UNIVERSIDADE FEDERAL DO ABC

Centro de Engenharia, Modelagem e Ciências Sociais Aplicadas
Trabalho de Graduação em Engenharia de Informação

Sistema de Recomendação de Cursos Baseado em Conteúdo

Ighor Ferreira de Melo

Santo André
Abril de 2024

Ighor Ferreira de Melo

Sistema de Recomendação de Cursos Baseado em Conteúdo

Trabalho de Graduação apresentado ao concluir a Graduação em Engenharia de Informação, como parte dos requisitos necessários para a obtenção do Título Bacharel em Engenharia de Informação.

Universidade Federal do ABC
UFABC

Orientador: Dr. Murilo Bellezoni Loiola

Santo André
Abril de 2024

Resumo

Nos últimos anos, houve um notável aumento na demanda e produção de cursos online, levando a uma ampla oferta de opções. No entanto, os estudantes muitas vezes enfrentam dificuldades para selecionar os cursos mais pertinentes às suas necessidades e interesses. Nesse contexto, sistemas de recomendação surgiram como uma solução para facilitar essa escolha de forma personalizada e eficaz. Este estudo concentra-se na criação de um sistema de recomendação de cursos, utilizando como base a descrição do curso e as habilidades que ele aborda. Para alcançar esse objetivo, foram explorados e comparados diversos algoritmos e modelos, avaliando a recomendação de cursos em uma base de dados do Coursera. Os resultados obtidos permitiram identificar as recomendações geradas por cada modelo individualmente, assim como a combinação deles, visando produzir recomendações finais que apresentem tanto alta similaridade quanto complementaridade em relação aos cursos previamente realizados pelos usuários.

Palavras-Chave: Recomendação baseada em conteúdo. Similaridade de Texto. Recomendação de cursos.

Abstract

In the past decade, the demand for and production of online courses significantly increased. With abundant courses available, students often encounter challenges in selecting the most relevant ones tailored to their interests and needs. In this context, recommender systems have emerged in personalized and efficient course selection. This study aims to develop a course recommender system based on course descriptions and the skills they offer. To achieve this goal, various algorithms and models were compared in a Coursera database. Consequently, the recommendations generated by each model and their combination were examined to produce final recommendations that encompass both high similarity and complementarity to courses previously undertaken.

Keywords: Content Based Recommendation. Text Similarity. Course Recommendation.

Sumário

1 Objetivo **3**

2 Introdução **4**

3 Trabalhos Relacionados **6**

4 Referencial Teórico **7**

4.1 Sistemas de recomendação 7

4.1.1 Recomendação baseada em conteúdo 7

4.1.2 Recomendação colaborativa 9

4.1.3 Abordagem Híbrida 10

4.2 Similaridade de Textos 11

4.2.1 Distância de Cosseno 12

4.2.2 Distância Euclidiana 12

4.2.3 Maior Subsequência Comum (LCS) 12

4.2.4 Distância de Levenshtein 13

4.2.5 TF-IDF 14

4.2.6 Abordagens semânticas 16

4.3 Matriz de Confusão 17

5 Metodologia **20**

5.1 Seleção dos dados 20

5.2 Pré-processamento e transformação dos dados 20

5.3 Implementação 21

5.4 Avaliação dos resultados 22

6 Resultados e Discussões **24**

6.1 Resultados do Coursera 24

6.2 Modelos Implementados 26

6.2.1 Maior Subsequência Comum 26

6.2.2 TF-IDF e Distância de Cosseno 28

6.2.3 Transformação Semântica e Distância de Cosseno 29

6.2.4 Busca Semântica 31

6.3 Análise Comparativa dos Modelos 33

7 Conclusão e Trabalhos Futuros **37**

7.1 Conclusão 37

7.2 Trabalhos Futuros 37

1 Objetivo

O objetivo deste trabalho é implementar e analisar a recomendação de cursos baseada em conteúdo como uma solução para o problema da seleção de cursos online. Será utilizada uma base de dados de cursos do Coursera como forma de abordar os conceitos e técnicas relevantes à recomendação baseada em conteúdo, bem como seus principais desafios.

Adicionalmente, serão explorados diferentes algoritmos de similaridade entre textos. Por meio dessa análise, busca-se destacar o potencial da recomendação baseada em conteúdos na melhoria da experiência de aprendizado de estudantes em cursos online.

2 Introdução

MOOCs (*Massive Online Open Courses*) são cursos oferecidos por instituições de ensino na Internet, cujo objetivo é oferecer a qualquer um com conexão uma oportunidade de aprendizado independente de localização geográfica ou requisitos mínimos educacionais[1]. Estima-se que 950 universidades ao redor do mundo disponibilizam um total de 19400 cursos em plataformas educacionais. Até o ano de 2011, aproximadamente 300 mil alunos estavam inscritos nos apenas 3 cursos disponíveis na internet pela universidade de Stanford. Uma década depois, mais de 200 milhões de alunos se inscreveram em MOOCs[2].

Pode-se destacar o Coursera, edx, FutureLearn e Swayam como as maiores plataformas educacionais disponíveis, sendo o Coursera o maior deles, com um total de 97 milhões de alunos inscritos em 6000 cursos[2].

Devido a essa abundância de MOOCs disponíveis, os estudantes muitas vezes enfrentam o desafio de selecionar os cursos mais relevantes e adequados aos seus interesses e necessidades. Nesse contexto, sistemas de recomendação têm emergido como uma abordagem para auxiliar na seleção de MOOCs de forma personalizada e eficiente.

Os sistemas de recomendação têm por objetivo sugerir itens, produtos e conteúdos, de forma automática baseando-se nas preferências do usuário, semelhança entre conteúdos acessados, avaliação, entre outros critérios. Ou seja, um sistema recomendação guia o usuário para algum objeto de seu interesse baseado em informações anteriores[3].

Os algoritmos de recomendação podem ser classificados em diferentes tipos, sendo os principais: recomendação baseada em conteúdo e recomendação colaborativa[3].

A recomendação colaborativa se apoia em dados históricos do comportamento de outros usuários do sistema [4]. No entanto, nem sempre é viável contar com uma base de dados abrangente e confiável. Nesse contexto, a recomendação baseada em conteúdo emerge como uma alternativa promissora para superar essa limitação.

No presente trabalho, será implementado e analisado a recomendação de cursos baseada em conteúdo como uma solução ao problema de seleção de cursos online. Essa implementação será feita tendo como referência um curso anteriormente concluido pelo usuário. A partir desse curso, as recomendações serão feitas com base no conteúdo.

A recomendação baseada em conteúdo, se fundamenta nas características do

conteúdo para determinar as recomendações [5]. Essa abordagem tem sido aplicada em diferentes contextos na literatura. Por exemplo, na recomendação de notícias, na qual o texto da notícia é implementado como fonte da recomendação[6].

Uma abordagem semelhante à implementada no presente estudo foi realizada no trabalho de pós-graduação do Antônio E. R. de Souza [7], em que foi construída uma recomendação dos cursos de uma universidade com base no perfil profissional do aluno. No entanto, essa recomendação foi feita a partir da abordagem colaborativa, em que as informações do histórico de outros alunos são referências para a recomendação.

3 Trabalhos Relacionados

Na literatura a recomendação com base em conteúdo tem sido uma área de pesquisa ativa, abordando desafios como sobrecarga de informações e personalização do aprendizado. Um estudo realizado por Antonio Eduardo Rodrigues de Souza propôs uma metodologia baseada no processo *Knowledge Discovery in Databases* (KDD) para recomendar cursos de especialização aos candidatos, utilizando técnicas de filtragem colaborativa e mineração de dados para adequar os cursos aos perfis profissionais dos candidatos [7].

Além disso, a sobrecarga de informações, particularmente evidente nos portais de notícias na web, tem sido abordada por Michal Kompan e Mária Bieliková através de uma abordagem de recomendação rápida de notícias baseada em conteúdo. O sistema proposto empregou a distância do cosseno para sugerir notícias relevantes, visando aliviar a sobrecarga de informações e melhorar a experiência do usuário na obtenção de notícias relevantes [6].

No contexto dos sistemas de aprendizagem, a personalização do caminho de aprendizagem é essencial. Chung-Ho Su propôs um sistema de recomendação adaptativo de aprendizagem baseado em estilos de aprendizagem modelados por um Modelo Estrutural Interpretativo (ISM). Este sistema utilizou métodos de recomendação baseada em conteúdo e modelos de Grafos para recomendar caminhos de aprendizagem personalizados, resultando em uma melhoria significativa na satisfação do usuário e no desempenho de aprendizagem [8].

Estes estudos destacam diferentes abordagens para lidar com desafios comuns na recomendação de cursos e conteúdo, fornecendo insights valiosos para o desenvolvimento de sistemas mais eficazes e personalizados em diversas áreas de aplicação.

No presente trabalho, serão empregadas diferentes estratégias de similaridade de texto e recomendação para analisar recomendações de cursos baseadas em conteúdo. Isso se alinha com os avanços recentes na área, como evidenciado pelos estudos apresentados, que exploraram abordagens similares para recomendação em diferentes contextos. Ao integrar essas estratégias de similaridade de texto e recomendação, espera-se aprimorar a precisão e a relevância das recomendações de cursos, contribuindo para uma experiência de aprendizado mais personalizada e eficaz para os usuários.

4 Referencial Teórico

As seções subsequentes apresentam o referencial teórico necessário para o desenvolvimento deste trabalho.

