Cadernos do IME - Série Matemática
DOI: 10.12957/cadmat.2020.54666
***e-ISSN:** 2236-2797 (online)*
Recebido em 24 Set. 2020 | Aceito em 26 Out. 2020

MINERAÇÃO DE TEXTO NA IDENTIFICAÇÃO DE PALAVRAS-CHAVE NO CONTEXTO DO COVID-19 NA MODELAGEM *FUZZY*

TEXT MINING TO IDENTIFICATION FOR KEYWORDS IN COVID-19 CONTEXT IN FUZZY MODELLING

CARLA CRISTINA P. CRUZ [a]
REGINA S. LANZILLOTTI [b]

Resumo

A mídia, geradora de informações, impõe a necessidade da aplicação de técnicas para análise textual que capta opiniões sobre temáticas diversas. A Mineração Textual refere-se a dados não-estruturados tratados pelo *KDT* que consiste em um conjunto de procedimentos para crítica e permissão da continuidade do processo analítico. Objetiva-se inferir palavras-chave no contexto da conhecimento da COVID-19. A opção metodológica ao *Fuzzy C-Means* agrupou os termos em função da similaridade semântica e inferiu termos considerados "objeto típico", palavras-chave. A aplicação valeu-se de dois textos que permitiram criar histogramas dos termos e determinar o limiar para corte dos considerados irrelevantes. O gráfico de dispersão foi construído pelas frequências relativas dos termos remanescentes e a alocação no plano cartesiano indicou a quantidade de grupos vizinhos. Ao aplicar o método *fuzzy* a palavra-chave de destaque foi "infecção", que espelha a letalidade do COVID-19, uma vez que liderou um dos grupos.

Palavras-chave: Mineração de Texto, *Fuzzy C-Means*, COVID-19.

[a]IME/CComp/UERJ, Rio de Janeiro, RJ, Brasil; ORCID: https://orcid.org/0000-0003-3656-4492 **E-mail: carlapassos2889@gmail.com**

[b]Instituto de Matemática e Estatística - IME, Universidade do Estado do Rio de Janeiro - UERJ, Rio de Janeiro, RJ, Brasil; ORCID: https://orcid.org/0000-0001-7789-6843 **E-mail: reginalanzillotti@gmail.com**

Abstract

The media, which generates information, imposes the need to apply techniques for textual analysis that captures information on different topics. Text Mining refers to unstructured data treated by KDT that consists of a set of procedures for criticizing and allowing the continuity of the analytical process. The objective is to infer keywords in the context at COVID-19. The methodological option to Fuzzy C-Means grouped the terms according to the semantic similarity and inferred terms considered "typical object", keywords. The application made use of two texts that allowed to create histograms of the terms and to determine the threshold for cutting those considered irrelevant. The dispersion graph was constructed by the relative frequencies of the remaining terms and the allocation in the Cartesian plane indicated the number of neighboring groups. When applying the fuzzy method the spotlight keyword was "infection", which reflects the lethality of COVID-19, since it led one of the groups.

Keywords: Text Mining, Fuzzy C-Means, COVID-19.

MSC2010: 03B52, 03E72, 94D05, 62A86

1 Introdução

A associação homem/máquina envolve interações tais como postagens nas redes sociais, pagamento de contas via celular, compras em sites, cadastramento de clientes, dentre outros. Nos últimos anos, 90% dos dados criados foram decorrentes da adesão das grandes empresas (redes sociais ou adesão de empresas as mesmas) e aplicativos de dispositivos móveis [21].

Deste modo, é desafiador coletar, limpar, organizar, correlacionar, vincular e transformar esses dados em informações relevantes. Das muitas informações criadas diariamente, menos de 10% consegue ser minerada e organizada, fazendo com que enormes quantidades de informações se tornem lixo eletrônico digital [47].

Na linha de tempo em relação ao estado da arte no que tange a mineração de texto, surge a necessidade da aplicação de técnicas que possibilitem a extração e análise de dados para propiciar opiniões que abranjam temáticas textuais e/ou análise de sentimentos. Este procedimento contempla desde o uso de histogramas, dendrogramas, técnicas estatísticas multivariadas até alcançar os métodos *fuzzy*. Os dados gerados podem se apresentar como estruturados, semi-estruturados e não estruturados, sendo que 80% das informações vigentes são não-estruturadas e 80% destas se encontram no formato textual [46] [7].

Dentre as principais áreas de conhecimento que compõem e contribuem com a Mineração de Textos (MT) se encontram o Aprendizado de Máquina (AM) [29], Processamento de Linguagem Natural (PLN), Estatística Inferencial [31], Inteligência Computacional (IC), Recuperação da Informação (RI), Mineração de Dados e *Web Mining* [5]. Cada uma dessas áreas do conhecimento, ou sua intercessão, é usada de forma a viabilizar o esquema de processamento computacional [25].

A informação no formato textual é representada por termos linguísticos que trazem consigo a incerteza, envolvidos na resolução de um problema, que podem ser decorrentes de alguma informação deficiente ou que possui mais de uma solução [19]. Dessa forma, a modelagem *fuzzy* é uma abordagem amplamente recomendada para aplicações cujo domínio esteja caracterizado por incerteza ou imprecisão da informação [8], pois se aproxima da forma com que o raciocínio humano, diante da incerteza, relaciona as informações, buscando respostas aproximadas aos problemas. Além disso, é uma ferramenta capaz de representar os termos utilizados na linguagem natural [41].

Dentre as abordagens utilizadas, destaca-se o agrupamento ou *cluster*, método utilizado para identificar relacionamentos entre objetos, facilitando a identificação de classes. No caso de textos, o agrupamento identifica conteúdos similares, caso não se tenha definição dos assuntos tratados em cada texto e se deseja separá-los por assunto [48], sendo classificados em *hard*, quando os dados pertencem a apenas um agrupamento, e em *soft*, cujos dados podem se enquadrar em mais de um agrupamento. Assim, o agrupamento *soft* leva ao agrupamento *fuzzy*, pois a cada elemento é associado um grau de pertinência que expressa o grau de pertencimento no agrupamento. A precisão é obtida na etapa de Recuperação da Informação utilizando como medida o *fuzzy clustering* [35], processo de particionamento não-hierárquico no qual se busca alocar um conjunto de elementos em grupos homogêneos através do algoritmo proposto.

Dentre os algoritmos *fuzzy* utilizados há o *fuzzy C-Means*, que serve para agrupar os documentos e termos em categorias [18] para propiciar o reconhecimento de padrões com o propósito de procurar, detectar e explicitar estruturas associadas às regularidades ou propriedades presentes em um conjunto de dados [27] que descreve cenários nas áreas do conhecimento tecnológico, biomédico e sociais. Neste trabalho, a motivação está associada a cenários biomédicos e a mineração de textos que podem vir a contribuir na busca de informação no contexto da pandemia de COVID-19.

A COVID-19 é uma doença causada pelo coronavírus, denominado SARS-CoV-2, que apresenta um espectro clínico variando de infecções assintomáticas a quadros graves. De acordo com a Organização Mundial de Saúde (OMS), a maioria (cerca de 80%) dos pacientes com COVID-19 podem ser assintomáticos, e aproximadamente 20% dos casos detectados requer atendimento hospitalar por apresentarem dificuldade respiratória [32].

O objetivo geral deste artigo é, através da Mineração de Textos, contribuir na criação de agrupamentos de termos julgados relevantes para a obtenção de palavras-chave com o intuito de agilizar buscas em textos sob o tema da COVID-19. De forma específica, consiste na adoção dos algoritmos de agrupamento *fuzzy* com o uso do Princípio de Extensão *fuzzy C-Means* pela Distância Euclidiana.

