conhecimento, evidenciando a conectividade dos mesmos. Essa linearização será foco de estudo nesse trabalho com o uso da técnica de mineração de texto, utilizando os conceitos de clusterização por meio do *software* modelo Cassiopeia (GUELPELI, 2012).

Nesse sentido, o conceito de mineração de texto (*Text Mining*) está se tornando cada vez mais popular como um método para exploração de informações. A mineração de textos (MT) é um conjunto de métodos usados para navegar, organizar, achar e descobrir informações em bases textuais. Pode ser vista como uma extensão da área de *Data Mining*, focada na análise de textos. (ARANHA; PASSOS, 2006, p. 2).

Logo, a MT é o processo de extrair informação útil (conhecimento) de dentro de um documento de texto que não está estruturado. Para isso, utiliza-se de diversas outras ferramentas conhecidas como Processamento de Linguagem Natural ou Descoberta de Conhecimento em Base Textuais (BARION; LAGO, 2018).

A técnica da mineração de texto que será utilizada é a clusterização, que consiste em conjunto de técnicas utilizadas para reunir um conjunto de objetos que apresentam características semelhantes em grupos distintos (ARORA; DEEPALI; VARSHNEY, 2015).

A ideia básica da clusterização é que elementos pertencentes a um mesmo grupo devem apresentar alta similaridade, contudo, devem ser deveras distintos de objetos de outros *clusters*. De acordo com (TAN *et al*., 2015), quanto maior a homogeneidade dentro de cada *cluster* e quanto maior a heterogeneidade entre *clusters*, melhor e mais distinta é a classificação.

Nesse contexto, o objetivo geral deste trabalho é relacionar a clusterização aplicada em artigos científicos de Química, encontrar palavras similares em artigos científicos de área, que possam demonstrar uma ligação entre alguns conceitos abordados no EM. Por meio da Mineração de Texto, utilizando a técnica de clusterização com a utilização do modelo Cassiopeia (GUELPELI, 2012).

2 METODOLOGIA

Após a leitura de alguns artigos científicos referentes a revista Química Nova na Escola, http://qnesc.sbq.org.br/edicoes.php, que aborda artigos relacionados a educação em Química, observa-se uma similaridade entre algumas palavras dentre alguns conteúdos da disciplina de Química, abordados no Ensino Médio (AMARIZ; GUELPELI, 2023).

Este estudo visa utilizar a mineração de textos, por meio da técnica de clusterização, utilizando o modelo Cassiopeia, para demonstrar a existência de uma ligação entre as palavras existentes nesses artigos, podendo relacionar os conteúdos relacionados a química (AMARIZ; GUELPELI, 2023).

O interesse em aprender os conceitos químicos está conectado à concepção de que os conhecimentos abordados permitem um olhar de um mundo mais articulado e menos fracionado. Contribuindo para que o cidadão se veja como participante de um mundo em constante transformação.

A construção de um *corpus* linguístico apresenta fatores que podem certamente auxiliar pesquisadores a obter e organizar informações para criar sua própria base de textos que auxiliem no processo de mineração de textos.

Foram coletados 120 artigos científicos disponíveis no idioma português da revista Química Nova na Escola. Essa revista foi escolhida devido se tratar exclusivamente do conteúdo de Química voltada a educação. Os artigos foram coletados de forma aleatória e a coleta teve como objetivo formar uma base de dados textuais, denominada "*Corpus*". Esse artigos científicos são referentes aos anos de 1978 a 2021.

O pré-processamento do *corpus* foi dividida em duas partes, primeiramente a conversão do formato Portable document format (PDF) para o formato que armazena texto simples (TXT), devido ao programa Get Finecount e o modelo Cassiopeia processarem esse formato.

E foi realizada a análise estatítisca do *corpus*, onde foram realizados cálculos de amplitude de palavras, médias das palavras, desvio padrão das palavras e o coeficiente de variação.

A clusterização consiste em dividir em grupos, denominados *clusters*, de modo que eles sejam mais semelhantes a outros pontos no mesmo grupo do que os de outros grupos, podendo utilizar cálculos estatísticos para cada grupo desenvolvido.

Este trabalho foi realizado pelo modelo Cassiopeia, que consiste em agrupar textos hierárquicos, para comprovar a ligação existente nos conteúdos químicos abordados no EM.

Segundo Soares (2013), a definição para cada etapa da mineração de texto é:

2.1 COLETA DOS DADOS

A premissa deste trabalho é a coleta dos textos, ou seja, a busca de artigos científicos relacionados à Química. A coleta tem como objetivo formar uma base de dados textuais, denominada "*Corpus*". Pode ser efetuada de várias maneiras, porém todas necessitam de grandes esforços, a fim de conseguir material de qualidade satisfatória e que sirva de matéria prima para a continuidade do processo e para a aquisição de conhecimento.

2.2 PRÉ-PROCESSAMENTO

Possui como objetivo preparar os documentos coletados de maneira a obter uma forma para o melhor processamento dos dados. Todo o sistema a ser desenvolvido depende de uma filtragem dos textos, ou seja, uma redução da quantidade de palavras para obter uma informatividade efetiva, que pode proporcionar um ganho qualitativo e quantitativo para o processamento dos artigos químicos.

2.3 INDEXAÇÃO

É responsável por estabelecer índices com o objetivo de estabelecer maior rapidez e agilidade para a recuperação dos documentos e seus termos.

2.4 MINERAÇÃO

O processamento utiliza o agrupamento de texto hierárquicos e um algoritmo para juntar os textos com similaridades. A clusterização é realizada pelo modelo Cassiopeia, que identifica as características das palavras nos artigos científicos, utilizando a frequência relativa, que defini a importância dos termos, de acordo com as periodicidades em cada texto utilizado. O modelo Cassiopeia, proporciona a remoção das *stopwords*, palavras que não fazem contexto a essa pesquisa.

2.5 ANÁLISE

Consiste na avaliação dos dados obtidos. Nessa etapa o modelo agrupa os artigos científicos, relacionados à Química, por similaridade possibilitando uma melhor avaliação dos dados, com eficiente grau de informatividade.

A mineração de texto, por meio das técnicas de clusterização, utilizando o modelo Cassiopeia (Guelpeli, 2012) organiza os artigos científicos voltados para Química, conforme a similaridade entre as palavras existentes em cada artigo clusterizado.

