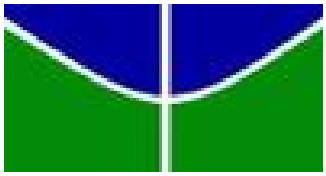

Universidade de Brasília

Instituto de Ciências Exatas - IE

Departamento de Estatística - EST

Sistema de recomendação de notícias do TJDFT baseado em filtragem por conteúdo

Douglas Gomes da Silva

Relatório para obtenção do título

de Bacharel em Estatística

Brasília

2024

Douglas Gomes da Silva

Sistema de recomendação de notícias do TJDFT baseado em filtragem por conteúdo

Orientador:
Prof. Dr Eduardo Monteiro de Castro Gomes

Relatório para obtenção do título
de Bacharel em Estatística

Brasília
2024

Dedido este trabalho a todos que me incentivaram, de modo especial
a minha querida esposa Vandressa

Resumo

Sistemas ou algoritmos de recomendações são ferramentas que provêem automaticamente sugestões de itens a serem desejados pelos usuários, onde estes itens se referem à conteúdos, produtos, serviços, dentre outros, desde que sejam passíveis de serem recomendados. Existem diversos tipos de sistemas de recomendações, para este trabalho foi implementado a técnica de filtragem colaborativa baseada no conteúdo para gerar recomendação de notícias aos usuários do TJDFT - Tribunal de Justiça do Distrito Federal e Territórios. Para isto, foram coletados os dados do site, realizado a vetorização para tranformar os dados textuais em numéricos, reduzido a dimensionalidade da matriz vetorizada, analisado os clusters por meio de técnicas como o K-means, testado um classificador para notícias novas e também foi ultilizado uma métrica de avaliação para medir o desempenho da filtragem. O sistema em si obteve um desempenho bom, porém há possibilidades de aprimoramento há serem exploradas como por exemplos uma limpeza mais meticulosa ou uma filtragem mais rigorosa para reduzir a quantidade palavras pouco relevantes do banco de dados. Outro aspecto a se pontuar foi o baixo desempenho do sistema quando considerado apenas notícias novas já classificadas e que futuramente poderá ser revertido esses resultados aumentando a quantidade de notícias a serem analisadas.

Palavras-chave: Sistema de recomendação, filtragem baseada no conteúdo, notícias, K-means, similaridade, TJDFT.

Sumário

Lista de tabelas

Lista de ilustrações

1 INTRODUÇÃO

Os sistemas de recomendações surgiram na tentativa de ajudar o usuário com a sobrecarga de informações e revolucionou a forma com que interagimos com elas. Por meio desse sistema é possível personalizar a experiência do usuário para apresentar as informações que despertem seu interesse, aumentando seu engajamento e satisfação (REIS, 2012).

A vantagem do sistema de recomendação também se estende às empresas que aumentam suas vendas e engajamento do público com seus E-commerces. Grande exemplo é a antiga locadora de DVDs, a Netflix, que hoje possui um poderoso site de streaming (armazenamento de vídeos na nuvem para visualização dos usuários) de filmes e séries (PELLIZZARO et al., 2016). Com o intuito de atrair mais visualizações, a Netflix faz recomendações baseadas nos filmes que os clientes assistiram e daqueles classificados como "gostei". Segundo um estudo da Netflix, 80% do tempo de uso da plataforma deve-se às recomendações, sendo fundamental para o sucesso do aplicativo (GOMEZ-URIBE; HUNT, 2015).

Mediante a importância destes sistemas, o presente trabalho visa utilizar desse mecanismo para criar um sistema de recomendação das notícias veiculadas no site do Tribunal de Justiça do Distrito Federal e Territórios - TJDFT, afinal, muitas das notícias tem relação quanto à temática ou cronologia, logo, torna-se interessante propor notícias mais preferíveis ou interessantes para cada leitor do site. Contudo, para desenvolver tal tipo de sistema é preciso utilizar uma série de técnicas estatísticas de modo que quando combinadas, transformem um sistema de recomendação eficiente.

E para escolher o tipo de sistema de recomendação em cada caso é preciso avaliar bem as características dos dados disponíveis e também do objtetivo do sistema. Como o intuito deste trabalho é a desenvolver um sistema de recomendação das notícias do TJDFT e nelas se têm informações completas e detalhadas de cada notícia, a filtragem baseada em conteúdo é a mais adequada para este tipo de trabalho.

Utilizando esse tipo de filtragem foi necessário fazer uma grande limpeza nos dados, justamente para poder separar as informações que tem e as que não tem relavância nos textos das notícias. Após esse procedimento, foi utilizado uma vetorização para quantificar a relevância dos termos que aparecia no banco de dados, de forma que seja mensurado a relevância dos termos presentes em cada notícia.

Como uma vetorização de dados textuais (ainda mais de notícias completas) gera uma matriz de alta dimensão, uma ferramenta estatística importantíssima para a redução de dimensão é a análise de componentes principais (ACP), e esta técnica foi primordial para a existência do sistema. Outras técnicas importantes utilizadas para gerar as 5 notícias mais relacionadas, com base na matriz reduzida através da ACP, forão as medidas de similaridade que calculam o quão similar um dado é de outro, neste caso, o quão associada uma notícia está de outra.

Outro procedimento importante neste tipo de sistema é a análise de cluster, onde nela é utilizada um conjunto de técnicas que fazem uma busca geral do banco de dados com o objetivo

de fazer agrupamentos automáticos dos dados segundo seu grau de semelhança.

E por fim foi utilizado o classificador supervisionado K-Nearest Neighbors (KNN) de modo a analisar e avaliar o comportamento do sistema frente à notícias novas.

2 REFERENCIAL TEÓRICO

2.1 PRÉ-PROCESSAMENTO DOS DADOS

O pré-processamento dos dados é a primeira etapa do processamento de dados, aquela em que será extraído os dados, é um processo semi-automático pois depende do analista para avaliar quais são os problemas presentes nos dados, assim como a origem desses problemas e quais técnicas realizar para corrigí-los (BATISTA et al., 2003).