4.1 Sistemas de recomendação

Os sistemas de recomendação têm por objetivo sugerir itens, produtos e conteúdos, de forma automática baseando-se em diferentes critérios e diferentes algoritmos. Ou seja, um sistema de recomendação guia o usuário para algum objeto de seu interesse baseado em informações anteriores [3]. A forma como a recomendação é feita e quais critérios são utilizados pelos algoritmos determinam como esses algoritmos são classificados. As técnicas mais referenciadas na literatura são descritas nas seções a seguir.[3]

4.1.1 Recomendação baseada em conteúdo

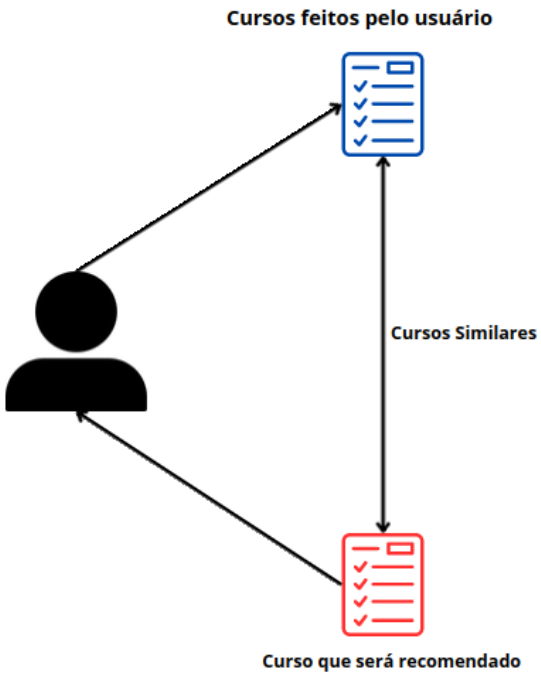

Figura 1: Reomendação de Cursos baseado no conteúdo. (Fonte: Autor)

A técnica de recomendação baseada em conteúdo consiste em analizar a descrição de um item para identificar características que possam interessar ao usuário[5]. Por exem-

plo, na Figura 1 é possível observar um sistema de recomendações de cursos com base no conteúdo. Seu objetivo é recomendar um curso para um usuário analizando a características do curso (professor, descrição, habilidades desenvolvidas, preço). A partir da análise de características, determina-se qual a relação desse curso com cursos passados que o usuário fez ou avaliou como interessante. Desse modo, apenas os cursos com uma relação alta são recomendados.

Afim de determinar essa relação, diferentes técnicas de similaridade de texto podem ser empregadas, pois a recomendação baseada em conteúdo é implementada majoritamente em dados não estruturados, como textos[9]. Na seção 3.2 são apresentadas algumas métricas de similaridade utilizadas nesse trabalho.

Além disso, é necessário que o sistema tenha informações sobre os itens a serem recomendados e sobre o usuário que está interagindo com o sistema[3]. Essas informações podem ser informadas pelo usuário em sua primeira interação através de perguntas [5], ou o sistema pode analizar o que já foi consumido pelo usuário e realizar a recomendação.

Desse modo, as limitações nessa técnica estão associadas à necessidade do esforço do usuário e à dificuldade de recomendar itens quando o interesse muda[5]. Por exemplo, um usuário pode ter feito 10 cursos de programação durante determinado período, mas decidiu que deseja estudar música no momento. Nesses casos, o algoritmo recomendará majoritariamente cursos de programação, pois não há no histórico do usuário cursos de música.

Como forma de solucionar esse problema, alguns sistemas implementam um baixo nível de aleatoriedade nas recomendações. O uso de algoritmos genéticos nesses casos é muito comum[3].

4.1.2 Recomendação colaborativa

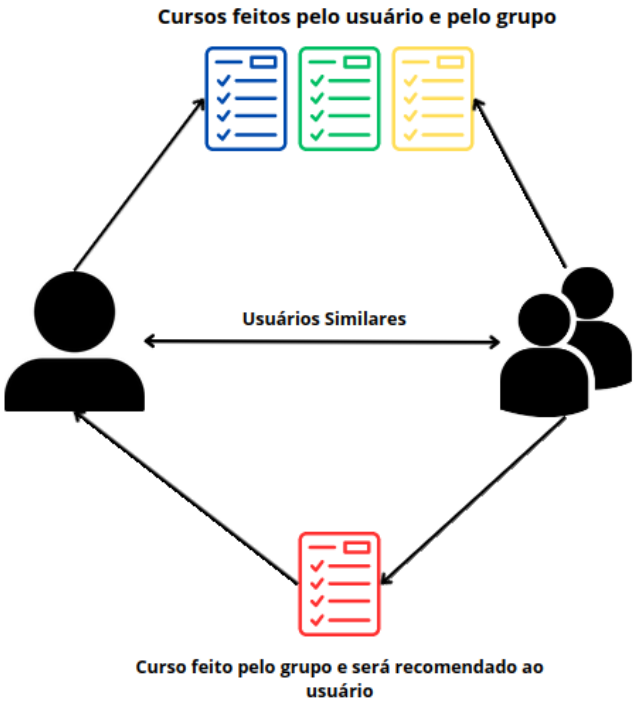

Figura 2: Reomendação de Cursos baseado na estratégia colaborativa. (Fonte: Autor)

A recomendação colaborativa, ou filtragem colaborativa, é baseada no processo de filtrar e avaliar itens utilizando a avaliação de outros usuários como parâmetro [3]. Esse processo utiliza como premissa um comportamento muito antigo da humanidade: considerar a avaliação de outras pessoas como parâmetro de escolha [4].

Na Figura 2 é possível visualizar um exemplo de uma recomendação de cursos baseada na filtragem colaborativa. A recomendação será baseada nos cursos que um grupo de usuários similares fizeram.

Em um sistema de recomendação colaborativa de uma plataforma de filmes que tem por objetivo recomendar um filme para o usuário X, por exemplo, o sistema buscará um conjunto de usuários que avaliaram determinado filme de maneira parecida com o usuário X e recomendará outros filmes que esses usuários também avaliaram positivamente.

Deste modo, a recomendação passa a ser baseada exclusivamente na avaliação de outros usuários. Sendo, portanto, desnecessário reconhecer as características do item recomendado.

Suas principais limitações consistem em saber o que recomendar a um novo usuário e como recomendar um novo item. Essas limitações podem ser parcialmente resolvidas em sistemas recomendação híbrida apresentados na sequência [6].

4.1.3 Abordagem Híbrida

A abordagem híbrida procura combinar duas ou mais técnicas de recomendação afim de evitar limitações relacionadas a uma única abordagem[10]. Essa técnica pode ser implementada de diferentes maneiras [11]:

1. Implementando separadamente;
2. Incorporando características da recomendação baseada em conteúdo na recomendação colaborativa;
3. Incorporando características da recomendação colaborativa na recomendação baseada em conteúdo;
4. Construindo um modelo unificado que implementa as duas técnicas conjuntamente.

Para o primeiro caso, a implementação pode ser feita combinando o resultado de cada uma das recomendações, ou pode-se implementar um sistema que escolha qual técnica deve ser utilizada segundo algum critério de qualidade [11].

Para o segundo caso, pode-se utilizar a recomendação baseada em conteúdo para determinar a semelhança entre os usuários e recomendar o item baseando-se nas avaliações desses usuários [7].

Para o terceiro caso, a implementação mais comum é a de redução de dimensionalidade, afim de criar uma visão colaborativa de um grupo de perfis baseados em conteúdo [7].

Por fim, para o último caso, existem diversas pesquisas sendo desenvolvidas com o objetivo de unificar essas duas técnicas de recomendação. Algumas utilizam abordagens semânticas, outras modelos Bayesianos de regressão, entre outros métodos, afim de determinar os parâmetros probabilistícos que serão utilizados [11].

4.2 Similaridade de Textos

De acordo a teoria da informação, o conceito de similaridade pode ser definido como a razão entre as informações semelhantes em dois objetos e as informações necessárias para descrever completamente os dois objetos[12]. Desse modo, a simaliridade entre textos pode ser descrita como a razão entre aquilo que é comum aos dois textos, e os dois textos completos. Quanto maior a semelhança, maior será a similaridade.

A maioria dos estudos divide os métodos de medição de similaridade entre textos em dois principais grupos: similaridade baseada em estatística e similaridade baseada em conhecimento. No entanto, métodos mais recentes, utilizando inteligência artifical e grafos também estão sendo implementados[13]. Na Figura 3 é possível visualizar diferentes estratégias para determinar a similaridade entre dois textos.

Os métodos baseados em conhecimento utilizam conhecimentos semânticos para determinar a similaridade entre dois textos. Por exemplo, as palavras "Rainha"e "Rei"possuem baixa similaridade nos modelos tradicionais. Entretanto, nos modelos semânticos elas possuem uma similaridade maior, pois a relação semântica da palavra é considerada no cálculo da similaridade[13].