Dessa forma, o conjunto contendo os cento e vinte artigos de Química, foram submetidos ao modelo Cassiopeia. O modelo executou o processo de agrupamento e reagrupamento, com o conjunto de cento e vinte artigos trinta vezes, ou seja, o modelo Cassiopeia analisou o *corpus* dessa pesquisa repetidamente durante essas trinta interações.

Os resultados dessas clusterizações foram analisados por meio de análises qualitativas dos artigos que constituem cada cluster e quantitativo, por meio da métrica interna ou não supervisionada denominada Coeficiente de Silhouette, explicadas no capítulo anterior anterior.

A análise dos dados foi realizada por meio das frequências de palavras similares ocorrentes nos artigos, efetuando os cálculos e uma tabela de frequências de palavras, em seguida foram analisados os *clusters* gerados durante o processo de clusterização.

Os *clusters* gerados por meio da clusterização foram analisados e foi efetuada uma seleção dos melhores dados gerados, por meio de resultados quantitativos, que foram os resultados das métricas internas, durante o processo de mineração de texto, e por meio de reultados qualitativos, que foram a análise dos conteúdos similares geados em cada *cluster*.

3 RESULTADOS E DISCUSSÕES

O *corpus* foi convertido do formato PDF para o formato TXT para o processamento computacional, devido aos programas computacionais utilizados nessa pesquisa aceitarem o formato TXT. Além disso, nessa conversão foram retiradas as imagens, gráficos, tabelas, números de páginas e todas as anotações que não faziam parte do corpo do texto. Os arquivos foram renomeados seguindo uma ordem sequencial iniciando em 1 e terminando em 120 e foi utilizado um *software* de análise denominado *Get Finecount* 2.6 para a contagem das palavras.

O software *Get Finecount* 2.6 é uma ferramenta que fornece análises de um documento. Analisa um texto e conta a quantidade de palavras, caracteres, repetições, espaços, espaços redundantes, linhas, frases e páginas de um arquivo ortográfico. Foi importante para a verificação da quantidade de palavras existentes em cada artigo que constitui o *corpus* dessa pesquisa.

No *corpus* foi realizado o cálculo da amplitude, por meio da quantidade de palavras existentes nos artigos. A amplitude é uma medida de dispersão que determina o grau de variação dos números, essa medida é determinada pela diferença entre o valor máximo encontrado e o valor mínimo, conforme a Equação matemática 1.

$$R = X_{máximo} - X_{mínimo}$$

Onde X máximo é o valor máximo de palavras encontrado e X mínimo é o valor mínimo de palavras encontrado.

O artigo que contém a maior quantidade de palavras apresentou 5.566 e o artigo que apresentou a menor quantidade de palavras apresentou 2.428 palavras. A amplitude encontrada no *corpus* de química foi 3.138 palavras (AMARIZ; GUELPELI, 2023).

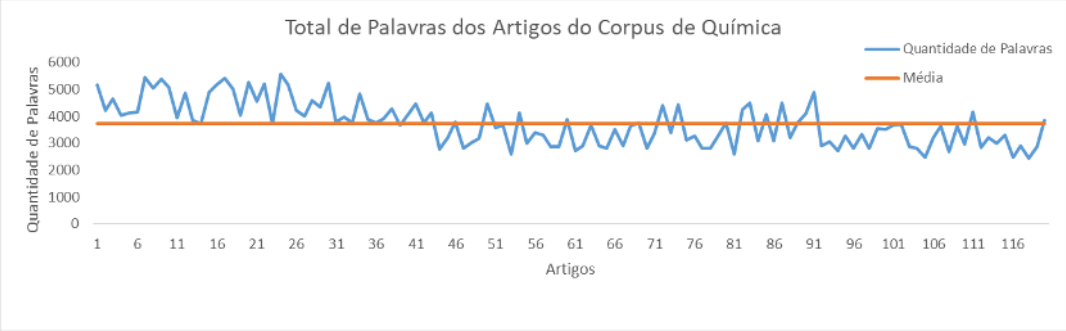
Também foi realizado o cálculo da média das palavras existente nos artigos do *corpus*. A Média faz parte dos conceitos da Estatística. No que concerne à média, esta pode ser aritmética (simples ou ponderada), geométrica, harmônica, quadrática, cúbica ou biquadrática. Tratando especificamente da média aritmética, esta é considerada como “o conceito mais básico da Estatística e da Ciência experimental, é também o mais utilizado na vida cotidiana das pessoas” (MAGINA; CAZORLA; GITIRANA; GUIMARÃES, 2010, p. 61-62), a Equação 2 é utilizada para o cálculo da média.

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

Onde, Xi é a quantidade total de palavras existentes em todos os artigos e n a quantidade de artigos selecionados.

Realizado esse cálculo observou-se que a média de palavras para os artigos de Química do *corpus* foi 3.730 palavras. A quantidade de palavras em relação à média está evidenciada na Figura 6. Esse gráfico evidência a dispersão da quantidade de palavras em relação à média (AMARIZ; GUELPELI, 2023).

Figura 6: Dispersão da quantidade de palavras em relação à média do *Corpus*.



Fonte: Próprio autor

Com o valor da média obtido calculou-se o desvio padrão existente nesse *corpus*. O desvio padrão é o protótipo das medidas de dispersão em virtude de suas propriedades matemáticas e de seu uso na teoria da amostragem" (OLIVEIRA, 2017, p. 8). É uma medida de dispersão, que indica o quanto o conjunto de dados é uniforme. Conforme Equação matemática 3.

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

Onde, X é o valor individual de cada palavra de cada artigo, $\overline{X}$ é a média dos dados obtidos e n número total de artigos do *corpus*.

O desvio padrão para os dados do *corpus* de Química foi 800 palavras. Com os dados obtidos da média e do desvio padrão pode-se calcular o coeficiente de variação dos dados.

Experimentos confiáveis requerem a avaliação dos resultados pela verificação da precisão deles próprios, que pode ser realizada pelos valores dos coeficientes de variação, CV, (NESI *et al*., 2010; STORCK *et al*., 2011). Conforme Steel *et al*. (1997), o CV permite a comparação de resultados de

diferentes experimentos, envolvendo uma mesma variável ou espécie, permitindo, assim, quantificar a precisão de suas pesquisas.

