Otimizando a produção acadêmica: resumos em Políticas Públicas com LLM

Sumário Executivo

SICSS FGV/ECMI BRASIL 2024

Solução desenvolvida como produto do curso de Ciência Social Computacional realizado de 01 a 11.07.2024

Grupo: Claudia Monteiro ▪ Daniel Oliveira ▪ Felipe Lamarca
Gabriellen Carmo ▪ Lauriano Benazzi ▪ Marcelle Amaral ▪ Talita Ribeiro.

Project Details: o ABSTRACT

A ideia

Resumos, em inglês “abstracts”, muitas vezes não passam do plano abstrato e pouco resumem o artigo. Mesmo assim, é o ponto de partida de qualquer acréscimo marginal às Ciências. Uma boa revisão de literatura invariavelmente começa pelo resumo do artigo.

A ideia da equipe foi o desenvolvimento de *scraper* para a base de dados **SciELO**, com a seleção de artigos por temas, a partir de modelo *LLM* (*Large Language Models*) que faz o resumo das seções do artigo, compilando-as ao final.

The elevator talk

Um resumo é uma versão condensada do artigo, com um número pré-estabelecido de palavras. Como em uma rápida conversa de elevador, o autor precisa resumir “o que faz”, “como faz”, os resultados do que está fazendo e o impacto de seu trabalho. Segundo Annesley (2010), entre as características de um bom resumo, estão:

- É completo ao ponto de não ser necessário ir ao artigo;
- Identifica a hipótese, pergunta ou objetivo e responde-os;
- Possui os mesmos termos que o título e a introdução;
- Espelha a ordem do texto principal;

Em síntese, é a parte mais lida de um artigo. Porém, resumos não costumam seguir as sugestões listadas acima, e a falta de padrão torna-se uma celeuma para pesquisadores em suas incursões nos repositórios de produções científicas.

É um LLM é um revisor confiável?

Ao analisarem as habilidades do GPT-3.5 e do GPT-4 da OpenAI, Zhou, Chen e Yu (2024) concluem que, aplicadas à pesquisa científica, as ferramentas:

- Conseguem fornecer scores com significado e avaliar frases individualmente;
- Raramente estão totalmente corretos;
- Não são críticos o suficiente;
- Possuem dificuldade com textos longos ignorando detalhes técnicos importantes;
- Falham na análise da qualidade.

Scielo: literatura da “periferia”

Considerado como um dos principais repositórios científicos mundiais o mais respeitado e utilizado no Brasil, nas palavras de Meneghini (2003) o **SciELO** é uma “Base de dados dedicada à literatura científica excluída da literatura da corrente principal”. Alguns destaques:

- Criado em 1996, em momento inicial da web no Brasil;
- Aumentou significativamente a visibilidade da produção científica brasileira;
- Objetivos foi o de construir uma base de dados com indicadores para avaliar a produção nacional de conhecimento;
- Os critérios e políticas do **SciELO** estão disponíveis no repositório: versão Set. 2022.

Proposta inicial da equipe

- Desenvolver um modelo que utilizasse técnicas de raspagem de dados para extrair informações da base de dados SciELO, combinadas com algoritmos de *LLM,* para a geração de resumos.
- Resultados seriam organizados em uma tabela de Excel, com as seguintes categorias: título, abstract, texto completo, resumo da LLM.
- O usuário forneceria palavras-chave e aplicaria filtros para definir recorte temporal, tais como ordenamento e número de artigos a serem extraídos.
- O modelo forneceria os artigos sobre o tema em ordem crescente, de acordo com a escolha do usuário : “data”; “relevância”; “número de citações”; ou “número de acessos”.
- As métricas do **SciELO** selecionariam os primeiros "n" artigos relacionados à pesquisa feita pelo usuário.
- A partir da seleção dos artigos conforme os filtros utilizados, e processando através de *LLM*, o modelo geraria um texto sintetizado com as principais informações dos artigos.