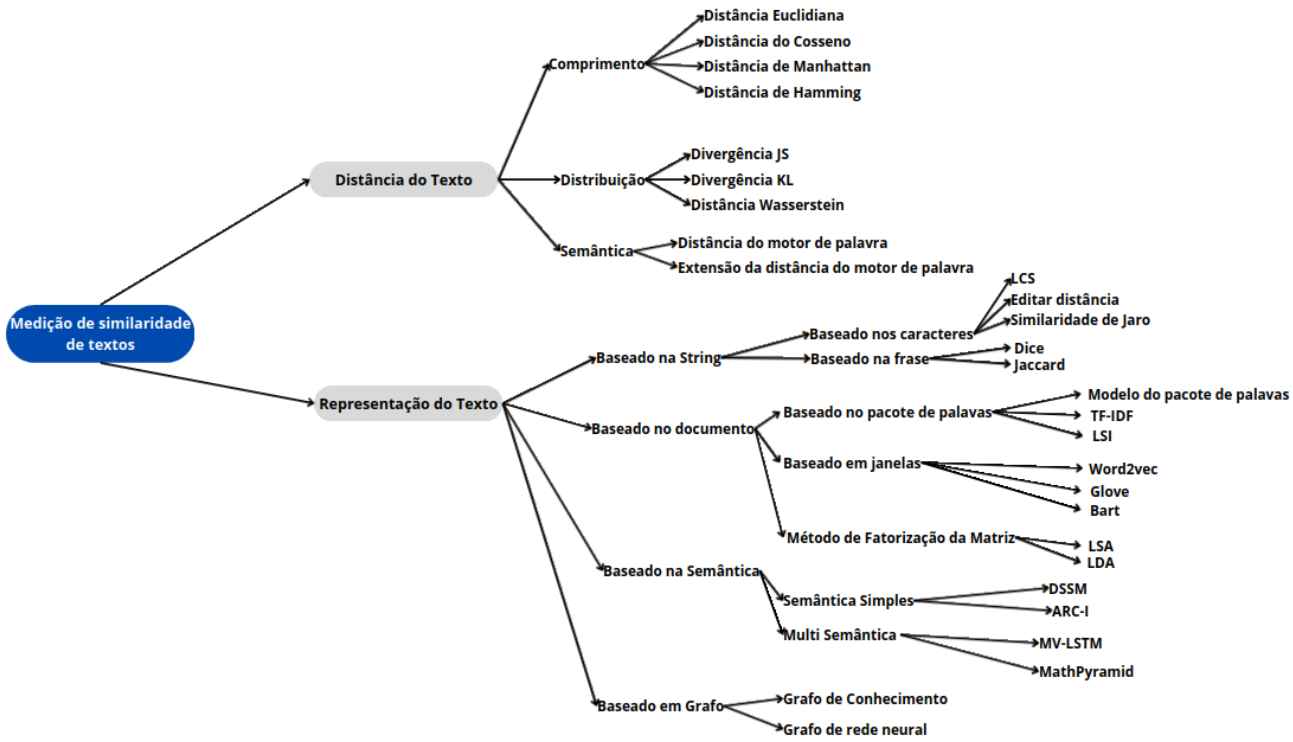

Figura 3: Estratégias de similaridade de texto[13]

As seções subsequentes apresenterão o referencial teórico de algumas técnicas utilizadas na literatura e neste trabalho.

4.2.1 Distância de Cosseno

O método de distância de cosseno utiliza características numéricas do texto para calcular o comprimento da distância[13].

Matematicamente, é possível representar qualquer texto como um vetor multidimensional. Na seção 3.2.5 é apresentado o *TF-iDF*, uma medida estatística que possibilita essa representação vetorial de um texto. Uma vez que o texto tenha sido convertido em um vetor, a similaridade entre dois textos pode ser determinada pelo cosseno do ângulo entre os vetores correspondentes. Quanto menor é o ângulo, maior é a similaridade [14]. A distância de cosseno pode ser calculada como:

$$cos(\alpha) = \frac{xy}{||x||||y||},$$

onde x e y são os vetores associados a cada texto, $|| \cdot ||$ representa a norma do vetor e α é o ângulo entre os vetores x e y.

Essa abordagem é muito utilizada em sistemas de recomendação baseados em conteúdo, como recomendação de notícias [15].

4.2.2 Distância Euclidiana

A distância euclidiana é a forma matematicamente mais simples de calcular a similaridade entre dois textos. Assim como na distância de cosseno, uma vez que os textos tenham sido convertidos em vetores, torna-se viável calcular a similaridade por meio da distância entre dois pontos[13]. Tal medida de distância pode ser determinada como:

$$d(x, y) = \sqrt{\sum_{i=1}^{\mathrm{n}} (x_i - y_i)^2}$$

onde x e y são os vetores associados a cada texto, x_i é o valor do ponto na horizontal, y_i o valor do ponto na vertical e n o número de pontos considerados no cálculo.

4.2.3 Maior Subsequência Comum (LCS)

A abordagem do LCS (*Longest Common Subsequence*) consiste em encontrar a maior subsequência comum entre duas strings originais S1 e S2 [16].Todas as subsequências

obedecem estritamente a ordem dos caracteres presentes nas strings originais. Assim, sejam X e Y duas strings, defini-se LCS(X,Y) como a maior subsequência comum entre X e Y[16].

Ao comparar as palavras “ABCBI” e “ABACAXI”, percebe-se que a maior subsequência comum entre as duas palavras é ABCI. Portanto, o LCS é igual a 4. Situação análogo ocorre para as palavras “monkey” e “money” cujo LCS obtido é igual a 5.

4.2.4 Distância de Levenshtein

A Distância de Levenshtein é um algoritmo que determina a similaridade entre duas strings S1 e S2 por meio do cálculo do menor número de edições necessárias para transformar a string S1 na string S2 [17]. Para realizar esse cálculo, o algoritmo observa três operações de custo igual a 1 [18]:

- Acrescentar um caractere;
- Remover um caractere;
- Substituir um caractere por outro.

O processo se dá com o algoritmo transformando cada string em uma matriz. Na Figura 4 é possível visualizar um exemplo para as palavras “monkey”e “money”, onde cada posição da matriz representa o custo necessário para transformar uma palavra na outra. Ou seja, é comparando cada caractere de S1 com todos os caracteres S2 e o custo de cada operação realizada na posição específica [17, 18].

A Figura 5 ilustra o funcionamento do algoritmo Levenshtein, em que a comparação entre as palavras “monkey” e “money” gera como resultado 1. Em outras palavras, apenas a remoção da letra k é necessária para tornar as duas palavras iguais.

		m	o	n	k	e	y
	0	1	2	3	4	5	6
m	1	0	1	2	3	4	5
o	2	1	0	1	2	3	4
n	3	2	1	0	1	2	3
e	4	3	2	1	1	1	2
y	5	4	3	2	2	2	1

Figura 4: Exemplo de representação utilizando o Levenshtein (Fonte: https://vinayakgarg.wordpress.com/2012/12/10/edit-distance-using-dynamic-programming/)

M	O	N	K	E	Y
M	O	N	E	Y	

	ALTERAÇÃO
DIS(lev)=1	

Figura 5: Exemplo de comparação utilizando o Levenshtein (Fonte: Autor)

4.2.5 TF-IDF

TF-IDF (*Term-Frequency - Inverse Document Frequency*) é uma medida estatística utilizada como forma de indicar a importância de um termo em um documento. Seu funcionamento consiste em identificar quais palavras aparecem com mais frequência em documentos que contenham aquela palavra[13].

Desse modo, o documento passa a ser representado como um vetor de probabilidades, em que cada posição do vetor representa o peso daquela palavra no documento[14].

Essa técnica é fundamental na recomendação baseada em conteúdo, sendo combinada com outras técnicas para determinar a similaridade entre dois textos[14].

O cálculo do *TF-IDF* é representado da seguinte maneira:

$$TF(t,d) = \frac{\text{número de vezes que t aparece em d}}{\text{número total de termos em d}}$$

$$IDF(t) = \log_{10} \frac{\text{número total de documentos}}{\text{número de documentos que contenham t}}$$

$$TF - IDF(t,d) = TF(t,d) * IDF(t)$$

onde t representa cada palavra e d representa cada documento.

Considere dois documentos como exemplo: “O cachorro está no telhado”e o “O cachorro está na rua”. O primeiro passo é calcular o TF de cada palavra para cada documento:

- Documento 1
  - TF(“o”) = 1/5
  - TF(“cachorro”) = 1/5
  - TF(“está”) = 1/5
  - TF(“no”) = 1/5
  - TF(“telhado”) = 1/5
- Documento 2
  - TF(“o”) = 1/5
  - TF(“cachorro”) = 1/5
  - TF(“está”) = 1/5
  - TF(“na”) = 1/5
  - TF(“rua”) = 1/5

O segundo passo é calcular o IDF:

- IDF(“o”) = log(2/2) = 0
- IDF(“cachorro”) = log(2/2) = 0
- IDF(“está”) = log(2/2) = 0
- IDF(“no”) = log(1/2) = 0.3
- IDF(“telhado”) = log(1/2) = 0.3
- IDF(“na”) = log(1/2) = 0.3
- IDF(“rua”) = log(1/2) = 0.3

O último passo é múltiplicar o TF com o IDF e obter a representação vetorial de cada documento. No exemplo acima a representação vetorial será:

- "O cachorro está no telhado": [0, 0, 0, 0.06, 0.06]
- "O cachorro está na rua": [0, 0, 0, 0.06, 0.06]

4.2.6 Abordagens semânticas

A similaridade semântica pode ser conceituada como uma métrica que se fundamenta na proximidade de significados, em oposição à semelhança de caracteres[13]. Em outras palavras, ao invés de se basear na correspondência de sequências de caracteres idênticas, a similaridade semântica busca identificar a proximidade entre os sentidos ou significados subjacentes das palavras, expressões ou frases em análise. Essa abordagem proporciona uma visão mais profunda e contextualizada da relação entre os elementos linguísticos, facilitando a compreensão de seus vínculos semânticos [13].

A presente abordagem demonstra uma característica intrinsecamente humana, destacando a capacidade inata do ser humano de reconhecer semelhanças entre palavras, mesmo quando suas grafias não são aparentemente similares. Por exemplo, "rei"e "príncipe", "ver"e "assistir", "correr"e "andar", "carro"e "automóvel", entre outras.

Ao empregar modelos de inteligência artificial, é viável estender essa habilidade à máquina, permitindo a percepção equivalente das relações entre termos. Desse modo, o uso da inteligência artificial pode contribuir para um sistema recomendação mais eficiente.

No presente trabalho será implementado um modelo semântico para calcular a similaridade entre textos.

O modelo utilizado é um modelo de transformação de texto chamado *all-MiniLM-L6-v2*, disponível no site da *Hugging Face*[19]. Esse é um modelo menor de Grandes Modelos de Linguagem (LLMs, do inglês *Large Language Models*), como o GPT-3 e o Bert.

