Filtragem Inteligente de Notícias: Uma Abordagem Baseada em Clusterização*

Luíza Diapp[1], Lisiane Reips[1], Aurora T. R. Pozo[1], Carmem S. Hara[1]

[1]Departamento de Informática, Curitiba-PR
Universidade Federal do Paraná

{luiza.diapp, lisiane.reips, aurora.pozo, carmemhara}@ufpr.br

Abstract. *With the growing volume of news available online, efficient tools are needed to help users quickly find relevant information. The ENoW (Web News Extractor) tool was designed to automatically collect news articles based on user-defined keywords, enabling data storage and applying an intelligent filtering system to highlight relevant content. However, the initial filtering process requires users to manually select the most relevant news from a randomly selected sample of collected articles. As a result, users often need to request multiple new samples to find relevant content, which makes the process both time-consuming and exhausting. To address this issue, this paper proposes the aplication of K-Means clustering algorithm to refine the filtering process, ensuring that the initial sample better represents the different extracted topics. The results showed a significant reduction in the number of articles users needed to browse in order to identify relevant content. This improvement was subsequently integrated into the ENoW tool, enhancing the overall user experience in news filtering.*

Resumo. *Com o grande volume de notícias disponíveis na Web, tornou-se essencial o uso de ferramentas que facilitem a busca por informações relevantes. A ferramenta ENoW (Extrator de Notícias da Web) foi desenvolvida para coletar automaticamente notícias com base em palavras-chave definidas pelo usuário, permitindo o armazenamento dos dados e aplicando um sistema de filtragem para exibir conteúdos de interesse. No entanto, o processo de filtragem exige que o usuário selecione manualmente as notícias mais relevantes para ele, dentro de uma amostra da coleta. Como essa amostra era obtida aleatoriamente, muitas vezes era necessário solicitar várias novas amostras até encontrar conteúdos pertinentes, tornando o processo demorado e exaustivo. Para mitigar esse problema, este artigo propõe a aplicação do algoritmo de clusterização K-Means para aprimorar a escolha da amostra, garantindo que ela seja mais diversificada e representativa dos diferentes tópicos extraídos. Os resultados mostraram uma redução significativa na quantidade de notícias analisadas pelo usuário, tornando a identificação de conteúdos relevantes mais rápida e eficiente. A abordagem foi incorporada à ferramenta ENoW, otimizando a experiência do usuário na filtragem de notícias.*

*Este trabalho foi possível graças à parceria e confiança da Professora Carmem Hara e da Lisiane Reips, às quais agradeço pela oportunidade e pelo apoio.

1. Introdução

Atualmente, o acesso eficiente a informações do cotidiano tornou-se indispensável. Diante da vasta quantidade de notícias disponíveis na *Web*, torna-se essencial o uso de ferramentas capazes de filtrar e organizar esses dados de maneira eficiente [Reips et al. 2023]. A extração automatizada de dados, aliada à filtragem inteligente, destaca-se como uma solução inovadora para esse desafio, pois possibilita não apenas a coleta estruturada de informações, mas também a seleção otimizada de conteúdos relevantes. A ferramenta ENoW (Extrator de Notícias da *Web*) realiza a raspagem gratuita de dados com base nas *strings* de pesquisa escolhidas pelo usuário, possibilitando o armazenamento de notícias juntamente com seus metadados e oferecendo filtragem inteligente baseada nas preferências dele [Reips 2023].

Durante o início do processo de filtragem, o usuário precisa selecionar manualmente, dentro de uma amostra aleatória inicial, quais notícias são de seu interesse. Caso nenhuma notícia relevante seja encontrada nessa primeira amostra, uma nova amostra é apresentada, e assim sucessivamente. Entretanto, era comum que o usuário precisasse analisar diversas amostras até encontrar uma notícia relevante, o que tornava o processo de busca demorado e exaustivo.

Nesse cenário, neste artigo é proposta a utilização de um algoritmo de clusterização das notícias extraídas para a geração da primeira amostra. O objetivo é aprimorar a apresentação da amostra inicial e minimizar a quantidade de notícias que o usuário precisa analisar. A vetorização TF-IDF [Aizawa 2003] foi utilizada para converter as notícias em representações numéricas, assim permitindo a aplicação do algoritmo *K-Means* [Xu and Wunsch 2005], da biblioteca *scikit-learn* de *Python*. O modelo empregado busca garantir que a amostra inicial contenha uma representação diversificada dos diferentes tópicos abordados nas notícias extraídas, reduzindo a necessidade de o usuário percorrer múltiplas amostras para selecionar conteúdos relevantes [Barbosa et al. 2021].

