UNIVERSIDADE FEDERAL DO CEARÁ

CAMPUS RUSSAS

CURSO DE GRADUAÇÃO EM ENGENHARIA DE SOFTWARE

JOÃO CARLOS ALVES BORGES

APLICAÇÃO DO ALGORITMO K-MEANS PARA GERAÇÃO DE UM SISTEMA DE RECOMENDAÇÃO DE PRODUÇÕES DE UMA PLATAFORMA DE STREAMING

RUSSAS

2022

JOÃO CARLOS ALVES BORGES

APLICAÇÃO DO ALGORITMO K-MEANS PARA GERAÇÃO DE UM SISTEMA DE RECOMENDAÇÃO DE PRODUÇÕES DE UMA PLATAFORMA DE STREAMING

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Engenharia De Software do Campus Russas da Universidade Federal do Ceará, como requisito parcial à obtenção do grau de bacharel em Engenharia De Software.

Orientador: Prof. Dra. Tatiane Fernandes Figueiredo

RUSSAS

2022

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Sistema de Bibliotecas
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

B732a Borges, João Carlos Alves.
Aplicação do algoritmo k-means para geração de um sistema de recomendação de produções de uma plataforma de streaming / João Carlos Alves Borges. – 2022.
27 f.

Trabalho de Conclusão de Curso (graduação) – Universidade Federal do Ceará, Campus de Russas, Curso de Engenharia de Software, Russas, 2022.
Orientação: Prof. Dr. Tatiane Fernandes Figueiredo.

1. Agrupamento de dados. 2. Amazon Prime Video. 3. Mineração de texto. 4. Sistema de recomendação. I. Título.

CDD 005.1

JOÃO CARLOS ALVES BORGES

APLICAÇÃO DO ALGORITMO K-MEANS PARA GERAÇÃO DE UM SISTEMA DE RECOMENDAÇÃO DE PRODUÇÕES DE UMA PLATAFORMA DE STREAMING

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Engenharia De Software do Campus Russas da Universidade Federal do Ceará, como requisito parcial à obtenção do grau de bacharel em Engenharia De Software.

Aprovada em: 15 de Dezembro de 2022

BANCA EXAMINADORA

Prof. Dra. Tatiane Fernandes Figueiredo (Orientador)
Universidade Federal do Ceará (UFC)

Prof. Dr. Bonfim Amaro Júnior
Universidade Federal do Ceará (UFC)

Prof. Dr. Marcio Costa Santos
Universidade Federal do Ceará (UFC)

AGRADECIMENTOS

Gostaria de agradecer a todos os meus familiares que me apoiaram durante essa jornada, especialmente à minha mãe Antonia Alves Monteiro, que me ligava todos os dias para saber como eu estava, mal sabia ela o combustível que me dava para resolver os desafios que apareciam ao longo da graduação.

Agradeço ao meu pai João Simeão Borges, por ser um porto seguro sempre que precisei. Agradeço às minhas irmãs Michelle, Mirla e Mikaely por serem meu abrigo contra tempestades e às vezes a própria tempestade. Agradeço aos meus sobrinhos Bernardo, Neto, Nicolas, Milena e meu irmão Josué por me fazerem querer ser alguém que sirva de exemplo para eles.

Agradeço à minha orientadora Tatiane Fernandes por ter me dado a minha primeira oportunidade em um projeto na UFC, o TILAPIA, que foi uma experiência enriquecedora sobre empreendedorismo, e por continuar acreditando no meu potencial mesmo quando nem eu acreditava.

Agradeço aos professores Pablo Soares e Márcio Costa pela oportunidade de fazer parte do projeto CAGECE. Agradeço a todos os professores pelo conhecimento que me foi passado, principalmente aos professores Bonfim Amaro, Pablo Soares, Osvaldo Mesquita e Tatiane Fernandes, por terem me proporcionado as melhores aulas da graduação.

Agradeço a todos os meus colegas que fizeram parte desta jornada. Agradeço a minha madrasta por me acolher como um filho e agradeço ao meu mestre de muay thai, por me ajudar a controlar a raiva e ter disciplina.

“Nunca deixe ninguém dizer que você não pode ser o que quiser.”

(Rio Morales)

RESUMO

Com a popularização do uso da internet nos últimos anos, o mercado cinematográfico precisou se reinventar, dando origem à criação de novos produtos tecnológicos, sendo o principal deles as plataformas de streaming. O tempo de vida dos assinantes de plataformas de streaming está diretamente relacionado à quantidade de vezes em que os mesmos assistiram um vídeo e gostaram. Se os assinantes não conseguem encontrar filmes e séries que os interessem e os envolvam, eles tendem a abandonar a plataforma. Visto isso, essa monografia apresenta o desenvolvimento de um sistema de recomendação de filmes e séries do catálogo do serviço de *streaming Amazon Prime Video*, utilizando o algoritmo *K-means* para realizar o agrupamento das produções de acordo com a similaridade das sinopses. Após o agrupamento dos títulos, realizou-se uma avaliação manual dos 37 grupos criados, levando em consideração a similaridade dos gêneros dos filmes e séries que estavam no mesmo *cluster* e suas sinopses. Os clusters em sua grande maioria continham produções com gêneros e sinopses semelhantes, pode se citar por exemplo o agrupamento que contém todos os filmes da saga do anime *Evangelion*, tendo apenas um grupo onde se concentrou os títulos que não tem uma similaridade de gênero e sinopses entre se.