Por meio dessa fase é possível explorar os termos de entedimento, seleção,limpeza e transformação de dados para a descoberta de conhecimento útil (NEVES, 2003). Algumas técnicas de pré-processamento são:

- Remoção de acentos: remove os acentos das palavras, inclusive retira a cedilha da letra c;
- Texto em mínusculo (text lower): retorna o resultado da conversão de todos os caracteres de text em minúsculos;
- Eliminação de stopwords: remoção de termos que são frequentemente considerados irrelevantes para a indexação de conteúdo. Essas palavras são geralmente preposições, artigos, pronomes e outras palavras comuns que são usadas com frequência na língua portuguesa. O objetivo das stop words é melhorar a eficiência dos mecanismos de busca, excluindo palavras que não são relevantes para a análise;
- Lematização: A lematização é o ato de representar as palavras através do infinitivo dos verbos e masculino singular dos substantivos e adjetivos. Ao lematizar palavras, podemos agrupar documentos de maneira mais eficaz com base em seus conteúdos;
- Tokenização: Consiste em dividir um texto em “tokens”, ou seja, em palavras, frases ou símbolos individuais. Esses tokens são então usados como unidades de análise para várias tarefas, como contagem de palavras, classificação de sentimento e análise de tópicos.

2.2 VETORIZAÇÃO TF-IDF

Após a coleta de dados, a segunda etapa é a mapear os dados em um vetor de características. Para cada atributo se estabelece um peso, nesse trabalho será usado o peso do termo TF-IDF, sendo Term Frequency (TF) e Inverse Document Frequency (IDF) (SILVA, 2016).

Com o cálculo da frequência de termos (TF) é possível calcular o número de vezes que cada palavra do vocabulário aparece nesse documento. Logo, quanto maior a frequência, maior a importância da palavra naquele documento específico. Já com o cálculo da frequência inversa de documentos (IDF) é possível mensurar o inverso do logaritmo da frequência de um termo em todos os documentos. Assim, palavras que aparecem em muitos documentos têm um IDF baixo, enquanto palavras raras têm um IDF alto. O TF-IDF mensura o grau de relevância de uma palavra (ou termo) em um documento em relação a todo o corpus dos documentos.

A ponderação TF-IDF será usada para obter a representação de itens por meio de palavras-chaves extraídas em cada notícia.

A fórmula do TF-IDF é a seguinte (SILVA, 2016):

$$TF_{t,n} = \sqrt{count(t,n)}, \tag{1}$$

em que t é o termo/palavra que se busca calcular a frequência na notícia n e *count* a função que realiza a contagem dos termos t na notícia n. O cálculo é feito contando as ocorrências do termo na notícia verificada. No Inverse Document Frequency(IDF) t é o termo/palavra que se busca calcular a frequência inversa no conjunto de notícias *news*. Para realizar esse cálculo é necessário definir o conjunto de termos presentes em todas as notícias. A variável n define a quantidade de notícias da coleção e $df_{t,news}$ é a quantidade de notícias em que o termo t está present, logo o IDF fica da seguinte forma:

$$IDF_{t,news} = \sqrt{\log(\frac{n}{df_{t,news}})}, \tag{2}$$

O produto entre o TF e o IDF resulta na métrica TF-IDF, que avalia a importância de um termo em relação ao seu uso em um documento e ao conjunto de documentos. Se uma palavra tem alta frequência em um documento específico, mas é rara em outros documentos do corpus, ela terá um alto valor TF-IDF, indicando que é importante para aquele documento. Isso ajuda a identificar palavras relevantes para a distinção de tópicos entre diferentes documentos.

Essa técnica é amplamente utilizada na vetorização de textos, permitindo que documentos sejam representados em um espaço vetorial, facilitando a aplicação de algoritmos de aprendizado de máquina para tarefas como classificação, clustering e busca de informações.

Em resumo, o TF-IDF, conforme apresentado por (LESKOVEC; RAJARAMAN; ULLMAN, 2020), é uma ferramenta eficaz para a análise de grandes volumes de dados textuais, destacando termos importantes em documentos e minimizando o impacto de palavras comuns que não contribuem para a semântica geral.

2.3 ANÁLISE DE COMPONENTES PRINCIPAIS

Como os documentos textuais podem estar em representações com milhares de dimensões, utiliza-se a Análise de Componentes Principais (ACP), ou Principal Component Analysis (PCA), que é uma técnica estatística amplamente utilizada para redução de dimensionalidade, especialmente quando há múltiplas variáveis. Esta técnica tem como princípio a ideia de reduzir a massa dos dados, mas sem perca significativa das informações (HONGYU; SANDANIELO; JUNIOR, 2016).

Em muitos casos, as variáveis em um conjunto de dados apresentam correlações elevadas, resultando em redundância de informação. A ACP resolve esse problema ao reduzir o número de variáveis, mantendo apenas os componentes principais que explicam a maior parte da variância. Isso não apenas simplifica a análise, mas também melhora a eficiência de algoritmos de

aprendizado de máquina, ao reduzir a dimensionalidade do espaço de atributos sem sacrificar significativamente a qualidade da informação (JOLLIFFE, 2002).

Por exemplo, foi-se criado um dataframe com 5 variáveis onde cada ponto representa uma observação, conforme gráfico de dispersão 3D abaixo:

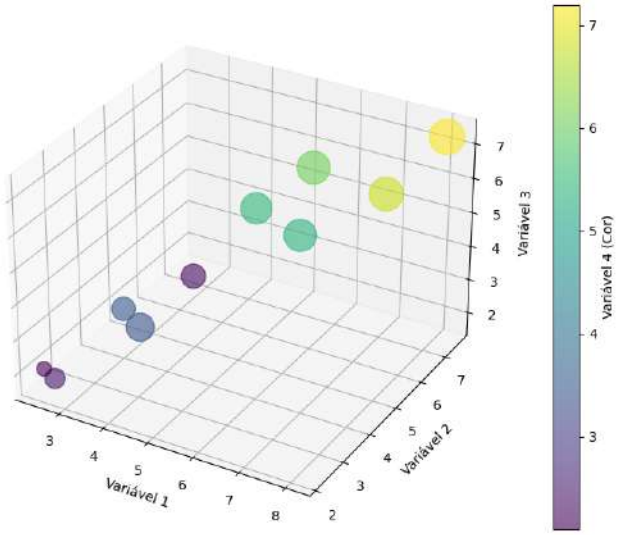

Figura 1 – Gráfico de dispersão 3D

Sendo que a variável 5 é representada pelo tamanho dos pontos do gráfico. Em seguida foi aplicado a PCA para obter os dois primeiros componentes principais, que capturam a maior parte da variância dos dados. Os resultados foram plotados em um gráfico de dispersão, onde os eixos representam os componentes principais:

