Recomendação de Cursos para Qualificação em Tecnologia

Projeto Aplicado III
TURMA 201825166.000.04A

Gabriela de Lima Freitas
RA - 10416055
Gustavo Guarizzo Esteves
RA - 10424491
Luis Fernando do Lago Attarian
RA - 10089158
Maria Fernanda Salles Vasconcellos

Universidade Presbiteriana Mackenzie
Faculdade de Computação e Informática
Tecnologia em Ciência de Dados

São Paulo, 28 de maio de 2025

Sumário

1 Introdução

A qualificação profissional em tecnologia tornou-se essencial em um mundo cada vez mais digital e orientado por dados. Com o avanço das plataformas de ensino online, é possível aprender novas habilidades de forma autônoma, sem depender de cursos tradicionais. No entanto, a ampla oferta de conteúdos pode tornar difícil a escolha do caminho ideal para a capacitação, tornando necessária a criação de mecanismos que auxiliem os alunos a estruturarem seus estudos de maneira eficiente.

1.1 Contexto do Trabalho

Nos últimos anos, a área de tecnologia tem se expandido rapidamente, impulsionada pela crescente demanda por profissionais qualificados em desenvolvimento de software, ciência de dados, cibersegurança, inteligência artificial, entre outras especializações. Diferente de outras áreas do conhecimento, a tecnologia da informação permite que a formação profissional ocorra de forma autônoma, sem a necessidade de cursos presenciais formais, graças à ampla oferta de conteúdos disponíveis online.

Plataformas de ensino, tutoriais, vídeos, artigos e documentações técnicas fornecem um vasto acervo de conhecimento acessível a qualquer pessoa interessada em aprender. No entanto, a grande quantidade de materiais pode gerar desafios, como a dificuldade em selecionar os cursos mais adequados, a falta de um direcionamento estruturado e a sobrecarga de informações. Muitos estudantes e profissionais acabam se perdendo nesse processo, tornando o aprendizado ineficiente e desorganizado.

1.2 Motivação

A rápida evolução do setor de tecnologia exige que os profissionais estejam sempre atualizados, acompanhando novas tendências, linguagens de programação e metodologias de desenvolvimento. No entanto, sem uma orientação clara, muitas pessoas encontram dificuldades para definir um plano de estudos coerente com seus objetivos profissionais.

Além disso, a diversidade de plataformas e cursos disponíveis pode tornar a escolha confusa. Enquanto algumas plataformas oferecem cursos completos e bem estruturados, outras podem apresentar conteúdos desatualizados ou superficiais. Sem um critério bem definido para selecionar os materiais de estudo, há o risco de investir tempo e esforço em conteúdos que não agregam valor ao aprendizado.

Com isso, surge a necessidade de um sistema que auxilie na escolha de cursos de qualificação em tecnologia, proporcionando uma trilha de aprendizado estruturada e personalizada para cada perfil de usuário.

1.3 Justificativa

A principal justificativa para o desenvolvimento deste estudo é a necessidade de otimizar o processo de qualificação em tecnologia, tornando-o mais eficiente e acessível. Um sistema de recomendação baseado em ciência de dados pode contribuir significativamente para a formação de profissionais, ajudando-os a tomar decisões mais assertivas sobre os cursos a serem seguidos.

A ausência de um roteiro estruturado e a dificuldade de encontrar cursos de qualidade podem levar ao abandono do aprendizado ou a uma formação fragmentada. Com a implementação de um modelo de recomendação, espera-se facilitar o acesso a conteúdos

relevantes, levando em consideração fatores como experiência prévia, objetivos de carreira e preferências de aprendizado.

1.4 Objetivo geral e específicos

Desenvolver um sistema de recomendação de cursos de qualificação profissional em tecnologia, com foco em atender às necessidades de capacitação de indivíduos em áreas tecnológicas emergentes, contribuindo para a promoção da inclusão digital e o desenvolvimento sustentável conforme os Objetivos de Desenvolvimento Sustentável da ONU, especialmente no que diz respeito à educação de qualidade, ao trabalho decente e crescimento econômico e à redução das desigualdades.

1. Explorar e analisar uma base de dados de cursos, identificando padrões e características relevantes para a recomendação.

2. Criar um modelo de recomendação baseado em filtros colaborativos e/ou técnicas de aprendizado de máquina para sugerir cursos mais adequados ao perfil do usuário.

3. Desenvolver um protótipo funcional do sistema de recomendação, permitindo a entrada de dados do usuário e a sugestão de cursos personalizados.