Palavras-chave: Agrupamento de dados. Amazon Prime Video. Mineração de texto. Sistema de recomendação.

ABSTRACT

With the popularization of internet use in recent years, the film industry needed to reinvent itself, giving rise to the creation of new technological products, the main one being streaming platforms. The lifetime of subscribers to streaming platforms is directly related to the number of occasions they have watched a video and enjoyed it. If subscribers cannot find movies and series that interest and engage them, they tend to abandon the platform. Given this, this monograph presents the development of a recommendation system for movies and series from the Amazon Prime Video streaming service, using the K-means algorithm to group the productions according to the similarity of their synopses. After grouping the titles, a manual evaluation of the 37 clusters created was performed, based on the similarity of the genres of the movies and series that were in the same cluster and their synopses. The clusters in its great majority contained productions with similar genres and synopses, it can be cited for example the cluster that contains all the films of the Evangelion anime saga, with only one group where it was concentrated the titles that do not have a similarity of genre and synopses between them.

Keywords: Amazon Prime Video. Data clustering. Recommendation system. Text mining.

LISTA DE ABREVIATURAS E SIGLAS

LSA *Latent Semantic Analysis*

SRI Sistema de Recuperação de Informações

SUMÁRIO

1 INTRODUÇÃO

Com a popularização do uso da internet nos últimos anos, o mercado cinematográfico precisou se reinventar, dando origem a criação de novos produtos tecnológicos, sendo o principal deles as plataformas de *streaming*. Após o sucesso da plataforma *Netflix*, lançada no Brasil em 2011, muitos canais de TV e produtoras também lançaram suas plataformas, como exemplo pode-se citar: a *Amazon Prime*, lançada no Brasil em 2016, enquanto a *Disney Plus* e *HBO Max* foram ambas lançadas no Brasil em 2020.

Com o objetivo de agregar mais opções de entreterimento aos seus usuários e assim se tornar mais competitiva no mercado, as plataformas de *streaming* buscam disponibilizar um número gradioso de produções através de parcerias com diversos estúdios. Porém, se por um lado disponibilizar muitas produções pode aumentar a diversidade de usuários interessados em se tornar assinantes de uma plataforma de *streaming*, um grande número de produções também dificulta a busca e escolha de um usuário. Para resolver este problema e melhorar a experiência de seus usuários, as plataformas de *streaming* tem investido em sistemas de recomendação baseado em produções já assistidas por um usuário.

Embora existam sistemas de recomendação em todas as principais plataformas de *streaming* no mercado, a literatura ainda apresenta poucos trabalhos sobre este tema. Acredita-se que esta lacuna se deve principalmente pela ausência de bases de dados públicas, pois a grande maioria das plataformas de *streaming* ainda são muito recentes. Por estar presente no mercado a mais tempo, os poucos artigos científicos existentes apresentam resultados utilizando em suma dados da plataforma *Netflix*. Desta forma, Nakka e Prasad (2020) propuseram um sistema de recomendações para a plataforma *Netflix* baseada em métrica de confiança entre os usuários, enquanto Bennett e Lanning (2007) descrevem os resultados obtidos no desafio da *Netflix* disponibilizado no site *Kaggle*, onde o problema proposto era realizar previsões para de conjunto de qualificação de filmes em uma base de dados diponibilizada pela própria empresa *Netflix*. Para tentar contornar a ausência de banco de dados disponíveis, Remigio, Bobadilla e Gutiérrez (2018) criam uma base de dados para desenvolvimento de sistemas de recomendações, utilizando dados de artigos científicos sobre ciência da computação e inteligência artificial.

Buscando estudar e propor soluções utilizando dados de outras plataformas de *sreaming*, esta monografia apresenta os resultados obtidos da criação de um sistema de recomendação utilizando mineração de textos de sipnoses de filmes e seriados disponíveis na plataforma *Amazon Prime*. Para tal, após a limpeza e tratamento da base de dados, aplicou-se o algoritmo de

clusterização *k-means* separando os dados das produções pelo seus gêneros. Para validação do sistema de recomendação, foi realizada uma análise dos resultados obtidos para o gênero animação, onde pode-se concluir a eficácia da metodologia apresentada.

A estrutura deste trabalho encontra-se da seguinte forma: No Capítulo 2 é apresentado o objetivo geral e os específicos; no Capítulo 3 são apresentados os tópicos chave da pesquisa; no Capítulo 4 são apresentados trabalhos encontrados na literatura que são similares a este; no Capítulo 5 é apresentada metodologia utilizada para realização desta pesquisa, assim como os resultados obtidos. Por fim, o Capítulo 6 apresenta as conclusões gerais e trabalhos futuros.