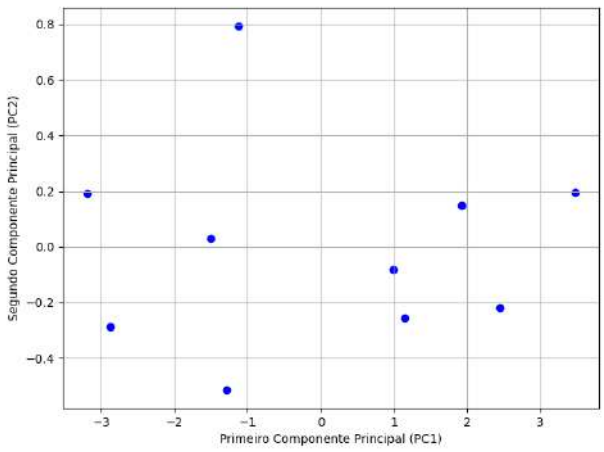

Figura 2 – Análise de componentes principais

Este gráfico permite observar a estrutura dos dados e identificar padrões ou agrupamentos. Por fim, foi calculado a variância explicada por cada um dos componentes principais, que foram

respectivamente 95 e 2%, permitindo entender que 97% da informação dos dados originais foram retida na nova representação. Este processo é essencial para a análise exploratória de dados, pois ajuda a identificar a variabilidade e as relações subjacentes em conjuntos de dados complexos.

Entretanto, o PCA tradicional apresenta desafios quando aplicado em datasets de grande escala, uma vez que requer o cálculo da decomposição da matriz de covariância de todo o conjunto de dados simultaneamente, o que pode exigir uma quantidade significativa de memória. Para superar essa limitação, o IPCA (Incremental Principal Component Analysis) foi proposto como uma variação mais eficiente em termos de memória, permitindo o processamento de dados em partes (lotes) menores.

O IPCA processa o conjunto de dados em pequenos lotes, ajustando-se incrementalmente às novas observações à medida que elas são introduzidas, sem a necessidade de carregar todo o conjunto de dados na memória. Essa abordagem é bastante útil em situações onde a quantidade de dados é muito grande para ser processada de uma só vez, como em sistemas de recomendação ou grandes bases de textos.

Conforme explicado por (BISHOP; NASRABADI, 2006), a técnica mantém as principais direções de variabilidade ao longo do processamento, atualizando progressivamente as componentes principais. Mesmo com essa abordagem incremental, o método garante uma redução de dimensionalidade eficiente e precisa, preservando a maior parte da variância explicada dos dados.

Essa técnica tem sido amplamente adotada em cenários onde o uso de PCA tradicional se torna inviável devido a limitações de memória, permitindo análises escaláveis e de alta performance em grandes volumes de dados. O IPCA oferece, portanto, uma solução eficiente para problemas de redução de dimensionalidade em big data, mantendo a robustez do PCA tradicional enquanto mitiga seus desafios computacionais.

2.4 MEDIDAS DE SIMILARIDADE

Medidas de similaridade estão interligadas com sistemas de recomendação, afinal é necessário mensurar a relação entre dois objetos para poder reconhecer padrões. No livro (BISHOP; NASRABADI, 2006) o autor discute diversas técnicas para calcular similaridade, particularmente no contexto de análise de dados multidimensionais. Essas medidas indicam o quão similares são dois objetos e o tipo medida de similaridade usar depende de vários fatores, entre eles a natureza da variável, escalas de mensuração e conhecimento do tema.

2.4.1 Distância Euclidiana

A distância euclidiana é baseada no teorema de Pitágoras, que estabelece, que em um triângulo retângulo, o quadrado da hipotenusa é a soma dos quadrados dos catetos. Segundo (BISHOP; NASRABADI, 2006) ela é definida na teoria dos vetores e é utilizada para medir a distância "em linha reta"entre dois pontos em um mesmo espaço multidimensional. A distância euclidiana entre dois pontos X e Y é definida por:

$$d(X,Y) = \sqrt{\sum_{i=1}^{n}(X_i - Y_i)^2}, \tag{3}$$

em que o X é o ponto $(x_1, x_2, ..., x_n)$ e Y é o ponto $(y_1, y_2, ..., y_n)$ e eles estão representados em um espaço euclidiano n-dimensional.

2.4.2 Cosseno do Vetor

A similaridade do cosseno é amplamente utilizada em diversas áreas, como processamento de linguagem natural e recuperação de informações, para medir a proximidade entre dois vetores em espaços de alta dimensionalidade. Esse método se destaca por focar na direção dos vetores, ignorando suas magnitudes, o que o torna ideal para comparar documentos ou objetos com tamanhos diferentes. Segundo (JURAFSKY; MARTIN, 2014), a similaridade do cosseno é essencial em modelos de representação vetorial, como o TF-IDF , onde a semelhança entre textos ou palavras é calculada com base no ângulo entre seus respectivos vetores.

A fórmula da similaridade de cosseno é expressa como:

$$similaridade = \cos(\theta) = \frac{\mathbf{X}.\mathbf{Y}}{\|\mathbf{X}\|.\|\mathbf{Y}\|} = \frac{\sum_{i=1}^{n} X_i Y_i}{\sqrt{\sum_{i=1}^{n} X_i^2}\sqrt{\sum_{i=1}^{n} Y_i^2}}, \tag{4}$$

Em que:

- $\mathbf{X}.\mathbf{Y}$ representa o produto escalar entre os vetores $\mathbf{X}$ e $\mathbf{Y}$.
- O produto $\|\mathbf{X}\|.\|\mathbf{Y}\|$ são as normas(ou magnitudes) dos vetores $\mathbf{X}$ e $\mathbf{Y}$, respectivamente.