O modelo *all-MiniLM-L6-v2* foi submetido a um extenso processo de treinamento, utilizando um conjunto de dados contendo mais de 1 bilhão de frases. Seu principal propósito consiste em realizar o mapeamento de sentenças e parágrafos para um espaço vetorial de 384 dimensões, visando extrair o significado semântico das sentenças fornecidas[19].

O modelo implementado nesse trabalho (*all-MiniLM-L6-v2* de 2023) não realiza o cálculo de similaridade de forma intrínseca. Entretanto, seus resultados podem ser empregados em consonância com outros métodos para a clusterização de texto e busca semântica. Essa abordagem combinada possibilita a exploração de sinergias e a potencial melhoria do desempenho dos algoritmos aplicados nessas tarefas.

Afim de realizar a busca semântica pode-se utilizar o modelo *mfaq*, também disponível no site da *Hugging Face*[20]. Esse modelo foi treinado utilizando o dataset *Multiligual FAQ* com o propósito de ranquear respostas a partir de uma pergunta recebida[20]. Esse modelo utiliza internamente a implementaçao do *all-MiniLM-L6-v2*.

Além disso, é importante ressaltar que os modelos semânticos implementados nesse estudo não possuem parametrização, ou seja, apenas uma frase é inserida como parâmetro e através de seu treinamento o modelo retorna a transformação ou recomendação.

A implementação de modelos semânticos (LLMs) representa um desafio significativo no contexto acadêmico devido ao elevado custo computacional associado. Esses modelos requerem uma quantidade substancial de recursos de memória e energia tanto para o treinamento quanto para a implementação[21].

O modelo do GPT-3, um dos mais poderosos modelos baseados em transformação, foi treinado com aproximadamente 300 bilhões de *tokens*. Sendo assim, um pequeno ajuste no GPT-3 só pode ser feito com o auxílio de GPUs de alto desempenho e densidade computacional, como Google Cloud Platform (GCP), Microsoft Azure, Amazon Web Services (AWS), entre outros [21]. No presente trabalho, os modelos já estão previamente treinados, sendo necessária apenas a implementação.

4.3 Matriz de Confusão

Uma abordagem para a avaliação do desempenho de algoritmos preditivos e de classificação na literatura acadêmica é por meio da utilização da matriz de confusão. Essa ferramenta se fundamenta na criação de uma matriz que registra os resultados corretos e incorretos obtidos pelo algoritmo em análise. A partir dessa matriz, é possível extrair medidas objetivas para análise e compreensão do desempenho do algoritmo em questão.[22] A matriz de confusão fornece uma visão detalhada sobre a capacidade do algoritmo em classificar corretamente as amostras, permitindo aferir sua eficácia e identificar eventuais

áreas de melhoria. Diante disso, essa foi a abordagem escolhida para avaliar as recomendações.

Na Figura 6 é possível visualizar 4 medidas:

- Verdadeiro Positivo: Resultados retornados pelo algoritmo e são de fato os resultados corretos[23];
- Verdadeiro Negativo: Resultados não retornados pelo algoritmo e de fato não são resultados corretos[23];
- Falso Positivo: Resultados incorretos retornados pelo algoritmo[23];
- Falso Negativo: Resultados não retornados pelo algoritmo, mas deveriam ter sido retornados[23].

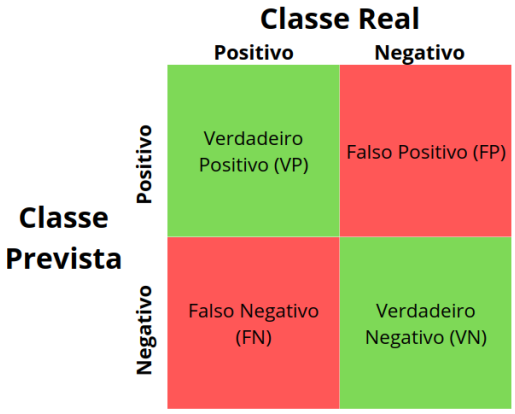

Figura 6: Matriz de Confusão (Fonte: Autor)

A partir dessa matriz é possível obter 3 principais medidas:

- *Recall*: Medida que indica a taxa de verdadeiros positivos [23]. Essa métrica expressa a proporção entre o número de elementos corretamente identificados e o número total de elementos que deveriam ser corretamente identificados. A fórmula para calcular a taxa de verdadeiros positivos é o número de verdadeiros positivos dividido pelo número total de elementos corretos.

$$= \frac{(VerdadeiroPositivo)}{(VerdadeiroPositivo + FalsoNegativo)}$$

- Precisão: Medida que indica a taxa de verdadeiro positivo em relação ao total do que foi classificado como positivo[23]. Em outras palavras, representa quanto dos elementos retornados correspondem aos elementos requeridos.

$$= \frac{(VerdadeiroPositivo)}{(VerdadeiroPositivo + FalsoPositivo)}$$

- *F-score*: Medida que combina Precisão e *Recall* com o objetivo de retornar um número único para avaliação do algoritmo. Essa medida é balanceada quando $\beta = 1$ [24].

$$= \frac{[(\beta^2+1) * Precisão * Recall]}{[\beta^2 * (Precisão + Recall)]}$$

5 Metodologia

5.1 Seleção dos dados

Para a realização do presente trabalho foi escolhido um conjunto de dados de Cursos do Coursera em 2021[25], que estão disponíveis na plataforma Kaggle. Na Figura 7 é possível visualizar a estrutura dos dados utilizados.

Esse conjunto de dados foi escolhido pelo fato do Coursera representar a maior plataforma no mercado de MOOCs, com o maior número de inscritos e cursos no ano de 2021[2].

Course Na...	University	Difficulty L...	Course Ra...	Course URL	Course De...	Skills
Write A Feature Length Screenplay For Film Or Television	Michigan State University	Beginner	4.8	https://www.coursera.org/learn/write-a-feature-length-screenplay-for-film-or-television	Write a Full Length Feature Film Script In this course, you will write a complete, feature-length s...	Drama Comedy peering screenwriting film Document Review dialogue creative writing Writing u...
Business Strategy: Business Model Canvas Analysis with Miro	Coursera Project Network	Beginner	4.8	https://www.coursera.org/learn/canvas-analysis-miro	By the end of this guided project, you will be fluent in identifying and creating Business Model Can...	Finance business plan persona (user experience) business model canvas Planning Business projec...
Silicon Thin Film Solar Cells	◆cole Polytechnique	Advanced	4.1	https://www.coursera.org/learn/silicon-thin-film-solar-cells	This course consists of a general presentation of solar cells based on silicon thin films. It is t...	chemistry physics Solar Energy film lambda calculus Electrical Engineering electronics energy...
Finance for	IESE Business	Intermediate	4.8	https://www.cou	When it comes	accounts

Figura 7: Conjunto de dadados de cursos do Coursera em 2021

5.2 Pré-processamento e transformação dos dados

Como forma de preparação para a aplicação dos algoritmos de recomendação os dados foram limpos e transformados. Essa limpeza incluiu a remoção de ruído e pontuação, lematização, e seleção de quais dados serão utilizados para recomendação.

1. Remoção de ruído e pontuação: Nesta etapa, são removidos caracteres que não são relevantes para a análise textual, como pontuação, números, símbolos, entre outros. Isso ajuda a limpar o texto e a evitar a contagem de palavras irrelevantes[26].

2. Lematização: A lematização é o processo de reduzir as palavras a sua forma base, ou seja, reduzir as palavras flexionadas a uma única forma. Por exemplo, as palavras

"correm", "corria"e "correndo"seriam reduzidas à forma base "correr". Isso ajuda a reduzir o conjunto de palavras utilizadas, uma vez que muitas palavras relacionadas podem ser agrupadas em uma única forma base[26].

3. Seleção de atributos relevantes: A seleção de features relevantes envolve a escolha das características mais importantes do texto para análise. Isso pode ser feito de diversas maneiras, como a análise de frequência de palavras, a análise de correlação com a variável de saída ou a análise de importância de recursos. Isso ajuda a reduzir a dimensão do espaço de características, eliminando palavras irrelevantes ou redundantes [26]. No presente trabalho, foi calculada a frequência de cada palavra e posteriormente as palavras mais frequentes que poderiam ser suprimidas sem prejudicar o sentido foram removidas.

Essas etapas são importantes para o pré-processamento de dados de texto, já que ajudam a reduzir a dimensão do espaço de características e a limpar o texto de informações irrelevantes. O objetivo é deixar apenas as informações mais relevantes e úteis para a análise de dados e para o modelo de recomendação.

5.3 Implementação

A implementação no presente trabalho foi feita na linguagem *Python* e as recomendações foram construídas a partir do usuário fornecer um curso de referência que ele já concluiu.

Na base de dados escolhida foram identificadas 3 estratégias diferentes para recomendação:

1. Recomendação baseada na descrição do curso;
2. Recomendação baseada no conjunto de habilidades desenvolvidas pelo curso;
3. Recomendação baseada nos dois parâmetros.

Devido ao fato da recomendação baseada em conteúdo fundamentar-se na utilização das informações provenientes de outros cursos feitos pelo usuário, emerge um cenário onde o propósito se desloca além de uma simples recomendação, passando a assumir a natureza complexa de avaliar a similaridade entre dois textos distintos.

As estratégias de similaridade podem ser divididas em dois principais grupos: cálculo da distância entre textos e formas de representar os textos.[26]

No presente trabalho, a partir de um curso de referência foi realizada a recomendação. Os cursos escolhidos como referência para a recomendação foram: *Python Programming Essentials* e *Business Strategy*. Ou seja, uma vez que o usuário fez o curso *Business Strategy* é recomendado uma lista de cursos que também podem ser interessantes.