O CV é uma medida importante sobre a variabilidade dos resultados experimentais, podendo ser útil na definição do número de repetições do ensaio, necessário para detectar uma diferença entre médias de tratamentos com uma dada probabilidade (PIMENTEL-GOMES, 2009; NESI *et al*., 2010).

De acordo com Storck *et al*. (2011), a distribuição de CV possibilita estabelecer faixas de valores que orientam os pesquisadores sobre a validade de seus experimentos. Dessa forma, pode-se dizer que o coeficiente de variação é uma forma de expressar a variabilidade dos dados excluindo a influência da ordem de grandeza da variável. O coeficiente de variação é igual ao desvio-padrão dividido pela média aritmética, multiplicado por 100% (LEVINE et al., 2014). A Equação 4 evidência como é realizado o cálculo para o CV.

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

Onde, S é o desvio padrão e $\overline{X}$ é a média dos dados obtidos

Como o coeficiente de variação analisa a dispersão em termos relativos, ele será dado em porcentagem. Quanto menor for o valor do coeficiente de variação, mais homogêneos serão os dados, ou seja, menor será a dispersão em torno da média (LEVINE et al., 2014). De uma forma geral, se o CV for menor ou igual a 15% o resultado apresenta uma baixa dispersão dos dados, dados homogêneos. Se o cálculo dos dados ficarem entre 15 e 30% os mesmos apresentam uma média dispersão. E se for maior que 30% apresentam uma alta dispersão, dados heterogêneos (LEVINE et al., 2014).

Portanto, ao se calcular o coeficiente de variação dos dados do *corpus* de Química obteve-se 21,45%, o que indica uma média homogeneidade dos artigos que compõem esse *corpus*. Entretanto esse valor obtido está mais próximo de 15% do que de 30% o que indica uma maior tendência a homogeneidade do que uma tendência para a heterogeneidade.

Todos os cálculos, amplitude, média aritmética, desvio padrão e coeficiente de variação, efetuados nesse *corpus* estão evidenciados na Tabela 1.

Tabela 1: Cálculos realizados no *Corpus* de Química.

Tabela Estatística	
Corpus	Quantidade
Média Aritmética	3.730 palavras
Desvio Padrão	800 palavras
Amplitude	3.138 palavras
Coeficiente de Variação	21,45%

Fonte: Próprio autor

Ao lidar com dados textuais, é preciso manter em mente que, para que se possa compreender de uma forma ágil e simplificada o conhecimento implícito, é necessário encontrar maneiras que possam representá-lo para transmitir algum conhecimento (SARGIANI et al., 2018). Algumas ferramentas gráficas, como histogramas e núvens de palavras podem ser utilizados no processo para avaliar documentos contendo textos não estruturados e para exploração de conhecimento oculto nos textos (BRUNO,2016).

O modelo Cassiopeia coloca todas as letras em minúsculas, além de outros cuidados, como descarte de todas as figuras, tabelas, marcações existentes e a remoção ou não de *stopwords*. A função denotada para as *stopwords* pode ser configurada pelo usuário.

Além disso, o modelo Cassiopeia identifica as características das palavras no documento, utilizando a frequência relativa, que define a importância de um termo, de acordo com a frequência com que é encontrado no documento. Quanto mais um termo aparecer em um documento, mais importante é, para aquele documento (GUELPELI, 2012). Tendo como base os pesos das palavras, obtidos na frequência relativa, é calculada a média sobre o total de palavras no documento.

A técnica de mineração de texto utilizada nessa pesquisa é a clusterização, por meio do modelo Cassiopeia, que separou os artigos em *cluster*, conjuntos de objetos que possuem semelhança entre si. Sendo assim, os artigos que possuem semelhanças entre as palavras ficaram em um mesmo *cluster*.

O modelo Cassiopeia realizou o processo de agrupamento e reagrupamento trinta vezes, ou seja, trinta interações foram realizadas, em que, cada artigo foi analisado 30 vezes sendo assim, foram totalizadas três mil e seiscentas análises. Com isso, o modelo realizou o cálculo do Coeficiente de Silhouette (CS) para a realização da análise quantitativa dos *clusters*.

Segundo Guelpeli (2012), o Coeficiente Silhouette baseia-se na ideia de quanto um objeto é similar aos demais membros do seu grupo, e de quanto este mesmo objeto é distante dos de um outro grupo. Assim, essa medida combina as medidas de coesão e acoplamento. A Equação 5 evidência como é realizado o cálculo para o Coeficiente de Silhoette.

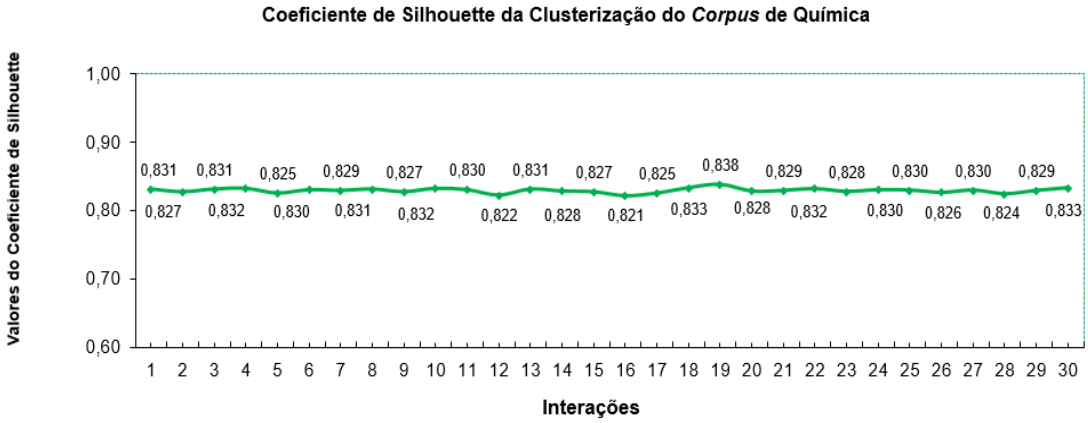
$$CS = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

Onde a(i) é a distância média entre o i-ésimo elemento do grupo e os outros do mesmo grupo. O b(i) é o valor mínimo de distância entre o i-ésimo elemento do grupo e qualquer outro grupo, que não contém o elemento, e max é a maior distância entre a(i) e b(i).

