ISSN: 2319-0124
josif 2023
15ª JORNADA CIENTÍFICA E TECNOLÓGICA
12º SIMPÓSIO DE PÓS-GRADUAÇÃO
DO IFSULDEMINAS
REALIZAÇÃO INSTITUTO FEDERAL Sul de Minas Gerais
APOIO FAPEMIG CNPq

UTILIZANDO TÉCNICAS DE CLUSTERING E NLP PARA RECOMENDAR FILMES A PARTIR DAS PREFERÊNCIAS DO USUÁRIO

Guilherme da S. NUNES[1]; João G. F. de AZEVEDO[2]; Diego SAQUI[3]

RESUMO

Este trabalho apresenta um estudo onde técnicas de programação em cluster, apoiadas por NLP, são usadas para prever possíveis filmes. O método envolve a categorização de dados de filmes em diferentes grupos ou "clusters" com base em tags anexadas aos filmes.

O objetivo deste estudo é identificar e prever filmes que possam despertar o interesse do usuário, para então fazer recomendações personalizadas com base nas tags pesquisadas.

Palavras-chave:
Cluster; Filmes; Recomendação; NLP.

1. INTRODUÇÃO

De acordo com Bertella (2016), é perceptível que o avanço da tecnologia nos últimos anos desempenhou um papel crucial no crescimento dos sistemas de *streaming*. Porém a ascensão dos serviços de *streaming* também foi impulsionada pela conveniência e flexibilidade que oferecem aos usuários. Parte desta conveniência está na implementação de sistemas de busca eficientes.

Dita a importância de sistemas de busca e o porquê se tornaram cada vez mais evidentes. O propósito deste projeto é criar um sistema de busca, baseado em palavras-chave, que permita aos usuários encontrar informações relevantes de forma rápida e precisa.

Nesse contexto, é fundamental olhar para outras abordagens que aprimoram a análise de dados. Uma dessas é o Clustering, técnica explicada por Azank e Corrêa (2022) que permite a análise detalhada dos dados ao separá-los em grupos com base em suas semelhanças. Existem métodos para identificar o número ideal de clusters, como a estratégia do cotovelo (Elbow). Segundo Temporal (2019), essa estratégia se torna mais sutil à medida que o número de clusters aumenta, utilizando a métrica WCSS.

Pensando ainda em Clustering e na necessidade de organizar fluxos de dados, é importante considerar o uso do Processamento de Linguagem Natural (NLP), descrito por Pinheiro (2021) e Ramadhan (2021), que utiliza técnicas como o TF-IDF para entender a linguagem humana. Para trazer esses dados ao usuário de forma eficaz, é igualmente vital ressaltar o MLOps, conforme

[1]Discente do Superior de Ciência da Computação, IFSULDEMINAS – Campus Muzambinho. E-mail:dev.gnunes@gmail.com.
[2]Discente do Superior de Ciência da Computação, IFSULDEMINAS – Campus Muzambinho. E-mail:12201001252@muz.ifsuldeminas.edu.br.
[3]Docente do Superior de Ciência da Computação, IFSULDEMINAS – *Campus* Muzambinho. E-mail: diego.saqui@muz.ifsuldeminas.edu.br

explicado por Bruel (2020), integrando Machine Learning, DevOps e Engenharia de Dados.

2. MATERIAL E MÉTODOS

Essa seção tem como finalidade descrever as etapas para construção da aplicação para recomendar filmes com o auxílio do algoritmo K-Means, utilizando o dataset MovieLens contendo diversas informações de vários filmes.

2.1 Dataset, Extração dos dados e Normalização

Os dados utilizados para este estudo foram obtidos do banco de dados MovieLens[4], disponibilizado pelo GroupLens (2023), segundo eles trata-se de um conjunto de dados contendo 10,5 milhões de pontuações computadas de relevância de filme de tag de um conjunto de 1.084 tags aplicadas a 9.734 filmes. Lançado em 2021, possui um tamanho aproximado de 1,8 GB.

A análise dos dados foi realizada com o auxílio do Google Colaboratory, um ambiente disponível em nuvem que possibilita a execução de códigos Python, linguagem utilizada durante toda a implementação. Para facilitar o manuseio e a análise dos dados, empregamos a biblioteca Pandas, uma poderosa ferramenta do Python voltada para a manipulação e análise de dados.

Foi construído uma nuvem de palavras com o auxílio da biblioteca wordcloud, sua finalidade é visualizar as principais tags e com qual frequência elas estavam vinculadas, isso para verificar e remover palavras irrelevantes.

A próxima etapa foi analisar e reconstruir a tabela de filmes descartando informações incompletas ou irrelevantes, além de estruturar os dados visando a aplicação do *clustering*. Para cada filme do dataset, foi destacado o id (identificador), título, direção, elenco, avaliação, id do imdb, e tags. Também foi removido filmes que não continham tags ou avaliação, pois entende-se que não seriam relevantes para o usuário final, além de potencialmente prejudiciais para o agrupamento dos dados.

2.2 K-Means Clustering

Para aplicar o algoritmo K-Means, é necessário processar a linguagem natural e vetorizar o texto contendo as tags para cada registro de filme, isso foi possível graças ao TfidfVectorizer da biblioteca sklearn.

Foi executado o método Elbow respectivamente para 10, 50, 100 e 1000 iterações, sendo interrompido durante as 1000 iterações devido ao tempo excessivo e limitações do próprio Google Colaboratory. Optou-se por seguir com 100 clusters, apesar do conhecimento da possibilidade de melhorá-lo.