Além disso, as estratégias de similaridade escolhidas foram:

1. Maior Subsequência Comum (LCS);
2. *TD-IDF* para conversão da *string* e *Distância de Cosseno* para cálculo de distâncias;
3. Transformação Semântica para conversão da *string* e *Distância de Cosseno* para cálculo de distâncias;
4. Busca Semântica.

5.4 Avaliação dos resultados

Conforme descrito na Seção 3.3, uma abordagem para a avaliação da capacidade preditiva dos modelos na literatura acadêmica é por meio da utilização da matriz de confusão. Essa ferramenta se fundamenta na criação de uma matriz que registra os resultados corretos e incorretos obtidos pelo algoritmo em análise. A partir dessa matriz, é possível extrair medidas objetivas para análise e compreensão do desempenho do algoritmo em questão[22]. A matriz de confusão fornece uma visão detalhada sobre a capacidade do algoritmo em classificar corretamente as amostras, permitindo aferir sua eficácia e identificar eventuais áreas de melhoria. Diante disso, essa foi a abordagem escolhida para avaliar as recomendações.

Neste trabalho, utilizou-se a plataforma Coursera como para obter os valores de referência para a matriz de confusão. Esses valores foram obtidos a partir da busca do Coursera, onde foi buscado os cursos *Bussiness Strategy* e *Python Programming Essentials*.

No entanto, esse modelo de avaliação é limitado para o contexto de recomendação, visto que os a base de dados do Coursera é de 2021 e a os resultados obtidos na busca do Coursera são de uma base mais atualizada. Além disso, esses resultados podem não refletir necessariamente as melhores recomendações possíveis. Por exemplo, um curso de

Python pode ter sido recomendado pelo modelo em questão, mas não ter sido sugerido na busca realizada diretamente no site. É fundamental considerar essas restrições ao interpretar e analisar os resultados apresentados neste estudo, visando uma abordagem mais abrangente e confiável no processo de recomendação.

No contexto da avaliação de modelos de recomendação, é importante considerar um desafio adicional que reside na ideia de que cursos altamente similares nem sempre são a melhor escolha para a recomendação. É possível que um curso extremamente similar a outros não desperte o interesse do usuário, tornando-se inadequado para a recomendação. Em tais situações, a indicação de cursos complementares pode se revelar mais atrativa e pertinente para atender às necessidades e preferências do usuário. Essa consideração ressalta a importância de explorar abordagens mais sofisticadas para a seleção e apresentação dos cursos recomendados, levando em conta não apenas a similaridade entre eles, mas também a diversificação das sugestões para melhor atender aos usuários.

Diante disso, os modelos utilizados para avaliação foram os modelos utilizando os parâmetros de descrição e de habilidades desenvolvidas no curso, pois foi percebido que os modelos apresentaram maior similaridade na comparação das habilidades do que na descrição. Isso possibilita que a combinação de descrição e habilidades forneça cursos com uma grande similaridade (habilidade) e cursos não tão similares (descrição).

Por fim, além da matriz de confusão, os resultados dos modelos foram comparados entre si. Isso possibilitou que os resultados fossem analisados sem considerar os resultados do Coursera. Desse modo, aplicando uma análise mais abrangente na recomendação.

6 Resultados e Discussões

Essa seção apresenta os resultados do presente trabalho em concordância com a metodologia (Seção 4). Todas as implementações foram feitas a partir de um dos dois cursos feitos pelo usuário:

- Business Strategy
- Python Programming Essentials

Esses cursos foram escolhidos, pois são áreas não correlacionadas com um conjunto grande de opções semelhantes na plataforma do Coursera. Essa abordagem visa a disponibilização de um maior volume de informações passíveis de serem empregadas para fins analíticos.

O propósito desta análise reside na compreensão da interligação de complementaridade, delineada por cursos que, embora não compartilhem uma semelhança flagrante, apresentam conhecimento complementar; assim como a dimensão de similaridade existente entre os cursos sugeridos. Isso é especialmente relevante, considerando o cenário de recomendação, em que cursos excessivamente similares podem não prover um valor substancial ao estudante, uma vez que o estudando pode não querer aprender o mesmo conteúdo que já aprendeu em outro curso.

Adicionalmente, foi realizada uma análise comparativa dos resultados obtidos pelos diferentes modelos. O propósito dessa análise foi avaliar a viabilidade de combinar determinados modelos, de modo a construir recomendações que incorporem elementos de complementaridade e similaridade de forma equilibrada e otimizada.

6.1 Resultados do Coursera

Afim de obter os resultados que serão utilizados nos cálculos da matriz de confusão foi necessário buscar no Coursera os cursos *Business Strategy* e *Python Programming Essentials*.

O objetivo dessa busca era identificar quais cursos o próprio Coursera recomendaria para um usuário interessado em aprender esses temas. Na Figura 8 são apresentados os resultados para a busca no Coursera referentes ao curso *Business Strategy*, enquanto na Figura 9, são exibidos os resultados para o curso *Python Programming Essentials*.

É importante ressaltar que a base de dados utilizada nesse estudo é de 2021 e alguns cursos que foram retornados não estão nessa base, por conta disso, esses resultados foram removidos.

Indice	Nome do Curso
3046	Corporate Strategy
1818	Foundations of Business Strategy
1685	The Strategy of Content Marketing
1979	Financial Markets
3322	Introduction to Negotiation: A Strategic Playb...
3238	Driving business towards the Sustainable Devel...
3154	Design-Led Strategy: Design thinking for busin...
3386	Successful Negotiation: Essential Strategies a...
2731	Strategy Implementation
2559	Business Growth Strategy
3339	Advanced Valuation and Strategy - M&A, Private...
1407	Creative Thinking: Techniques and Tools for Su...
947	Strategic Planning and Execution
3220	Innovation Management
2313	Pricing Strategy
2285	Corporate Strategy
1826	Advanced Business Strategy
2063	Strategic Management
1195	Competitive Strategy
1623	Bridging the Gap between Strategy Design and D...

Figura 8: Resultados na busca de “Business Strategy”no Coursera (Fonte: Autor)

Indice	Nome do Curso
2003	Data Analysis with Python
3165	Computer Science: Programming with a Purpose
2012	HTML, CSS, and Javascript for Web Developers
2961	Foundations of Data Science: K-Means Clusterin...
2797	Introduction and Programming with IoT Boards
1435	Design Computing: 3D Modeling in Rhinoceros wi...
1773	AI for Medical Diagnosis
2033	Python Data Visualization
1778	AI Workflow: Feature Engineering and Bias Dete...
23	Preparing for the Google Cloud Professional Da...
1753	MongoDB Aggregation Framework
3377	Optimizing Machine Learning Performance
674	Meaningful Predictive Modeling
2273	Python Data Analysis
725	Python Data Representations

Figura 9: Resultados na busca de "Python Programming Essentials"no Coursera (Fonte: Autor)

6.2 Modelos Implementados

6.2.1 Maior Subsequência Comum

A primeira implementação foi a utilização da biblioteca *fuzzywuzzy* para implementar a estratégia de Maior Subsequência Comum.

A descrição e habilidades foram unificadas em uma única recomendação, onde foram mescladas em ordem de similaridade cinco recomendações baseadas em habilidades e cinco recomendações baseadas em descrição. O objetivo dessa implementação foi mesclar resultados com maior similaridade, obtidos a partir das habilidades, com resultados com maior complementariedade, obtidos a partir da descrição.

Nas Figuras 10 e 11 é possível visualizar os cursos recomendados para os cursos de *Business Strategy* e *Python Programming Essentials*, respectivamente. Nessas figuras também é possível visualizar a fonte de recomendação (Habilidades ou Descrição).

	Course	Index	Similarity	University	Category
0	Impact from digital transformation: A Nano course	671	0.68	EIT Digital	Skills
1	Corporate Strategy	3046	0.66	University of London	Skills
2	Design Thinking for Innovation	1993	0.65	University of Virginia	Skills
3	Sales Strategy	3091	0.65	Funda►►o Instituto de Administra►►o	Skills
4	Strategy Formulation	2998	0.65	Copenhagen Business School	Skills
5	Entrepreneurial Strategic Management	951	0.46	University of New Mexico	Description
6	Value Chain Management Capstone	3184	0.45	University of Illinois at Urbana-Champaign	Description
7	Design your own trading strategy ► Culminating...	2235	0.41	Indian School of Business	Description
8	Astrobiology and the Search for Extraterrestri...	3078	0.41	The University of Edinburgh	Description
9	Cryptography	2868	0.40	University of Maryland, College Park	Description

Figura 10: Recomendações curso Business Strategy utilizando o LCS (Fonte: Autor)

	Course	Index	Similarity	University	Category
0	Introduction to Programming in Swift 5	1228	0.64	LearnQuest	Skills
1	Python Programming: A Concise Introduction	537	0.64	Wesleyan University	Skills
2	Learn to Program: The Fundamentals	1616	0.64	University of Toronto	Skills
3	Programming for Everybody (Getting Started wit...	452	0.63	University of Michigan	Skills
4	Problem Solving, Python Programming, and Video...	2163	0.63	University of Alberta	Skills
5	Astrobiology and the Search for Extraterrestri...	3078	0.48	The University of Edinburgh	Description
6	Learn to Program: The Fundamentals	1616	0.47	University of Toronto	Description
7	Genomic Data Science with Galaxy	2758	0.45	Johns Hopkins University	Description
8	C++ For C Programmers, Part B	1368	0.43	University of California, Santa Cruz	Description
9	Cryptography	2868	0.42	University of Maryland, College Park	Description

Figura 11: Recomendações curso de Python utilizando o LCS (Fonte: Autor)

É possível observar que tanto para o curso *Business Strategy*, quanto para o curso *Python Programming Essentials*, a similaridade das habilidades foi maior que a descrição. Isso pode ser observado pelo fato da tabela ter sid ordenada pela métrica de similaridade.