O Coeficiente Silhouette de um grupo é a média aritmética dos coeficientes calculados para cada elemento pertencente ao grupo, o valor de CS situa-se na faixa de 0 a 1 (GUELPELI,2012).

Para melhor organização dos resultados dos Coeficientes de Silhoette gerados pela clusterização, por meio do modelo Cassiopeia, obtida ao longo das 30 interações. Os resultados estão apresentados na Figura 7.

Figura 7: Coeficiente de Silhouette da Clusterização do *Corpus* de Química.



Fonte: Próprio autor

O coeficiente ou índice Silhouette (CS) é um valor que mede o quão similar um objeto é em relação ao seu próprio cluster (coesão) em comparação com os demais clusters (acoplamento).

Segundo Guelpeli (2012), o coeficiente varia entre 0 a 1, em que valores próximos de 1 indica que o objeto está bem relacionado ao seu *cluster* e valores próximos de 0 indicam que o objeto não esta bem relacionado com seu *cluster*.

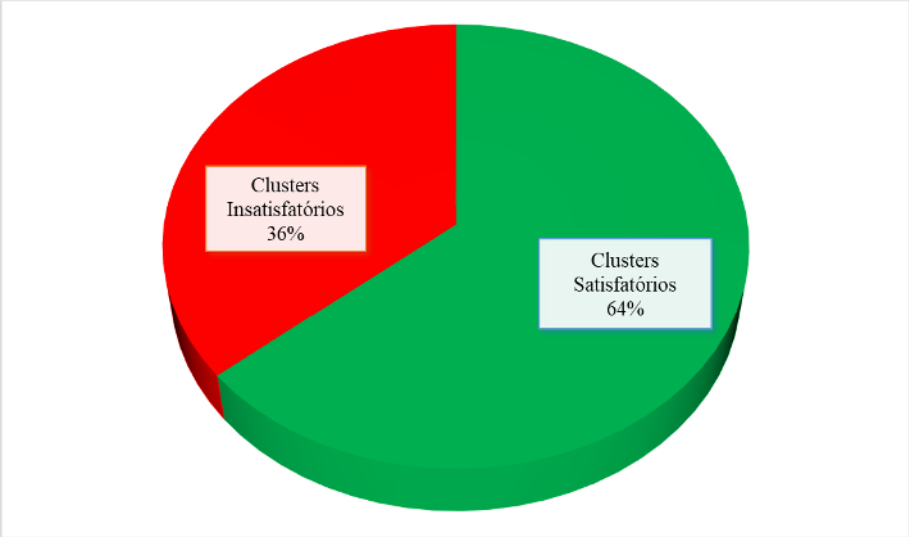
Em todas as interações observa-se valores para o Coeficiente de Silhouette compreendidos entre 0,821, valor mínimo encontrado e 0,838, valor máximo encontrado na clusterização, valores esses bem próximos de 1 o que indica que os artigos científicos que compõem o *corpus* dessa pesquisa estão bem relacionados dentro de seus respectivos *clusters*.

Após a análise quantitativa dos *clusters* foram realizadas as análises qualitatitas e observa-se que a clusterização gerou 36 *clusters*, ou seja, trinta e seis conjuntos de artigos com palavras semelhantes.

Nesse conjunto observa-se que alguns continham poucos conteúdos de Química agrupados. Enquanto outros apresentavam muitos conteúdos agrupados em um mesmo *cluster* e também ocorreram *clusters* com o mesmo conteúdo agrupado, isso ocorreu devido a similaridade entre as palavras.

Diante desses conjuntos de *clusters* gerados, os mesmos foram classificados em *clusters* satisfatórios e *clusters* insatisfatórios, conforme a Figura 8.

Figura 8: Análise dos *Clusters*.



Fonte: Próprio Autor

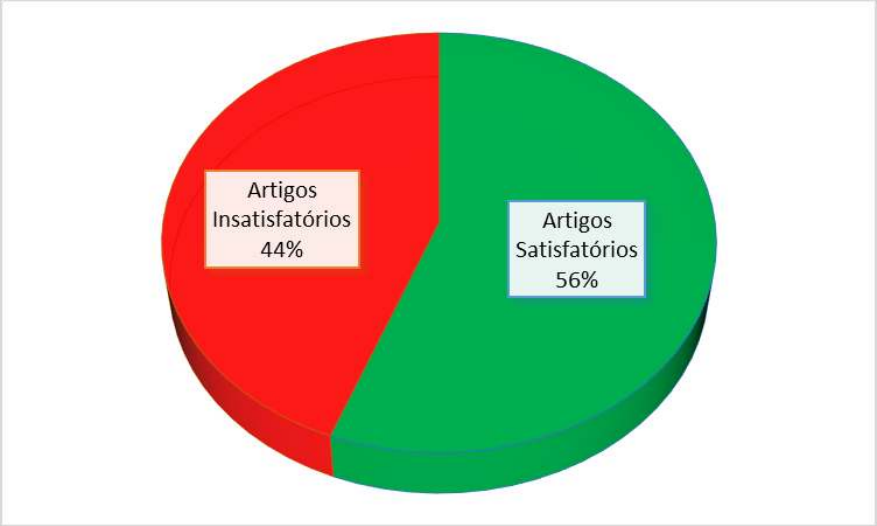
Conforme a Figura 8, a análise gerou 64% de *clusters* satisfatórios, ou seja 23 *clusters*, que agruparam artigos que continham de dois a três conteúdos distintos de química abordados no EM. Esses *clusters* apresentaram uma frequência de palavras similares em seus artigos, por isso foram classificados como satisfatórios.

Assim sendo, foram gerados 13 *clusters* insatisfatórios, computando 36%, que não apresentaram uma frequência considerável de palavras similares nos seus artigos e agruparam vários conteúdos de química em um mesmo *cluster*.

Essa análise ocorreu devido a clusterização realizada pelo modelo Cassiopeia, que indica que é satisfatório uma análise com maior frequência de palavras similares em um *clusters*, do que variados conteúdos em um mesmo *cluster*.