A implementação dessa abordagem reduziu a quantidade de notícias apresentadas ao usuário na etapa inicial do processamento. Diante dos resultados, a solução foi incorporada à ferramenta ENoW.

O restante do artigo está organizado da seguinte forma. A Seção 2 apresenta trabalhos relacionados à abordagem de clusterização apresentada na seção seguinte. A Seção 3 descreve a abordagem proposta para a obtenção de uma amostra de notícias, bem como os resultados de um estudo de caso. Na Seção 4 são apresentados os trabalhos futuros.

2. Trabalhos Relacionados

Na presente Seção, são apresentados os estudos correlatos que empregam a clusterização dos dados de notícias *online*. Destacam-se, além da clusterização de grandes volumes de artigos de notícias baseados em similaridade de conteúdo [Bouras and Tsogkas 2012], pesquisas referentes à otimização em clusterização de textos de notícias utilizando a técnica TF-IDF [Zhou et al. 2020] e abordagens sobre medição de similaridade de textos, usando clusterização [Lan 2022].

A técnica de clusterização para notícias com base na similaridade de conteúdo foi adaptada pelo *PeRSSonal* [Bouras and Tsogkas 2012]. Esse sistema engloba as etapas de

coleta de artigos, pré-processamento, aplicação de métodos de *clustering*, recuperação de informações (RI) e rotulagem dos grupos gerados, seguindo um fluxo semelhante ao do ENoW. O método de *clustering* adotado baseia-se no algoritmo *K-Means*, também empregado no ENoW. No entanto, a principal distinção do *PeRSSonal* está na incorporação do *WordNet*, que fornece informações externas e estabelece relações semânticas entre palavras, aprimorando a técnica utilizada. Com isso, o *PeRSSonal* aprimora a eficácia da RI e organiza as informações de forma estruturada, enquanto o ENoW adota uma abordagem sem essa integração semântica.

Um método otimizado para a clusterização de tópicos em textos de notícias foi desenvolvido por [Zhou et al. 2020], empregando a técnica de vetorização TF-IDF sobre a plataforma de *Spark*. Assim como no ENoW, a abordagem utiliza o TF-IDF, porém com o diferencial de estar integrada a um ambiente de processamento distribuído. Seu principal objetivo é melhorar a eficiência no processamento de grandes volumes de dados. Além do TF-IDF, a metodologia inclui o *CountVectorizer* e o modelo *Latent Dirichlet Allocation* (LDA) para identificação de tópicos. A principal diferença para o ENoW está na adoção dessas técnicas e no uso da plataforma *Spark*, enquanto o ENoW emprega exclusivamente o *K-Means*, para agrupar as notícias por similaridade.

[Lan 2022] propõe outra abordagem que combina a vetorização do TF-IDF e integra informações semânticas extraídas da base *HowNet* para comparação de artigos científicos. Embora compartilhe o uso do TF-IDF com o ENoW, essa proposta se diferencia pela incorporação da estrutura *Term Similarity Weighting Tree* (TSWT), utilizada para ponderação da similaridade entre termos. Enquanto o ENoW se baseia no *K-Means* para a clusterização de notícias, essa abordagem busca aprimorar a precisão na medição de similaridade textual em RI e classificação de textos com um modelo semântico.

O diferencial do ENoW reside em sua etapa de filtragem semi-supervisionada pós-coleta, que permite a adaptação dinâmica aos interesses do usuário através de feedback interativo - contrapondo-se às abordagens comparadas, que se limitam à organização automática de textos. Essa integração entre extração automatizada de notícias e filtragem inteligente das mesmas, baseadas no interesses do usuário, configura a originalidade da proposta [Reips and Hara 2022].

3. Geração de uma Amostragem das Notícias

Para que a ferramenta ENoW (Extrator de Notícias da *Web*) execute uma coleta de notícias, é necessário seguir algumas etapas. Primeiramente, o usuário cria um projeto, vinculando palavras-chave e selecionando os *sites* de jornal de notícias que lhe interessam. Em seguida, a ferramenta realiza a coleta das notícias associadas ao projeto. Embora essa coleta esteja restrita às palavras-chave e aos *sites* definidos, o conjunto de notícias coletado normalmente é bastante diverso, abordando múltiplos temas, incluindo notícias que não são de interesse do usuário.