2 Mineração de Texto

A Mineração Textual é um processo de Descoberta de Conhecimento que utiliza técnicas de análise e extração de dados a partir de diferentes tipos de textos, além de envolver a aplicação de algoritmos computacionais que processam os textos, identificando informações úteis e implícitas que não poderiam ser recuperadas por métodos tradicionais de consulta, uma vez que a informação obtida se encontra em formato não-estruturado [33]. As etapas que compõem o processo baseia-se no *KDT*, ilustradas na Figura 1 [46] [10] [33] e são:

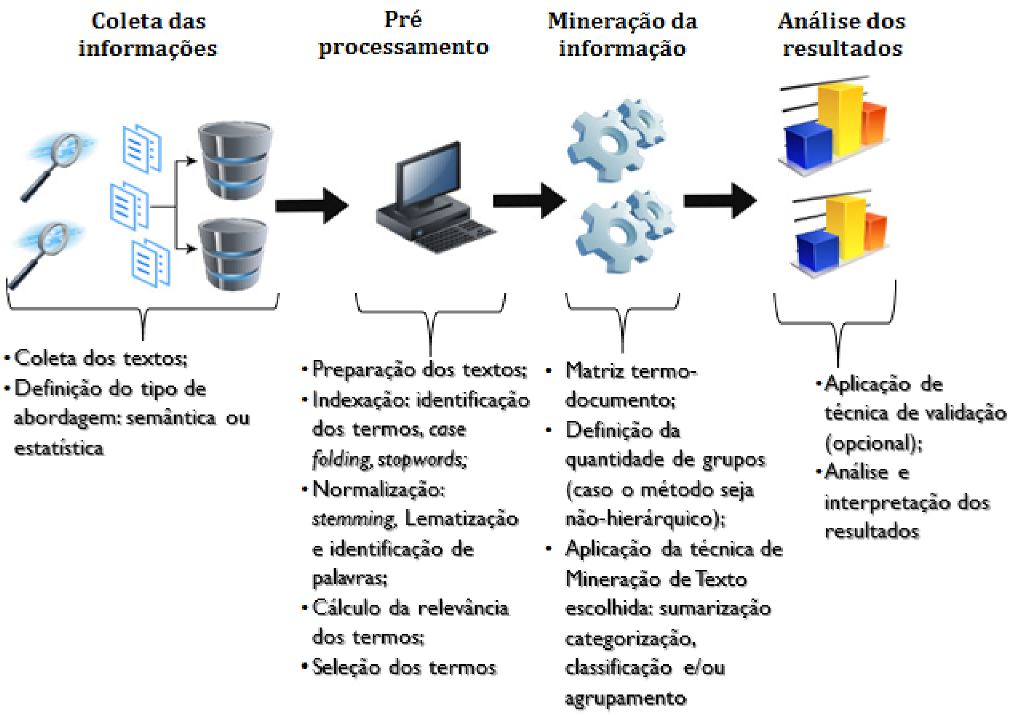
a **Coleta das informações.** Etapa de busca, coleta e armazenamento dos dados que serão analisados. É conhecida na literatura como corpus ou corpora. Também é nesta etapa em que se define qual o tipo de abordagem a ser aplicada nos textos: semântica ou estatística [26], para que se possa realizar a etapa seguinte;

b **Pré-processamento.** É o conjunto de ações tomadas sobre os documentos textuais a fim de torná-los manipuláveis para a extração do conhecimento, sendo considerada a parte mais importante do processo. Consiste na preparação dos textos, a definição do tipo de abordagem dos dados, sua preparação, indexação, normalização, cálculo/relevância (dentre as usadas, está a frequência relativa, utilizada neste trabalho) e seleção dos termos;

c **Mineração da informação.** Nesta etapa, uma vez que os textos selecionados foram transformados para dados estruturados, entra-se em um estágio em que os dados ficam compatíveis para o uso de técnicas de Mineração de Textos [45]. Assim, buscam-se partes relevantes de um texto em um documento e extrai-se informações específicas [25]. Em outras palavras, é a etapa na qual se aplicam os métodos de mineração de texto. Neste artigo será o algoritmo de agrupamento *fuzzy C-Means*. Antes de aplicar o algoritmo, faz-se necessário padronizar os valores, em unidades do desvio-padrão;

d **Análise dos resultados**. Trata-se da etapa final tendo como objetivo a interpretação e análise dos achados (avaliação das descobertas) obtidos na fase anterior, etapa na qual podem ser utilizadas técnicas de validação dos resultados para verificar as taxas de erro e de acurácia do método (opcional).

Figura 1: Etapas do *KDT*



Fonte: Autoral, 2019.

Dentre as principais técnicas [5] [48] encontra-se o agrupamento [45] que será utilizado. Também conhecido como *Clustering*, é uma técnica estatística multivariada usada para encontrar grupos de textos similares nos documentos de acordo com suas características ortográficas. É uma importante técnica exploratória, uma vez que, ao estudar a estrutura natural de grupos, possibilita avaliar a dimensionalidade dos dados, identificar *outliers* e levantar hipóteses relacionadas à estrutura (associações) dos objetos [23].

Em Mineração de Textos objetiva-se agrupar textos em classes de acordo com as características de cada documento, sem que haja a necessidade de alguma definição pelo usuário [43], pois o *cluster* (agrupamento) identifica co-relacionamentos e associações entre os elementos, o que vem facilitar a identificação das classes ao criar o conjunto de textos por assunto [48]. No entanto, um dos maiores problemas é a identificação dos grupos de documentos mais coesos de forma a mantê-los durante a utilização do sistema, pois todo documento inserido ou modificado deve ser reanalisado a fim de ser colocado no grupo correto [48]. Para a aplicação da técnica, sugere-se os passos a seguir [40]:

1. Análise das variáveis e dos objetos a serem agrupados como seleção de variáveis, identificação de *outliers* e padronização;

2. Seleção da medida de similaridade entre cada par de observações;

3. Seleção do algoritmo de agrupamento: método hierárquico ou não-hierárquico;

4. Definição da quantidade dos agrupamentos formados;

5. Interpretação e validação dos agrupamentos (opcional).

Uma medida de distância no conjunto $\Re^t$ é uma função $d : \Re^t \times \Re^t \rightarrow \Re^+$ que associa a cada par ordenado de elementos $\mathbf{v}$, $\mathbf{y} \in \Re^t$ um número real $d(v, y)$, chamado distância entre $\mathbf{v}$ e $\mathbf{y}$. Para uma análise dos grupos a serem formados, faz-se necessária à aplicação de medidas de distância para avaliar o quão próximos (parecidos) ou distantes (diferentes) estão os elementos amostrais.

Existem diversas medidas de distância presentes na literatura, sendo que neste trabalho foi usada a distância Euclidiana, também conhecida como Distância Métrica, que representa a distância linear entre dois pontos, em um espaço T-dimensional. Sejam $\mathbf{v}$ e $\mathbf{y}$ dois vetores (elementos) amostrais, a distância euclidiana entre eles é definida como [23]:

$$d(v, y) = \sqrt{(v - y)^{'} (v - y)} \tag{1}$$

Para a construção de agrupamentos é necessária à utilização de métodos que podem ser divididos em Hierárquicos e Não-Hierárquicos, sendo este último usado neste artigo. Os métodos Não-Hierárquicos precisam satisfazer dois critérios: semelhança interna e separação dos agrupamentos [31]. Como resultado, procedimentos não-hierárquicos tendem a possuir maior eficiência computacional, fazendo com que algoritmos desta natureza tenham alta aplicabilidade quando se analisa grandes conjuntos de dados [23].

3 Lógica *Fuzzy* e Princípio de Extensão *Fuzzy*

A Lógica *Fuzzy* (LF) caracteriza-se por ser contextual e descritiva. Desta forma um texto pode ser decodificado em funções de pertinência e sua operacionalidade seria realizada através de operadores lógicos. Os estudos sobre lógica *fuzzy* começaram em 1965 com Zadeh [50], cujas primeiras observações surgiram através dos recursos tecnológicos disponíveis à época, que eram incapazes de automatizar as atividades relacionadas a problemas que compreendessem situações ambíguas.

Essas observações foram baseadas na lógica multivalorada do polonês Jan Lukasiewicz [44] que muito contribuiu para a lógica *fuzzy* adotando funções de pertinência de conjuntos *fuzzy* [30], na qual uma variável assume graus de pertinência no intervalo escalar $[0, 1]$, onde o zero e o um indicam a exclusão e pertinência completa, respectivamente.