Dessa forma, foram clusterizados um *corpus* contendo 120 artigos, que foram divididos em duas classes, sendo elas, *clusters* satisfatórios e *clusters* insatisfatórios. Foram realizadas análises para identificar a quantidade de artigos que estão presentes nos *clusters* satisfatórios e nos c*lusters* insatisfatórios. Essas divisões estão evidênciadas na Figura 9.

Figura 9: Análise das divisões dos artigos dentro dos *Clusters*.



Fonte: Próprio Autor

O *corpus* apresenta 56% de seus artigos nos *clusters* satisfatórios, ou seja 67 artigos do *corpus*, estão agrupados nos *clusters* de interesse para essa pesquisa. Esses artigos apresentaram uma frequência de palavras similares dentro de um mesmo *cluster.*

Entretanto, o *corpus* apresenta 44% dos artigos dentro dos *clusters* insatisfatórios, computando 53 artigos, que apresentaram uma frequência considerável de palavras similares dentro dos seu *cluster*, porém agruparam vários conteúdos de química em um mesmo *cluster* ou conteúdos semelhantes aplicados no EM.

Com os dados adquiridos pela clusterização, realizou-se uma análise de cada *cluster* relacionando os artigos e seus conteúdos abordados, com os anos em que os mesmos são ministrados durante o Ensino Médio, conforme Tabela 2.

Tabela 2: Relação entre os conteúdos de cada *Cluster* e os anos do EM.

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

Fonte: Próprio autor

Com isso, observa-se que, por meio da clusterização, o *corpus* dessa pesquisa apresenta os conteúdos abordados no EM em 26 *clusters* contendo conteúdos do 1º ano e do 2º ano do EM e 16 *clusters* contendo os conteúdos ministrados no 3º ano do EM.

Os 23 *clusters* satisfatórios gerados nessa pesquisa, estão evidenciados na Tabela 3, que indica a relação entre entre os conteúdos de cada *clusters* satisfatórios e os anos do EM em que são abordados.

Tabela 3: Relação entre os conteúdos dos *Cluster* satisfatórios e os anos do EM.

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

Fonte: Próprio autor

Verifica-se que em 12 *clusters* satisfatórios ocorreram a presença de conteúdos que podem ser ministrados no 1º ano do EM, 15 *clusters* apresentaram conteúdos que podem ser ministrados no 2º ano do Em e 9 *clusters* evidenciaram conteúdos do 3º ano do EM. Com esses resultados foram selecionados cinco *Clusters* satisfatórios para serem abordados.

O *cluster 3,* gerado pelo modelo Cassiopeia, apresenta similaridade entre os conteúdos de eletroquímica e tabela periódica, conteúdos ministrados em períodos diferentes do EM, pois apresentaram palavras semelhantes a estes conteúdos.

Isso ocorreu devido a análise das palavras encontradas no conjunto de artigos deste *cluster*. Palavras como: eletrólise, solução, pilhas, cátodo, oxidação, redução, elétrons, eletrodo, metal e eletronegativo estão presentes em todos os artigos desse conjunto textual, com uma determinada frequência relativa, conforme Tabela 4.

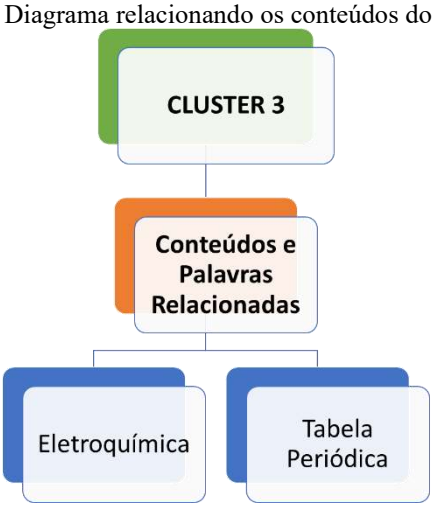
Tabela 4: Frequência relativa das palavras do Cluster 3

Tabela de frequência relativa das palavras do *Cluster 3*		
	Frequencia Relativa (%)	
Palavras	Artigo 44	Artigo 56
Eletrólise	70,6	29,4
Valência	48,9	51,1
Pilhas	43,9	56,1
Cátodo	60,9	39,1
Oxidação	46,7	53,3
Redução	61,8	38,2
Elétrons	45,0	55,0
Eletrodo	73,1	26,9
Metal	30,6	69,4
Eletronegativo	66,7	33,3

Fonte: Próprio autor

A frequência relativa das palavras que mais apareceram nos artigos desse *cluster* evidência a relação entre os conteúdos de eletroquímica e tabela periódica, conforme Figura 10.

Figura 10: Diagrama relacionando os conteúdos do *Cluster 3*.



Fonte: Próprio Autor

Palavras como valência, pilhas oxidação, elétrons apresentaram frequências relativas consideráveis no artigo 44 e no artigo 56, que integram esses *cluster*. As palavras que mais aparecem nos artigos e caracterizam o conteúdo de eletroquímica são eletrólise, pilhas, cátodo, oxidação, redução, elétrons e eletrodo. Assim como, as palavras elétrons, metal, valência e eletronegativo são muito abordadas no conteúdo de tabela periódica.

O *cluster* 14, gerado pelo modelo Cassiopeia, apresenta similaridade entre os conteúdos modelos atômicos e ligações químicas, esses conteúdos são ministrados no 1º ano do EM. Esse *cluster* é composto por dois artigos, o artigo 7 e o artigo 17 do *corpus* que constituem essa pesquisa.

Isso ocorreu devido a análise das palavras encontradas no conjunto de artigos que formam esse *cluster*. Palavras como: Dalton, leis, átomo, ponderais, ligação, elétrons, molécula, geometria,

polaridade e iônica estão presentes em todos os artigos desse conjunto textual, com uma determinada frequência relativa, conforme Tabela 5.

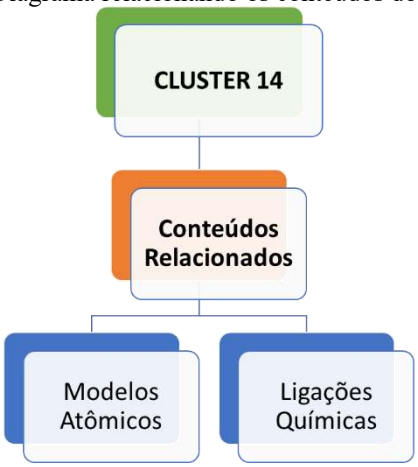
Tabela 5: Frequência relativa das palavras do *Cluster 14*.