Nesse sentido, para atender aos interesses específicos de cada usuário, torna-se necessária uma etapa de filtragem de notícias. Como os critérios de relevância podem variar entre os indivíduos, a ferramenta não pode determinar, por si só, quais notícias são mais adequadas sem um ponto de referência definido pelo próprio usuário [Chawla and Karakoulas 2005]. Assim, é necessário que ele classifique manualmente um conjunto amostral de notícias, indicando quais são relevantes e quais não são. Esse con-

junto servirá como base para que a ferramenta aprenda e se adapte às preferências individuais do usuário. A qualidade dessa amostra tem um impacto direto na precisão das recomendações futuras, pois é com base nela que o sistema ajustará seus critérios para identificar notícias relevantes.

Na versão anterior do ENoW, essa amostra inicial era selecionada de maneira aleatória. Isso fazia com que o usuário, muitas vezes, precisasse avaliar diversas amostras iniciais (cada uma com 10 notícias) antes de encontrar alguma que realmente fosse de seu interesse. Essa abordagem prolongava a etapa, tornando-a demorada e cansativa, o que poderia comprometer a experiência do usuário [Barbosa et al. 2021] e desmotivar a utilização da ferramenta.

Para resolver essa questão, este artigo propõe a utilização do método de clusterização *K-Means* para a geração da amostra inicial. O algoritmo subdivide as notícias coletadas em grupos temáticos distintos [Zhou et al. 2020].

A proposta é ilustrada na Figura 1. Após a coleta das notícias (1) e a consequente formação da base bruta de dados (A), a ferramenta realiza a extração de atributos (2) de um subconjunto de 1.000 notícias da base bruta. Nesta etapa, elementos considerados irrelevantes para a análise, como a URL da notícia e a data da última atualização do *site*, são removidos, garantindo a formação da base estruturada de notícias (B) que contém exclusivamente atributos relevantes.

Com essa base estruturada, ocorre a seleção das amostras iniciais (3), conduzida por meio do método de clusterização *K-Means*, resultando na geração de um conjunto de amostras balanceado (C). Este conjunto é obtido visando maximizar a representatividade temática, para que cada amostra contenha pelo menos uma notícia de cada tópico identificado na base bruta. Assim, o usuário faz uma seleção inicial das notícias mais relevantes para si (4), formando uma base rotulada (D).

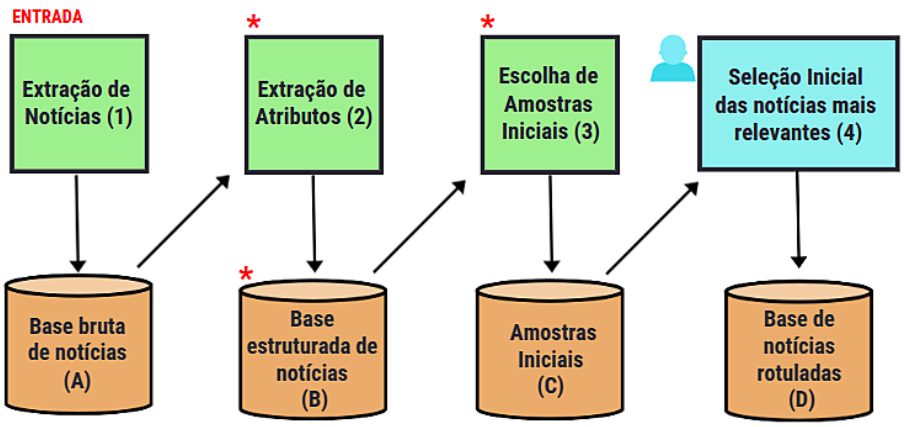


Figura 1. Escolha de Amostras Iniciais - Versão Atual

Essa abordagem permite que a amostra inicial mostrada para o usuário não seja aleatória, mas sim baseada em uma diversidade representativa dos temas identificados no conjunto de notícias coletadas. Com essa estratégia, em vez de avaliar uma lista aleatória de notícias, o usuário passa a classificar uma amostra mais equilibrada, contendo pelo menos uma notícia de cada grupo temático identificado.

Os asteriscos na figura representam as mudanças que foram feitas nessa nova versão da ferramenta. Antes, a etapa *2* não era implementada, não havia a base *B* de notícias e a etapa *3*, como já mencionado, era feita de maneira aleatória e não clusterizada.

Esse ajuste torna a seleção do usuário, ou seja a etapa de rotulagem inicial, mais eficiente, pois reduz a necessidade de percorrer inúmeras notícias irrelevantes antes de encontrar conteúdos interessantes, permitindo que a ferramenta aprenda mais rapidamente as preferências do usuário e garantindo uma melhor representatividade dos temas disponíveis.