Se a similaridade de cosseno é 1, os vetores são idênticos; se é -1, são opostos; e se é 0, são ortogonais (sem similaridade angular). Quanto mais próximo a 1, maior a similaridade angular.

Essa métrica é amplamente utilizada em algoritmos de classificação e agrupamento, como o K-means, além de ser uma das técnicas mais robustas para a comparação de vetores de texto, especialmente quando as representações vetoriais são construídas a partir de modelos como o TF-IDF (SINGHAL et al., 2001). Por meio da similaridade do cosseno, é possível medir a similaridade sem ser afetado pela magnitude dos vetores, o que é particularmente útil para comparar documentos de diferentes comprimentos.

2.5 ANÁLISE DE CLUSTER

Um dos principais desafios na análise de dados é resumir as informações coletadas. Geralmente, esta análise é feita com base em grande quantidade de observações dentro do banco de dados, por tanto é interessante criar grupos para facilitar esses resumos. Isto é chamada de análise de cluster (DONI, 2004).

A análise de cluster é uma técnica estatística usada para classificar elementos em grupos, de forma que elementos dentro de um mesmo cluster sejam muito parecidos, e os elementos em diferentes clusters sejam distintos entre si (VALLI, 2012).

Para definir a semelhança – ou diferença – entre os elementos é usada uma função de distância, que precisa ser definida considerando o contexto do problema em questão. Logo, baseando-se na similaridade entre os dados é possível classificar elementos em grupos por suas características e é aplicável a um grande volume de dados (DONI, 2004).

2.5.1 Método K-médias (K-means)

O método K-médias (K-means) é um algoritmo de clusterização não supervisionado. De acordo com (MACQUEEN, 1967) o algoritmo visa particionar um conjunto de objetos em k grupos, de modo que os objetos dentro de cada grupo sejam mais próximos uns dos outros do que dos objetos de outros grupos. O K-means visa minimizar as distâncias quadradas entre cada ponto e o centróide de seu grupo e busca maximizar a variabilidade entre os grupos.

O algoritmo K-means segue um método fácil para classificar um conjunto de dados através de um número específico de clusters (grupos), para (BISHOP; NASRABADI, 2006) o algoritmo funciona da seguinte forma:

1. Definir k centroides iniciais aleatórios, um para cada cluster;
2. Atribuir cada ponto de dados ao centroide mais próximo, ou seja, o que tiver com menor distância baseando-se na distância euclidiana (confrome vimos na subseção 2.4.1);
3. Atualizar o centroide de cada cluster para ser o ponto médio de todos os pontos no cluster;
4. O método se encerra quando chega ao momento de convergência, ou seja, repetem-se os passos 2 e 3 até que os centroides não se alterem significativamente entre as iterações ou até que um número máximo de iterações seja alcançado.

O K-means é uma técnica poderosa e frequentemente usada, mas sua eficiência depende de um correto início de centróides, da escolha do número de clusters e da adequação dos dados ao modelo de clusterização. Em termos de vantagens desta técnica podem ser citadas a eficiência e simplicidade em termos computacionais e também a facilidade de interpretação de seus resultados. Por outro lado, a técnica é um sensível a escolha dos centróides iniciais e também ela necessita que o número de clusters k previamente determinado seja específico. Para contornar esta desvantagem, aplica-se o método do cotovelo de modo a encontrar o número k ideal de clusters.

2.5.2 Método do Cotovelo (Elbow Method)

A determinação do número adequado de clusters é um dos desafios fundamentais no uso do algoritmo K-means, uma vez que o algoritmo requer que o valor de K seja especificado

previamente. Um dos métodos mais populares para essa tarefa é o método do cotovelo (elbow method), que oferece também uma abordagem visual e simples para identificar o número ideal de clusters em um conjunto de dados. O objetivo é identificar um ponto onde o acréscimo de mais clusters não resulta em uma melhoria significativa na variabilidade explicada.

De acordo com (KETCHEN; SHOOK, 1996), o método do cotovelo é uma técnica amplamente utilizada na análise de clusters, particularmente em pesquisas na área de gestão estratégica. Ele envolve executar o K-means para uma série de valores de k (por exemplo, de 1 a 12), calcular e plotar a soma das distâncias quadradas intra-cluster (within-cluster sum of squares, WCSS) em relação a diferentes valores de K . O WCSS é calculado da seguinte forma:

$$WCSS = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2, \tag{5}$$

Em que:

- k é o número de clusters;
- C_i é o conjunto de ponto pertencentes ao cluster i;
- x é um ponto de dados;
- μ_i é o centróide do cluster i;
- $||x - \mu_i||^2$ é a distância euclidiana ao quadrado entre o ponto x e o centroide μ_i do cluster.

À medida que o número de clusters aumenta, a WCSS diminui, pois os pontos ficam mais próximos dos centróides de seus respectivos clusters. No entanto, após um certo ponto, aumentar o número de clusters não resulta em uma redução significativa da WCSS. Esse ponto de inflexão na curva é chamado de "cotovelo", e corresponde ao número ideal de clusters, pois equilibra a simplicidade do modelo com a qualidade do agrupamento. Este processo é fácil de aplicar e interpretar. A visualização da relação entre K e a variabilidade explicada acaba sendo clara. Em contrapartida, a identificação visual do cotovelo nem sempre seja clara, o que pode dificultar sua aplicação em algumas situações.

Supondo que deseja-se encontrar o número ideal de clusters em um banco de dados com 250 observações aleatórias. Para fazer a aplicação do método do cotovelo, aplica-se o K-means para um intervalo de valores de k pré-definido e em seguida calcula-se os WCSS para cada valor de k, conforme no gráfico abaixo:

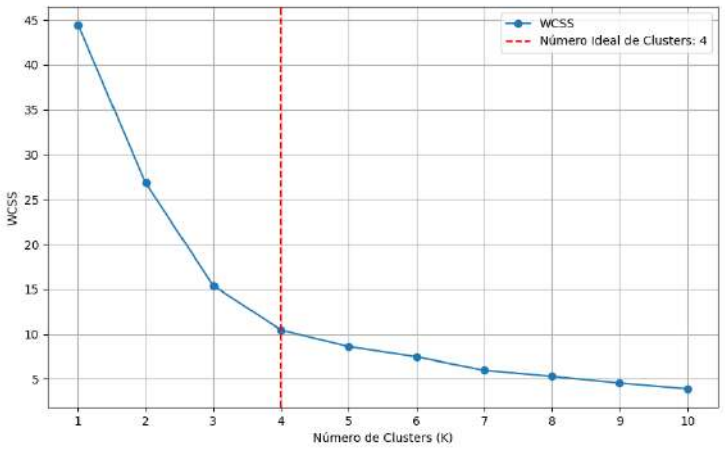

Figura 3 – Exemplo do método do cotovelo

O gráfico mostrará como o WCSS diminui à medida que o número de clusters aumenta, e também é possível perceber que quando k=4 aparece o "cotovelo"do gráfico, onde representa que a redução dos WCSS não foi tão significativa.

