Relatório de Análise de Dados Netflix

1. Introdução

Este relatório detalha a análise exploratória e a aplicação de técnicas de mineração de dados no vasto dataset de filmes e séries da Netflix. O objetivo primordial deste estudo foi desvendar padrões ocultos e agrupar conteúdos com características intrínsecas semelhantes, empregando o robusto algoritmo de K-Means Clustering. A relevância desta abordagem reside na capacidade de transformar grandes volumes de dados brutos em insights acionáveis, que podem otimizar a experiência do usuário, refinar estratégias de recomendação e aprimorar a organização do catálogo da plataforma.

2. Dataset e Origem

O dataset empregado nesta análise, denominado `netflix_titles.csv`, é uma compilação abrangente de informações sobre os títulos disponíveis na plataforma de streaming Netflix. Cada registro no dataset oferece uma riqueza de detalhes, incluindo o tipo de conteúdo (seja um filme ou uma série de TV), o título original, o nome do diretor, o elenco principal, o país de produção, a data em que o título foi adicionado ao catálogo, o ano de lançamento original, a classificação indicativa, a duração (para filmes) ou o número de temporadas (para séries), os gêneros aos quais o conteúdo está associado (`listed_in`), e uma descrição textual concisa. A proveniência deste dataset é o Kaggle, uma plataforma amplamente reconhecida e respeitada por disponibilizar conjuntos de dados públicos de alta qualidade, garantindo a confiabilidade e a integridade dos dados utilizados nesta pesquisa.

3. Metodologia

A metodologia adotada para esta análise foi estruturada em etapas sequenciais, garantindo a robustez e a validade dos resultados obtidos. Cada fase foi cuidadosamente planejada para otimizar a aplicação do algoritmo de clustering e maximizar a extração de insights significativos.

3.1. Pré-processamento de Dados

A qualidade dos dados é um pilar fundamental para qualquer análise de mineração de dados. Consequentemente, um rigoroso processo de pré-processamento foi implementado para tratar inconsistências e garantir a completude do dataset. Especificamente, as colunas `description` e `listed_in`, que são cruciais para a caracterização textual dos conteúdos, foram inspecionadas. Quaisquer valores nulos (`NaN`) presentes nessas colunas foram sistematicamente substituídos por strings vazias. Esta medida preventiva assegura que todos os registros, independentemente da presença de descrições ou gêneros, possam ser processados sem interrupções ou erros, evitando a exclusão de dados valiosos. Adicionalmente, para enriquecer a representação de cada título, as informações contidas nas colunas `description` e `listed_in` foram concatenadas em uma nova coluna sintética, denominada `combined_features`. Esta agregação permite que o algoritmo de clustering considere simultaneamente a narrativa textual e as categorias temáticas, proporcionando uma compreensão mais holística do conteúdo.

3.2. Vetorização de Texto (TF-IDF)

A natureza textual das `combined_features` exige uma transformação para um formato numérico que possa ser interpretado por algoritmos de aprendizado de máquina. Para este fim, a técnica de TF-IDF (Term Frequency-Inverse Document Frequency) foi empregada. O TF-IDF é uma medida estatística que reflete a importância de uma palavra em relação a um documento em um corpus. Ele atribui pesos mais altos a termos que são frequentes em um documento específico, mas raros em todo o conjunto de documentos, e vice-versa. Isso permite que palavras distintivas, que realmente caracterizam um conteúdo, recebam maior relevância. Para otimizar o processo e focar nos termos mais informativos, o vetorizador TF-IDF foi configurado para remover *stop words* (palavras comuns como 'o', 'a', 'e', 'de' que não agregam significado) na língua inglesa e para limitar o número máximo de features (termos) a 5000. Esta limitação ajuda a reduzir a dimensionalidade dos dados, mitigando o risco de *overfitting* e melhorando a eficiência computacional.

3.3. K-Means Clustering

Com os dados textuais devidamente vetorizados, o algoritmo K-Means Clustering foi aplicado à matriz TF-IDF resultante. O K-Means é um algoritmo de agrupamento não supervisionado que particiona n observações em k clusters, onde cada observação pertence ao cluster cujo centroide é o mais próximo. Para esta análise, optou-se por definir k (o número de clusters) como 5. Esta escolha inicial foi baseada em um equilíbrio entre a granularidade da análise e a interpretabilidade dos resultados, servindo como um ponto de partida eficaz para a exploração dos dados. Para garantir a robustez e a consistência dos resultados, o algoritmo foi configurado com `init='k-means++'`, um método de inicialização que seleciona os