2 OBJETIVOS

2.1 Objetivo geral

Criar um sistema de recomendação baseado em filmes e series disponíveis na plataforma de streaming *Amazon Prime Video*.

2.2 Objetivo específicos

- Obter uma base de dados relacionada a temática;
- Aplicar as fases de pré-processamento na base de dados, objetivando a realização de melhorias na mesma;
- Aplicar o algoritmo *k-means* para geração de agrupamento de filmes e séries por similaridade de sinopses;
- Analisar a qualidade dos resultados obtidos.

3 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo serão apresentados os conceitos básicos necessários para realização desta pesquisa. Na Seção 3.1 é apresentado uma descrição introdutória sobre a área tema deste trabalho de conclusão de curso. Na seção 3.2 é descrito o processo para mineração de texto.

3.1 Mineração de Dados

De acordo com Galvão e Marin (2009), a mineração de dados é uma das alternativas mais eficazes para extrair conhecimento a partir de grandes volumes de dados, descobrindo relações ocultas, padrões e gerando regras para predizer e correlacionar dados, que podem ajudar instituições em suas tomadas de decisões.

3.2 Mineração de texto

A Mineração de Textos faz parte das técnicas de processo de descoberta de conhecimento da Mineração de Dados, sendo utilizado para tal, técnicas de análise e extração de dados a partir de textos completos, frases ou apenas palavras. O garimpo desses textos envolve o uso de algoritmos computacionais responsáveis por identificar informações úteis, que normalmente não poderiam ser encontradas utilizando métodos tradicionais de consulta. De forma geral, as etapas do processo de mineração de textos podem ser descritas conforme apresentado por Morais e Ambrósio (2007):

- seleção de documentos;
- definição do tipo de abordagem dos dados (como por exemplo, análise semântica ou estatística);
- preparação dos dados;
- indexação e normalização;
- cálculo da relevância dos termos;
- seleção dos termos e pós-processamento.

A seguir é apresentado uma breve descrição sobre cada uma das etapas mencionadas.

3.2.1 Seleção dos documentos e dados

Nessa etapa é selecionado os documentos e dados considerados interessantes e que possam gerar ou inferir alguma *informação relevante*, conhecida na literatura como *insight*. Essa é uma etapa crítica do processo de mineração, pois, documentos selecionados de forma precipitada ou imprecisa, podem acabar influenciando de forma negativa nos resultados obtidos pelos algoritmos de mineração.

3.2.2 Definição dos tipos de abordagens de dados

Pode-se dividir as abordagens de análise de dados textuais em dois tipos: *semântica* e *estátistica*. Na análise *semântica*, procura-se identificar qual a importância das palavras dentro da estrutura dos textos. Por ser fundamentada em técnicas de processamento de linguagem natural, é necessário a aplicação de conhecimentos morfológicos, sintáticos, semânticos e pragmáticos do discurso. Dentre os métodos para extração e representação de significado semântico de palavras, destaca-se o método de *Análise Semântica Latente*, do inglês, *Latent Semantic Analysis* (LSA), onde seu modelo de indexação semântica é baseado na coocorrência de palavras em um texto, partindo da suposição de que palavras que tendem a ocorrer juntas dentro de um mesmo documento, representam similaridade semântica.

Na análise *estátistica*, a importância das palavras é dada pela quantidade de vezes que a mesma se repete ao longo do texto. Nessa abordagem é utilizado um processo de aprendizagem estatística a partir de dados, que normalmente faz o uso das seguintes etapas: codificação dos dados, estimativas dos dados e modelos de representação de documentos. Dentre os modelo de representação de documentos mais utilizados, destaca-se a abordagem de *bag of words* (saco de palavras), onde se ignora a ordem com que as palavras aparecem nos textos, assim como qualquer informação de pontuação ou de estrutura, guardando apenas a quantidade de vezes que cada palavra se repete .

3.2.3 Preparação dos Dados

Na preparação dos dados, busca-se selecionar um núcleo de textos que melhor representa o conteúdo dos textos obtidos na seleção de documentos. Os objetivos principais desta etapa são: obter uma redução dimensional dos textos e identificar similaridades em relação à estrutura em que se encontram as palavras ou do significado das mesmas nos textos originais.

3.2.4 Indexação e Normalização

Na etapa de indexação e normalização dos textos, as características dos documentos são identificadas e adicionadas ao Sistema de Recuperação de Informações (SRI), que é um sistema desenvolvido para indexar e recuperar documentos do tipo textual. O índice de um documento é composto por um conjunto de termos correspondentes ao vocabulário da área, nesse caso utilizaremos o vocabulário da área de companhias de água e esgoto. Em mineração de textos o processo de indexação é automático, a seguir é apresentado as suas principais fases.