4. Testar o sistema com usuários reais ou simulações para validar a usabilidade e a relevância das recomendações geradas.

2 Referencial Teórico

O uso de sistemas de recomendação tem ganhado destaque no contexto educacional, principalmente em plataformas de ensino online, como forma de personalizar a experiência de aprendizagem e orientar os usuários na escolha de conteúdos relevantes. Esses sistemas se baseiam em diferentes técnicas, desde métodos heurísticos até modelos baseados em aprendizado de máquina e processamento de linguagem natural (PLN).

2.1 Sistemas de Recomendação

Sistemas de recomendação são algoritmos que visam sugerir itens relevantes aos usuários com base em suas preferências, histórico ou perfil. Segundo Ricci et al. (2011), esses sistemas podem ser classificados em três categorias principais:

Baseados em conteúdo (Content-based filtering);

Baseados em filtragem colaborativa (Collaborative filtering);

Híbridos, que combinam ambas as abordagens.

No contexto educacional, sistemas de recomendação têm sido aplicados para sugerir cursos, trilhas de aprendizagem, materiais de leitura, entre outros (MANRIQUE et al., 2019). A principal vantagem está na personalização e no aumento do engajamento do aluno, tornando a jornada de aprendizado mais eficiente e direcionada.

A avaliação foi conduzida com dois conjuntos de dados rotulados manualmente, aplicando métricas como nDCG@10 (Normalized Discounted Cumulative Gain) e testes A/B com usuários reais. Os resultados mostraram melhora significativa na precisão das recomendações em relação a métodos tradicionais. O uso da métrica nDCG@K permitiu

quantificar não apenas a presença de itens relevantes nas recomendações, mas também sua posição na lista, priorizando itens relevantes nas primeiras posições.

Como resultado, o método apresentou ganhos significativos em relação a abordagens tradicionais, tanto em precisão quanto em posicionamento de itens relevantes. No entanto, o uso de sumarizadores avançados e conjuntos de dados anotados manualmente pode representar uma limitação prática para replicação em cenários com poucos recursos ou sem infraestrutura de anotação especializada.

2.2 Recuperação de Informação com BM25

O BM25 (Best Matching 25) é um modelo probabilístico amplamente utilizado em tarefas de recuperação de informação textual, como busca em documentos e mecanismos de recomendação. Ele foi proposto por Robertson et al. (1994) e é baseado na frequência dos termos em documentos e consultas. Sua principal característica é atribuir pesos diferenciados a termos raros e comuns, ajustando também pela extensão dos documentos, o que o torna eficaz na recuperação inicial de documentos potencialmente relevantes.

2.3 Reranqueamento com Modelos de Linguagem

Com os avanços em modelos de linguagem natural, especialmente com a arquitetura Transformer (VASWANI et al., 2017), surgiram novas possibilidades para melhorar a ordenação de documentos recuperados inicialmente por métodos lexicais como o BM25. Modelos como o T5 (Text-to-Text Transfer Transformer), proposto por Raffel et al. (2020), são capazes de interpretar consultas e documentos de forma semântica, superando limitações dos métodos lexicais.

O RankT5, derivado do T5, é uma especialização voltada para tarefas de reranqueamento supervisionado. Estudos recentes, como o de Ramesh e Bhandwal (2024), mostram que combinar uma etapa de recuperação inicial (BM25 ou GTR) com uma etapa de reranqueamento neural (RankT5) pode aumentar significativamente a precisão de sistemas de recomendação, inclusive em contextos educacionais.

2.4 Aplicações em Plataformas de Ensino

Diversos trabalhos têm investigado o uso de recomendadores em plataformas educacionais. Tang e McCalla (2005) exploram sistemas que adaptam recomendações conforme o progresso e preferências do aluno. Já Liu et al. (2019) mostram a eficácia de modelos baseados em aprendizado profundo para sugerir conteúdos em MOOCs (Massive Open Online Courses).

Esses sistemas enfrentam desafios como:

escassez de dados rotulados,

ruído nas descrições dos cursos,

dificuldade em modelar preferências implícitas dos usuários.

Para lidar com essas questões, abordagens que integram PLN, aprendizado supervisionado e estratégias híbridas têm se mostrado promissoras.

y

3 Metodologia

3.1 Descrição Sistemática das Técnicas Utilizadas

Este trabalho propôs um sistema de recomendação de cursos em tecnologia estruturado em um pipeline de duas etapas principais: recuperação inicial e reranqueamento supervisionado. A seguir, descrevemos sistematicamente as técnicas empregadas em cada fase.