2.6 K-NEAREST NEIGHBORS

O K-Nearest Neighbors (KNN) é um dos algoritmos mais simples e eficazes para classificação e regressão, amplamente utilizado em aprendizado de máquina. Trata-se de um método baseado em instâncias, ou seja, o algoritmo armazena o conjunto de treinamento e adia o processo de generalização até o momento da consulta. Isso significa que, ao receber um novo dado, o KNN identifica os K vizinhos mais próximos no conjunto de dados de treinamento, utilizando uma métrica de distância, como a distância euclidiana, para determinar a proximidade entre os pontos (MINING, 2006). Em seguida classifica com base na maioria das classes dos k vizinhos mais próximos.

O KNN é um classificador onde o aprendizado é baseado “no quão similar” um ponto é de outro. Ele leva em consideração os "K vizinhos mais próximos"de uma observação. Ele necessita de um conjunto de dados rotulado para treinamento e o pré-processamento dos dados é fundamental para que as distâncias sejam calculadas corretamentes. As etapas seguintes de um algoritmo KNN são:

1. Recebe um dado não classificado;
2. Mede a distância (Euclidiana por exemplo) do novo dado com todos os outros dados que já estão classificados;
3. Obtém as k menores distâncias;
4. Contabilize a quantidade de vezes que cada classe que apareceu;

5. Toma como resultado a classe que mais apareceu dentre os dados que tiveram as menores distâncias;

6. Classifica o novo dado com a classe tomada como resultado da classificação.

Percebe-se que o KNN é um algoritmo simples, mas poderoso, que pode ser aplicado com sucesso em várias áreas, desde que a escolha de K e da métrica de distância seja adequada às características dos dados. Contudo o algoritmo também necessita de grande memória e sua performance em grandes conjuntos de dados nem sempre é eficaz.

2.7 SISTEMA DE RECOMENDAÇÃO

Sistema de Recomendação (SR) ou Algoritmo de Recomendação é uma combinação de técnicas para selecionar algo personalizado a um indivíduo, com a ajuda de predições probabilísticas de que ele vai gostar de tal indicação. Este mecanismo é capaz de compreender e analisar dados dos usuários e ou dos itens (que são o que se quer recomendar) para se fazer recomendações relevantes (REIS, 2012).

Segundo (JANNACH, 2010) os sistemas de recomendação podem ser definidos como ferramentas que ajudam os usuários a encontrar itens de interesse em grandes conjunto de dados. Esses sistemas utilizam informações prévias sobre as preferências dos usuários ou as características dos itens para fazer previsões sobre o que o usuário pode querer.

Os sistemas de recomendação funcionam basicamente de duas formas. A primeira é sugerindo com base na similaridade entre os usuários, sejam seus gostos, características, opiniões ou etc; por exemplo, há duas pessoas A e B, sendo que o indivíduo A acompanha notícias sobre os temas X,Y e Z, e o indivíduo B que acompanha apenas notícias das temáticas X e Z, logo recomenda-se o tema Y para o usuário B. A segunda é sugerindo com base na similaridade dos itens, sejam produtos ou serviços semelhantes, por exemplo, uma pessoa que em seu Instagram começa a seguir uma conta sobre Fórmula 1 o algoritmo da rede social vai começar a lhe indicar outras contas que falam sobre Fórmula 1.

2.7.1 Filtragem Baseada no Conteúdo

A recomendação baseada no conteúdo parte do princípio de que se um usuário gosta de um determinado item, ele também irá gostar de itens semelhantes (SILVA, 2016).

Resumidamente, o sistema calcula matematicamente a importância dos termos utilizados para descrever os itens. Por exemplo, se um leitor pesquisa no Google "Notícias sobre a seleção brasileira de futebol", embora os termos "sobre", "a"e "de"sejam termos mais frequentes nas pesquisas, a importância do termo "seleção brasileira"é bem maior. Assim, o cálculo deve ponderar o efeito de palavras de alta frequência na determinação da importância de um item.

Essa abordagem oferece algumas vantagens e desvantagens importantes, conforme exposto por (VIEIRA; NUNES, 2012).

Vantagens:

- Explicabilidade das recomendações: Os atributos de sugestão são específicos, tornando mais fácil para o usuário entender o porquê de determinada recomendação;
- Novos usuários: A técnica funciona bem para novos usuários, já que não depende de um histórico de interações para fazer as primeiras recomendações;
- Novos itens: A abordagem também é eficaz para recomendar novos itens, desde que suas características sejam descritas de forma adequada.

Desvantagens:

- Qualidade das descrições: Depende diretamente da qualidade das descrições dos itens, podendo recomendar informações menos relevantes caso as descrições forem imprecisas ou incompletas;
- Dificuldade em capturar nuances: Pode apresentar dificuldade em capturar nuances de preferências;
- Superficialidade: Podem ser superficiais já que não consideram o contexto de uso ou as interações sociais do usuário;
- Limitada a atributos descritos: Como é descrito atributos, caso um ponto importante não for considerado, as recomendações podem ser incompletas.

3 MATERIAIS E MÉTODOS

3.1 BANCO DE DADOS

Para a aplicação neste trabalho, foi realizado uma raspagem de dados das notícias do Tribunal de Justiça do Distrito Federal e dos Territórios (TJDFT) no site oficial do órgão: tjdft.jus.br/institucional/imprensa/noticias. Foram recolhidos informações referente ao título, link e o texto de 3.000 notícias no período entre Abril/2022 e Maio/2024 por meio da linguagem de programação Python utilizando técnicas de web scraping.