Na recomendação do curso *Business Strategy* apenas um curso (*Corporate Strategy* linha 1) foi retornado tanto pelo modelo quanto na busca do Coursera, o que gerou um f-score de 6.45%.

Na recomendação do curso *Python Programming Essentials* nenhum curso foi retornado em comum entre o modelo e a busca do Coursera, fato que impossibilitou o cálculo do f-score.

É interessante observar que as duas recomendações retornaram os cursos *Astrobiology and the Search for Extraterrestrial Life* e *Cryptography* na recomendação por descrição. Isso ocorreu porque as descrições desses cursos são muito curtas e, por isso,

qualquer termo em comum implica em uma similaridade alta.

6.2.2 TF-IDF e Distância de Cosseno

Nessa implementação foi utilizado o *TF-IDF* como estratégia de vetorização do texto juntamente com *Distância de cosseno* para cálculo da similaridade entre os textos.

Nas Figuras 12 e 13 é possível visualizar as recomendações para os cursos *Business Strategy* e *Python Programming Essentials*, respectivamente. Nas figuras também é possível visualizar a fonte de recomendação (Habilidades ou Descrição).

	Course	Index	Similarity	University	Category
0	Strategic Management	2063	0.862009	Copenhagen Business School	Skills
1	Strategy Formulation	2998	0.795023	Copenhagen Business School	Skills
2	Competitive Strategy	1195	0.783014	Ludwig-Maximilians-Universit$\diamond$t M$\diamond$nchen (LMU)	Skills
3	Advanced Business Strategy	1826	0.748118	University of Virginia	Skills
4	Bridging the Gap between Strategy Design and D...	1623	0.735760	Brightline Initiative	Skills
5	Strategic Leadership and Management Capstone	2342	0.386648	University of Illinois at Urbana-Champaign	Description
6	Corporate Strategy	2285	0.330316	University of Illinois at Urbana-Champaign	Description
7	Managerial Economics and Business Analysis Cap...	2304	0.313125	University of Illinois at Urbana-Champaign	Description
8	Accounting for Business Decision Making: Strat...	1021	0.309880	University of Illinois at Urbana-Champaign	Description
9	Managerial Accounting: Tools for Facilitating ...	2978	0.289466	University of Illinois at Urbana-Champaign	Description

Figura 12: Recomendações Business Strategy a partir do TF-IDF (Fonte: Autor)

	Course	Index	Similarity	University	Category
0	Learn to Program: The Fundamentals	1616	0.718544	University of Toronto	Skills
1	Python Data Representations	725	0.718436	Rice University	Description
2	Python Programming: A Concise Introduction	537	0.692379	Wesleyan University	Skills
3	Python Basics	1825	0.673100	University of Michigan	Skills
4	Python Data Structures	297	0.672841	University of Michigan	Skills
5	Python Data Analysis	2273	0.660865	Rice University	Description
6	Programming for Everybody (Getting Started wit...	452	0.657731	University of Michigan	Skills
7	Compose and Program Music in Python using Ears...	1182	0.477377	Coursera Project Network	Description
8	Python Functions, Files, and Dictionaries	731	0.466325	University of Michigan	Description
9	Crash Course on Python	978	0.432145	Google	Description

Figura 13: Recomendações Python Programming Essentials a partir do TF-IDF (Fonte: Autor)

Assim como no modelo utilizando a maior subsequência comum, os cursos com maior similaridade foram os cursos com base nas habilidades. Isso pode ser explicado pelo fato de que as habilidades não são um texto corrrido, mas sim um conjunto de termos que são pré-definidos na plataforma.

Por exemplo: Um curso de python terá nas suas habilidades: “Programação”, “Python”, etc... Enquanto sua descrição terá um texto que pode conter diferentes palavras. Desse modo, a similaridade entre 2 cursos de *Python* será maior quando for utilizado como referência as habilidades. Entretanto, um curso de Python, *Python Data Representations* (linha 1 da Figura 13) teve uma similaridade maior na descrição que nas habilidades. Isso ocorreu pelo fato desse curso ser uma continuação direta do curso buscado, *Python Programming Essentials*, e sua descrição repetir termos do curso anterior, enquanto apresenta habilidades diferentes.

O modelo recomendado para os cursos relacionados com *Business Strategy* apresentou um *f-score* de 33.34% e o modelo para os cursos relacionados com *Python Programming Essentials* apresentou um *f-score* de 16%. Isso implica que esse modelo de recomendação apresentou resultados superiores ao resultado da Maior Subsequência Comum.

No entanto, é interessante observar que apesar dos resultados retornados para o curso de *Python* não serem os mesmos retornados pela busca do Coursera, os cursos *Python Basics* (linha 3), *Python Functions, Files, and Dictionaries* (linha 8) poderiam ter sido retornados pela busca. Por conta disso, avaliar o modelo apenas utilizando a busca do Coursera não representa a melhor estratégia.

6.2.3 Transformação Semântica e Distância de Cosseno

Nessa implementação foi utilizada a estratégia de transformação semântica para vetorização do texto juntamente com o *Distância de Cosseno* para cálculo da similaridade entre os textos.

Nas Figuras 14 e 15 é possível visualizar as recomendações para os cursos de *Business Strategy* e *Python Programming Essentials*, respectivamente.

	Course	Index	Similarity	University	Category
0	Strategic Management	2063	0.937182	Copenhagen Business School	Skills
1	Strategy Formulation	2998	0.911756	Copenhagen Business School	Skills
2	Corporate Strategy	2285	0.876259	University of Illinois at Urbana-Champaign	Description
3	Advanced Business Strategy	1826	0.848010	University of Virginia	Skills
4	Bridging the Gap between Strategy Design and D...	1623	0.826512	Brightline Initiative	Skills
5	Design-Led Strategy: Design thinking for busin...	3154	0.792424	The University of Sydney	Skills
6	Managerial Accounting: Cost Behaviors, Systems...	2541	0.739233	University of Illinois at Urbana-Champaign	Description
7	Managerial Accounting: Tools for Facilitating ...	2978	0.738911	University of Illinois at Urbana-Champaign	Description
8	Accounting for Business Decision Making: Strat...	1021	0.734672	University of Illinois at Urbana-Champaign	Description
9	Accounting for Business Decision Making: Measu...	829	0.714990	University of Illinois at Urbana-Champaign	Description

Figura 14: Recomendações Business Strategy a partir da transformação semântica (Fonte: Autor)

	Course	Index	Similarity	University	Category
0	Learn to Program: The Fundamentals	1616	0.891536	University of Toronto	Skills
1	Programming for Everybody (Getting Started wit...	452	0.871239	University of Michigan	Skills
2	Python Basics	1825	0.832220	University of Michigan	Skills
3	Crash Course on Python	978	0.824918	Google	Description
4	An Introduction to Interactive Programming in ...	2753	0.822881	Rice University	Skills
5	Python Programming: A Concise Introduction	537	0.812542	Wesleyan University	Skills
6	An Introduction to Interactive Programming in ...	3157	0.811543	Rice University	Description
7	Python Data Representations	725	0.805483	Rice University	Description
8	Create Your First Python Program	719	0.780337	Coursera Project Network	Description
9	Problem Solving, Python Programming, and Video...	2163	0.768588	University of Alberta	Description

Figura 15: Recomendações Python Programming Essentials a partir da transformação semântica (Fonte: Autor)

O modelo recomendado para os cursos relacionados com *Business Strategy* apresentou um *f-score* de 34.48% e o modelo para os cursos relacionados com *Python Programming Essentials* apresentou um *f-score* de 8.33%.

Apesar da métrica de *f-score* não ter apresentado um ganho significativo em relação ao modelo anterior, é possível observar que as similaridades retornadas foram significativamente maiores. Por exemplo, o curso *Corporate Strategy* (linha 2 da Figura 14) que está presente nas duas recomendações e na busca do Coursera, teve um ganho de similaridade de mais de 50%. Isso evidencia que a estratégia de transformação semântica do texto possibilitou uma similaridade maior entre os cursos.

Além disso, o curso *Corporate Strategy* não apenas apresentou uma similaridade maior, como também ganhou um destaque em relação as recomendações baseadas em des-

crição. Desse modo, podendo indicar que esse curso apresenta uma relação de similaridade maior do que de complementaridade.

O curso *Crash Course on Python* (linha 3 da Figura 15) também apresentou uma diferença significativa na sua similaridade, pois no modelo anterior esse curso ocupava a última posição com similaridade de 43% e na transformação semântica ocupa a 4º posição com similaridade de 82%.

Esse resultado indica que, assim como o curso *Corporate Strategy*, o curso *Crash Course on Python* apresenta uma similaridade maior do que sua complementaridade com o *Python Programming Essentials*.

Além disso, o fato dessas duas recomendações terem ganhado destaque entre as 5 mais similares pode indicar que a transformação semântica consegue calcular bem a similaridade tanto em textos longos (descrição), quanto em palavras (habilidades). Fato também observado no curso *Python Data Representations* (linha 7), onde apesar da similaridade ter sido maior, de 71% para 80%, ele não se destacou tanto quanto no modelo anterior. Essa constatação indica o destaque do modelo semântico em calcular a similaridade de maneira mais eficaz.

Dado que o objetivo da recomendação nem sempre é garantir a maior similaridade, o uso da transformação semântica pode resultar em recomendações excessiavamente similares, fato que pode não ser interessante ao usuário.

6.2.4 Busca Semântica

Nessa implementação foi utilizado um modelo de busca semântica para realizar as recomendações. Esse modelo, utiliza a transformação semântica para transformar o texto e determina a similaridade diretamente no modelo.