Por fim, a partir de uma base rotulada inicial, o sistema aplica o método de similaridade de cosseno [Park et al. 2020] para filtrar o resto da base bruta de notícias coletadas (5), avaliando a proximidade entre as notícias não rotuladas e as previamente rotuladas. Em seguida, a ferramenta novamente fornece para o usuário um conjunto de 10 notícias, visando assegurar que metade delas sejam relevantes. Ele realiza então uma nova etapa de seleção, expandindo a base de notícias rotuladas. Esse processo pode ser repetido quantas vezes for necessário. Esse ciclo de aprendizado do ENoW pode ser observado na Figura 2.

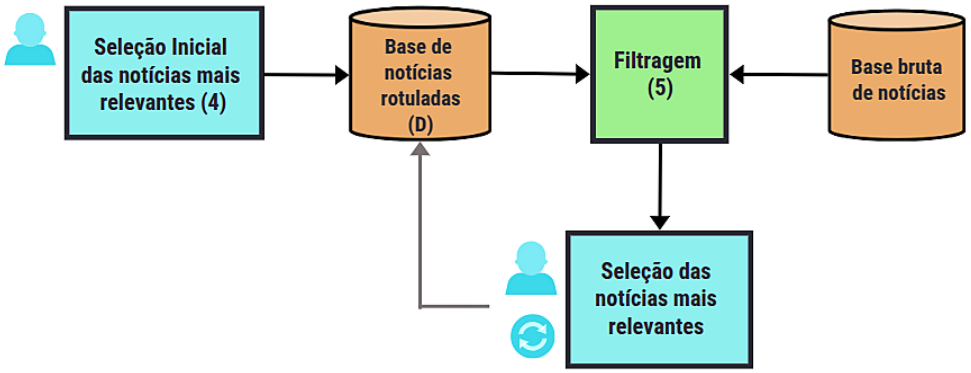


Figura 2. Processo de Aprendizado Iterativo ENoW

Dessa maneira, conforme o usuário continua a interagir com o sistema, a ferramenta ajusta e melhora continuamente suas recomendações. Progressivamente, proporcionando um conjunto de notícias mais alinhado ao tema de interesse do usuário.

3.1. Nova Amostra Inicial de Notícias

O processo de clusterização, empregado para identificar diferentes temas nas notícias coletadas, utiliza técnicas de aprendizado não supervisionado, tendo como base o algoritmo *K-Means*. Esse algoritmo começa com a escolha do número de *clusters* (K) e a seleção aleatória de K centróides. Cada artigo é atribuído ao centróide mais próximo, e os centróides são recalculados com base na média dos pontos atribuídos a cada *cluster*. Esse processo se repete até que os centróides se estabilizem ou um número máximo de iterações seja atingido, garantindo a formação de grupos coerentes [Madhulatha 2012].

A técnica apresenta limitações, como a necessidade de definir previamente o número de *clusters*, sensibilidade à inicialização dos centróides e dificuldades com *clusters* de tamanhos variados e a presença de *outliers* [Madhulatha 2012]. No entanto, foi es-

colhida pela implementação simples, alta velocidade de processamento e capacidade de lidar eficientemente com grandes volumes de dados.

Para aplicar a clusterização na ferramenta ENoW, desenvolveu-se um código capaz de processar diferentes conjuntos de informações extraídas das notícias no banco de dados. Como se pode observar na Figura 1, depois da extração de atributos (2) da base bruta de notícias se obtém uma base estruturada de mil notícias (B). Com esta base foram formados três grupos: o primeiro contém o título e o resumo das notícias (grupo 1), o segundo contém o título e o corpo do texto (grupo 2), e o terceiro contém o título, o resumo e corpo do texto (grupo 3). Após testes e análises, cujos resultados serão discutidos na seção 3.2, a versão final do sistema foi configurada utilizando o grupo 1 como base para a clusterização e o número de *clusters* foi definido para 10.

O pseudocódigo do processo de clusterização utilizado no ENoW é exibido na Figura 3. A entrada do algoritmo consiste em três parâmetros: *B*, a base estruturada contendo a divisão prévia dos grupos 1, 2 ou 3; *n*, a quantidade de notícias a serem extraídas de cada *cluster*; e *k*, o número total de *clusters* desejados. Na linha 1, ocorre a seleção do grupo a ser utilizado na clusterização, conforme a divisão previamente estabelecida em *B*.