1. **Identificação de termos:** nesta fase ocorre a identificação dos termos contidos nos textos, sejam eles simples ou compostos. Na etapa de identificação também são eliminados símbolos e espaços múltiplos, também sendo realizada a removeção de caracteres de formatação de arquivos, realização de correção ortográfica, assim como conversão de letras maiúsculas em minúsculas e padronização de datas.
2. **Remoção de *stopwords*:** nesta etapa são removidas as palavras irrelevantes para a análise dos textos, por não traduzirem a sua essência. Normalmente estas palavras são preposições, pronomes, artigos, advérbios, e outras classes de palavras auxiliares. Além dessas, também pode-se remover palavras que aparecem com muita frequência em todos os textos, pois as mesmas acabam não sendo capazes de inferir nada sobre um documento;
3. **Normalização Morfológica:** aqui eliminamos as variações morfológicas das palavras. Primeiramente, identificamos o radical da palavra, então os seu prefixos e sufixos são retirados, restando apenas seu radical que é adicionado ao índice do documento. Também é comumente removido as características de gênero e número de grau das palavras.

3.2.5 Cálculo da Relevância e Seleção de Termos

Nessa fase calculamos a relevância dos termos em cada documento. Na literatura, comumente utiliza-se o cálculo de frequência relativa. Esta técnica considera a quantidade de vezes que uma palavra ocorre em um documento para calcular a relevância de um termo. A seguir é apresentada a fórmula matemática utilizada para calcular a frequência relativa(F_{rel}) de uma palavra x em um documento qualquer, que é dada pela sua frequência absoluta (F_{abs}) dividida pelo número total de palavras no mesmo documento (N).

$$f(x) = f_{abs}(x)/N$$

3.2.6 Pós-processamento

Nessa etapa é executado os algoritmos que recebem como entrada os dados obtidos pelos procedimentos descritos anteriormente, assim como os documentos tratados da mineração textual. Como exemplos de algoritmos utilizados na literatura para classificação de dados temos o *Naive Bayes* e *Decision Tree* (Árvore de Decisão). Para clusterização de dados, pode-se citar o *k-Means*, algoritmo utilizado nesta monografia.

Nunes (2016) define *clustering* com uma partição de um conjunto de dados em subconjuntos, onde os elementos de cada *cluster* apresentam alguma semelhança entre si e se diferem dos elementos que se encontram nos outros *clusters*. O *K-means* aplica uma abordagem gananciosa para encontrar o *clustering* que minimiza a soma dos erros quadrados, que serve para avaliar a dispersão dos elementos dentro de um subconjunto.

Inicialmente o algoritmo distribui aleatoriamente os pontos do conjunto dados D em k *clusters* e calcula os centróides através da média dos pontos do *cluster* C_i para todo $i = \{1, ..., k\}$. Em seguida aplica-se duas fases interativamente: atribuição de *clusters*, onde cada ponto $x_j \in D$ é associado a um *cluster* C_i que tem o centróide z_i mais próximo do ponto x_j; e a atualização dos centróides, onde atualiza-se cada centroides z_i através da média de todos os pontos que se encontram no *cluster* C_i. Essas duas fases se repetem até que os centróides não se modifiquem durante uma interação.

Um decisão importante a ser tomada quando se utiliza o algoritmo $k-means$ é o número de cluster a serem criados. Dentre as técnicas utilizadas para definir o número de cluster, destaca-se o método do cotovelo que é um método para determinar o número ideal de k clusters. Em suma, esse método examina a variação dos dados em termos do número de clusters. Desta forma, o valor ideal de k é aquele que tem a menor "*Within Sum Of Squares*"que é a soma de todos os dados até o centroide do cluster que ele pertencem, de forma que também possuia o menor número de clusters possível. Dá-se a esse conceito o nome de curva de cotovelo, pois não há tanta flutuação quanto haveria no ponto “cotovelo”. O número ideal de k-clusters seria exatamente onde o cotovelo estaria (ROUSSEEUW, 1987).

4 TRABALHOS RELACIONADOS

Nesta seção está descrito os trabalhos da literatura mais relevantes para a contextualização do problema proposto nesta monografia.

4.1 Criação de um sistema de recomendação utilizando dados da plataforma de *streaming Netflix*

Nakka *et al.* (2020) propuseram um sistema de recomendações para a plataforma de *streaming Netflix* baseada em uma métrica de confiança entre os usuários. De acordo com os autores oferecer recomendações baseado apenas em métricas de similaridade entre os usuários podem gerar recomendações de baixa qualidade, uma vez que os dados disponíveis para análise são escassos.

A métrica de confiança definida pelos autores considera a quantidade de confiança que um usuário i tem sobre um outro usuário i, que é obtida pelo número de produções avaliadas em comum por ambos usuários, dividido pelo número de produções avaliado pelo usuário j. Inicialmente, Nakka *et al.* (2020)) realizaram uma análise da base de dados disponível com o objetivo compreender as associações que existiam entre os dados. Após a análise, foi desenvolvido o conjunto de regras principais do algoritmo de recomendação. Por fim, para validação e comparação dos resultados obtidos, os autores dividiram os dados disponíveis em lotes, e executaram além do algoritmo de recomendação proposto, um algoritmo de recomendação por filtragem existente na literatura.

O valor do erro quadrático médio para todos os lotes de dados do algoritmo utilizando a métrica de confiança proposta, foi menor que o valor obtido pelo algoritmo por filtragem existente, o que mostra que as classes de saída previstas correspondem às classes de saída real em maior extensão. Por outro lado, o tempo para se calcular a similaridade entre os usuários se mostrou um ponto crítico, visto que, levou se um tempo considerável para se realizar os cálculos.