A fim de criar um dataframe teste classificador KNN, também foi feito uma raspagem de dados contendo 120 notícias do período de 21/08/2024 à 06/09/2024 do mesmo site e que será explicado nos resultados.

3.2 MÉTODOS

A proposta do presente trabalho em criar um sistema de recomendação para veiculadas no site do Tribunal de Justiça do Distrito Federal e Territórios – TJDFT foi feita considerando a filtragem baseada em conteúdo. Tal metodologia foi utilizada pois a filtragem colaborativa possui como limitação a dificuldade em fazer recomendações para novos usuários ou novos itens, pois não há histórico de interações.

Assim, após a coleta dos dados foi realizado um pré-processamento na variável dos textos das notícias, conforme listado na subseção 2.1, sendo aplicadas diferentes técnicas de processamento textual para um melhor refinamento e análise das informações das notícias.

Em seguida foi realizada a ponderação TF-IDF, vide seção 2.2, para transformar os dados textuais em dados numéricos de acordo com a frequência de cada palavra nos documentos (notícias), assim os valores da matriz serão os pesos TF-IDF, que medem a relevância de cada palavra em uma notícia específica, considerando todas as notícias. Esta matriz é de grande dimensionalidade, ou seja, pode conter muitas colunas. Da matriz TF-IDF (matriz de termos ponderados pelas frequências inversas de documentos) também foram desconsiderados as palavras com menos de 5% de frequência em todas notícias e as palavras com mais de 85% de frequência. Tal filtragem foi necessária para que estas palavras não dominassem o sistema. Após esta filtragem foi aplicado o PCA Incremental com pequenos lotes de 500 observações, vide subseção 2.3, para transformar essa matriz em um conjunto de 431 componentes principais que explicam 95% da variância da matriz TF-IDF. Foi preferível utilizar o PCA Incremental ao invés do PCA pois o banco de dados era de grande escala e não havia memória para realizar o PCA. Ao fazer isto, reduziu-se o número de colunas da matriz, o que facilitou a aplicação das técnicas de clusterização K-Means e KNN.

As técnicas de clusterização K-means e KNN foram utilizadas para a criação do sistema de recomendação baseada no conteúdo, ou seja, nas notícias. Antes de aplicar essas técnicas realizou-se o método do cotovelo pois ele consiste em determinar qual o número ideal de clusters

para o conjunto de notícias. Adiante foi utilizado a técnica K-means para agrupar as notícias de acordo com sua temática que o algoritmo foi determinando, essa quantidade de grupos foi definida de acordo com o número ideal de cluster que o método do cotovelo mostrou.

E por fim, foi realizado o treinamento do KNN usando os clusters do K-means. Na seção dos resultados será demonstrado um exemplo do classificador KNN com os dados teste de notícias.

4 RESULTADOS

4.1 ANÁLISE DESCRITIVA E ALGORITMO DE RECOMENDAÇÃO

Filtragem Baseada em Conteúdo: Para iniciar o algoritmo, foi-se realizado o pré-processamento dos dados das notícias, onde foram retirados, após visualização prévia dos dados, todos os termos e palavras que apresentavam ser irrelevantes para a análise dos dados.

Na sequência foi-se realizado a vetorização TF-IDF para calcular quais termos melhor representavam cada uma das notícias. No meio deste processo foi necessário realizar a Análise de Componentes Principais Incremental (IPCA) para gerar a matriz TF-IDF reduzida.

Adiante foi-se utilizado o método de clusterização K-means para agrupar as notícias por temas, calculando as distâncias euclidianas entre as notícias. E para realizar tal método foi preciso encontrar o número ideal de clusters (grupos) por meio do método do cotovelo (elbow method).

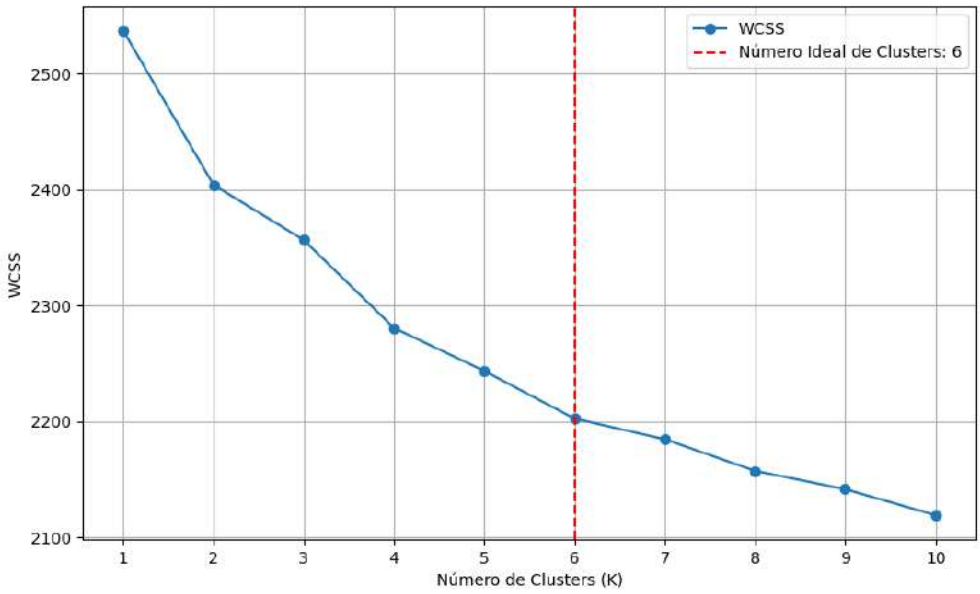

Figura 4 – Gráfico do cotovelo

Conforme ilustrado na Figura 4, conclui-se que o número ideal de clusters é 6, pois é onde a diferença dos WCSS não foi tão significativa, e graficamente é onde aparece o "cotovelo". E para uma melhor visualização desta conclusão, segue abaixo a Tabela 1 com os valores das soma das distâncias quadradas dentro dos clusters (os WCSS) para os diferentes valores de K:

Número de Clusters (K)	WCSS
1	2536,95
2	2404,05
3	2356,40
4	2280,61
5	2243,48
6	2202,42
7	2184,58
8	2157,33
9	2141,49
10	2119,09
11	2113,95
12	2098,92

Tabela 1 – WCSS para cada valor de K no método do cotovelo

Pela Tabela 1, é possível perceber que a redução da WCSS do k=6 para o k=7 foi bem menor que as diferenças do WCSS nos k iniciais, e essa diferença voltou a ser um pouco maior quando tiveram 8 clusters.

