

Use of the PAM algorithm in k-medoids for clustering mental health textual data

Utilização do algoritmo pam em k-medóides para agrupamento de dados textuais de saúde mental

Bruno G. Silva¹, Anderson C. S. Oliveira², Lia H. M. Morita¹, Thiago M. Brito²

¹ Departamento de Estatística – Universidade Federal do Mato Grosso (UFMT)

² Departamento de Psicologia – Universidade Federal do Mato Grosso (UFMT)

anderson.oliveira@ufmt.br

Abstract. *The aim of this study was to investigate the application of the PAM algorithm for clustering using the k-medoid method, applied to open questions in a questionnaire on mental health of university students. The PAM algorithm was used with the Euclidean distance, based on the matrix of documents and terms. The results revealed that the PAM algorithm, with two, three and four initial k-medoids, analyzed 427 open responses, with a volume of 2101 words, with processing time of 40.05, 40.39 and 48.69 seconds respectively. The PAM algorithm demonstrated good efficiency to perform cluster analysis on textual data.*

Keywords: *Text mining; Document-term matrix; Mental health perception; Unstructured data; Questionnaires.*

Resumo. *O objetivo deste estudo foi investigar a aplicação do algoritmo PAM para o agrupamento utilizando o método k-medóide, aplicado em perguntas abertas de um questionário sobre saúde mental de estudantes universitários. O algoritmo PAM foi empregado com a distância euclidiana, tendo como base a matriz de documentos e termos. Os resultados demonstraram que o algoritmo PAM, com dois, três e quatro k-medóides iniciais, analisou 427 respostas abertas, com um volume de 2101 palavras, com tempo de processamento de 40.05, 40.39 e 48.69 segundos respectivamente. O algoritmo PAM demonstrou uma boa eficiência para realizar análises de clusters em dados textuais.*

Palavras-chave: *Mineração de texto; Matriz de documentos e termos; Percepção de saúde mental; Dados não estruturados; Questionários.*

1. Introdução

A clusterização de texto, uma técnica também reconhecida como agrupamento, desempenha um papel fundamental ao classificar documentos similares em categorias distintas. Nesse processo, a complexidade dos dados textuais é simplificada e padrões emergem à luz. Esta abordagem é de suma importância para a identificação de informações cruciais e estruturas subjacentes nos textos analisados [Ariff et al. 2018, García et al. 2020, Vishwakarma et al. 2017].

Um método de agrupamento não-hierárquico amplamente utilizado é o *k*-medóides, que viabiliza a identificação de pontos centrais representativos em grupos, reduzindo sua vulnerabilidade a valores atípicos e proporcionando uma abordagem robusta para a tarefa de agrupamento de dados. No algoritmo de particionamento em torno de medóides (PAM), analisa-se um conjunto de dados X contendo n objetos, cada um deles com f atributos. Esse método é direcionado a um número predeterminado de grupos k . Cada objeto selecionado formará, em conjunto com outros objetos, um novo agrupamento. Os $n - k$ objetos restantes precisam ser associados ao medóide mais próximo, utilizando uma medida de distância que avalia a similaridade ou dissimilaridade [Brito et al. 2010, Ariff et al. 2018, Vishwakarma et al. 2017]

O objetivo deste estudo é investigar a utilização do algoritmo PAM para agrupar dados textuais, fazendo uso de informações de um questionário sobre saúde mental. Além disso, busca-se contribuir para o campo da mineração de texto, da leitura de dados não estruturados e das possibilidades de visualização, compreensão e interpretação de um volume considerável de respostas abertas em questionários.