Em seguida, na linha 2, é criada uma representação textual onde é realizada uma etapa de limpeza e normalização [Chapman 2005]. Esse procedimento inclui a remoção de caracteres especiais, a conversão para letras minúsculas e a eliminação de palavras irrelevantes (*stopwords*), permitindo que apenas os termos mais informativos sejam considerados na clusterização.

Escolha de Amostras (B, k, n)

```
1 grupoNoticias = B["grupo1"];
2 baseNormalizada = normalizacao(grupoNoticias);
3 baseVetorizada = TF-IDF(baseNormalizada);
4 kGrupos = KMeans(baseVetorizada, k);
5 A = {};
6 para i = 1 ate k
7     A = A U "n elementos mais proximos do centroide";
8 retorna A;
```

Figura 3. Pseudocódigo Representativo do Método de Clusterização

Assim, na linha 3, é possível realizar a vetorização dos textos por meio da técnica TF-IDF (*Term Frequency-Inverse Document Frequency*). Esse método transforma os textos em representações numéricas, ponderando a importância de cada termo dentro do conjunto de documentos [Lan 2022]. Para aprimorar a qualidade da vetorização, são empregados parâmetros responsáveis por excluir palavras excessivamente frequentes ou extremamente raras, de modo a preservar apenas os termos mais relevantes para a análise [Zhou et al. 2020]. Com a matriz TF-IDF gerada, na linha 4, o algoritmo *K-Means* é aplicado para segmentar os textos em *clusters* distintos.

Uma vez formados os *clusters*, o grupo de amostras é criado (linha 5) e o sistema procede à seleção das notícias mais representativas de cada grupo (linha 7). Essa escolha é baseada na distância entre os documentos e o centro do *cluster* correspondente. Assim, para cada *cluster* identificado, são selecionadas as notícias mais próximas ao centro,

assegurando uma amostra diversificada que contempla os principais temas identificados.

O resultado do processo é uma lista de notícias representativas (linha 8) dos diferentes *clusters* obtidos, o que permite obter uma amostra abrangente dos temas presentes no conjunto de dados analisado. Caso o procedimento não encontre textos válidos, uma mensagem de erro é retornada, evitando que o modelo opere sobre um conjunto vazio.

3.2. Análise Experimental

A implementação do código de clusterização foi realizada na linguagem *Python*, utilizando bibliotecas especializadas em aprendizado de máquina e processamento de linguagem natural. O pré-processamento textual foi conduzido com *spaCy*, responsável pela *tokenização*, remoção de *stopwords* e lematização. A vetorização dos textos foi feita por meio da técnica TF-IDF, utilizando a classe *TfidfVectorizer* do *scikit-learn*. Para segmentar as notícias em grupos semanticamente semelhantes, aplicou-se o algoritmo *K-Means*, também do *scikit-learn*. A seleção das notícias mais centrais em cada *cluster* foi realizada com suporte da *NumPy*, que possibilitou a aplicação de operações vetoriais, incluindo o cálculo de distâncias e a identificação dos pontos mais representativos em cada grupo.

Para avaliar a eficácia do processo de clusterização, foi criado um projeto denominado 'RioIguaçu' dentro da ferramenta ENoW. A palavra-chave escolhida para esse projeto foi 'Rio Iguaçu', visando coletar notícias relacionadas à poluição da água desse rio no estado do Paraná. Optou-se por restringir a busca à *string* 'Rio Iguaçu' para evitar a coleta de notícias irrelevantes. A inclusão de termos mais genéricos, como 'poluição', poderia resultar na obtenção de artigos sobre diferentes tipos de poluição, como do ar ou sonora, que não estariam necessariamente relacionados ao Rio Iguaçu.

Em seguida, foi feita a coleta de 9.232 notícias sobre o assunto, sendo 349 da Folha de São Paulo e 8.883 da Gazeta do Povo. A coleta demorou em torno de 9 horas. Os testes foram realizados em um computador Lenovo IdeaPad S145-15IWL com processador Intel Core i5-8265U (4 núcleos, 8 threads), 8 GB de RAM e 447 GB de armazenamento. O sistema operacional foi o Linux Mint 21 "Vanessa", com kernel 5.15.0-130-generic.

Foram realizados sete testes no ENoW. O primeiro seguiu a abordagem original, utilizando o modo aleatório para a geração de amostra de notícias, enquanto os demais empregaram técnicas de clusterização.