3.1.1 Visão Geral

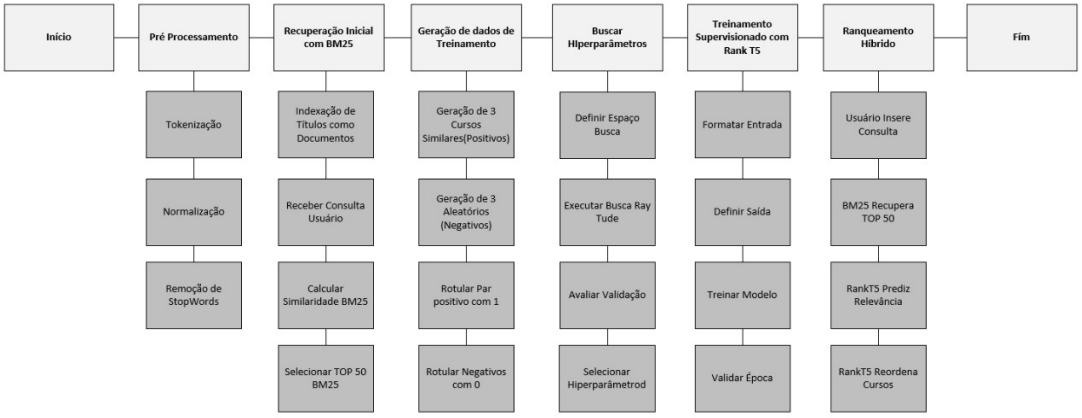

Figura 1: Visão geral metodologia.

3.1.2 Pré-processamento de Textos

Antes da aplicação dos algoritmos de recuperação e reranqueamento, foi realizada uma limpeza textual dos títulos dos cursos:

- **Tokenização**: As palavras dos títulos foram segmentadas utilizando a função `word_tokenize` da biblioteca NLTK.
- **Normalização**: Os textos foram convertidos para letras minúsculas (*lowercasing*).
- **Remoção de *Stopwords***: Foram eliminadas palavras irrelevantes utilizando uma lista customizada de *stopwords*, ampliada para termos frequentes em cursos online (como "course", "beginner", "tutorial", entre outros).

3.1.3 Recuperação Inicial com BM25

O primeiro estágio do pipeline utilizou o algoritmo BM25 (Best Matching 25), um modelo de recuperação de informações baseado em termos de busca:

- Cada título de curso foi indexado como um documento.
- Dada uma consulta textual do usuário, o BM25 calculou uma pontuação de similaridade para todos os cursos, baseada na frequência dos termos, no comprimento dos documentos e na raridade dos termos.
- Foram selecionados os 50 cursos com maiores pontuações BM25 para cada consulta, compondo uma lista inicial de candidatos.

3.1.4 Geração Automática de Dados de Treinamento

Visando a criação de um conjunto supervisionado para treinamento do modelo de reranqueamento:

- Para cada título, os três cursos mais similares (excetuando o próprio) foram selecionados como pares **positivos** (relevantes).
- Três cursos aleatórios com baixa similaridade (além da 50ª posição no ranking BM25) foram selecionados como pares **negativos** (não relevantes).
- Cada par (consulta, curso) foi formatado como um exemplo textual no formato ‘‘`Query: [consulta] Course: [curso]`’’, rotulado com “1” (relevante) ou “0” (não relevante).

3.1.5 Divisão de Dados em Treinamento, Validação e Teste

O conjunto de exemplos gerados foi dividido em três subconjuntos:

- **Treinamento** (70%): Utilizado para ajustar os pesos do modelo RankT5.
- **Validação** (15%): Utilizado para monitorar o desempenho durante o treinamento e prevenir overfitting.
- **Teste** (15%): Reservado para a avaliação final de desempenho dos modelos BM25 e RankT5.

3.1.6 Busca pelos Melhores Hiperparâmetros

Com o objetivo de otimizar o desempenho do modelo RankT5, foi realizada uma busca pelos melhores hiperparâmetros:

- A busca foi conduzida utilizando a técnica de **Hyperparameter Search** com suporte do backend Ray Tune.
- Parâmetros ajustados:
  - Taxa de aprendizado (**learning rate**).
  - Tamanho do batch (**batch size**).
  - Número de épocas de treinamento (**epochs**).
  - Fator de decaimento de peso (**weight decay**).
  - Número de passos de aquecimento (**warmup steps**).
- Critério de seleção: **minimização da perda de validação (validation loss)**.