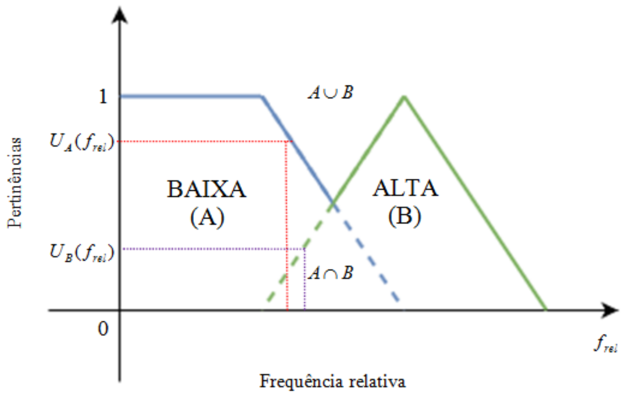
Os Conjuntos *Fuzzy* representam termos linguísticos através do uso de formas geométricas e são uma generalização da Teoria dos Conjuntos, pois se afere a possibilidade de um determinado elemento poder pertencer a dois conjuntos simultaneamente (área de interseção) com um respectivo grau de pertinência, o que difere da lógica *booleana*, que só admite valores zero e um, traduzidos por pertence ou não pertence, verdadeiro ou falso [39]. Logo não há conjunto vazio, mas conjuntos cujos elementos possuem relevância de pertinência [29].

Ao transportar a utilização para a mineração textual, a ideia proposta pela teoria *fuzzy* parte do princípio que, caso não haja a possibilidade de determinar os limites exatos da contingência de um texto por estarem definidos ambiguamente, faz-se necessário buscar uma escala que permita tomar a decisão sobre a relevância inerente ao termo no texto, o que vem a permitir estabelecer valores de pertinência e respectivas designações categóricas que estão afeitas a possibilidade [2].

A Figura 2 permite visualizar dois conjuntos *fuzzy*, o trapezoidal A (azul) e o triangular B (verde), que podem ser categorizados pelos termos linguísticos BAIXA e ALTA, sem negligenciar a sua interseção ($A \cap B$). A união ($A \cup B$) representa o envoltório dos conjuntos *fuzzy* trapezoidal e triangular. Na área de interseção [27], deve-se adotar a menor pertinência. Caso não esteja em áreas superpostas, o valor da pertinência encontra-se no eixo da ordenada.

Em alguns casos, quando se considera uma análise multivariada em Lógica *Fuzzy*, pode haver uma "explosão de regras". Dessa forma, recorre-se ao Princípio de Extensão e não aos Conjuntos *Fuzzy* para determinar as pertinências, principalmente nos casos em que o problema impõe muitos Conjuntos *Fuzzy*, pois os mesmos implicam em uma multiplicidade de regras *fuzzy* que são proporcionais à combinação

Figura 2: Função de pertinência dos Conjuntos *Fuzzy* A e B

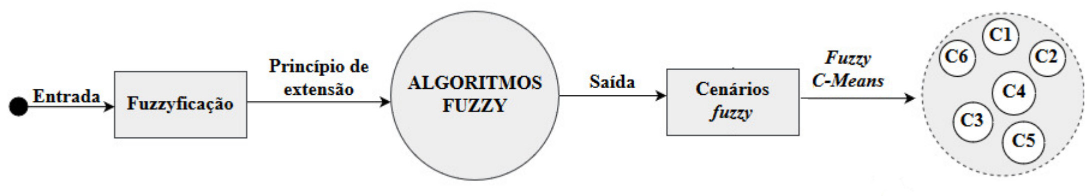


Fonte: Adaptado de [9].

da quantidade dos conjuntos, o que pode tornar inviável o tratamento operacional das regras no reconhecimento de padrões ou em sistemas de controle [28]. A alternativa consubstancia-se no Princípio de Extensão *Fuzzy*, ferramenta que sustenta a extensão das expressões matemáticas do domínio clássico ao domínio *fuzzy* [50], permitindo calcular a imagem de um objeto inferindo o grau de pertinência da Teoria *Fuzzy* [51].

A Figura 3 ilustra um problema que usa o princípio de extensão *fuzzy*. A entrada é composta pelos dados já tratados pelo *KDT* que resultam nas frequências relativas padronizadas em unidades do desvio padrão. Em seguida, passam por um processo de fuzzificação pelo Princípio de Extensão *Fuzzy*, que será representado pelo algoritmo *fuzzy C-Means*. A saída do Sistema *Fuzzy* são agrupamentos com termos linguísticos que inferem cenários que traduzem as palavras-chave da COVID-19.

Figura 3: Arquitetura do princípio de extensão *fuzzy*



Fonte: Autoral, 2020.

O *fuzzy C-Means* (FCM) [11] [4], indica que cada agrupamento tem um agrupamento central representativo de um objeto típico e o valor de associação expresso segundo pertinência do objeto em relação à proximidade ao centro do agrupamento. Desta forma, se o valor da associação for alto, o elemento será semelhante ao elemento central do agrupamento e, caso seja baixo, o elemento tem pouca semelhança com o agrupamento [20].

O FCM requer uma pré-especificação do número de grupos [31] e, em seu processo de partição em um conjunto de dados, $X = \{x_i | i = 1, 2, ..., k\}$, onde x_k é um vetor de características $x_k = \{x_{k1}, x_{k2}, ..., x_{kp}\} \in \Re^p$ para todo $k \in \{1, 2, ..., n\}$ e $\Re^p$ o espaço P-dimensional. O problema da aglomeração *fuzzy* é encontrar uma pseudopartição que represente a estrutura dos dados da melhor forma possível. Os termos dos textos permitem obter a distribuição de frequência em valores relativos, sendo que estes foram padronizados em unidades do desvio padrão, observando a diferença entre cada frequência relativa em relação a frequência relativa média dividida pelo desvio padrão. Os valores assim padronizados foram os valores de entrada no sistema *fuzzy c-means*, sendo que o algoritmo gerou uma família de k subconjuntos onde o vetor X é denotado pelo vetor $U = \{u_1, u_2, ..., u_k\}$, onde cada elemento $u_i = 1, ..., k$ é chamado de pertinências *fuzzy* e foram obtidos pelo critério de normatização (padronização) que satisfaz a condição:

$$\sum_{k=1}^{n_k} \sum_{j=1}^{p} u_{ij} = 1 \qquad (2)$$

Em outras palavras, busca-se minimizar a função objetivo dada por:

$$J(U, V) = \sum_{i=1}^{n} \sum_{j=1}^{g} u_{ij}^m d(x_i, c_j) \qquad (3)$$

em que:

- n, quantidade de elementos (termos);
- g, número de grupos pré-determinado;
- m, parâmetro de fuzzificação, provavelmente no intervalo $[1.25; 2.50]$, sendo o ponto médio $m = 2$, a escolha mais usual na literatura [34];
- c_j, centro do agrupamento *fuzzy* para cada agrupamento obtida segundo uma média ponderada em função dos valores padronizados correspondentes aos termos, cujas pertinências correspondem aos ponderadores:

$$c_j = \frac{\sum_{j=1}^{p} u_{kj}^m x_{kj}}{\sum_{j=1}^{p} u_{kj}^m} \qquad (4)$$

- u_{kj}, pertinências *fuzzy* geradas, inicialmente, de forma aleatória e posteriormente atualizada conforme fórmula:

$$u_{kj} = \frac{\sum_{k=1}^{n} \left(\frac{1}{d(z_{kj}, c_{ok})}\right)^{\left(\frac{2}{m-1}\right)}}{\sum_{k=1}^{n} u_{kj}^{m}} \tag{5}$$

- $d\left(x_i, c_j\right)$ distância euclidiana entre o elemento da amostra (termo) x_i e o centro do agrupamento c_j.