2. Materiais e Métodos

2.1. Dados

Neste estudo, a fonte de dados compreendeu uma questão aberta extraída de um questionário aplicado a estudantes de graduação dos campi Cuiabá e Várzea Grande da Universidade Federal de Mato Grosso. A pesquisa foi conduzida em conformidade com os padrões éticos estabelecidos e passou por análise pelo Comitê de Ética em Pesquisa (CEP/Humanidades/UFMT), obtendo aprovação sob o número de protocolo CAAE 53235421.5.0000.5690. A questão é apresentada a seguir:

- Q1 - No momento, como descreve sua saúde mental (sentimentos, pensamentos, emoções, propósitos de vida, relações sociais)?

2.2. Pre-processamento

A questão passou pela etapa de pré-processamento com o objetivo de remover informações irrelevantes. As etapas envolvidas nesse estágio incluíram a exclusão de números, pontuações, acentos e espaços em branco duplicados. Além disso, as stopwords, que são palavras comuns, mas com pouco valor semântico, foram eliminadas para focar as palavras mais significativas presentes nas respostas. Como parte do processo de normalização, todo o texto foi convertido para letras minúsculas.

Após essa etapa, foi empregado o processo de tokenização para fragmentar cada resposta em unidades individuais conhecidas como "tokens" ou palavras. Posteriormente, foi construída uma matriz de documentos-terms para representar os textos de maneira matemática. Nessa matriz, cada linha corresponde à resposta de um estudante, enquanto cada coluna representa uma palavra específica. O valor em cada célula indica a frequência da palavra na respectiva resposta. Isso permite análises e comparações baseadas na frequência das palavras.

O processo de pré-processamento foi executado utilizando o pacote `tm` [Feinerer and Hornik 2023] do software R [R Core Team 2023].

2.3. Processo de agrupamento

A partir da matriz de documentos e termos, foi realizado o processo de agrupamento. Nesse contexto, a medida de similaridade adotada entre os documentos (respostas da questão) foi a distância euclidiana.

Dessa forma, o particionamento foi executado com o objetivo de minimizar a função de custo representada por:

$$J = \sum_{i=1}^k \sum_{\forall x_j \in med_i} d_{ij}$$

em que $d_{i,j}$ é a distância euclidiana, x_j é a observação j pertencente ao medóide i .

O processo de agrupamento foi executado utilizando o pacote cluster [Maechler et al. 2023] do software R [R Core Team 2023].

3. Resultados

A Tabela 1 apresenta os tempos de execução do algoritmo PAM para dois, três e quatro clusters, sendo 40,05, 40,39 e 48,69 segundos, respectivamente. Esse desempenho rápido evidencia a eficácia do algoritmo, que produz agrupamentos precisos em um tempo inferior a um minuto. Isso destaca a capacidade do PAM em lidar eficientemente com tarefas complexas de agrupamento, permitindo análises ágeis e tomadas de decisão rápidas.

Tabela 1. Tempo de execução do algoritmo PAM, realizado na análise de agrupamento

Quantidade de Clusters	Tempo de processamento em segundos
2	40,05
3	40,39
4	48,69

Na literatura, os tempos de execução do algoritmo PAM são amplamente discutidos, especialmente em contextos que envolvem dados textuais. A medida de similaridade escolhida desempenha um papel crucial nesse aspecto, sendo que a utilização da distância euclidiana na matriz de frequência de palavras após a etapa da Matriz de Documentos e Termos se destacou como uma abordagem eficiente. Essa estratégia otimizou o tempo de execução, resultando em agrupamentos significativos e consistentes, evidenciando a relevância da metodologia adotada para um desempenho eficaz do algoritmo PAM [Vishwakarma et al. 2017].

Embora a configuração com quatro clusters tenha exigido mais tempo de execução, ela exibiu uma melhoria na homogeneidade interna dos agrupamentos. Assim, os resultados serão apresentados apenas para essa configuração.

Na Figura 1 são apresentadas as nuvens de palavras correspondentes a cada um dos quatro clusters. Cada nuvem de palavras proporciona uma representação visual das palavras mais recorrentes ou pertinentes dentro de cada agrupamento levando em consideração a análise com quatro k-medoides. Além disso, a palavra em destaque nas nuvens de palavras (azul) denota o medóide.