4.2 Os desafios existentes na criação de sistemas de recomendação utilizando dados da plataforma de *streaming* Netflix

De acordo com Bennett *et al.* (2007), o tempo de vida dos assinantes de plataformas de *streaming* está diretamente relacionado ao número de produções que o mesmo assiste e avalia positivamente. Se os assinantes não conseguem encontrar filmes e series que os interessem e

os envolvam, eles tendem a abandonar a plataforma. Desta forma, a *Netflix* lançou o desafio de desenvolver um sistema de recomendação com uma precisão melhor que a do sistema em uso pela empresa - o *Cinematch*. Neste artigo, Bennett *et al.* (2007) descrevem os dificuldades encontrados durante a sua participação neste desafio e analisam seus resultados obtidos e dos demais participantes.

Neste desafio, inicialmente, os competidores deveriam analisar um grande conjunto de dados de classificação de produções disponíveis na plataforma *Netflix*, tendo como objetivo final o desenvolvimento de sistema de recomendação. Os resultados obtidos das previsões realizadas por cada sistema criado foi analisado utilizando o *Root Mean Squared Error*, sendo esta a métrica definida para criação do *ranking* de participantes.

Até julho de 2007, mais de 20 mil equipes se inscreveram neste desafio. Dessas, 2 mil enviaram sistemas de recomendação válidos, atualmente a mais de 13 mil envios. Somente cerca de 650 equipes excederam a precisão do *Cinematch*, e apenas 90 equipes ultrapassaram uma melhoria de precisão de 5%. O sistema de recomendação com maior precisão excedeu o *Cinematch* em aproximadamente 8%. Porém, o objetivo do desafio era exceder ao menos 10%.

4.3 Criação uma base dados pública para estudo e criação de sistemas de recomendação

Ortega *et al.* (2018) criam uma base de dados para desenvolvimento de sistemas de recomendações, utilizando para tal artigos científicos sobre ciência da computação e inteligência artificial e também apresentaram um sistema de recomedação para a base criada utilizando uma abordagem de recomendação filtragem colaborativa.

Para criação da base, os autores utilizaram o *scopus* como fonte de dados. De posse da base de dados criada, Ortega *et al.* (2018) geraram um conjunto de dados denominado SD4AI (Scientific Documentation for Artificial Intelligence) com a estrutura necessária para realizar filtragens colaborativas. Para tal, foi utilizado o framework CF4J (*Collaborative Filtering for Java*) para a aplicação dos filtros, sendo executados 4 métodos, correlação de Pearson, cosseno, medidas de similaridade baseadas na memória atual e o PMF (*Probabilistic Matrix Factorization*). De todos os métodos, o PMF foi a filtragem que obteve melhor resultado.

5 METODOLOGIA

Este capítulo apresenta a metodologia proposta nesta monografia para geração de um sistema de recomendação. Todas as etapas foram codificadas utilizando a linguagem de programação Python, versão 3.8.16. Os códigos foram executados na plataforma *Google Collaboratory*, serviço de *Jupyter Notebook* em nuvem da Google, em um máquina com 12GB de Memória RAM e processador Intel(R) Xeon(R) CPU @ 2.20GHz com arquitetura x86_64. A seguir é descrito detalhadamente cada um das etapas realizadas.

5.1 Compreensão dos dados

A base de dados utilizada neste trabalho foi obtida através do Kaggle, que é uma comunidade online de cientistas de dados e entusiastas do aprendizado de máquina. A base consiste num conjunto de registros contendo informações sobre filmes e séries do serviço de streaming Prime Vídeo, disponibilizados nos Estados Unidos em maio de 2022. A base de dados contém 9871 registros, sendo 8514 registros sobre filmes e 1357 sobre séries. A Tabela 1 mostra as informações sobre as colunas.

Tabela 1 – Informações sobre os dados disponíveis na base de dados estudada.

Nome	Descrição	Tipo de dado
id	Identificador do filme/série no JustWatch.	String
title	Nome do filme/série.	String
type	Informa se o vídeo é de uma série ou é um filme.	String
description	Sinopse do filme/série.	String
release_year	Ano de lançamento do filme/série.	Int
age_certification	Idade indicativa do filme/série.	String
runtime	Duração do filme/série em minutos.	Int
genres	Gêneros em que o filme/série estão classificados.	String
production_countries	Países em que o filme/série foi filmado.	String
seasons	Quantidade de temporadas de uma série	Float
imdb_id	Identificador do filme/série no IMDB.	String
imdb_score	Avaliação do filme/série no IMDB.	Float
imdb_votes	Quantidade de votos do filme/série no IMDB.	Float
tmdb_popularity	Popularidade do filme/série no TMDB	Float
tmdb_score	Avaliação do filme/série no TMDB	Float

5.2 Preparação dos dados

Nesta fase foi realizado uma análise inicial na base de dados, com o objetivo de determinar quais dados seriam utilizados pelo algoritmo de clusterização. Portanto, definiu-se que a priori seria utilizado apenas a coluna *description* com o objetivo de clusterizar produções que contenham uma sinopse semelhante.