Nas Figura 16 e 17 é possível visualizar os cursos recomendados para *Business Strategy* e *Python Programming Essentials*, respectivamente.

	Course	Index	Similarity	University	Category
0	Operations Management: Analysis and Improvemen...	3371	0.993282	University of Illinois at Urbana-Champaign	Description
1	Operations Management: Strategy and Quality Ma...	1480	0.992808	University of Illinois at Urbana-Champaign	Description
2	Marketing Management I	1490	0.990933	University of Illinois at Urbana-Champaign	Description
3	Global Strategy II: Doing Business in The Glob...	1599	0.990244	University of Illinois at Urbana-Champaign	Description
4	Foundations of Everyday Leadership	2292	0.989825	University of Illinois at Urbana-Champaign	Description
5	Critical Perspectives on Management	603	0.980301	IE Business School	Skill
6	Developing a Google SRE Culture	319	0.979474	Google Cloud	Skill
7	Two Speed IT: How Companies Can Surf the Digit...	2813	0.978869	CentraleSup$\hat{e}$lec	Skill
8	Diversity and inclusion in the workplace	3193	0.978153	ESSEC Business School	Skill
9	Building the Business Model for Corporate Entr...	982	0.978118	University of Maryland, College Park	Skill

Figura 16: Recomendações Business Strategy a partir da busca semântica (Fonte: Autor)

	Course	Index	Similarity	University	Category
0	Python Functions, Files, and Dictionaries	731	0.994354	University of Michigan	Description
1	Python Data Analysis	2273	0.992736	Rice University	Description
2	Introduction to Data Science in Python	1022	0.992217	University of Michigan	Description
3	Python and Statistics for Financial Analysis	2859	0.992143	The Hong Kong University of Science and Techno...	Description
4	Python Classes and Inheritance	2374	0.991814	University of Michigan	Description
5	Crash Course on Python	978	0.980011	Google	Skill
6	Python Programming: A Concise Introduction	537	0.978522	Wesleyan University	Skill
7	Python Data Structures	297	0.977406	University of Michigan	Skill
8	Create Your First Python Program	719	0.976760	Coursera Project Network	Skill
9	Python Basics	1825	0.976032	University of Michigan	Skill

Figura 17: Recomendações Python Programming Essentials a partir da busca semântica (Fonte: Autor)

O modelo recomendado para os cursos relacionados com *Business Strategy* não apresentou nenhum curso retornado pela busca do Coursera. Portanto, não é possível calcular o *f-score*. Entretanto, o modelo para os cursos relacionados com *Python Programming Essentials* apresentou um *f-score* de 8%, semelhante aos modelos anteriores.

Apesar do modelo para *Business Strategy* não ter nenhum resultado correspondente a busca do Coursera, é possível analisar que os cursos retornados possuem uma maior complementaridade que os cursos retornados em outros modelos.

É interessante observar que o curso *Corporate Strategy*, recomendado em outros modelos, não está entre os resultados. Além disso, os cursos com recomendação baseada em descrição possuiram uma similaridade maior que os baseados em habilidade. Isso pode ter ocorrido, pois o modelo de busca semântica não calcula similaridade para retornar os resultados. Baseando-se em seu treinamento, o modelo procura identificar qual descrição

ou habilidade está mais relacionada ao curso buscado. Sendo assim, as medidas de similaridade não são determinadas baseando-se na semelhança entre as *strings* buscadas.

O curso *Crash Course on Python* em todos os modelos retornou como uma recomendação baseada em descrição. No entanto, na busca semântica, ele foi retornado como uma recomendação baseada em habilidade. Isso também pode ser justificado pela estratégia de similaridade da busca semântica, pois a similaridade do texto não é a que o algoritmo usa para recomendar os resultados.

Finalmente, o presente modelo recuperou um curso comum tanto em sua descrição quanto em sua habilidade, intitulado *Python Basics* (linha 9 da Figura 17). Contudo, a inclusão desse resultado foi omitida a fim de evitar redundância. No entanto, tal constatação pode sugerir uma notável afinidade com o curso denominado *Python Programming Essentials*. Importante ressaltar que esta referida recomendação foi identificada em todos os modelos examinados.

Portanto, é possível afirmar que a Busca Semântica, em contraposição ao modelo anterior, apresentou uma melhor relação de complementaridade que de similaridade. Essa constatação implica que a combinação desses dois modelos pode garantir tanto cursos similares, quanto cursos complementares na recomendação.

6.3 Análise Comparativa dos Modelos

Nesta seção, será conduzida uma comparação dos resultados obtidos por todos os modelos analisados. O propósito é sintetizar os cursos retornados por cada modelo e apresentar uma análise que resuma as contribuições individuais de cada modelo.

Figura 18: Recomendações consolidadas de Business Strategy (Fonte: Autor)

Na Figura 18, é evidenciado que sete cursos estão em comum entre o modelo baseado em TFIDF+Cosseno e o modelo que incorpora a transformação semântica seguida da *Distância de Cosseno*. Considerando que o método de cálculo de similaridade (Distância Cosseno) permanece inalterado em ambos os casos, conclui-se que a diferença nos cursos apresentados é determinada pelo método de transformação empregado.

Portanto, devido ao maior custo computacional associado à transformação semântica, sua implementação não se mostra relevante em comparação com o método baseado em TFIDF.

Na figura 19, é ilustrada a recomendação resultante da combinação de todos os modelos, visando obter uma recomendação completa.

Index	Nome	TFIDF + Cosine	LCS	Transformação Semântica + Cosine	Busca Semântica
2063	Strategic Management	☑		☑	
2998	Strategy Formulation	☑		☑	
1195	Competitive Strategy	☑			
1826	Advanced Business Strategy	☑		☑	
1623	Bridging the Gap between Strategy Design and Delivery	☑		☑	
2342	Strategic Leadership and Management Capstone	☑			
2285	Corporate Strategy	☑		☑	
2304	Managerial Economics and Business Analysis Capstone	☑			
1021	Accounting for Business Decision Making: Strategy Assessment and Con	☑		☑	
2978	Managerial Accounting: Tools for Facilitating and Guiding Business Decis	☑		☑	
671	Impact from digital transformation: A Nano course		☑		
3046	Corporate Strategy		☑		
1993	Design Thinking for Innovation		☑		
3091	Sales Strategy		☑		
2998	Strategy Formulation		☑		
951	Entrepreneurial Strategic Management		☑		
3184	Value Chain Management Capstone		☑		
2235	Design your own trading strategy ' Culminating...		☑		
3078	Astrobiology and the Search for Extraterrestri...		☑		
2868	Cryptography		☑		
3154	Design-Led Strategy: Design Thinking for business			☑	
829	Accounting for Business Decision Making: Measure...			☑	
2541	Managerial Accounting: Cost Behaviors...			☑	
3371	Operations Management: Analysis and Improvement				☑
1480	Operations Management: Strategy and Quality Ma...				☑
1490	Marketing Management I				☑
1599	Global Strategy II: Doing Business in the Global...				☑
2292	Foundations of Everyday Leadership				☑
603	Critical Perspectives on Management				☑
319	Developing a Google SRE Culture				☑
2813	Two Speed IT: How Companies Can Surf the Digit...				☑
3193	Diversity and Inclusion in the workplace				☑
982	Building the Business Model for Corporate Entre...				☑

Figura 19: Recomendações consolidadas de BusinessDistância Strategy (Fonte: Autor)

Dado o propósito da recomendação de cursos visando garantir similaridade, complementaridade e uma dose adequada de aleatoriedade, uma lista de cursos integrando todos os modelos seria benéfica para esse procedimento.

No entanto, é importante ressaltar que a combinação dos modelos pode impactar

a performance da recomendação, pois os modelos semânticos apresentam um maior custo computacional. Devido a isso, é importante que o objetivo do usuário seja considerado no modelo, uma vez que para uma maior similaridade a implementação dos modelos semânticos não se destaca.

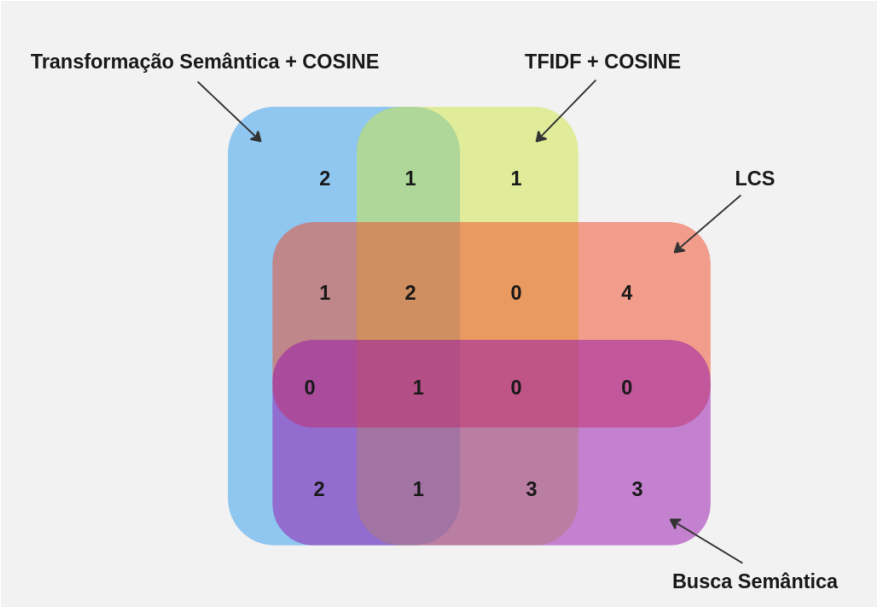

Figura 20: Recomendações consolidadas Python Programming Essentials (Fonte: Autor)

Na figura 20 é possível observar que no curso *Python Programming Essentials* a intersecção entre os modelos foi muito maior, e 1 curso (*Python Programming a Concise Introduction*) foi retornado em todas as implementações. Isso evidencia que esse curso pode possuir muita similaridade com o *Python Programming Essentials*.