---

[4] https://grouplens.org/datasets/movielens

Com o algoritmo devidamente treinado e a partir das tags vetorizadas, criou-se um método capaz de receber uma sentença, predicar qual o seu grupo ou cluster correspondente e retornar os 10 resultados mais próximos medindo a similaridade entre dois vetores em um espaço vetorial.

2.3 MLOps

Com o auxílio da biblioteca pickle, foi possível exportar o algoritmo treinado, além de todos os dados necessários para sua execução, compactando todo o conteúdo em um arquivo de extensão. Uma vez criado, basta importá-lo no seu projeto e descompactá-lo também através do pickle.

A interface responsável pela interação entre o usuário e o modelo foi criada em Python, utilizando a biblioteca Flask, junto da framework de estilização Bootstrap 5. Consiste em um formulário onde será inserido um texto descrevendo as preferência do usuário e submetido ao modelo, que irá resultar na exibição de uma tabela contendo informações de título, direção, elenco e avaliações dos filmes recomendados.

3. RESULTADOS E DISCUSSÃO

Com base na Figura 1, observa-se que o usuário terá a capacidade de inserir um texto contendo "tags" e obter os filmes e suas características principais (título, diretor, elenco e também a avaliação dos usuários para aquele filme) de acordo com o esperado.

Title	Directed by	Starring	Rating
The Fight Within (2016)	Michael William Gordon	John Major Davis,Lelia Symington,Matt Leddo,Mike H. Taylor,Wesley Williams	2.5
Riders of Destiny (1933)	Robert N. Bradbury	John Wayne, Cecilia Parker, George 'Gabby' Hayes	2.33333
The Art of Seduction (2005)	Ki-hwan Oh	Song Il-gook,Son Ye-Jin,Sun-yeong Ahn,Noh Joo-hyun	2.75
Carry On at Your Convenience (1971)	Gerald Thomas	Sid James,Kenneth Williams,Charles Hawtrey,Joan Sims,Hattie Jacques	3.04167
Khushi (2003)	SJ. Surya	Kareena Kapoor,Fardeen Khan,Amrish Puri	0.875
Have Dreams, Will Travel (2007)	Brad Isaacs	Cayden Boyd, Lara Flynn Boyle, Matthew Modine, AnnaSophia Robb	3.38462
Mouna Raagam (1986)	Mani Ratnam	Mohan,Revathi,Karthik Muthuraman,Poornam Viswanathan,V. K. Ramasamy	3.66667
My Brother Talks to Horses (1947)	Fred Zinnemann	Butch Jenkins, Peter Lawford, Beverly Tyler	2.5
Zoo in Budapest (1933)	Rowland V. Lee	Loretta Young, Gene Raymond, O.P. Heggie, Wally Albright	3.0
Otto - The Movie (1985)	Xaver Schwarzenberger	Otto Waalkes,Elisabeth Wiedemann,Sky du Mont,Jessika Cardinahl,Peter Kuiper	3.36364

Figura 1: Exemplo de funcionamento da aplicação

Fonte: Acervo pessoal, 2023.

4. CONCLUSÃO

No estudo em questão, o emprego conjunto do algoritmo k-means e da técnica TfidfVectorizer se mostrou fundamental para a classificação e recomendação de filmes com base em tags.

Para fins de validação, foi implementada uma interface web simples. Mesmo com sua simplicidade, a interface cumpriu seu propósito, permitindo que os usuários interajam com o

sistema e testem sua eficácia.

Os resultados alcançados enfatizam a eficiência dessa abordagem, sublinhando o valor de se adotar métodos robustos na busca constante por aprimorar a experiência do usuário em ambientes de recomendação.

REFERÊNCIAS

AZANK, F.; CORRÊA, G. Clustering — Conceitos básicos, principais algoritmos e aplicações. Turing Talks. 1 de maio de 2022. Disponível em: https://medium.com/turing-talks/clustering-conceitos-básicos-principais-algoritmos-e-aplicação-ace572a062a9. Acesso em: 29 jun. 2023.

BERTELLA, G. S. A era do streaming: Uma análise da interação, produção, distribuição e consumo de conteúdo. 2016. 65 f. Monografia (Bacharel em Publicidade e Propaganda) - Curso de Publicidade e Propaganda, Universidade de Passo Fundo, Passo Fundo, 2016.

BREUEL, Cristiano. ML Ops: Machine Learning como Disciplina de Engenharia. 2020. Disponível em: https://medium.com/@cbreuel/ml-ops-machine-learning-como-disciplina-de-engenharia-a058770b93dc. Acesso em: 1 jul. 2023.

GroupLens. (2023). MovieLens Datasets. Disponível em https://grouplens.org/datasets/movielens. Acesso em: 28 jun. 2023.

PINHEIRO, N. M. Introdução ao Processamento de Linguagem Natural — Natural Language Processing (NLP). Data Hackers, 2021. Disponível em: https://medium.com/data-hackers/introdução-ao-processamento-de-linguagem-natural-natural-language-processing-nlp-be907cd06c71. Acesso em: 30 jun. 2023.

RAMADHAN, Luthfi. TF-IDF Simplified: A short introduction to TF-IDF vectorizer. Towards Data Science, 20 jan. 2021. Disponível em: https://towardsdatascience.com/tf-idf-simplified-aba19d5f5530. Acesso em: 28 jun. 2023.

TEMPORAL, J. Como definir o número de clusters para o seu KMeans. Pizzadedados, 10 de abril de 2019. Disponível em: https://medium.com/pizzadedados/kmeans-e-metodo-do-cotovelo-94ded9fdf3a9. Acesso em: 28 jun. 2023.