Na etapa de limpeza da base, foram removidos registros com valores faltantes. Logo após, foi mapeado a lista de gêneros das produções para colunas, adicionando o valor número um às colunas em que o vídeo tinha o gênero, e o valor númerico zero nas que ele não se enquadrava. Após a etapa de limpeza e mapeamento, foi realizada a etapa de tratamento da base de dados. Por se tratar de uma base textual, foi aplicado a metodologia para mineração de textos como mencionado no Capítulo 3.

Inicialmente, efetuou-se uma transformação do texto para conter apenas letras minúsculas. Depois foram removidos todos os caracteres especiais e numéricos, assim como efetuado uma transformação das palavras para o padrão de representação de texto Unicode, que faz parte da ISO 10646, e consegue representar mais de 143 mil caracteres. Também foram removidas as principais *stop words* da língua inglesa. Por fim, foi efetuado uma transformação para manter apenas os radicais das palavras.

5.3 Modelagem

Como a clusterização realizada neste trabalho é apenas com dados textuais, foi utilizado o algoritmo *MiniBatchKMeans* da biblioteca *sklearn.cluster* da linguagem *Python* para realizar o agrupamento dos dados. O algoritmo *MiniBatchKMeans* recebe como parâmetro um valor k que determina a quantidade de grupos e um estado para definir o determinismo do algoritmo. Após a instanciação da classe, executou-se o método *fit*, para a criação do modelo de clusterização, a função *fit* recebe como parâmetro uma matriz, onde cada coluna representa um radical das palavras obtidas após a mineração de texto de todas as descrições, e as linhas representam a quantidade de vezes que cada palavra apareceu em cada sinopse. Com o modelo treinado, realizou-se a predição dos clusters para as descrições dos filmes e séries.

Para a decisão do valor de k, inicialmente foi aplicado o método do cotovelo. Porém, não foi possível obter bons resultados. Por ser uma base de dados textual, após a criação do *Bag of Words*, houve a criação de muitas colunas, resultando em alta dimensionalidade da base de

dados. Neste tipo de situação o método do cotovelo não é indicado. Assim, como segunda forma de analisar o melhor valor para k, realizou-se um teste empírico iniciando o valor de k igual a 2 e incrementando-o até obter um conjunto de cluster que contivesse as produções da base de dados de forma distribuída.

5.4 Análise dos resultados obtidos

Por ser uma agrupamento de dados textuais, a análise dos resultados foi realizada de forma manual, verificando se os gêneros subgêneros das produções que pertencem ao mesmo grupo são semelhantes, e se, os filmes e séries do cluster têm histórias parecidas. Por conta da grande quantidade de filmes e séries, foram analisadas apenas as produções do gênero animação, por conter uma dimensionalidade menor para se realizar o estudo. Os 434 filmes e séries do gênero de animação foram agrupados em 37 clusters. A Tabela 2 mostra a quantidade de cluster gerados possuindo a a mesma quantidade de produções. A seguir é apresentado uma análise mais precisa de alguns dos cluster gerados.

Tabela 2 – Quantidade de cluster gerados possuindo a a mesma quantidade de produções.

Quantidade de grupos	Quantidade de produções
5	2
8	3
3	4
3	5
1	6
2	7
4	8
3	9
1	10
1	12
1	13
1	14
1	21
1	27
1	39
1	158

5.4.1 Análise de um cluster com 5 filmes ou séries

Todos os filmes e séries deste cluster possuem comédia como subgenêro, sendo a maioria no estilo anime ou muito semelhantes a este estilo. A Tabela 3 apresenta os títulos e sinopses dos filmes agrupados. Os termos utilizados pelo algoritmo *k-means* para realizar o agrupamento foram destacados em negrito.

Tabela 3 – Título e sinopse dos 5 filmes clusterizados.

Nome do Filme/Série	Sinopse
Fritz the Cat	A swinging hypocritical **college student** cat raises hell in a satirical vision of the ...
Izzie's Way Home	A constantly picked on aquarium fish escapes her yacht home unaware of the dangers that await her in the open **ocean** with the help of other misfit sea creatures she learns not only how to brave the perils of the deep but how to be true to herself.
Grand Blue	A **college student** joins the local **diving club** after meeting some rowdy upperclassmen new adventures in booze and the **ocean** await.
DIVE!!	The series revolves around the Mizuki **diving club** mdc which is on the verge of closing down after having financial troubles the club's new coach persuades the club's parent company to stay open on one condition that the club sends one of its members to next year's olympics as part of japan's olympic team.
Drop Kick on My Devil!!	Jashin chan a devil from hell was abruptly summoned to the human world by yurine hanazono a stoic **college student** who lives in a run down apartment in jinbocho they're forced to become roommates since yurine doesn't know how to send jashin chan back but according to Jashin chan she could return by killing yurine so she takes action ?! a viperous roomie comedy that keeps you on your toes!