Ambos os cursos são projetados para iniciantes completos em Python, indicando que se um aluno completou um deles, não haveria necessidade de recomendar o outro. No entanto, devido à abordagem baseada em similaridade do modelo de recomendação, esse curso foi incluído em todas as recomendações.

Dada a significativa interseção entre os cursos retornados pelos diferentes modelos, é desafiador determinar qual modelo ofereceria os melhores resultados. Na Figura 21, é apresentada uma listagem de todos os cursos se os modelos fossem combinados.

Index	Nome	TFIDF + Cosine	LCS	Transformação Semântica + Cosine	Busca Semântica
1616	Learn to Program: The Fundamentals	☑	☑	☑	
725	Python Data Representations	☑		☑	
537	Python Programming: A Concise Introduction	☑	☑	☑	☑
1825	Python Basics	☑		☑	☑
297	Python Data Structures	☑			☑
2273	Python Data Analysis	☑			☑
452	Programming for Everybody (Getting Started with Python)	☑	☑	☑	
1182	Compose and Program Music in Python using Earsketch	☑			
731	Python Functions, Files, and Dictionaries	☑			☑
978	Crash Course on Python	☑		☑	☑
1228	Introduction to Programming in Swift 5		☑		
2163	Problem Solving, Python Programming, and Video...		☑	☑	
3078	Astrobiology and the Search for Extraterrestri...		☑		
2758	Genomic Data Science with Galaxy		☑		
1368	C++ For C Programmers, Part B		☑		
2868	Cryptography		☑		
2753	An Introduction to Interactive Programming in...			☑	
3157	An Introduction to Interactive Programming in...			☑	
719	Create your First Python Program			☑	☑
1022	Introduction to Data Science in Python				☑
2859	Python and Statistics for Financial Analysis				☑
2374	Python Classes and Inheritance				☑

Figura 21: Recomendações consolidadas Python Programming Essentials (Fonte: Autor)

7 Conclusão e Trabalhos Futuros

7.1 Conclusão

O propósito primordial deste estudo foi a implementação de diversas estratégias de similaridade textual para facilitar a recomendação de cursos com base no conteúdo, especificamente a descrição e as habilidades associadas.

Observou-se que a utilização de diferentes estratégias de análise de similaridade textual e recomendação tem um impacto direto nos resultados dos cursos recomendados. Isto é evidenciado pela disparidade no comportamento da recomendação entre os cursos de *Python* e *Business*.

No entanto, dado que o objetivo da recomendação de cursos não se limita apenas à indicação daqueles com a maior similaridade, a integração dos resultados advindos de cada modelo revelou-se vantajosa. Dessa forma, torna-se possível combinar os resultados que se assemelham ao curso feito pelo aluno com cursos similares, complementares e até mesmo cursos que possam não ter qualquer relação.

Considerando as limitações inerentes à base de dados empregada, a avaliação da qualidade das recomendações torna-se um desafio. Seria necessário implementar os modelos e analisar o comportamento dos usuários da plataforma após a aplicação das recomendações.

7.2 Trabalhos Futuros

Como extensão deste estudo, poderia ser implementada a utilização de modelos que incorporem informações de múltiplos cursos como referência. Isso apresentaria um desafio, uma vez que a similaridade não seria calculada com base apenas na descrição ou habilidade de um curso individual, mas sim em um conjunto de cursos.

Além disso, seria viável implementar o modelo considerando outras características do curso, como avaliações dos alunos ou o nível do curso. Integrar essas funcionalidades garantiria que o curso recomendado não apenas seja relevante para o aluno, mas também tenha uma alta classificação de qualidade.

Por fim, seria interessante combinar os resultados da recomendação baseada em conteúdo com uma base de dados dos usuários. Isso resultaria em um modelo híbrido de recomendação, o qual não apenas auxiliaria na avaliação da qualidade dos resultados,

mas também na personalização das recomendações. Por exemplo, um curso recomendado com base no conteúdo acessado por usuários semelhantes poderia ser uma métrica valiosa de avaliação.

Referências

[1] Ruiqi Deng, Pierre Benckendorff, and Deanne Gannaway. Progress and new directions for teaching and learning in moocs. *Computers Education*, 129:48–60, 2019.

[2] Dhawal Shah. By the numbers: Moocs in 2021, Dec 2021.

[3] Robin Burke. Hybrid recommender systems: Survey and experiments. *User Modeling and User-Adapted Interaction*, 12(4):331–370, nov 2002.

[4] J. Ben Schafer, Dan Frankowski, Jon Herlocker, and Shilad Sen. *Collaborative Filtering Recommender Systems*, pages 291–324. Springer Berlin Heidelberg, Berlin, Heidelberg, 2007.

[5] Michael J. Pazzani and Daniel Billsus. *Content-Based Recommendation Systems*, pages 325–341. Springer Berlin Heidelberg, Berlin, Heidelberg, 2007.

[6] Michal Kompan and Mária Bieliková. Content-based news recommendation. In *E-Commerce and Web Technologies: 11th International Conference, EC-Web 2010, Bilbao, Spain, September 1-3, 2010. Proceedings 11*, pages 61–72. Springer, 2010.

[7] A. E. R. Souza. Um modelo para recomendação de cursos de especialização baseada no perfil profissional do candidato. *Dissertação de Mestrado*, 2013.

[8] Chungho Su. Designing and developing a novel hybrid adaptive learning path recommendation system (alprs) for gamification mathematics geometry course. *Eurasia Journal of Mathematics, Science and Technology Education*, 13(6):2275–2298, 2017.

[9] Hang Xue and Dongmei Zhang. A recommendation model based on content and social network. In *2019 IEEE 8th Joint International Information Technology and Artificial Intelligence Conference (ITAIC)*, pages 477–481, 2019.

[10] Robin Burke. *Hybrid Web Recommender Systems*, pages 377–408. Springer Berlin Heidelberg, Berlin, Heidelberg, 2007.

[11] Gediminas Adomavicius and Alexander Tuzhilin. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Trans. on Knowl. and Data Eng.*, 17:734–749, June 2005.

[12] Dekang Lin. An information-theoretic definition of similarity. In Jude W. Shavlik, editor, *Proceedings of the Fifteenth International Conference on Machine Learning (ICML 1998), Madison, Wisconson, USA, July 24-27, 1998*, pages 296–304. Morgan Kaufmann, 1998.

[13] Jiapeng Wang and Yihong Dong. Measurement of text similarity: A survey. *Information*, 11(9), 2020.

[14] Safia Kanwal, Sidra Nawaz, Muhammad Kamran Malik, and Zubair Nawaz. A review of text-based recommendation systems. *IEEE Access*, 9:31638–31661, 2021.

[15] Jae-wook Ahn, Peter Brusilovsky, Jonathan Grady, Daqing He, and Sue Yeon Syn. Open user profiles for adaptive news systems: Help or harm? pages 11–20, 05 2007.

[16] Antonio Roberto Paoli. Um estudo avançado do problema da maior subsequência comum. *Trabalho de Conclusão de Curso (Graduação) - Ciência da Computação (IC).*

[17] Loh S. Crocco G. Agrupamento de textos por similaridade para sistema de clipping web: clusterclipping. *Monografia de Conclusão de Curso da ULBRA*, 2010.

[18] Hinrich Schütze, Christopher D Manning, and Prabhakar Raghavan. *Introduction to information retrieval*, volume 39. Cambridge University Press Cambridge, 2008.

[19] Hugging Face Community. Sentence transform model, Disponível em: whttps://huggingface.co/sentence-transformers/all-MiniLM-L6-v2. Acesso em 08 de Outubro de 2023.

[20] Maxime De Bruyn, Ehsan Lotfi, Jeska Buhmann, and Walter Daelemans. MFAQ: a multilingual FAQ dataset. In Adam Fisch, Alon Talmor, Danqi Chen, Eunsol Choi, Minjoon Seo, Patrick Lewis, Robin Jia, and Sewon Min, editors, *Proceedings of the 3rd Workshop on Machine Reading for Question Answering*, pages 1–13, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[21] Prakhar Ganesh, Yao Chen, Xin Lou, Mohammad Ali Khan, Yin Yang, Hassan Sajjad, Preslav Nakov, Deming Chen, and Marianne Winslett. Compressing Large-Scale Transformer-Based Models: A Case Study on BERT. *Transactions of the Association for Computational Linguistics*, 9:1061–1080, 09 2021.

[22] AK Santra and C Josephine Christy. Genetic algorithm and confusion matrix for document clustering. *International Journal of Computer Science Issues (IJCSI)*, 9(1):322, 2012.

[23] Jorge Luis Cavalcanti Ramos, João Silva, Leonardo Prado, Alex Gomes, and Rodrigo Rodrigues. Um estudo comparativo de classificadores na previsão da evasão de alunos em ead. In *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*, volume 29, page 1463, 2018.

[24] Marina Sokolova, Nathalie Japkowicz, and Stan Szpakowicz. Beyond accuracy, f-score and roc: a family of discriminant measures for performance evaluation. In *Australasian joint conference on artificial intelligence*, pages 1015–1021. Springer, 2006.

[25] *Coursera Courses Dataset 2021.*, Disponível em: www.kaggle.com/datasets/khusheekapoor/coursera-courses-dataset-2021. Acesso em 12 de Setembro de 2022. Versão 1.

[26] Louis Hickman, Stuti Thapa, Louis Tay, Mengyang Cao, and Padmini Srinivasan. Text preprocessing for text mining in organizational research: Review and recommendations. *Organizational Research Methods*, 25(1):114–146, 2022.