O Quadro 1 apresenta o passo a passo do algoritmo *fuzzy C-Means* [4] [49]:

Quadro 1: Algoritmo do Método *fuzzy C-Means*

Entrada	Z = matriz com os escores ranqueados padronizados em unidades do desvio-padrão correspondentes aos termos; k = número de agrupamentos; $n = \sum_{k=1}^{n_k} \sum_{j=1}^{p} n_{kj}$ total de termos da matriz Z.
Passo 1	Pertinências iniciais u_{0kj} geradas aleatoriamente para cada agrupamento, adotando o critério de normalização: $\sum_{k=1}^{n_k} \sum_{j=1}^{p} u_{kj} = 1$;
Passo 2	Centro inicial c_{0_k} para cada agrupamento gerado pela média ponderada dos valores padronizados tendo as pertinências como pesos: $c_{0k} = \frac{\sum_{j=1}^{p} u_{kj}^{m} z_{kj}}{\sum_{j=1}^{p} u_{kj}^{m}}$;
Passo 3	Distância Euclidiana entre os elementos e o centroide: $d_{0k}\left(z_{kj}, c_{0k}\right) = \|\|z_{kj} - c_{0k}\|\| = \sqrt{\sum_{j=1\ p} \left(z_{kj} - c_{0k}\right)^2}$;
Passo 4	Função objetivo inicial: $J_0(Z, C) = \sum_{j=1}^{p} \sum_{k=1}^{n_k} u_{ij}^{m} d(z_{kj}, c_{0k})$;
Passo 5	Atualização das pertinências u_{kj}, para cada $z_{kj} \in Z_k$: $u_{kj} = \frac{\sum_{k=1}^{n} \left(\frac{1}{d(z_{kj}, c_{ok})}\right)^{\left(\frac{2}{m-1}\right)}}{\sum_{k=1}^{n} u_{kj}^{m}}$;
Passo 6	Centroides atualizados pelo **Passo 2**: $c_j = (c_1, c_2, ..., c_p) \in \Re^p$, $v \in \Re^p$ para todo j;
Passo 7	Volte para os **passos 3** e **4**. **SE** a função objetivo estiver minimizada: **FIM**. **SENÃO**: volte para o **Passo 5**;
Saída	Agrupamentos segundo temas com respectivos termos inerentes a COVID-19.

Fonte: Autoral, 2019.

4 Materiais e Métodos

Nesta Seção serão descritos os conceitos utilizados neste artigo, necessários para a implementação do método *fuzzy C-Means*, sob a ótica do princípio de extensão *fuzzy* para a Mineração de Texto. Em todas as etapas da Mineração de Textos se utilizou o *software RStudio* [42].

4.1 Materiais

O banco de dados é composto por dois artigos que tratam do assunto da COVID-19, cujos termos são as palavras identificadas em cada documento. Esses dois itens podem ser considerados como covariáveis que permitem a construção de uma tabela de contingência, em que as células correspondem as frequências observadas dos termos em cada documento, tal qual unidades amostrais.

O método *KDT* inclui o pré-processamento de dados não estruturados para textos que consiste em transformar as palavras em atributos comportamentais. Objetiva-se fazer com que o texto possa ser interpretado como um cenário, cabendo verificar a coerência imposta pelo agrupamento. O *software* escolhido foi a IDE *RStudio* [42], livre e integrado para linguagem de programação R [36], utilizada para gráficos e cálculos estatísticos.

4.2 Métodos

Os textos foram lidos pelo *software RStudio* [42], através da função *read_pdf().* Posteriormente, foi feito o pré-processamento, mas para esta etapa, foram utilizados os pacotes *NLP* [22], responsável pelo Processamento da Linguagem Natural e *tm* [14] que fornece as funções para a conversão de letras maiúsculas em minúsculas, remoção de números, espaços extras e links.

Na remoção das *stopwords*, além dos pacotes *tm* [14] e *stopwords* [3], foi necessária complementação, pois não há uma lista completa, em virtude de um mesmo idioma ter variação. Para o artigo se utilizou a lista do site Ranks NL Webmaster Tools [38], citados pela maioria das referências deste artigo. Para a remoção de figuras, emojis e caracteres especiais criou-se uma função que fizesse a remoção. Quanto aos acentos, utilizou-se o pacote *stringi* [17], que faz a normalização para o formato unicode, que fornece um "número"único para cada caractere.

A seguir, os documentos e suas respectivas palavras passaram pelo processo de Normalização para o qual foi utilizado o pacote *rslp* [13], que faz o processo de *stemming* para a língua portuguesa. Quanto às palavras sinônimas, a junção é feita de forma manual, isto é, não se tem um pacote pronto no programa, tendo que ser construída uma função com esse propósito.

Depois, guardou-se os termos na matriz termo-documento através da função *DocumentTermMatrix()*, disponível no pacote *tm* [14]. Essa matriz já disponibiliza as frequências absolutas que permitiram calcular as frequências relativas simples e acumuladas que subsidiaram estabelecer um limiar de corte, adotando-se o critério de inclusão de termos para este artigo, aqueles que tivessem frequência relativa acima deste limiar, gerando uma nova matriz termo-documento, armazenando-a em uma tabela externa através da função *write.csv2()*.

Para a determinação da quantidade de grupos, calculou-se a frequência relativa e utilizou-s o gráfico de dispersão, com os valores observados na escala de razão. Quanto ao método *fuzzy*, foram utilizados os pacotes *fclust* [16] (para os cálculos do método), *ppclust* [6] e *factoextra* [24] para a visualização gráfica. Por fim, no pós-processamento, após a aplicação do algoritmo *fuzzy C-Means*, para visualização dos termos, utiliza-se a Nuvem de Palavras, geradas com o auxílio do pacote *wordcloud* [15].

5 Resultados

A proposta deste trabalho refere-se à Mineração Textual no sentido de agregar termos julgados relevantes para a obtenção de palavras-chave que possam agilizar a busca na área da COVID-19.

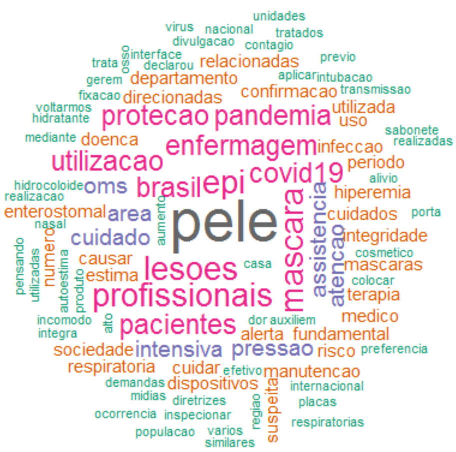
5.1 Análises Descritivas e Pré-processamento

Inicialmente, foram selecionados dois artigos que tratam do assunto da COVID-19, cujos títulos são "Lesão por pressão relacionada a dispositivo médico nos profissionais de saúde em época de pandemia" [37] e "Recomendação PREPI—COVID19" [1].

A etapa pré-processamento foi aplicada a cada documento, seguida da Normalização para reduzir a quantidade de termos com o mesmo radical e unir termos sinônimos. Após este procedimento, foi gerada a matriz inicial termo-documento que serviu para selecionar os prováveis termos que estariam presente na tabela de contingência, onde a linha e a coluna representam os termos a ser considerados finais e o respectivo documento.