No modo aleatório, foram necessárias quatro amostras iniciais — um total de 40 notícias, pois cada amostra possui 10 notícias — até que uma relacionada à poluição da água do Rio Iguaçu fosse encontrada. A Figura 4 é um *print* da quarta amostra inicial mostrada para o usuário, onde se encontrou a primeira notícia relevante.

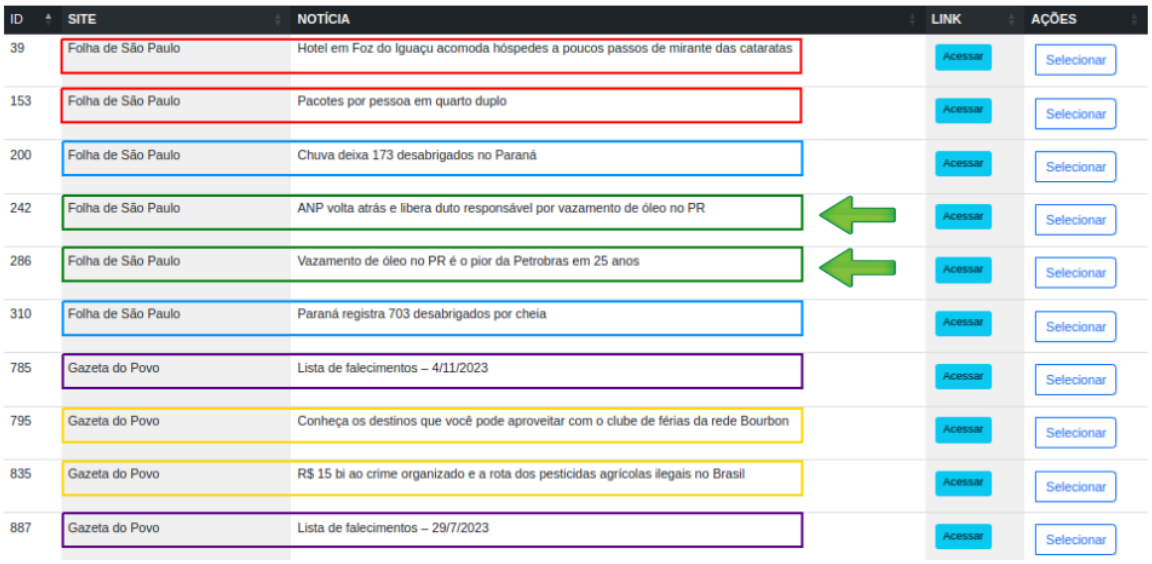
Os testes subsequentes foram organizados com base na clusterização das notícias em três grupos distintos. Como descrito na Seção 3.1, cada grupo utilizou uma combinação específica de atributos: (1) título e resumo, (2) título e corpo do texto e (3) título, resumo e corpo do texto. Para cada grupo, foram realizados dois testes: um com k=5 *clusters* (com duas notícias por *cluster*) e outro com k=10 *clusters* (com uma notícia por *cluster*).

Quando a amostra inicial foi gerada com o grupo 1, utilizando k=5 e k=10, ambas as configurações resultaram na identificação de duas notícias relevantes já na primeira iteração, cada uma abordando uma situação distinta. Ao aplicar a clusterização ao grupo

ID	SITE	NOTÍCIA	LINK	AÇÕES
271	Folha de São Paulo	Barreiras instaladas pela Petrobras não conseguem conter óleo	Acessar	Selecionar
693	Gazeta do Povo	Lista de falecimentos – 27/1/2024	Acessar	Selecionar
1218	Gazeta do Povo	Lista de falecimentos – 1/9/2022	Acessar	Selecionar
1678	Gazeta do Povo	Lista de falecimentos – 19/5/2021	Acessar	Selecionar
1682	Gazeta do Povo	Lista de falecimentos – 16/5/2021	Acessar	Selecionar
3015	Gazeta do Povo	Não durou muito: sol vai embora e dá espaço para névoa e frio	Acessar	Selecionar
4716	Gazeta do Povo	Economia sem descuidar da saúde	Acessar	Selecionar
7290	Gazeta do Povo	6ª rodada	Acessar	Selecionar
7299	Gazeta do Povo	Crise e Justiça travam projetos do PAC	Acessar	Selecionar
7565	Gazeta do Povo	Chega a 174 o número de mortos por dengue no RJ	Acessar	Selecionar

Figura 4. *Print* da 4ª Amostra Inicial Aleatória mostrada pelo ENoW

2, a primeira iteração também resultou na identificação de notícias relevantes. Com k=5, foram encontradas duas notícias, que eram sobre um mesmo acidente que impactou o rio. Para k=10, foi identificada uma única notícia relevante, sobre este mesmo evento.