5.4.2 Análise de um cluster com 4 filmes ou séries

Todos os filmes e séries deste cluster possuem ação e fantasia como subgenêro, sendo todos com o tema principal de luta. A Tabela 4 apresenta os títulos e sinopses dos filmes agrupados. Os termos utilizadas para realizar o agrupamento foram destacados em negrito.

Tabela 4 – Título e sinopse dos 4 filmes clusterizados.

Nome do Filme/Série	Sinopse
Saint Seiya: Legend of Crimson Youth	Athena receives the visit of phoebus abel her older brother and **god** of the corona he informs her that he has come to **destroy** humanity as punishment for their corruption **just** as it was done in ancient times he dismisses seiya and the bronze saints as she will now be guarded by abel's three corona saints atlas of carina jaow of lynx and berenike of coma berenices and the five resurrected gold saints who died in the sanctuary battle saga of gemini deathmask of cancer shura of capricorn camus of aquarius and aphrodite of pisces when athena rebels against abel's plan he attacks her sending her soul to elysion the final resting place from which there is no return the bronze saints immediately rush to the sanctuary to save her and ultimately overcome abel.
The Miracle Maker	A mother and father in search of help for their sick daughter cross paths with an extraordinary carpenter **named** jesus who has devoted his life to spreading **god's** word an amazing miracle brings to light the true meaning of christ and the sacrifices he endured for the deliverance of **mankind** a compelling story of faith trust and devotion.
Elfen Lied	The diclonius a mutated homo sapien that is said to be selected by **god** and will eventually **become** the **destruction** of **mankind** possesses two horns in their heads and has a "sixth sense"which gives it telekinetic abilities due to this dangerous power they have been captured and isolated in laboratories by the government lucy a young and psychotic diclonius manages to break free of her confines and brutally murder most of the guards in the laboratory only to get shot in the head as she makes her escape she survives and manages to drift along to a beach where two teenagers **named** kouta and yuka discovers her having lost her memories she was **named** after the only thing that she can now say "nyuu "and the two allow her to stay at kouta's home however it appears that the evil "lucy"is not dead **just** yet.
Ninja Scroll: The Series	fourteen years after defeating the immortal warrior himuro genma and thwarting the shogun of the dark's evil plans kibagami jubei continues to roam all over japan as a masterless swordsman during his journey he meets shigure a priestess who has never seen the world outside her village but when a group of demons **destroys** the village and kills everyone jubei **becomes** a prime target after acquiring the dragon jewel – a stone with an unknown origin meanwhile shigure – along with the monk dakuan and a young thief **named** tsubute – travels to the village of yagyu and with two demon clans now hunting down shigure dakuan must once again acquire the services of jubei to protect the priestess of light.

5.4.3 Análise de Cluster com 3 filmes ou séries

Todos os filmes e séries deste cluster misturam animação com realidade. É importante mencionar que o filme Mágico de Oz e a série do Mágico do Oz estão presentes neste cluster, o que deve ser esperado de um sistema de recomendação. Os subgêneros deste cluster também são todos iguais: família. A Tabela 5 apresenta os títulos e sinopses dos filmes agrupados. Os termos utilizadas para realizar o agrupamento foram destacados em negrito.

Tabela 5 – Título e sinopse dos 3 filmes clusterizados.

Nome do Filme/Série	Sinopse
Legends of Oz: Dorothy's Return	**Dorothy** wakes up in post tornado **Kansas** only to be whisked back to **Oz** to try to save her old **friends** the scarecrow the lion the tin man and glinda from a devious new villain the jester wiser the owl marshal mallow china princess and tugg the tugboat join **Dorothy** on her latest **magic**al **journey** through the colorful landscape of **Oz** to restore order and happiness to **Emerald city**.
Lost in Oz	When year old **Dorothy** gale discovers her mother's mysterious journal in her **Kansas home** she and her dog Toto are transported into a bustling modern **Emerald city** disoriented and determined to get home **Dorothy** embarks on an epic **journey** with west a young witch and ojo a giant munchkin to seek the **magic** she needs as **Oz** faces its greatest **magic** crisis based on **Frank** baum's **books**. .
Christmas Thieves	After a robbery goes wrong **Frank** and vince break into a **home** when two kids mistake them for their babysitters hoping to make them fall asleep so they can make their getaway **Frank** reads stories from a **magic book** that he stole during the robbery taking them to the enchanted world of arctic **friends** the two thieves have to endure the kid's hijinks as they make various absurd attempts to escape.

Outros clusters com 3 filmes ou séries também apresentaram opções de sequências, como o caso de um cluster que apresentou as três sequências de *Evangelion: 1.0 You Are (Not) Alone, 3.0 You Can (Not) Redo e 3.0+1.0 Thrice Upon a Time*.