3.1.7 Treinamento Supervisionado com RankT5

O modelo RankT5, baseado na arquitetura T5 (Text-to-Text Transfer Transformer), foi treinado supervisionadamente para aprender padrões de relevância:

- Entrada: Texto concatenado contendo a consulta e o título do curso.
- Saída esperada: Rótulo “1” (relevante) ou “0” (não relevante).
- Treinamento: Realizado por meio da biblioteca Hugging Face Transformers, utilizando otimização supervisionada e validação automática por época.

3.1.8 Reranqueamento Híbrido: BM25 + RankT5

Na fase de recomendação real:

1. **Recuperação**: O BM25 recupera os 50 cursos mais similares para a consulta do usuário.
2. **Reranqueamento**: O RankT5 prediz a relevância de cada par (consulta, curso), reordenando os cursos de acordo com as pontuações preditas.

Essa abordagem híbrida permite combinar a eficiência lexical do BM25 com a capacidade semântica do modelo RankT5.

4 Resultados

Nesta seção, são apresentados e analisados os principais resultados obtidos a partir do desenvolvimento e avaliação do sistema de recomendação proposto, com foco na comparação entre as técnicas BM25 e RankT5. Os resultados foram organizados com base nas métricas de avaliação aplicadas, identificando os pontos fortes e limitações de cada abordagem.

4.1 Desempenho do Modelo RankT5 em Relação ao BM25

A avaliação comparativa entre os métodos BM25 e RankT5 foi realizada utilizando a métrica nDCG@5 (Normalized Discounted Cumulative Gain), que leva em conta tanto a relevância quanto a posição dos cursos recomendados em relação à consulta. O modelo RankT5 demonstrou desempenho superior na maioria das consultas analisadas.

O modelo RankT5 apresentou melhores resultados principalmente em consultas genéricas ou ambíguas, graças à sua capacidade de capturar relações semânticas. Já o BM25 teve desempenho mais competitivo em consultas com forte correspondência lexical direta, como nomes de linguagens de programação.