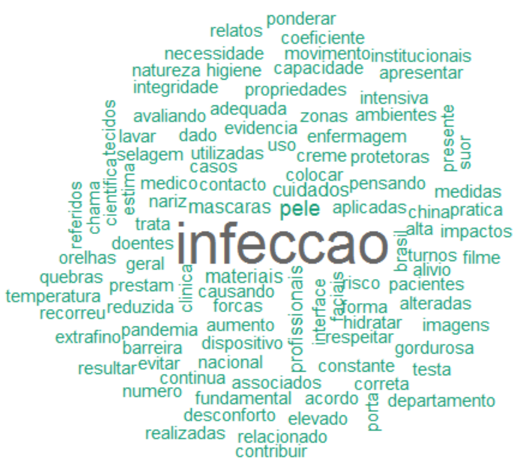
As quantidades de termos obtidos foram de 221 e 489 para o Documento 1 e Documento 2, respectivamente, totalizando 710 termos. A representação pictográfica

Figura 4: Nuvens de Palavras



(a) Documento 1

(b) Documento 2



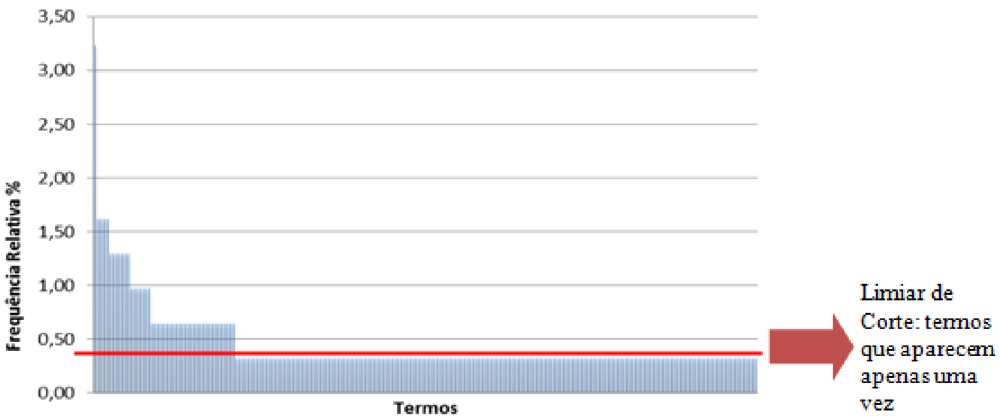
(c) Nuvem Geral

Fonte: Autoral, 2020.

em nuvem de palavras *wordcloud*, para cada um dos documentos e dos dois agregados, Figura 4, foi elaborada em função da frequência de ocorrência dos termos, sendo que os mais frequentes são representados em dimensão de destaque, mas por questões de visualização, as nuvens exibiram os 100 termos mais frequentes.

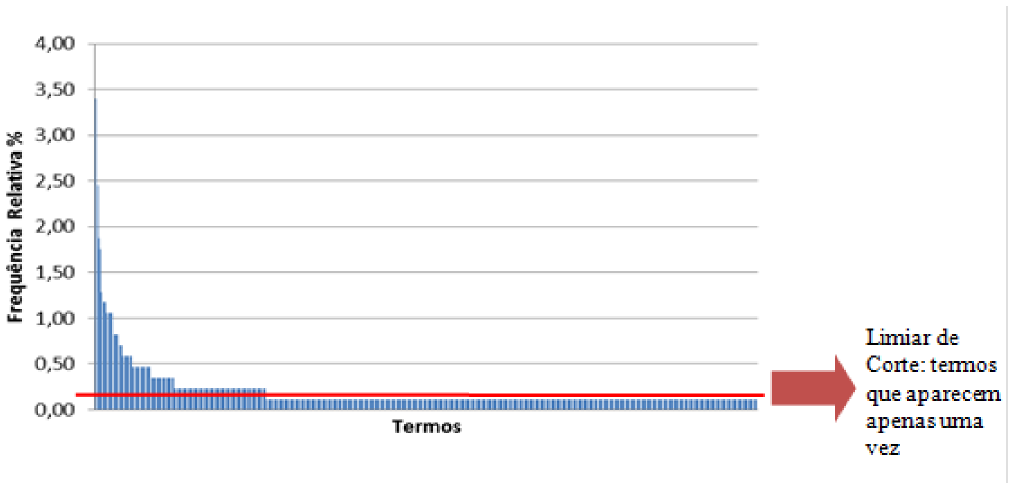
As Figuras 5 e 6 são os histogramas das frequências relativas para os Documentos 1 e 2, tendo sido adotado como limiares de corte percentual as frequências relativas 0,32% e 0,12%, respectivamente, sendo que em termos absolutos indicam as ocorrências de 175 e 361 termos que apareceram apenas uma única vez nestes documentos, totalizando 536 termos. Nesta proposta, optou-se por 143 termos na aplicação do *fuzzy C-Means*.

Figura 5: Histograma do Documento 1



Fonte: Autoral, 2020.

Figura 6: Histograma do Documento 2

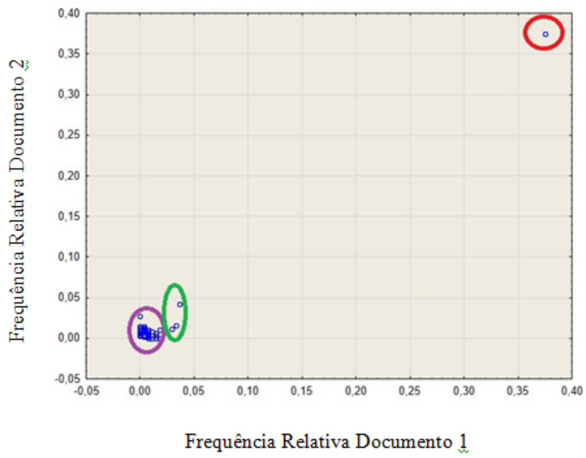


Fonte: Autoral, 2020.

5.2 Processamento e Pós-Processamento

Após a seleção dos termos, os valores das suas frequências passaram pelo processo de normalização para expressá-las em unidades do desvio padrão para aplicação do método que exige conhecer a quantidade de grupos e para tal, foi utilizado o gráfico de dispersão com base nas frequências relativas dos dois textos que orienta para esta seleção em função do agregamento, Figura 7. Observa-se que um dos termos distanciou-se dos demais pelas altas frequências e os restantes delinearam mais dois grupos com frequências até 0,05 para os dois eixos.

Figura 7: Quantidade de agrupamentos *fuzzy C-Means*



Fonte: Autoral, 2020.

O algoritmo *fuzzy C-Means* usa a distância euclidiana como medida de similaridade, e para a sua implementação adotou-se o parâmetro de fuzzificação igual a dois, sugestão proposta na Seção 3. Na mineração de texto optou-se por verificar se os três agrupamentos indicados via gráfico de dispersão viabilizaria a minimização da função objetivo do *fuzzy C-Means*. Primeiramente, os valores das pertinências iniciais u_{kj}, pesos aleatórios para obter os centroides, foram gerados segundo valores aleatórios para cada agrupamento, havendo imposição da soma ser um.

Em seguida, o centroide inicial c_{0k}, para cada agrupamento, foi gerado por uma média ponderada em função dos valores padronizados dos termos, tendo como ponderadores as pertinências iniciais. A distância euclidiana entre os elementos e o centroide do agrupamento $d(z_{kj}, c_{ok})$ foi calculada para ser inserida na função objetivo inicial $J_o(Z, C)$, que assumiu valor de $12,8676$ em 65 iterações, permitindo desta forma, obter os valores das pertinências atualizadas dos termos cada agrupamento. Quanto aos centroides, Tabela 1, o de maior expressividade corresponde no Documento 1 no Agrupamento 2 tendo valor de $11,5928$.

Tabela 1: Valores dos Centroides no *fuzzy C-Means*

Agrupamento	**Documento 1**	**Documento 2**
1	-0,1477	-0,0912
2	11,5928	11,5764
3	0,0301	-0,0764

Fonte: Autoral, 2020.

Em seguida, foram atualizadas as pertinências e na Tabela 2 estão os agrupamentos do *fuzzy C-Means* com as respectivas pertinências máximas diferenciadas para os Agrupamentos 1, 2 e 3, com os valores de 0,7676, 0,9829 e 0,8943, respectivamente. Nota-se que a maioria dos termos ficaram alocados no Agrupamento 1 e que o Agrupamento 2 apresentou apenas o termo infecção, que ficou bem distante dos demais.