5.4.4 Análise de Cluster com 2 filmes ou séries

O cluster em análise apresenta duas séries de ficção científica, onde a sinopse é bastante semelhante com dois protagonistas que sofreram injustiças no decorrer de sua jornada. É importante mencionar que os filmes são de anos bem diferentes (2017/2011), indicando

que uma escolha da coluna *Year* para realizar clusterizações não seria indicado. A Tabela 6 apresenta as sinopses e títulos dos filmes agrupados. Novamente, os termos considerados similares pelo algoritmo *k-means* foram destacados em negrito. Ademais, alguns dos outros clusters gerados com 2 filmes e séries apresentaram sequencias que, claramente, poderiam ser opções de recomendação por similaridade, como é o caso de *Cartoon Classics - Vol. 1: 25 Favorite Cartoons - 3 Hours* e *Cartoon Classics - Vol. 2: 25 Favorite Cartoons - 3 Hours*

Tabela 6 – Título e sinopse dos 2 filmes clusterizados.

Nome do Filme/Série	Sinopse
Detentionaire	No **student** likes having to spend time in detention so you can only imagine how lee ping feels the freshman at a nigma high has been sentenced to a year in detention after being accused of pulling off the biggest prank in high school history the problem is that lee is innocent now in order to clear his name lee must escape from the highly fortified detention room every day infiltrate a new social clique and unravel another piece of the gigantic prank puzzle to try to figure out who **actually** pulled off the epic stunt.
Fighter of the Destiny	Chun chang sheng was abandoned in a flowing river and plucked up by a taoist monk he's **actually** the fourth prince of the chen's royal bloodline he's plagued with an incurable illness fated not to live past the age of to find a cure he leaves his temple armed with a promise of marriage scroll to become a **student** at a famous academy he meets xu you rong and they slowly fall in love after hopping through the trials and tribulations of his journey. .

5.4.5 Problemas encontrados e possíveis melhorias

Como pode-se notar pela Tabela 2, há um cluster com 158 filmes. Este cluster agrupou todos os filmes que não possuiam nenhum termo em comum em suas sinopses. Em trabalhos futuros, espera-se utilizar outros dados da base na tentativa de separar este cluster em agrupamentos com algum grau de similaridade. Como exemplo, pode-se citar a coluna denominada *Ators* e a própria coluna *Title*.

6 CONCLUSÕES E TRABALHOS FUTUROS

Este capítulo apresenta um breve resumo dos resultados e questões aprendidas a cerca da aplicação da metodologia descrita neste trabalho, assim como uma discussão sobre os resultados obtidos.

6.1 Considerações gerais

Esta trabalho apresentou uma análise, tratamento e aplicação do algoritmo *k-means* em conjuntos de dados de filmes e séries da plataformas de *streaming Amazon Prime*. Para realizar o agrupamento foi utilizada a coluna *description* que apresenta a sinopse dos filmes e séries analisados. Por se tratar de uma coluna com dados textuais, aplicou-se uma metodologia para mineração de textos, realizando uma limpeza, tokenização e remoção de *stop words* da base de dados.

Após o tratamento da base, aplicou-se o algoritmo de clusterização *k-means* tendo-se obtido 36 clusters com filmes e séries agrupados por similaridade e um cluster contendo todos os filmes onde não foi possível encontrar termos similares. Após realizar uma análise dos clusters obtidos, constatou-se que em sua grande maioria houve de fato o agrupamento eficiente, havendo muitos clusters com filmes e séries bastante similares. Alguns clusters apresentaram filmes e séries sequenciais, por exemplo.

Para resolver a questão do cluster com filmes e séries sem termos similares em suas sinopses, espera-se em trabalhos futuros adicionar novos dados que possam ajudar o algoritmo *k-means* agrupar este subconjunto de dados. Uma possibilidade seria adicionar as colunas *Title* e *Actor* que possuem dados textuais, relacionados aos títulos e atores dos filmes e séries à base de treinamento. Também espera-se em trabalhos futuros utilizar outros algoritmos de clusterização ou criar modelos híbridos, que utilizem tanto dados textuais quantos dados númericos para realização dos agrupamentos.

REFERÊNCIAS

BENNETT, J.; LANNING, S. *et al.* The netflix prize. In: NEW YORK. **Proceedings of KDD cup and workshop**. [S.l.], 2007. v. 2007, p. 35.

GALVÃO, N. D.; MARIN, H. d. F. Técnica de mineração de dados: uma revisão da literatura. **Acta Paulista de Enfermagem**, SciELO Brasil, v. 22, p. 686–690, 2009.

MORAIS, E. A. M.; AMBRÓSIO, A. P. L. Mineração de textos. **Relatório Técnico–Instituto de Informática (UFG)**, 2007.

NAKKA, D. R.; PRASAD, G.; KUMAR, R. K. Offering recommendations on netflix dataset by associations among users as trust metric. 2020.

NUNES, D. H. F. **Um breve estudo sobre o algoritmo K-means**. Dissertação (Mestrado) — Universidade de Coimbra, 2016.

ORTEGA, F.; BOBADILLA, J.; GUTIÉRREZ, A.; HURTADO, R.; LI, X. Artificial intelligence scientific documentation dataset for recommender systems. **IEEE Access**, IEEE, v. 6, p. 48543–48555, 2018.

ROUSSEEUW, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. **Journal of computational and applied mathematics**, Elsevier, v. 20, p. 53–65, 1987.