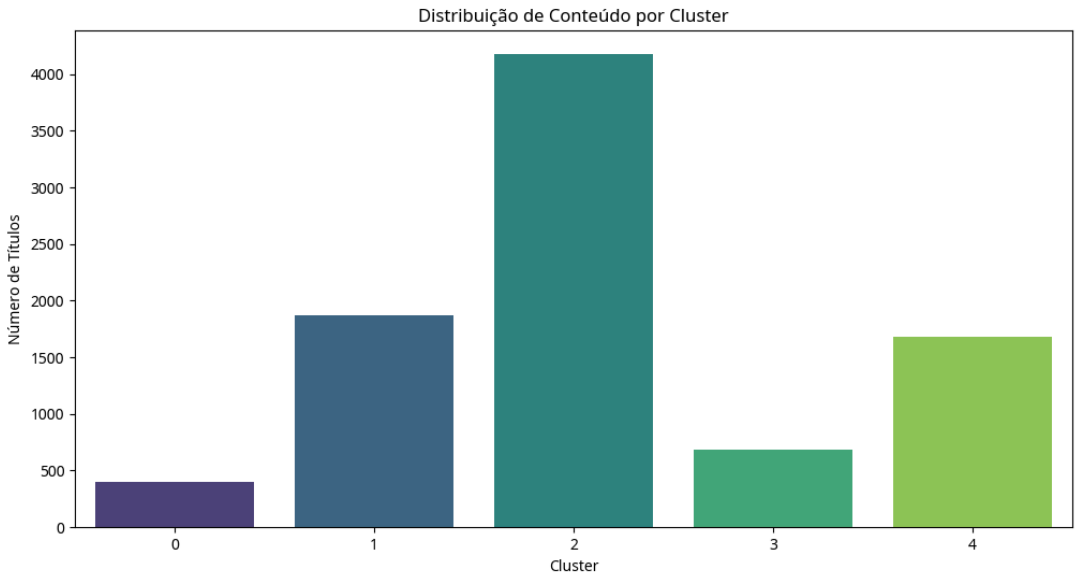
centroides iniciais de forma inteligente para acelerar a convergência e evitar mínimos locais subótimos. Além disso, foram realizadas 10 inicializações (`n_init=10`), e o algoritmo foi permitido a iterar no máximo 300 vezes (`max_iter=300`) para cada inicialização. A inclusão de um `random_state` fixo (42) é crucial para a reprodutibilidade dos resultados, garantindo que a execução do script sempre produza os mesmos clusters, facilitando a validação e a comparação de análises futuras. O resultado do K-Means é a atribuição de cada título a um cluster específico, armazenado na nova coluna `cluster` do DataFrame.

4. Resultados Obtidos

A execução do script de análise gerou uma série de resultados quantitativos e visuais que oferecem uma compreensão aprofundada da estrutura do catálogo da Netflix, conforme segmentado pelos clusters identificados.

4.1. Distribuição de Conteúdo por Cluster

O gráfico `cluster_distribution.png` fornece uma representação visual clara da alocação de títulos em cada um dos 5 clusters. A análise deste gráfico permite identificar a densidade de conteúdo em cada grupo, revelando se há clusters mais populosos ou mais esparsos. Essa distribuição é fundamental para entender a proporção de diferentes tipos de conteúdo que a Netflix oferece e como eles se agrupam naturalmente.



4.2. Principais Termos por Cluster

Uma das saídas mais reveladoras do K-Means é a identificação dos termos mais representativos para cada cluster. Ao analisar os 5 principais termos (palavras-chave) que

mais contribuem para a definição de cada grupo, é possível inferir a temática predominante e as características distintivas de cada cluster. Estes termos atuam como descritores semânticos, permitindo uma interpretação qualitativa dos agrupamentos:

- **Cluster 0:** `stand`, `comedy`, `comedian`, `special`, `comic` - Este cluster é fortemente associado a conteúdos de comédia stand-up e especiais de comediantes, indicando uma segmentação clara para este gênero.
- **Cluster 1:** `documentaries`, `tv`, `kids`, `documentary`, `series` - Este grupo parece englobar documentários e séries de TV, com uma possível inclinação para conteúdos infantis ou educativos, dada a presença do termo 'kids'.
- **Cluster 2:** `movies`, `dramas`, `international`, `comedies`, `independent` - Este cluster é mais abrangente, reunindo filmes de diversos gêneros como dramas, comédias, e filmes independentes, com uma forte presença de produções internacionais.
- **Cluster 3:** `children`, `family`, `movies`, `comedies`, `save` - Claramente focado em conteúdo para crianças e família, incluindo filmes e comédias. O termo 'save' pode sugerir temas de aventura ou heroísmo comuns nesse tipo de produção.
- **Cluster 4:** `tv`, `shows`, `international`, `crime`, `dramas` - Este cluster é dominado por séries de TV, com uma ênfase em produções internacionais, dramas e, notavelmente, conteúdos relacionados a crimes, indicando séries policiais ou de suspense.