Tabela de frequência relativa das palavras do *cluster 14*		
	Frequencia Relativa (%)	
Palavras	Artigo 7	Artigo 17
Dalton	14,6	85,4
Leis	16,7	83,3
Átomo	60,0	40,0
Ponderais	33,3	66,7
Ligação	69,4	30,6
Elétrons	58,3	41,7
Molécula	61,5	38,5
Geometria	34,5	65,5
Polaridade	58,8	41,2
Iônica	47,1	52,9

Fonte: Próprio autor

A frequência relativa das palavras que mais aparecem nos artigos que constituem esse *cluster* evidência a relação entre os conteúdos modelos atômicos e ligações químicas. A Figura 11 apresenta um diagrama relacionando os conteúdos ao *cluster* 14.

Figura 11: Diagrama relacionando os conteúdos do *Cluster 14*.



Fonte: Próprio Autor

O *cluster* 28, gerado pelo modelo Cassiopeia, apresenta similaridade entre os conteúdos de estrutura atômica e polímeros, conteúdos ministrados em períodos diferentes do EM, pois apresentaram palavras semelhantes a esses conteúdos.

Assim, palavras como plástico, polímero, elétrons, blindagem, átomo, ionização, eletrônica e orbitais estão presentes no conjunto de artigos desse *cluster*, com uma determinada frequência relativa, conforme Tabela 6.

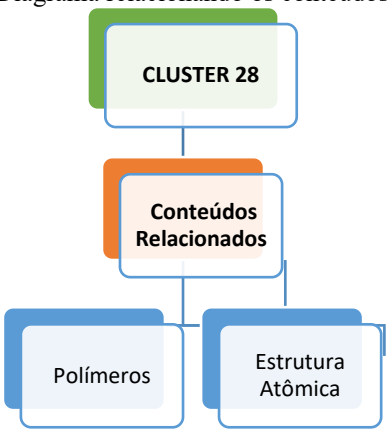
Tabela 6: Frequência relativa das palavras do *Cluster 28*

Tabela de frequência relativa das palavras do *cluster 28*					
	Frequencia Relativa (%)				
Palavras	Artigo 14	Artigo 91	Artigo 108	Artigo 119	Artigo 120
Plástico	4,9	8,7	36,9	35,0	14,6
Polímero	11,9	20,3	28,8	25,4	13,6
Elétrons	39,5	25,8	8,9	14,5	11,3
Blindagem	66,7	33,3	0,0	0,0	0,0
Átomo	31,4	14,7	17,6	15,7	20,6
Ionização	35,3	17,6	20,6	11,8	14,7
Eletrônica	75,0	25,0	0,0	0,0	0,0
Orbitáis	25,0	39,6	6,3	18,8	10,4

Fonte: Próprio autor

A frequência relativa das palavras que mais apareceram nos artigos desse *cluster* evidência a relação entre os conteúdos estrutura atômica e polímeros, conforme Figura 12.

Figura 12: Diagrama relacionando os conteúdos do *Cluster 28*.



Fonte: Próprio Autor

Palavras como polímeros, elétrons, átomo e ionização apresentaram frequências relativas consideráveis nos artigos que integram esse *cluster*. As palavras que mais aparecem nos artigos e caracterizam os conteúdo apresentados na Figura 12.

O *cluster* 32, gerado pelo modelo Cassiopeia, também apresenta similaridade entre alguns conteúdos relacionados ao ensino de química para o EM. Os conteúdos química ambiental e conceitos ácido-base estão evidenciados nesse *cluster*.

Logo, palavras como efeito, estufa, infravermelho, absorção, gases, ácido, base, íons, indicadores e reação estão presentes no conjunto de artigos que constituem esse *cluster*, conforme Tabela 7.

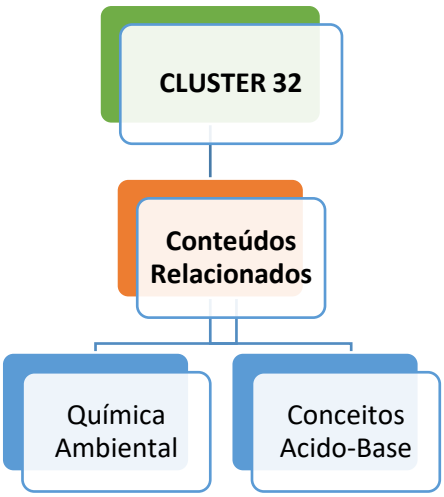
Tabela 7: Frequência relativa das palavras do *Cluster 32*.

Tabela de frequência relativa das palavras do *cluster 32*		
	Frequencia Relativa (%)	
Palavras	Artigo 22	Artigo 23
Efeito	61,2	38,8
Estufa	61,2	38,8
Infravermelho	61,5	38,5
Absorção	45,8	54,2
Gases	58,7	41,3
Ácido	28,0	72,0
Base	36,8	63,2
Íons	40,9	59,1
Indicadores	59,3	40,7
Reação	55,0	45,0

Fonte: Próprio autor

A frequência relativa das palavras que mais aparecem nos artigos que constituem esse *cluster* evidência a relação entre os conteúdos química ambiental e conceitos ácido-base. A Figura 13 apresenta um diagrama relacionando os conteúdos ao *cluster* 32.

Figura 13: Diagrama relacionando os conteúdos do *Cluster* 32.



Fonte: Próprio Autor

O *cluster* 32 apresenta o conteúdo conceitos ácido-base, que é atualmente ministrado no primeiro ano do EM e o conceito de química ambiental é ministrado no meio do EM, no segundo ano.

O *cluster* 36, gerado pelo modelo Cassiopeia, apresenta similaridade entre os conteúdos interações intermoleculares e compsotos orgânicos. Esse *cluster* é composto por dois artigos, o artigo 81 e o artigo 83 do *corpus* que constitue essa pesquisa.