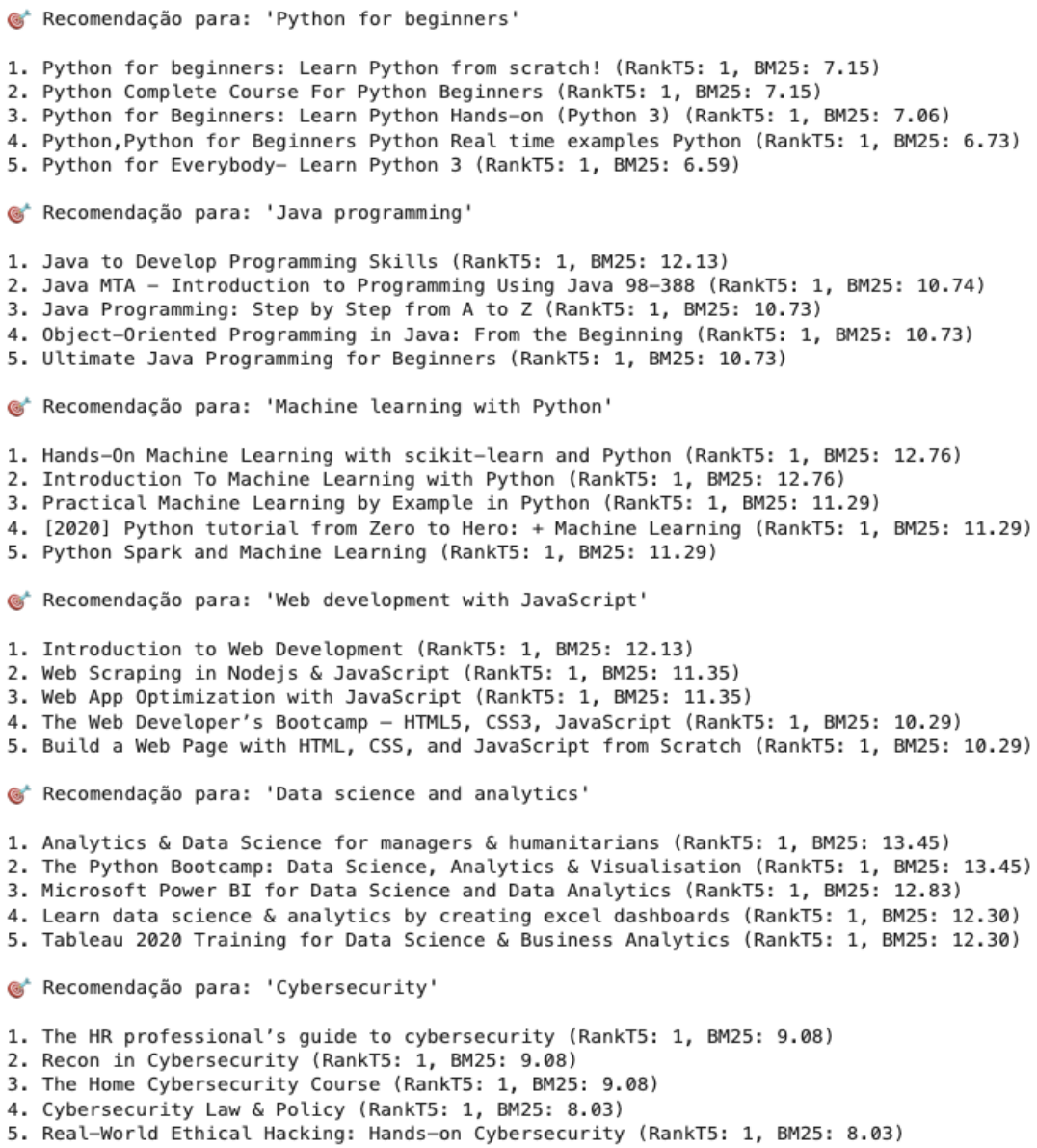

Figura 2: Exemplos de execução

```
=== Resultados ===
nDCG@5    - BM25   : 0.2819
nDCG@5    - RankT5 : 0.9934
MAP       - BM25   : 0.2330
MAP       - RankT5 : 0.9923
MRR       - BM25   : 0.2314
MRR       - RankT5 : 0.9942
```

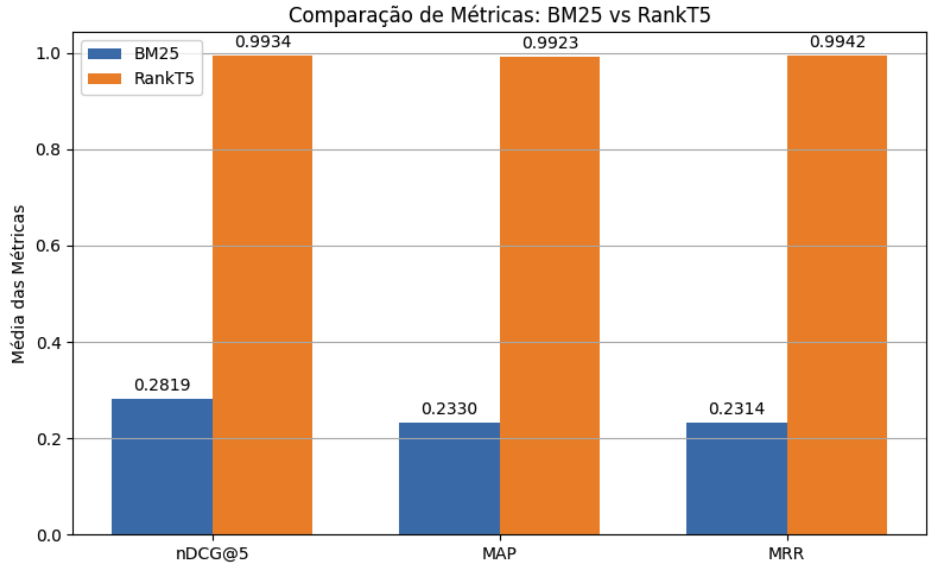

Figura 3: Comparativo entre as médias.

4.2 Pontos Positivos e Negativos das Técnicas Utilizadas

RankT5

Pontos Positivos:

- Alta capacidade de compreensão semântica.
- Superou o BM25 em todas as métricas avaliadas (nDCG@5, MAP, MRR).
- Maior precisão na ordenação dos cursos mais relevantes nas primeiras posições.

Pontos Negativos:

- Requer treinamento supervisionado, com maior custo computacional.
- Depende de um volume considerável de dados anotados (mesmo que gerados automaticamente).
- Apresenta variações de desempenho se mal ajustado, devido à sensibilidade aos hiperparâmetros.