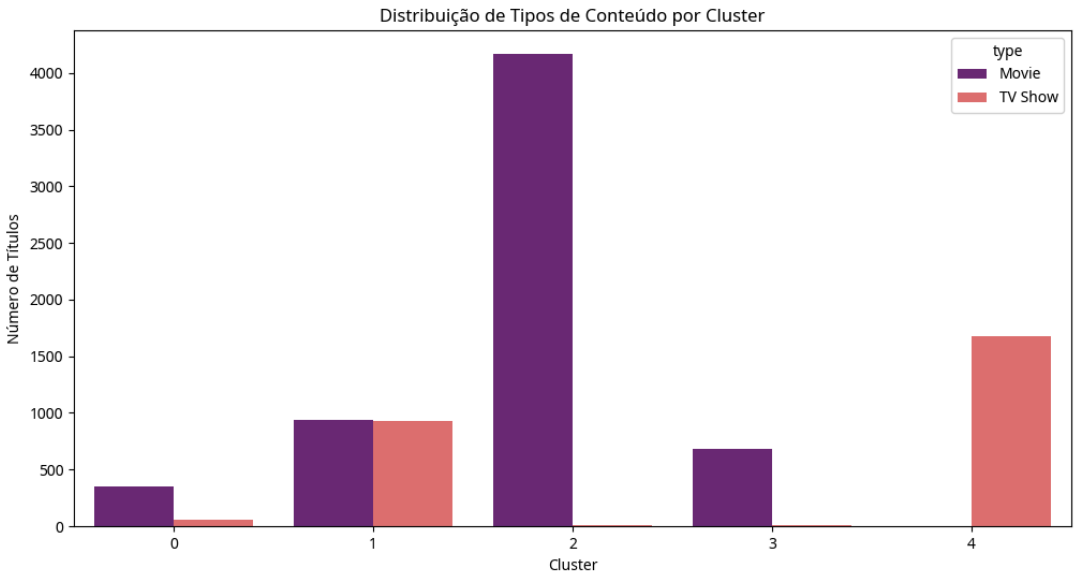
4.3. Exemplos de Títulos por Cluster

Para validar a coerência dos agrupamentos e facilitar a interpretação, foram selecionados aleatoriamente alguns títulos de exemplo de cada cluster. A inspeção desses títulos permite uma verificação empírica da qualidade do agrupamento e reforça a compreensão do perfil de cada cluster:

- **Cluster 0:** ["Wanda Sykes: Not Normal", "Jeff Dunham: Minding the Monsters", "Ladies Up", "Beyond Stranger Things", "Brian Regan: Nunchucks and Flamethrowers"]
- **Cluster 1:** ["Beyond All Boundaries", "Perfect Bid: The Contestant Who Knew Too Much", "Nazi Mega Weapons", "Mobile Suit Gundam UC", "Bread Barbershop"]
- **Cluster 2:** ["Posesif", "A Christmas Special: Miraculous: Tales of Ladybug & Cat Noir", "Jefe", "Hisss", "I'll See You in My Dreams"]
- **Cluster 3:** ["Pets United", "Krish Trish and Baltiboy: Part II", "Coraline", "LEGO Marvel Spider-Man: Vexed by Venom", "Cobra Kai - The Afterparty"]
- **Cluster 4:** ["Dueños del paraíso", "The Parisian Agency: Exclusive Properties", "It's Okay to Not Be Okay", "The Snitch Cartel: Origins", "Republic of Doyle"]