Isso ocorreu devido a análise das palavras encontradas no conjunto de artigos deste *cluster*. Palavras como polar, apolar, ligação, dipolo, hidrogênio, molécula, detergente e orgânica estão

presentes em todos os artigos desse conjunto textual, com uma determinada frequência relativa, conforme Tabela 8.

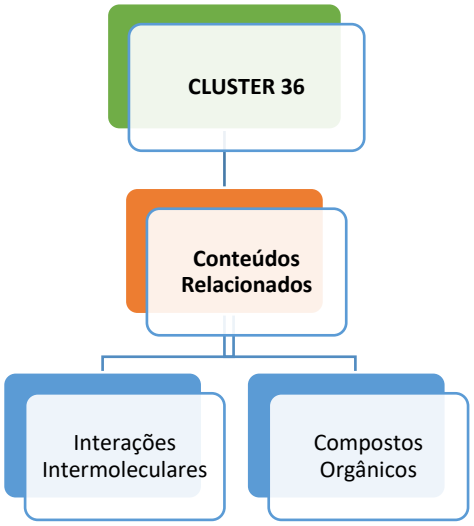
Tabela 8: Frequência relativa das palavras do *Cluster 36*.

Tabela de frequência relativa das palavras do *cluster 36*		
	Frequencia Relativa (%)	
Palavras	Artigo 81	Artigo 83
Polar	58,8	41,2
Apolar	46,7	53,3
Ligação	61,1	38,9
Dipolo	47,5	52,5
Hidrogênio	61,1	38,9
Molécula	60,9	39,1
Detergente	100,0	0,0
Orgânica	41,7	58,3

Fonte: Próprio autor

A frequência relativa das palavras que mais apareceram nos artigos desse *cluster* evidência a relação entre os conteúdos de eletroquímica e tabela periódica, conforme Figura 14.

Figura 14: Diagrama relacionando os conteúdos do *Cluster 36*.



Fonte: Próprio Autor

O *cluster* 36 apresenta o conteúdo interação intermolecular, que é atualmente ministrado no primeiro ano do EM e o conceito relacionado a compostos orgânicos é ministrado no ano final do EM, no terceiro ano.

Os 13 *clusters* não satisfatórios gerados nessa pesquisa, estão evidenciados na Tabela 9, indicando a relação entre entre os conteúdos de cada *clusters* não satisfatórios e os anos do EM em que são abordados.

Tabela 9: Relação entre os conteúdos dos *Clusters* não satisfatórios e os anos do EM.

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		

Fonte: Próprio autor

Esses *clusters* foram classificados como não satisfatórios devido a não apresentarem palavras similires entre os artigos que os compõem. Alem disso, demonstram grande quantidade de artigos em um mesmo *cluster*, proporcionando uma similaridade entre vários conteúdos químicos ao mesmo tempo. Essa similaridade entre vários artigos, que apresentam conteúdos variados, pode ser considerada como uma imprecisão na clusterização.

4 CONCLUSÃO

Devido a grande quantidade de textos disponíveis, principalmente na área da educação, surge a necessidade da mineração de texto, que pode facilitar na busca e recuperação de informações textuais. Com isso, essa pesquisa utilizou-se da clusterização, uma das técnicas empregadas na mineração de texto, como forma de contribuição na solução de alguns dos problemas evidênciados nesse trabalho, tal como o tratamento isolado de alguns conceitos abordados na disciplina de Química.

O *corpus* dessa pesquisa foi constituído de 120 artigos científicos, foi realizada a contagem de palavras em cada artigo que constitui o *corpus* e realizada uma análise estatística. A análise estatística determinou uma média das palavras do *corpus*, o que proporciona uma pequena variação entre a quantidade de palavras entre os artigos (AMARIZ; GUELPELI, 2023).

Com o valor da média obteve-se o desvio padrão das palavras dos artigos, observa-se que esse valor é pequeno em relação a quantidade de palavras dos textos de cada artigo do *corpus*. Isso indica que as palavras estão condensadas próximas da média, indicando uma tendência a homogeneidade do *corpus* (AMARIZ; GUELPELI, 2023).

Essa homogeneidade pode ser evidênciada, por meio do cálculo do coeficiente de variação, que comprova uma tendência mais homogênea com o o resultado efetuado para o *corpus* em 21,45%.

Obtendo assim, uma tendência mais homogênea do que heterogênea para o *corpus* químico desenvolvido nessa pesquisa.

Depois da análise estatística, realizou-se a mineração de texto, por meio do modelo Cassiopeia. O modelo Cassiopeia retirou todos os itens indesejáveis do texto, deixando apenas as palavras importantes para que a mineração de texto fosse utilizada.

A técnica de clusterização é utilizada nessa pesquisa para a realização da mineração de texto. Assim, retomando a questão central desta pesquisa, é possível concluir que o modelo Cassiopeia proporcionou a relação entre conceitos químicos, encontrando palavras similares nos artigos científicos que compõem o *corpus* desenvolvido nessa pesquisa. Isso pode ser mensurado quantitativamente por meio do coeficiente de Silhouette e qualitativamente pelas análises dos artigos constituintes de cada cluster.

Dessa forma, o principal objetivo dessa pesquisa foi relacionar conceitos de Química encontrando palavras similares em artigos científicos de área, que possam demonstrar uma ligação entre alguns conceitos abordados no Ensino Médio. Por meio da Mineração de Texto, utilizando a técnica de clusterização com a utilização de um *software* denominado Cassiopeia, em um *corpus* de artigos acadêmicos. A partir das avaliações dos artigos que compõem o *corpus*, análises estatísticas, que comprovaram a homogeneidade desse *corpus* e a clusterização, constatou-se que a hipótese da pesquisa foi confirmada.

Destaca-se como contribuição desta pesquisa a criação de um *corpus* relacionado ao conteúdo de Química, que pode ser utilizado por pesquisadores em trabalhos futuros. Além disso, destaca-se a relação existente entre palavras em diversos artigos do *corpus*, que demonstram a ligação de conteúdos de Química abordados no EM (AMARIZ; GUELPELI, 2023).