Na sequência foi iniciado a clusterização K-means com o número ideal de cluster, em outras palavras, cada notícia do dataset foi classificado em um dos 6 grupos.

Antes de demonstrar as nuvens de palavras de cada cluster, foi plotado a nuvem de palavras geral, de todas as notícias, conforme Figura 5. Estas nuvens de palavras são essenciais para a compreensão das notícias, pois elas são representações visuais da frequência e da importância das palavras em seus contextos.

Figura 5 – Nuvem de palavras de todas as notícias

Pela nuvem de palavras percebe-se temas mais intuitivos como justiça do Distrito Federal e outros menos intuitivos como violência doméstica, infância e juventude.

Na sequência foram gerados os gráfico de nuvens de palavras para cada um desses 6 clusters afim de compreender sobre quais temas estão sendo tratados nesses grupos de notícias:

Figura 6 – Nuvem de palavras - direito civil

Conforme analisado no gráfico de nuvem de palavras, Figura 6, o primeiro cluster contém notícias relacionadas ao direito civil, como processos relacionado à empresas, danos morais e etc.

Figura 7 – Nuvem de palavras - crianças e adolescentes

No segundo cluster, Figura 7, há notícias relacionadas à crianças e adolescentes, família, rede solidária, vara da infância e juventude do DF e demais assuntos associados.

Figura 8 – Nuvem de palavras - ex-presidente do TJDFT

Já no terceiro grupo, Figura 8, tem como temática notícias relacionado ao presidente do TJDFT do biênio 2022-2024, o desembargador Cruz Macedo.

Figura 9 – Nuvem de palavras - violência contra a mulher

No quarto cluster, Figura 9, contém notícias relacionadas a violência doméstica e contra a mulher, lei Maria da Penha e feminicídio.

Figura 10 – Nuvem de palavras - tribunal

No quinto cluster, Figura 10, têm-se notícias mais relacionadas especificamente ao próprio TJDFT como um todo, seus processos e eventos, o poder judiciário e notícias afins.

Figura 11 – Nuvem de palavras - direito criminal

No último cluster, Figura 11, há notícias interligadas ao direito criminal como crimes, prisão em flagrante entre outras.

Em relação a distribuição das notícias em cada cluster ficou conforme a Tabela 2:

Cluster	Total de Notícias
1	717
2	202
3	463
4	194
5	1057
6	367

Tabela 2 – Distribuição de notícias por cluster

Antes de criar e treinar o classficador KNN para as novas notícias, foi gerado uma taxa para avaliar se as notícias recomendadas faziam parte do mesmo cluster que a notícia indexada, e isto tudo foi obtido graças a matriz do cosseno, que é o produto da aplicação do cosseno do vetor na matriz TF-IDF reduzida.

Com base na poderação TF-IDF pode ser calculada a similiaridade de cosseno, vide subseção 2.4.2, das notícias do tribunal.

Essa matriz computa o quão similar uma notícia e de todas as outras. Assim, por meio dela foram extraídas as 5 notícias mais similares de cada notícia do conjunto de dados. E para avaliar se essas similaridades de cada notícia estava com bom desempenho foi calculado a taxa de recomendações no mesmo cluster, resumindo, isto implica descobrir a proporção de notícias que estão sendo recomendadas para notícias do mesmo grupo. Tais resultados podem ser verificados na Figura 12:

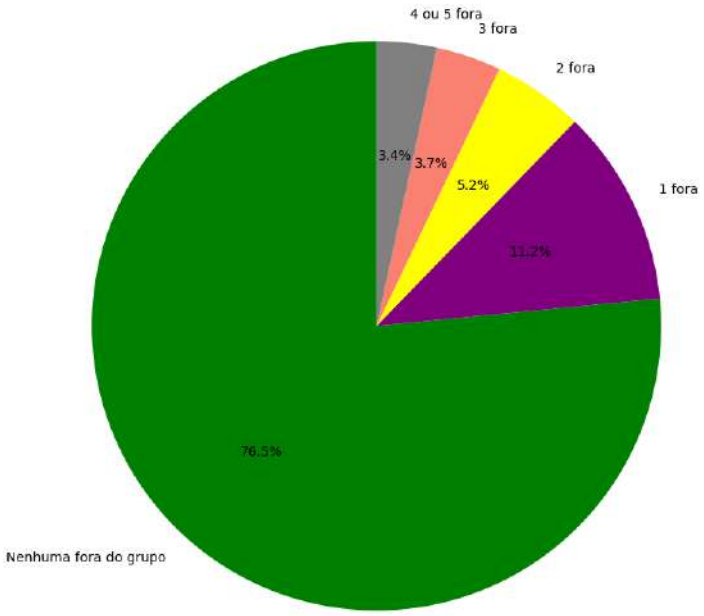

Figura 12 – Percentual das TOP-5 notícias no mesmo cluster

Cerca de 76,5% das TOP-5 notícias mais indicadas fazem parte do mesmo grupo da notícia indexida e aproximadamente apenas 6% das TOP-5 notícias mais indicadas estão com 3 ou mais notícias em clusters diferentes.

Com base nesses clusters obtidos do K-means e a matriz TF-IDF reduzida, foi criado um classificador KNN utilizando os 5 vizinhos mais próximos de cada nova notícia embutida no banco de dados, ou seja, para cada nova notícia o algoritmo classifica ela com base em seus 5 vizinhos mais próximos e aloca ela em um dos cinco clusters usando o modelo KNN treinado.

O teste foi feito com um banco de dados contendo 120 notícias novas e foram realizados os mesmos processos: calculando a similaridade de cosseno, calculando a taxa de notícias recomendadas que estão no mesmo cluster e realizando a análise gráfica deste teste.