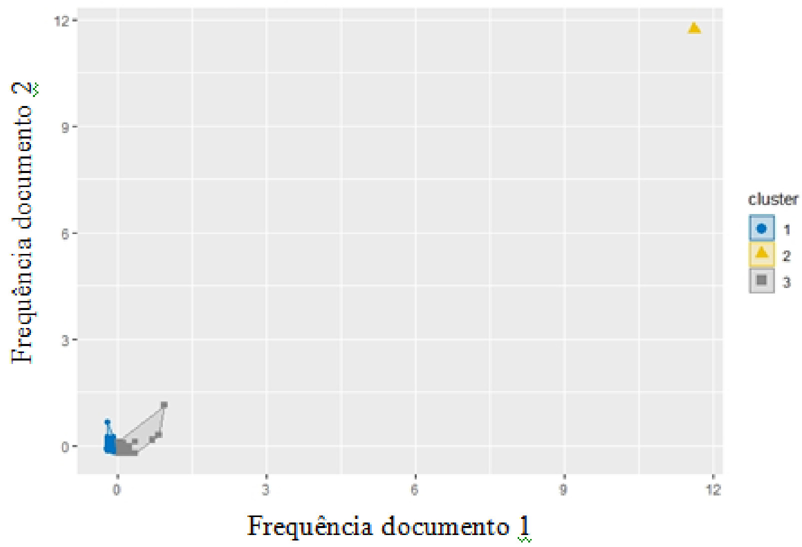
Tabela 2: Agrupamentos, pertinência média e respectivos Termos Agrupados

Agrupamento	**Pertinência**	**Termos Agrupados**
1	0,7676	acordo, alteradas, avaliando, barreira, científica, correta, elevado, gordurosa, lavar, movimento, orelhas, prática, quebras, tecidos, testa
2	0,9829	infecção
3	0,8943	imagens, nacional

Fonte: Autoral, 2020.

A visualização do resultado do método *Fuzzy C-Means*, Figura 8, indica a superposição dos Agrupamentos 2 e 3, significando que os termos envolvidos podem pertencer a qualquer um dos grupos, o que é esperado na metodologia *fuzzy*.

Figura 8: Visualização dos agrupamentos *Fuzzy* segundo *Fuzzy C-Means*



Fonte: Autoral, 2020.

A alocação dos termos que se relacionariam com os verdadeiramente inerentes de cada agrupamento está na Tabela 3 e estão grifados em negrito. Os 15 termos polarizadores do Agrupamento 1 correspondem a 14,56% dos demais relacionados, enquanto no Agrupamento 3 este índice é 5,13% e no Agrupamento 2 o termo "infecção"é absoluto.

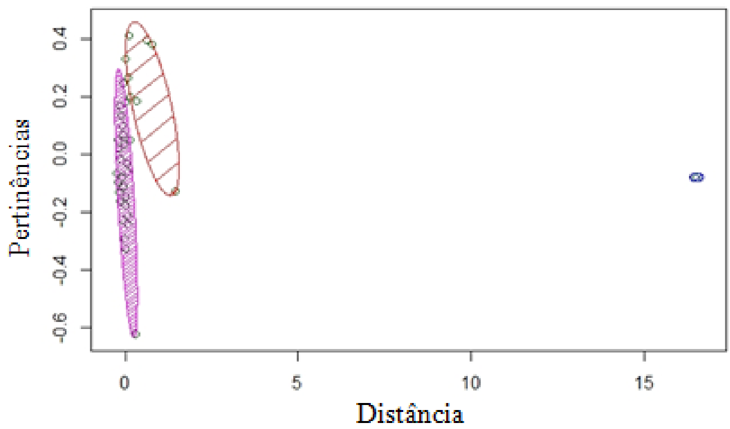
Tabela 3: Agrupamentos, pertinência média e respectivos Termos Agrupados

Agrupamento	Quantidade	Termos Agrupados
1	103	**acordo**, adequada, ajuste, alívio, **alteradas**, ambientes, apresentar, associados, aumento, **avaliando**, **barreira**, capacidade, chinesas, científica, clinica, coeficiente, colocar, concreto, confirme, constante, contacto, contínuo, contribuir, **correta**, corte, creme, dado, desconforto, desenvolvimento, dias, doentes, dor, **elevado**, encontrar, evidência, evitar, existe, extrafino, faciais, fatores, filme, forcas, forma, garantindo, geral, **gordurosa**, hidrocoloide, impactos, implementadas, individual, influenciado, institucionais, interface, **lavar**, limpeza, materiais, medidas, morte, **movimento**, nariz, natureza, necessidade, oclusivos, orelhas, particular, pensando, permita, ponderar, porta, **prática**, preferencia, presente, prestam, propriedades, protetoras, publicadas, qualidade, **quebras**, realizadas, recomenda, recorreu, recursos, reduzida, referidos, relatos, respeitar, resposta, resultar, secar, selagem, silicone, suor, sustentada, tabela, **tecidos**, temperatura, **testa**, troca, tubos, turnos, verificar, vida, zona
2	1	**infecção**
3	39	alerta, alta, aplicadas, brasil, casos, causando, chama, cuidados, departamento, direcionadas, dispositivo, enfermagem, equipamentos, estima, figura, fundamental, hidratar, higiene, hiperemia, **imagens**, integridade, intensiva, mascaras, medico, **nacional**, numero, pacientes, pandemia, pele, profissionais, reconhece, relacionado, risco, sociedade, suspeita, terapia, trata, uso, utilizadas

Fonte: Autoral, 2020.

A Figura 9 representa todos os termos que se agregaram aos polarizadores segundo o *fuzzy C-Means* e observa-se que o Agrupamento 2 – Infecção não se associa a outros termos, enquanto o Agrupamento 1 agrega muitos termos e destaca-se pela expressiva variabilidade das pertinências. O Agrupamento 3 – Imagens; Nacional apresenta a maioria dos seus termos tendo valores de pertinência entre 0,2 e 0,4, mas é um polarizador de menor poder agregativo.

Figura 9: Termos x polarizadores segundo *Fuzzy C-Means*



Fonte: Autoral, 2020.

6 Considerações e trabalhos futuros

O suporte computacional para a análise de grande volume de dados estruturados na busca de informações sobre um determinado tema torna-se viável pelo uso de algoritmos computacionais, particularmente em COVID-19, cujo acervo é vasto e disperso. Dessa forma, a Mineração de Texto, vertente incluída na Inteligência Artificial, auxilia o reconhecimento de padrões para estabelecer cenários prospectivos e associativos interativos que possam ser úteis para a produção de conhecimento.

O algoritmo proposto sob a ótica da Mineração de Texto *fuzzy* otimiza a descoberta de palavras-chave geradas pelo *fuzzy C-Means*, pois caracterizou agrupamentos cognitivos textuais, compostos por termo único ou por termos agregados que permitem perceber as indicações circunstanciais como instrumentos de interlocução para contextualizar o assunto e estimular a reflexão.

A palavra-chave "infecção" destacou-se dentre as demais, pois na COVID-19 as pessoas infectadas apresentam sintomas leves ou moderados, mas podem desenvolver quadros mais agressivos e levar a óbito, o que configura o cenário "Temerário", que só poderá ser atenuado pelo distanciamento social, higienização e o hábito do uso de máscaras ao sair de casa.

Os termos "imagens" e "nacional" são palavras-chave advindas da mineração de texto sob a ótica *fuzzy* que infere o cenário de um dos conhecimentos da Inteligência Artificial, Processamento de Imagem. Os pesquisadores do Centro Nacional de Pesquisa em Energia e Materiais (CNPEM/Campinas) realizaram os primeiros testes com o Sirius, acelerador nacional de partículas para uso de uma proteína imprescindível ao ciclo de vida do novo Coronavírus (Sars-Cov-2) [12].

O conjunto de termos "Acordo, alteradas, avaliando, barreira, científica, correta, elevado, gordurosa, lavar, movimento, orelhas, prática, quebras, tecidos, testa" não permitiu a discriminação de uma palavra-chave, mas indicou critérios inerentes a contaminação (gordurosa, elevado), a higienização corporal (lavar, orelhas), o distanciamento social (alteradas, barreira, movimento, quebras), o comprometimento institucional (acordo, correta, quebras) e o saber (científica, prática, tecidos, testa).

O modelo *fuzzy C-Means* tratado pelo Princípio de Extensão consubstanciado na Distância Euclidiana capacitou a indicação de palavras-chave da temática da COVID-19 para reconhecer dois cenários simbólicos: Infecção e Processamento de Imagens. Foi capaz de discernir sobre os contextos contaminação, higienização corporal, distanciamento social e comprometimento institucional.