REFERÊNCIAS

AMARIZ, D. S; GUELPELI, M. V. C. Linguistica de *corpus* aplicado a artigos científicos de química. International Journal of Development Research Vol. 13, Issue, 02, pp. 61813-61815, February, 2023

ARAUJO, Amanda Caroline Ferreira; DE OLIVEIRA FÉLIX, Maria Elisabeth; DA SILVA, Gilberlândio Nunes. RELATO DAS DIFICULDADES EM APRENDER QUÍMICA DE ALUNOS DA EDUCAÇÃO BÁSICA DE UMA ESCOLA PÚBLICA DE CAMPINA GRANDE.

ARANGANAYAGIL, S. and THANGAVEL, K. Clustering Categorial Data Using Silhouette Coefficient as a Relocating Measure. In International conference on computational Intelligence and multimedia Applications, ICCIMA, 2007, Sivakasi, India. Proceedings. Los Alamitos: IEEE 2007. p13-17.

ATKINS, Peter; JONES, Loretta; LAVERMAN, Leroy. Princípios de Química-: Questionando a Vida Moderna e o Meio Ambiente. Bookman Editora, 2007.

BIZERRA, Ayla Márcia Cordeiro et al. Dificuldades e motivações no ensino de química: uma análise da perspectiva docente. VI CONEDU - Vol 1... Campina Grande: Realize Editora, 2020. p. 1406-1420. Disponível em: <https://editorarealize.com.br/artigo/visualizar/65351>. Acesso em: 26/11/2022 13:43

CAVALCANTE, Danielle Baltazar. Análise do desempenho de parques eólicos por meio de clusterização de aerogeradores. 2020.

CAVALCANTE, Marcelo; DA COSTA, João Gomes. Considerações sobre planejamento experimental e adequabilidade do uso de testes estatísticos em Ciências Agrárias. Diversitas Journal, v. 6, n. 4, p. 3706-3723, 2021.

CRUZ, Luanna Azevedo. Modelo para recuperação de informação em repositórios institucionais utilizando a técnica de sumarização a partir da seleção de atributos do Cassiopeia. 2019.

DE AGUIAR, Luís Henrique G. et al. Uma coleção de artigos científicos de Português compondo um no domínio educacional. Plurais Revista Multidisciplinar, v. 2, n. 2, p. 107-119, 2017.

DE LIMA YAMAGUCHI, Klenicy Kazumy. Ensino de química inorgânica mediada pelo uso das tecnologias digitais no período de ensino remoto. Revista Prática Docente, v. 6, n. 2, p. e041-e041, 2021.

GIULIANI, Ricardo et al. Clusterização de trajetórias multiaspecto usando árvores de decisão. 2022.

GUELPELI, MARCUS VINICIUS CARVALHO. Cassiopeia: Um modelo de agrupamento de textos baseado em sumarização. Niterói: Tese (Doutorado em Computação) - Univerisade Federal Fluminense, 2012.

GUELPELI, Marcus VC; BRANCO, Antonio Horta; GARCIA, Ana Cristina B. 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. IEEE, 2009. p. 1-8.

KUNZ, T., BLACK, J.P.: Using Automatic Process Clustering for Design Recovery and Distributed Debugging. IEEE Trans. Software Eng.515-527,1995.

LEVIN, Jack; FOX, James Alan. Estatística para ciências humanas. In: Estatística para ciências humanas. 2004. p. xv, 497-xv, 497.

LUHN, H. P. The automatic creation of literature abstracts. IBM Journal of Research and Development, 2, pp. 159-165, 1958.

MAGINA, Sandra; FONSECA, Sônia. 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, v. 7, n. 1, 2016.

MAXIMIANO, Flavio Antonio. Princípios para o currículo de um curso de Química. Estudos Avançados, v. 32, p. 225-245, 2018.

MCHUGH, M.L. Descriptive statistics, Part II: Most commonly used descriptive statistics. J Spec Pediatr Nurs. 2003 Jul-Sep;8(3):111-6.

OLIVEIRA, Cassius Gomes et al. Desvio padrão e imprecisão de leitura: Paquímetro. Caderno de Graduação-Ciências Exatas e Tecnológicas-UNIT-SERGIPE, v. 5, n. 3, p. 27-27, 2019.

OLIVEIRA, Hércules Batista de. Framework Oráculo: camada de coleta e mineração de textos para o Twitter. 2019.

PINTO, Ivan de Jesus Pereira. *Corpus* para o Domínio Acadêmico: Modelos e Aplicações. 2021. Tese de Doutorado. PUC-Rio.

REIS, Laura Andrade et al. Comportamento ingestivo de ovinos em pastos heterogêneos de capim-marandu com mesma altura média. 2021.

SANTIAGO, Johan Carlos et al. Compostos orgânicos versus inorgânicos: um estudo sobre as propriedades físico-químicas entre essas duas classes de compostos. ENCICLOPEDIA BIOSFERA, v. 11, n. 21, 2015.

SCHMILDT, Edilson Romais et al. Coeficiente de variação como medida da precisão em experimentos de alface. Revista Agro@ mbiente On-line, v. 11, n. 4, p. 290-295, 2017.

SILVA, Rogerio de Aquino et al. Uma metodologia para criação de um *corpus* textual adequada ao reconhecimento de entidades nomeadas em português. 2021.

TAN, P. N.; STEINBACH, M.; and KUMAR, V. Introduction to Data Mining.Addison-Wesley, 2006.

TWYCROSS, A; SHIELDS, L. Statistics made simple. Part 1. Mean, medians and modes. Paediatr Nurs. 2004 May;16(4):32.

RODRIGUES, Célio Fernando de Sousa; LIMA, Fernando José Camello de; BARBOSA, Fabiano Timbó. Importância do uso adequado da estatística básica nas pesquisas clínicas. Revista brasileira de anestesiologia, v. 67, p. 619-625, 2017.

VEIGA, Márcia S. Mendes; QUENENHENN, Alessandra; CARGNIN, Claudete. 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, 2012.

VIEIRA, Hector Matheus Soares. O uso da mineração de textos para o incremento da segurança dentro de sistemas de recuperação da informação da área financeira. 2020. Tese de Doutorado. Mestrado em Sistemas de Informação e Gestão do Conhecimento.

ZOUBI, M. B. and RAWI, M. An Efficient Approach for Computing Silhouette Coefficients. Journal of Computer Science Volume 4 Page No.: 252 – 255, 2008.