ID	SITE	NOTÍCIA	LINK	AÇÕES
39	Folha de São Paulo	Hotel em Foz do Iguaçu acomoda hóspedes a poucos passos de mirante das cataratas	Acessar	Selecionar
153	Folha de São Paulo	Pacotes por pessoa em quarto duplo	Acessar	Selecionar
200	Folha de São Paulo	Chuva deixa 173 desabrigados no Paraná	Acessar	Selecionar
242	Folha de São Paulo	ANP volta atrás e libera duto responsável por vazamento de óleo no PR	Acessar	Selecionar
286	Folha de São Paulo	Vazamento de óleo no PR é o pior da Petrobras em 25 anos	Acessar	Selecionar
310	Folha de São Paulo	Paraná registra 703 desabrigados por cheia	Acessar	Selecionar
785	Gazeta do Povo	Lista de falecimentos – 4/11/2023	Acessar	Selecionar
795	Gazeta do Povo	Conheça os destinos que você pode aproveitar com o clube de férias da rede Bourbon	Acessar	Selecionar
835	Gazeta do Povo	R$ 15 bi ao crime organizado e a rota dos pesticidas agrícolas ilegais no Brasil	Acessar	Selecionar
887	Gazeta do Povo	Lista de falecimentos – 29/7/2023	Acessar	Selecionar

Figura 5. *Print* Amostra Inicial Clusterizada a partir do grupo 2 com 5 *clusters*

A Figura 5 é um *print* da primeira amostra apresentada ao usuário após a clusterização, realizada com o grupo 2 e k=5. As diferentes cores ao redor de cada notícia indicam os *clusters* formados, sendo que, neste caso, cada grupo contém duas notícias. O grupo verde é o grupo que apresentou notícias relevantes.

Por fim, ao gerar a amostra inicial a partir do grupo 3, notícias relevantes mais uma vez foram identificadas já na primeira iteração. Com k=5, duas notícias relacionadas ao mesmo evento foram encontradas, enquanto com k=10, novamente duas notícias relevantes foram identificadas, mas abordando acontecimentos diferentes.

Em todos os seis casos testados, a clusterização reduziu as amostras iniciais para uma, limitando a 10 o número de notícias a serem verificadas até a primeira relevante.

Isso reduziu a carga de trabalho do usuário e o tempo de processamento.

O gráfico da Figura 6 diz respeito à quantidade de notícias relevantes encontradas na primeira amostra inicial, ou seja, dentro de um conjunto de 10 notícias, nas diferentes formas de clusterização. O gráfico não inclui a amostragem aleatória, pois na primeira amostra inicial nenhuma notícia relevante foi encontrada.

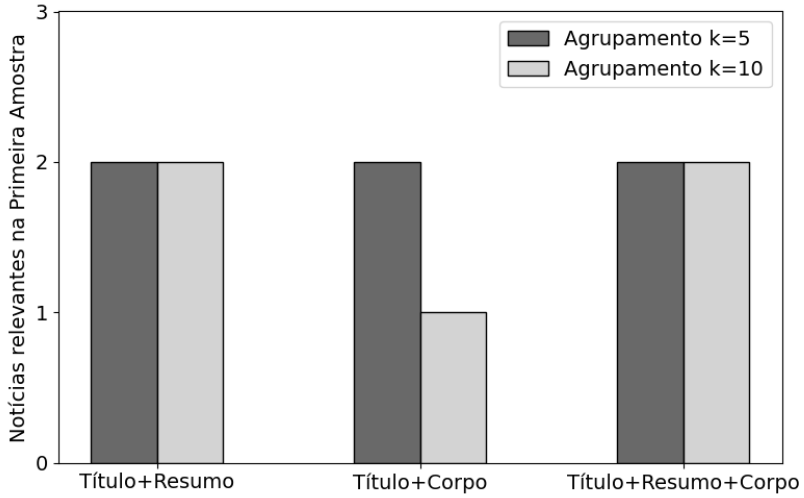


Figura 6. Gráfico de notícias relevantes apresentadas na 1ª amostra inicial em diferentes processos de clusterização

Um ponto importante é que, embora a ferramenta apresentasse notícias relevantes, muitas delas abordavam o mesmo acontecimento, especialmente com k=5 *clusters*. Para aumentar a diversidade da amostra inicial, optou-se por k=10. Além disso, como o desempenho dos grupos foi semelhante, o critério de escolha para a versão final baseou-se no tempo de clusterização e exibição. O grupo 1, que utilizou apenas título e resumo, gerou a amostra em aproximadamente 20 segundos, enquanto os demais levaram mais de 4 minutos. Assim, para a versão final do ENoW, escolheu-se o grupo 1.