Cabe ressaltar que há viabilidade na realização de confronto de algoritmos de clusterização *fuzzy*, como trabalhos futuros. Pode-se optar pelo algoritmo Gustafson-Kessel que adota a matriz de covariância no princípio de extensão, porém, se a matriz não admite inversa, cria-se o óbice. A alternativa mais recente indica o algoritmo *fuzzy C-Medoids* que perpassa esse entrave.

Referências

[1] ALVES, P.; *et al.* PREPI—COVID19. PRevenção de lesões cutâneas causadas pelos Equipamentos de Proteção Individual (Máscaras faciais, respiradores, viseiras e óculos de proteção). **Journal of Tissue Healing and Regeneration**. Suplemento da edição Outubro/Março XV, 2020. Disponível em: `https://www.researchgate.net/publication/340105316_RECOMENDACAO_PREPI_COVID19_PRevencao_de_lesoes_cutaneas_causadas_pelos_Equipamentos_de_Protecao_Individual_Mascaras_faciais_respiradores_viseiras_e_oculos_de_protecao`. Acesso em: 11 ago. 2020.

[2] BELLMAN, R. E.; ZADEH, L. A. Decision-Making in a Fuzzy Environment. **Management Science**, v. 17, n. 4, dez. 1970, p. 141-164. Disponível em: `http://www.dca.fee.unicamp.br/~gomide/courses/CT820/artigos/DecisionMakingFuzzyEnvironmentBellmanZadeh1970.pdf`. Acesso em: 12 ago. 2020.

[3] BENOIT, K.; MUHR, D.; WATANABE, K. *stopwords: Multilingual Stopword Lists*. R package version 1.0, 2019. Disponível em: `https://CRAN.R-project.org/package=stopwords`. Acesso em: 04 jul. 2020.

[4] BEZDEK, J. C. **Pattern Recognition with Fuzzy Objective Function Algorithms**. 1 ed. NewYork: Plenum Press, 1981.

[5] CARRILHO JUNIOR, J. R. **Desenvolvimento de uma Metodologia para Mineração de Textos**. 96f. Dissertação (Mestrado em Engenharia Elétrica) - Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2007. Disponível em: `https://www.maxwell.vrac.puc-rio.br/Busca_etds.php?strSecao=resultado&nrSeq=11675@1`. Acesso em: 07 set. 2020.

[6] CEBECI, Z. *et al.* ppclust: Probabilistic and Possibilistic Cluster Analysis. R package version 0.1.3, 2019. Disponível em: `https://CRAN.R-project.org/package=ppclust`. Acesso em: 09 mar. 2019.

[7] CHEN, H. **Knowledge management systems**: a text mining perspective. Arizona: Knowledge Computing Corporation, 2001.

[8] DAS, S. Pattern Recognition using the Fuzzy c-means Technique. **International Journal of Energy, Information and Communications**, v. 4, n. 1, p. 1-14, fev. 2013. Disponível em: `https://www.researchgate.net/publication/303150319_Pattern_Recognition_using_the_Fuzzy_c-means_Technique`. Acesso em: 11 set. 2020.

[9] DELGADO, M. R. D. B. da S. **Projeto Automático de Sistemas Nebulosos: Uma Abordagem Co-Evolutiva**. 204f. Tese (Doutorado em Engenharia Elétrica e de Computação) - Universidade Estadual de Campinas, Campinas, 2002. Disponível em: `http://repositorio.unicamp.br/jspui/handle/REPOSIP/260259`. Acesso em 24 ago. 2020.

[10] DIXON, M. **An Overview of Document Mining Technology**, 1997. Disponível em: `http://citeseerx.ist.psu.edu/viewdoc/download`. Acesso em: 09 jun. 2020.

[11] DUNN, J. C. A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters. **Journal of Cybernetics**, p. 32-57, 1973. Disponível em: `https://www.tandfonline.com/loi/ucbs19`. Acesso em: 28 ago. 2020.

[12] ESTADO DE MINAS. **Primeiras imagens da COVID-19 são reveladas por acelerador de partículas nacional**. Disponível: `https://www.em.com.br/app/noticia/nacional/2020/07/12/interna_nacional,1166179/primeiras-imagens-da-covid-19-sao-reveladas-por-acelerador-de-\.html`. Acesso em: 12 set. 2020.

[13] FALBEL, D. **rslp: A Stemming Algorithm for the Portuguese Language**. R package version 0.1.0, 2016. Disponível em: `https://CRAN.R-project.org/package=rslp`. Acesso em: 07 jun. 2019.

[14] FEINERER, I.; HORNIK, K.; MEYER, D. Text Mining Infrastructure in R. **Journal of Statistical Software**, v. 25, n. 5, p. 1-54, 2008. Disponível em: `http://www.jstatsoft.org/v25/i05/`. Acesso em: 03 ago. 2020.

[15] FELLOWS, I. **wordcloud: WordClouds**. R package version 2.6, 2018. Disponível em: `https://CRAN.R-project.org/package=wordcloud`. Acesso em: 20 mai. 2020.

[16] FERRARO, M. B.; GIORDANI, P. A toolbox for fuzzy clustering using the R programming language. **Fuzzy Sets and Systems**, v. 279, p. 1-16, nov. 2015. Disponível em: `http://dx.doi.org/10.1016/j.fss.2015.05.001`. Acesso em: 02 mai. 2020.

[17] GAGOLEWSKI, M. *et al.* **R package stringi: Character string processing facilities**, 2019. Disponível em: `http://www.gagolewski.com/software/stringi/`. Acesso em: 04 ago. 2020.

[18] GOSWAMI, S.; SHISHODIA, M. S. A fuzzy based approach to text mining and document clustering. **ArXiv Journal**, jun. 2013. Disponível em: `https://arxiv.org/abs/1306.4633`. Acesso em: 14 ago. 2020.

[19] GOULARTE, F. B. **Método *fuzzy* para a sumarização automática de texto com base em um modelo extrativo (FSumm)**. 119f. Dissertação (Mestrado em Ciência da Computação) - Universidade Federal de Santa Catarina, Florianópolis, 2015. Disponível em: `https://repositorio.ufsc.br/handle/123456789/132756`. Acesso em: 12 set. 2020.

[20] GREKOUSIS G.; FOTIS, Y. N. A fuzzy for detecting spatiotemporal outliers. **Geoinformatica**, p. 597-619, out. 2011. Disponível em: `https://dl.acm.org/citation.cfm?id=2159278`. Acesso em: 10 set. 2020.

[21] HILBERT, M.; LÓPEZ, P. The World's Technological Capacity to Store, Communicate, and Compute Information. **Sciencexpress Ressearch Article**, Isle of Man, v. 332, p. 60-65, 2011. Disponível em: `http://science.sciencemag.org/content/332/6025/60`. Acesso em: 14 set. 2020.

[22] HORNIK, K. **NLP: Natural Language Processing** Infrastructure.R package version 0.2-0, 2018. Disponível em: `https://CRAN.R-project.org/package=NLP`. Acesso em: 20 ago. 2020.

[23] JOHNSON, R. A.; WICHERN, D. W. **Applied Multivariate Statistical Analysis**. 6 ed. New Jersey: Pearson Pratice Hall, 2007.

[24] KASSAMBARA, A.; MUNDT, F. **factoextra: Extract and Visualize the Results of Multivariate Data Analyses**.R package version 1.0.5, 2017. Disponível em: `https://CRAN.R-project.org/package=factoextra`. Acesso em: 03 jul. 2020.

[25] MACHADO, A. P. *et al.* Mineração de texto em redes sociais aplicada à educação a distância. **Colabor@ - Revista Digital da CVA-RICESU**, v. 6, n. 23, jul. 2010. Disponível em: `http://pead.ucpel.tche.br/revistas/index.php/colabora/article/view/132`. Acesso em: 07 set. 2020.

[26] MADEIRA, R. de O. C. **Aplicação de técnicas de mineração de texto na detecção de discrepâncias em documentos fiscais**. 68f. Dissertação (Mestrado em Ciências, ênfase em Modelagem Matemática da Informação) - Escola de Matemática Aplicada da Fundação Getúlio Vargas, Rio de Janeiro, 2015. Disponível em: `http://bibliotecadigital.fgv.br/dspace/bitstream/handle/10438/14593/TEXTO%20DISSERTA%c3%87%c3%830%20VFINAL1.pdf?sequence=1&isAllowed=y`. Acesso em: 08 ago. 2020.