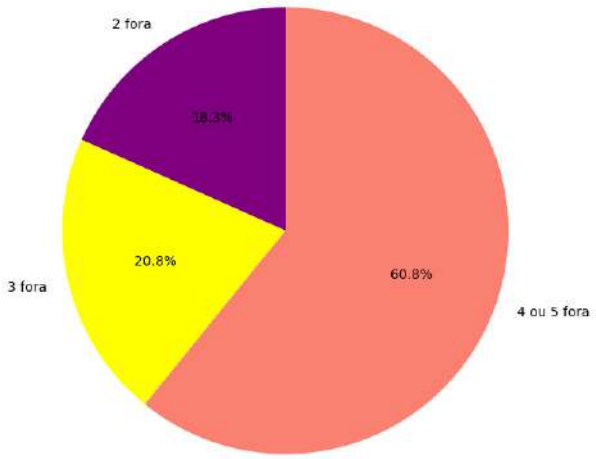

Figura 13 – Percentual das TOP-5 notícias no mesmo cluster - notícias teste

Nos dados testes o sistema de recomendação também funcinou. Classificou cada notícia em um dos clusters já existentes e calculou as taxas de recomendação no mesmo cluster (dentre as TOP-5 de cada notícias), porém, conforme Figura 13, não obteve um retorno bom, pois mais de 60% das TOP-5 notícias mais indicadas das novas notícias estavam com 4 ou as 5 notícias em clusters diferentes e em nenhuma dela teve as TOP-5 notícias mais relacionadas no mesmo cluster.

5 CONCLUSÃO

Apesar da distribuição de notícias em cada cluster ter sido desproporcional, o sistema de recomendação funcionou bem, retornando as 5 notícias mais relacionadas de cada notícia do banco e com um bom percentual de retorno, que era a proporção de notícias recomendadas estivessem no mesmo cluster que a notícia indexada, porém os pontos cruciais para melhorar esse desempenho sejam a limpeza dos dados e a escolha do intervalo da frequência percentual dos itens da matriz TF-IDF que seria considerado para realizar a análise de componentes principais, esses 2 fatores influenciam consideralmente em um sistema de recomendação, ainda mais com dados de textos.

Em relação ao banco-teste o seu retorno foi aquém ao esperado, tal fato decorre da limitação do número de observações já que no banco de dados original os clusters 2 e 4 tinha um baixo percentual das observações e esta amostra, que foi o banco de dados teste, não captou bem notícias destes grupos e as recomendações de notícias dos demais grupos ficaram heterogêneas.

O algoritmo de recomendação se torna uma ferramenta fundamental onde ele de fato retorna itens que manualmente e humanamente são esperados, porém para futuramente refinar melhor o algoritmo, ainda mais em dados textuais, seria interessante analisar com mais detalhes e fazer uma varredura com maior cuidado as palavras que não sejam tão relevantes e também aplicar outras métricas que avaliem o desempenho do algoritmo.

Referências

BATISTA, G. E. d. A. P. et al. **Pré-processamento de dados em aprendizado de máquina supervisionado**. Tese (Doutorado) — Universidade de São Paulo, 2003.

BISHOP, C. M.; NASRABADI, N. M. **Pattern recognition and machine learning**. [S.l.]: Springer, 2006. v. 4.

DONI, M. V. **Análise de cluster: métodos hierárquicos e de particionamento**. Trabalho de Conclusão de Curso — Universidade Presbiteriana Mackenzie, 2004.

GOMEZ-URIBE, C. A.; HUNT, N. The netflix recommender system: Algorithms, business value, and innovation. **ACM Transactions on Management Information Systems (TMIS)**, ACM New York, NY, USA, v. 6, n. 4, p. 1–19, 2015.

HONGYU, K.; SANDANIELO, V. L. M.; JUNIOR, G. J. de O. Análise de componentes principais: resumo teórico, aplicação e interpretação. **E&S Engineering and science**, v. 5, n. 1, p. 83–90, 2016.

JANNACH, D. **Recommender Systems: An Introduction**. [S.l.]: Cambridge University Press, 2010.

JOLLIFFE, I. T. **Principal component analysis for special types of data**. [S.l.]: Springer, 2002.

JURAFSKY, D.; MARTIN, J. H. **Speech and language processing. Vol. 3**. [S.l.]: Pearson London London, 2014.

KETCHEN, D. J.; SHOOK, C. L. The application of cluster analysis in strategic management research: an analysis and critique. **Strategic management journal**, Wiley Online Library, v. 17, n. 6, p. 441–458, 1996.

LESKOVEC, J.; RAJARAMAN, A.; ULLMAN, J. D. **Mining of massive data sets**. [S.l.]: Cambridge university press, 2020.

MACQUEEN, J. Some methods for classification and analysis of multivariate observations. In: **Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability/University of California Press**. [S.l.: s.n.], 1967.

MINING, W. I. D. Data mining: Concepts and techniques. **Morgan Kaufinann**, v. 10, n. 559-569, p. 4, 2006.

NEVES, R. d. C. D. d. **Pré-processamento no processo de descoberta de conhecimento em banco de dados**. Dissertação de Mestrado — Universidade Federal do Rio Grande do Sul, 2003.

PELLIZZARO, M. et al. Sistemas de recomendação: Um estudo de caso. In: 1º Congresso Nacional de Inovação e Tecnologia, 2016.

REIS, L. F. M. d. **Sistema de Recomendação Baseado em Conhecimento**. Dissertação (Mestrado), 2012.

SILVA, R. G. N. **Sistema de Recomendação baseado em conteúdo textual: avaliação e comparação**. Dissertação de Mestrado — Universidade Estadual de Feira de Santana, 2016.

SINGHAL, A. et al. Modern information retrieval: A brief overview. **IEEE Data Eng. Bull.**, v. 24, n. 4, p. 35–43, 2001.

VALLI, M. Análise de cluster. **Augusto Guzzo Revista Acadêmica**, v. 4, p. 77–87, 2012.

VIEIRA, F. J. R.; NUNES, M. A. S. N. Dica: Sistema de recomendação de objetos de aprendizagem baseado em conteúdo. **Scientia Plena**, v. 8, n. 5, 2012.