4. Conclusão

Neste artigo, propôs-se a utilização do algoritmo *K-Means* para aprimorar a seleção da amostra inicial de notícias na ferramenta ENoW, substituindo o método de escolha aleatória. A abordagem consistiu na vetorização dos textos com TF-IDF e posterior clusterização, buscando garantir uma distribuição mais representativa dos diferentes tópicos abordados nas notícias coletadas. Os resultados indicaram que a nova estratégia reduziu a quantidade de amostras que o usuário precisa percorrer até encontrar uma notícia relevante, limitando a busca inicial a um único conjunto de 10 notícias. Além disso, a escolha de k=10 *clusters* aumentou a diversidade na amostra inicial, diminuindo a repetição de notícias sobre o mesmo evento. Com base nesses resultados, conclui-se que a proposta atendeu ao objetivo de tornar a filtragem mais eficiente, reduzindo o tempo e a quantidade de interações necessárias para que o usuário encontre conteúdos de seu interesse, sendo por isso incorporada ao ENoW.

Entretanto, algumas limitações ainda podem ser exploradas em trabalhos futuros. Como próximos passos, propõe-se: (i) a substituição do TF-IDF por *embeddings*

avançados (BERT, LLMs) para melhor representação semântica; (ii) experimentação com algoritmos de clusterização alternativos (DBSCAN, métodos hierárquicos) e técnicas automáticas para definição de K (*elbow method*); (iii) avaliação sistemática da qualidade dos *clusters* via métricas de coesão e separabilidade; e (iv) integração de sistemas de recomendação baseados em grafos (GCNs). Ademais, a sensibilidade do *K-Means* à inicialização demanda a exploração de outros algoritmos, enquanto testes com dados multilíngues e em múltiplos domínios poderiam validar a generalidade do método para diferentes contextos de aplicação.

Referências

Aizawa, A. (2003). An information-theoretic perspective of tf–idf measures. *Information Processing & Management*, 39(1):45–65.

Barbosa, S. D. J., Silva, B. d., Silveira, M. S., Gasparini, I., Darin, T., and Barbosa, G. D. J. (2021). Interação humano-computador e experiência do usuario. *Auto publicação*.

Bouras, C. and Tsogkas, V. (2012). A clustering technique for news articles using wordnet. *Knowledge-Based Systems*, 36:115–128.

Chapman, A. D. (2005). *Principles and methods of data cleaning*. GBIF.

Chawla, N. V. and Karakoulas, G. (2005). Learning from labeled and unlabeled data: An empirical study across techniques and domains. *Journal of Artificial Intelligence Research*, 23:331–366.

Lan, F. (2022). Research on text similarity measurement hybrid algorithm with term semantic information and tf-idf method. *Advances in Multimedia*, 2022(1):7923262.

Madhulatha, T. S. (2012). An overview on clustering methods. *arXiv preprint arXiv:1205.1117*.

Park, K., Hong, J. S., and Kim, W. (2020). A methodology combining cosine similarity with classifier for text classification. *Applied Artificial Intelligence*, 34(5):396–411.

Reips, L. (2023). Enow - um extrator de notícias da web. Dissertação de mestrado, Universidade Federal do Paraná, Curitiba, Brasil. Orientadora: Carmem Satie Hara.

Reips, L. and Hara, C. (2022). Integração e rotulação automatizada de dados sobre o cnidário physalia physalis, usando a geolocalização como referência. In *Anais Estendidos do XXXVII Simpósio Brasileiro de Bancos de Dados*, pages 105–111, Porto Alegre, RS, Brasil. SBC.

Reips, L., Musicante, M., Vargas-Solar, G., Pozo, A. T., and Hara, C. S. (2023). Enow-extrator de dados de notícias da web. In *Anais Estendidos do XXXVIII Simpósio Brasileiro de Bancos de Dados*, pages 78–83. SBC.

Xu, R. and Wunsch, D. (2005). Survey of clustering algorithms. *IEEE Transactions on neural networks*, 16(3):645–678.

Zhou, Z., Qin, J., Xiang, X., Tan, Y., Liu, Q., and Xiong, N. N. (2020). News text topic clustering optimized method based on tf-idf algorithm on spark. *Computers, Materials & Continua*, 62(1).