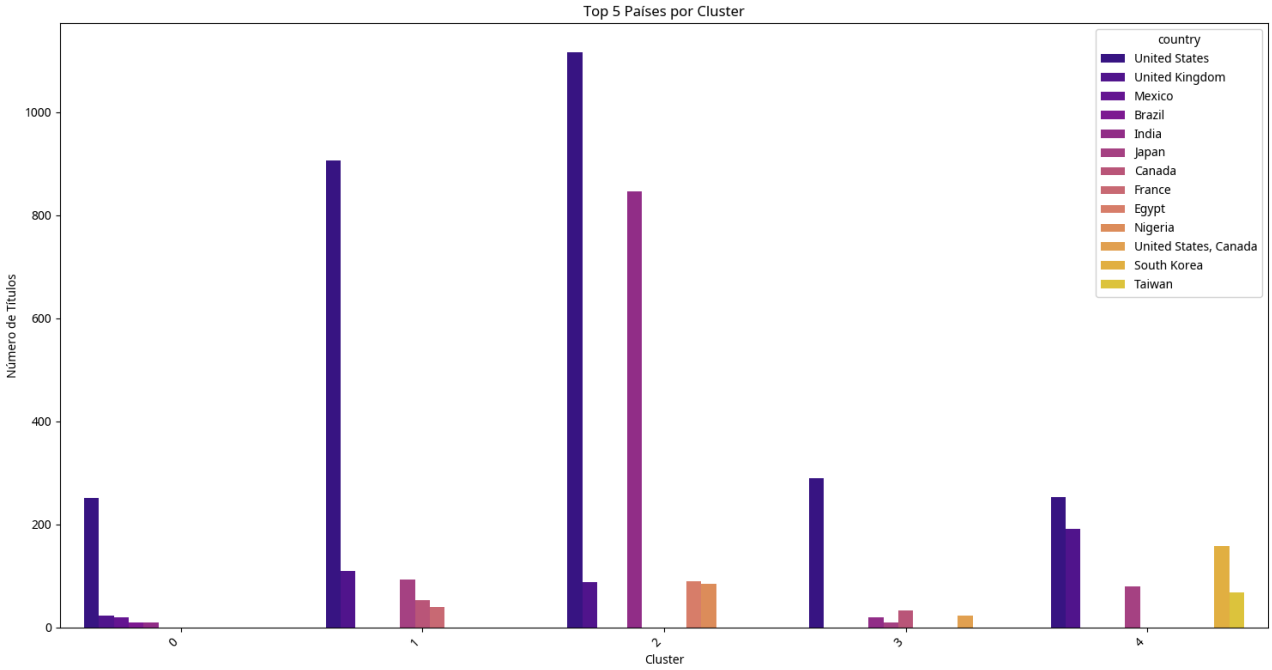
4.4. Distribuição de Tipos de Conteúdo por Cluster

O gráfico `type_distribution_by_cluster.png` oferece uma análise mais granular, discriminando a proporção de filmes e séries de TV dentro de cada cluster. Este visual é crucial para entender se os clusters são homogêneos em relação ao tipo de conteúdo ou se representam uma mistura. Por exemplo, um cluster predominantemente composto por filmes de comédia e outro por séries de drama internacional indicaria uma segmentação eficaz baseada não apenas no gênero, mas também no formato do conteúdo.



4.5. Top 5 Países por Cluster

O gráfico `country_distribution_by_cluster.png` explora a dimensão geográfica dos clusters, apresentando os 5 principais países de produção associados a cada grupo. Esta visualização pode revelar tendências regionais na produção de certos tipos de conteúdo. Por exemplo, um cluster dominado por filmes de Bollywood (Índia) ou por dramas coreanos (Coreia do Sul) evidenciaria a influência cultural e a especialização de determinados países na produção de conteúdos específicos que se agrupam de forma coesa.



5. Conclusão

A aplicação do K-Means Clustering ao dataset da Netflix demonstrou ser uma ferramenta poderosa para a identificação de grupos distintos de conteúdo. A análise dos termos-chave que definem cada cluster, juntamente com a inspeção de títulos de exemplo, confirmou a coerência e a interpretabilidade dos agrupamentos. As visualizações da distribuição por tipo de conteúdo (filme/série) e por país de produção enriqueceram ainda mais nossa compreensão, revelando nuances na composição de cada cluster. Este estudo não apenas valida a eficácia das técnicas de mineração de dados para organizar e categorizar grandes volumes de informações, mas também sublinha o potencial de tais análises para aplicações práticas. No contexto da Netflix, os insights derivados poderiam ser diretamente aplicados para aprimorar os sistemas de recomendação, otimizar a curadoria de conteúdo, personalizar a experiência do usuário e informar estratégias de aquisição de novos títulos. A capacidade de identificar e compreender esses segmentos de conteúdo é um diferencial competitivo que pode impulsionar o engajamento do usuário e a satisfação geral com a plataforma.

6. Referências

- Dataset Netflix: https://www.kaggle.com/datasets/shivamb/netflix-shows
- Scikit-learn (K-Means): https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html
- Scikit-learn (TF-IDF): https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html

- Matplotlib: https://matplotlib.org/
- Seaborn: https://seaborn.pydata.org/