[27] MAMDANI, E. H.; ASSILIAN, S. An experiment in linguistic synthesis with a fuzzy logic controller. **International Journal of Man-Machine Studies**, v. 7, p. 1–13, 1975. Disponível em: `https://www.sciencedirect.com/science/article/pii/S0020737375800022`. Acesso em: 24 ago. 2020.

[28] MARTINS, N. H. **Avaliação de "custo-benefício" sob a ótica da Inteligência Computacional com enfoque quali-quantitativo**. 81f. Exame de

Qualificação (Mestrado em Ciências Computacionais) - Universidade do Estado do Rio de Janeiro, Rio de Janeiro, 2015.

[29] MENDONÇA, J. H.; SANDRI, S. A.; DRUMMOND, I. N. Técnicas Baseadas em Redes Neurais Artificiais e Lógica Difusa para Mineração de Textos. *In*: 10 Brazilian Congress on Computacional Intelligence, 10, 2011, Fortaleza, CE. **Anais [...]**, Fortaleza: ABRICOM, 2011. Disponível em: `http://abricom.org.br/wp-content/uploads/2016/03/st_27.5.pdf`. Acesso em: 05 set. 2020.

[30] MERLI, R. F. ALMEIDA, L. M. W. de. Nem tudo é tão certo como parece ser: A Matemática Fuzzy como Linguagem. *In*: XI Encontro Paranaense de Matemática, 11, 2011, Apucarana, RS. **Anais [...]**, Apucarana: XI EPREM, 2011. Disponível em: `http://www.uel.br/grupo-pesquisa/grupemat/docs/CC07_eprem2011.pdf`. Acesso em: 07 ago. 2020.

[31] MINGOTI, S. A. **Análise de Dados Através de Métodos de Estatística Multivariada**: Uma Aborgadem Aplicada. 2 ed. Belo Horizonte: UFMG, 2013.

[32] MINISTÉRIO DA SAÚDE. **Sobre a Doença**: O que é COVID-19. Disponível em: `https://coronavirus.saude.gov.br/sobre-a-doenca`. Acesso em: 14 set. 2020.

[33] MORAIS, E. A. M.; AMBRÓSIO, A. P. L. **Mineração de Textos**. Goiânia: Universidade Federal de Goiás, 2007. Disponível em: `http://www.inf.ufg.br/sites/default/files/uploads/relatorios-tecnicos/RT-INF_005-07.pdf`. Acesso em: 21 jul. 2020.

[34] PAL, N. R.; BEZDEK, J. C. On Cluster Validity for the Fuzzy c-Means Model. **IEEE Transactions on Fuzzy Systems**, v. 3, p. 370-379, ago. 1995. Disponível em: `https://ieeexplore.ieee.org/document/413225`. Acesso em: 24 jul. 2020.

[35] PATIL, D. B.; DONGRE, Y. V. A fuzzy approach for text mining. **International Journal Mathematical Sciences and Computing**, Hong Kong p. 34-43, nov. 2015. Disponível em: `http://www.mecs-press.org/ijmsc/ijmsc-v1-n4/IJMSC-V1-N4-4.pdf`. Acesso em: 13 ago. 2020.

[36] R PROJECT. **The R Project for Statistical Computing**. *Software* para Análises Estatísticas. Disponível em: `https://www.r-project.org`. Acesso em: 02 set. 2020.

[37] RAMALHO, A. de O.; FREITAS, P. de S. S.; NOGUEIRA, P. C. Lesão por pressão relacionada a dispositivo médico nos profissionais de saúde em época de pandemia. **ESTIMA, Braz. J. Enterostomal Ther**, São Paulo, v. 18, 120 ed., 2020. Disponível em: https://www.revistaestima.com.br/index.php/estima/article/view/867/pdf_1. Acesso em: 10 ago. 2020.

[38] RANKS NL WEBMASTER TOOLS. **Brazilian Stopwords**. 1998. Disponível em: https://www.ranks.nl/stopwords/brazilian. Acesso em: 25 mar. 2020.

[39] RIGNEL, D. G. de S; CHENCI, G. P.; LUCAS, C. A. Uma Introdução a Lógica Fuzzy. **Revista Eletrônica de Sistemas de Informação e Gestão Tecnológica**, v. 1, p. 17-28, 2011. Disponível em: http://www.logicafuzzy.com.br/wp-content/uploads/2013/04/uma_introducao_a_logica_fuzzy.pdf. Acesso em: 18 ago. 2020.

[40] RODRIGUES, R. L. **Estatística Computacional**: Análise de Conglomerados. Recife, 2018. Universidade Federal Rural de Pernambuco. Disponível em: https://pt.slideshare.net/rodrigomuribec/aula-6-anlise-de-conglomerados?from_action=save. Acesso em: 26 ago. 2020.

[41] ROSS, T. J. **Fuzzy logic with engineering applications**. 3 ed. University of New Mexico: John Wiley & Sons, 2010.

[42] RSTUDIO TEAM. **RStudio**: Integrated Development for R. RStudio, Inc., Boston, MA, 2016. Disponível em: http://www.rstudio.com/.

[43] SILVA, G. L. A. da. **Text Mining, um estudo a partir da rede social *Twitter***. 33f. Monografia (Bacharelado em Estatística) - Universidade Federal do Rio Grande do Sul, Porto Alegre, 2013. Disponível em: https://lume.ufrgs.br/handle/10183/102012. Acesso em: 04 ago. 2020.

[44] SIMONS, P. Łukasiewicz and the Several Senses of Possibility. **European Review**, v. 23, p. 114-124, fev. 2015. Disponível em: https://www.cambridge.org/core/journals/european-review/article/lukasiewicz-and-the-several-senses-of-possibility/0DA9832152FD67886E69D63A3042ACEC. Acesso em: 24 ago. 2020.

[45] STAUDT JUNIOR, J. L. **Text Mining Utilizando o Software R: um estudo de caso de uma biblioteca Americana**. 49f. Trabalho de Conclusão

de Curso (Bacharelado em Estatística) – Universidade Federal do Rio Grande do Sul, Porto Alegre, 2016. Disponível em: `https://lume.ufrgs.br/handle/10183/149102`. Acesso em: 08 ago. 2020.

[46] TAN, A.-H. **Text Mining**: The state of the art and the challenges. Singapore: Nanyang Technological University, 1999. Kent Ridge Digital Labs. Disponível em: `http://www3.ntu.edu.sg/home/asahtan/papers/tm_pakdd99.pdf`. Acesso em: 10 set. 2020.

[47] TESSAROLO, P. H.; MAGALHÃES, W. B. **A era do big data no conteúdo digital**: os dados estruturados e não estruturados. Universidade Paranaense, Paranavaí, 2015. Disponível em: `http://web.unipar.br/~seinpar/2015/_include/artigos/Pedro_Henrique_Tessarolo.pdf`. Acesso em: 12 set. 2020.

[48] WIVES, L. K. **Tecnologias de Descoberta de Conhecimento em Textos Aplicadas à Inteligência Competitiva**. 116f. Tese (Doutorado em Computação) - Universidade Federal do Rio Grande do Sul, Porto Alegre, 2002. Disponível em: `https://seer.ufrgs.br/cadernosdeinformatica/article/view/v1n1p25-28`. Acesso em: 10 ago. 2020.

[49] YONAMINE, F. S. *et al.* **Aprendizado não-supervisionado em domínios *fuzzy* - algoritmo *fuzzy c-means***. São Carlos: Universidade Federal de São Carlos, 2002. Disponível em: `http://www2.dc.ufscar.br/~carmo/relatorios/RT_Fuzzy_Cmeans_Final.PDF`. Acesso em: 11 set. 2020.

[50] ZADEH, L. A. Fuzzy Sets. **Journal Information and Control**, p. 338-353, 1965.

[51] ZIMERMMANN, H.-J. **Fuzzy Set Theory – and Its Applications**. 4 ed. New York: Springer Science+Business, 2001.