BM25

Pontos Positivos:

- Simples e eficiente, com baixa demanda computacional.
- Boa performance em consultas com correspondência direta de termos.
- Adequado para recuperação inicial de candidatos.

Pontos Negativos:

- Limitado em consultas sem correspondência lexical exata.
- Ignora aspectos semânticos, o que compromete a qualidade das recomendações em tópicos mais subjetivos.
- Depende muito do pre-processamento remoção de stop words, lemantização.

4.3 Considerações sobre a Avaliação

A avaliação foi realizada com uma base de dados rotulada automaticamente, o que facilitou a escalabilidade do experimento, mas pode introduzir viés nos rótulos. Além disso, a métrica *nDCG* assume que os cursos mais bem ranqueados são os mais relevantes, o que pode não refletir com precisão a percepção do usuário final. Assim, os resultados quantitativos devem ser interpretados com cautela.

5 Conclusão e Trabalhos futuros

Este estudo teve como objetivo comparar técnicas de ranqueamento aplicadas à recomendação de cursos da categoria *IT & Software*, utilizando dados da plataforma Udemy. Foram implementados e avaliados dois métodos: o tradicional BM25 e o modelo supervisionado RankT5, mais recente e orientado por aprendizado profundo.

Os resultados indicaram que o RankT5 apresentou desempenho superior ao BM25 em todas as métricas consideradas (nDCG@5, MAP e MRR), demonstrando maior precisão na ordenação dos cursos mais relevantes nas primeiras posições. Isso evidencia o potencial de métodos baseados em aprendizado profundo para melhorar a experiência do usuário em sistemas de recomendação.

Contudo, uma das principais dificuldades enfrentadas foi a avaliação objetiva da qualidade do sistema de recomendação. Apesar das métricas quantitativas fornecerem uma base comparativa, elas não capturam nuances subjetivas, como a real satisfação do usuário ou a aplicabilidade dos cursos recomendados em diferentes contextos. Além disso, a rotulação automática dos dados pode introduzir viés e limitar a validade externa dos resultados.

A avaliação de sistemas híbridos e complexos, como o desenvolvido neste trabalho, representa um desafio adicional. Nesses casos, os efeitos de cada componente (modelo semântico, modelo lexical, heurísticas de agrupamento, entre outros) tendem a se sobrepor de forma não trivial, dificultando a interpretação isolada de seu impacto no resultado final. Isso exige abordagens de avaliação mais refinadas e multifacetadas.

5.1 Pesquisas futuras

- A inclusão de métodos de validação manual com **feedback direto de usuários**, como questionários de relevância ou testes A/B em ambiente controlado.
- O uso de **modelos generativos** para fornecer recomendações explicáveis e resumos personalizados.
- O desenvolvimento de estratégias de **personalização baseadas em perfis de usuário**, considerando comportamento e histórico.

- A investigação do impacto de diferentes representações textuais, como *contextual embeddings* (BERT, SBERT).
- A realização de testes voltados ao problema de **cold-start**, tanto para novos cursos quanto para novos usuários.
- A aplicação de **aprendizado por reforço**, em que o sistema evolui conforme interações contínuas com os usuários.
- A análise de **viés e justiça** nas recomendações geradas, garantindo diversidade e equidade nos resultados.
- A comparação com outros algoritmos supervisionados de ranking, como RankNet, LambdaMART e LightGBM-Rank.

Em suma, embora o RankT5 se destaque tecnicamente, a eficácia real de um sistema de recomendação depende de sua aceitação e utilidade prática, o que reforça a importância de avaliações contínuas, centradas no usuário e sensíveis à complexidade dos sistemas modernos.

6 GitHub

`https://github.com/lattarian/projeto_aplicado_III`

7 Referências Bibliográficas

Referências

[1] KAGGLE. Udemy Courses Dataset. Disponível em: `https://www.kaggle.com/datasets/andrewmvd/udemy-courses`. Acesso em: 4 mar. 2025.

[2] RAMESH, Shubham; BHANDWAL, Deepika. Efficient Course Recommendations with T5-based Ranking and Summarization. arXiv preprint arXiv:2406.19018, 2024. Disponível em: `https://arxiv.org/abs/2406.19018`. Acesso em: 4 abr. 2025.

[3] RAFFEL, Colin et al. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. Journal of Machine Learning Research, v. 21, n. 140, p. 1–67, 2020. Disponível em: `http://jmlr.org/papers/v21/20-074.html`. Acesso em: 4 abr. 2025.