Figura 1. Nuvem de palavras para os cluster formados para a pergunta



É observado que o medóide é uma palavra de baixa frequência (específicos, semana, morte e bolsista com frequências de dois, quatro, três e um respectivamente), entretanto, desempenha um papel fundamental na redução da distância entre as demais palavras presentes no texto. Isso implica que, mesmo com sua ocorrência menos frequente, o medóide assume a função de um ponto central que fortalece a coesão do agrupamento, assegurando a proximidade entre as palavras circundantes.

Esse resultado ressalta a importância estratégica do medóide no processo de agrupamento de palavras. Ele atua como uma representação condensada do cluster, exercendo um impacto direto na interconexão e distribuição das outras palavras dentro desse agrupamento.

Ao examinar o conteúdo das nuvens de palavras em cada cluster, não foi possível identificar uma relação imediata. Isso pode indicar que a análise de sentimentos ainda não foi integrada nessa etapa, o que poderia ter impacto na compreensão textual, sobretudo devido à natureza das informações ligadas à saúde mental.

É importante destacar, adicionalmente, que no âmbito da saúde mental, palavras de relevância surgem no cluster 1, tais como "graduação", "faculdade", "futuro" e "depressão". Além disso, no cluster 2, termos como "tempo", "vezes", "saúde", "vida", "relações sociais" e "sempre" ganham destaque. No cluster 3, emergem palavras como "momentos", "perspectivas", "pessoas" e "pandemia". Por último, no cluster 4, termos como "ansiedade", "crises", "amigos" e "cansado" se sobressaem. Esses agrupamentos de palavras proporcionam uma visão inicial sobre as temáticas subjacentes a cada cluster.

Portanto, apesar dos resultados positivos, é importante considerar a necessidade

de análises mais avançadas, como a análise de sentimentos e a incorporação de estruturas sintáticas mais complexas, para compreender plenamente as relações semânticas e a riqueza dos resultados obtidos.

4. Considerações Finais

O algoritmo PAM para agrupamento de dados textuais, demonstrou uma boa eficiência na realização de análises de clusters em conjuntos de dados textuais. Especificamente, em questionários que incluem respostas abertas, esse método se revelou altamente eficaz. Além de proporcionar uma visão diversificada das respostas dos participantes, ele também ofereceu uma abordagem ágil para lidar com dados categóricos. A capacidade do algoritmo PAM de identificar padrões semânticos e estruturas subjacentes nas respostas abertas tornou-o uma ferramenta valiosa para a exploração e compreensão dos dados textuais, contribuindo significativamente para o campo da mineração de texto e análise de dados não estruturados.

Referências

- Ariff, N. M., Bakar, M. A. A., and Rahmad, M. I. (2018). Comparative study of document clustering algorithms. *International Journal of Engineering Technology*, 7(4.11):246–251.
- Brito, J. A. M., Ochi, L. S., Brito, L. R., and Montenegro, F. M. T. (2010). Um algoritmo para o agrupamento baseado em k-medoids. *Revista Brasileira de Estatística*, 71(234):75–100.
- Feinerer, I. and Hornik, K. (2023). *tm: Text Mining Package*. R package version 0.7-11.
- García, R. G., Beltrán, B., Vilariño, D., Zepeda, C., and Martínez, R. (2020). Comparison of clustering algorithms in text clustering tasks. *Computación y Sistemas*, 24(2):499–437.
- Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., and Hornik, K. (2023). *cluster: Cluster Analysis Basics and Extensions*.
- R Core Team (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Vishwakarma, S., Nair, D. P. S., and Rao, D. S. (2017). Comparative study of k-means and k-medoid clustering for social media text mining. *INTERNATIONAL JOURNAL OF ADVANCE SCIENTIFIC RESEARCH AND ENGINEERING TRENDS*, 2(1):297–302.