Processo de Desenvolvimento

Realizamos uma análise da página inicial do **SciELO** para explorar como incluir os parâmetros de busca. Com isso, começamos a implementação:

1. Estruturação de script para raspar (*data scraping*) *d*os seguintes dados: “título”, “autores” e ”resumo”, mais o texto completo;
2. Desenvolvimento de script que puxava o texto integral. No entanto, como os modelos do *Huggingface* tinham limitação do número de tokens no input, não era possível usar o texto completo;
3. Optou-se então pela raspagem do texto em seções, para numa segunda etapa passar pela *LLM*. O objetivo era passar as seções individualmente, gerando pequenos resumos de cada tópico do texto, os quais seriam juntados no final;
4. Foram testados 3 modelos, mas nenhum apresentou resultados satisfatórios. A saber:

a. **facebook/bart-large-cnn;**
b. **stjiris/t5-portuguese-legal-summarization;**
c. **thaymison/flan-t5-portuguese-small-summarization**.

5. Migrou-se então para o *Ollama*, que dá acesso ao **Llama3 8b** (e outros mais potentes), modelo considerado o "estado-da-arte" quando se trata de *LLMs open-source*.

   Além da ferramenta permitir inputar o texto completo, retornou excelentes resumos.

Open Source

O projeto está disponível no GitHub e inclui instruções de uso:

https://github.com/felipelmc/SciELO-Summarizer

Exemplo de output: resumo original x resumo LLM

Artigo: SÁ, Evelin Naked de Castro; RABELLO, Maria Celina Guimarães. Administração por convênios, um instrumento gerencial de políticas: uma análise dos convênios da Secretaria de Estado da Saúde de São Paulo - 1987. **Revista de Saúde Pública**, São Paulo, v. 22, n. 2, p. 164-177, 1988. Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0034-89101988000200011&lang=pt. Acesso em: 11 jul. 2024.

Resumo original:

Foi realizada análise dos 1.141 convênios vigentes no primeiro semestre de 1987, celebrados pela Secretaria de Saúde do Estado de São Paulo, segundo os objetivos dos convênios e entidades conveniadas. Foram consideradas as potencialidades do convênio como instrumento de execução de políticas públicas de saúde e as necessidades de resposta organizacional da Secretaria para definição, acompanhamento, avaliação e controle dos convênios. Foram recomendadas algumas medidas para aumentar o rendimento de ações conveniadas para aperfeiçoar a expressão orçamentário-financeira da administração por convênios, com base num plano abrangente das ações de saúde.

Resumo LLM:

O texto discute o convênio-modelo de municipalização da saúde, que foi autorizado pelo Decreto n.º 27.140/87. O autor destaca que a autonomia política dos municípios pode levar a situações em que os municípios não desejem participar do convênio ou o rescindam. Além disso, há dificuldades organizacionais para a execução do convênio, pois muitos municípios não têm implantados seus Conselhos de Saúde e órgãos responsáveis previstos no acordo.

O autor sugere que a Secretaria de Estado da Saúde identifique unidades de comando que sejam responsáveis pela execução dos convênios e use as informações para um planejamento

integral. Além disso, é necessário criar uma entidade civil ou comercial para dar execução aos termos do convênio, o que permitiria manter o convênio com simples pacto de cooperação.

O texto também destaca a necessidade de avaliação das ações comunitárias de acompanhamento e a participação da comunidade na implementação dos serviços de saúde. As vantagens do convênio estão na utilização de uma estrutura formal e técnicas gerenciais dinâmicas, enquanto as desvantagens estão na diluição de responsabilidade distribuída pelos vários níveis de governo.

Em resumo, o texto discute os desafios da implementação do convênio-modelo de municipalização da saúde e sugere soluções para superar esses obstáculos.

Resultados

- O script de raspagem dos dados funciona bem para as páginas padrão do **SciELO**.
- Embora a proposta inicial fosse a de gerar um resumo de todos os artigos em um único texto, tal resultado ainda não foi possível no estágio atual do projeto, demandando mais algumas movimentações.
- Embora os primeiros modelos não tenham alcançado resultados satisfatórios, com o uso do **Llama3** o retorno foi de resumos (ou síntese dos resumos) com alta qualidade textual e semântica, com mais objetividade e frequentemente (em altos índices) mais informativos do que os resumos dos próprios artigos.
- Em um computador com uma GTX 1650, cada artigo teve a extração de dados e foi resumido em ~4 minutos. Em computadores mais potentes ou no Google Collab, por exemplo, é esperado que a execução seja mais rápida.

Melhorias futuras

Como melhorias futuras do projeto, destacamos:

- Estruturação de um resumo final que contemple todos os artigos extraídos.
- Elaborar rotinas para o tratamento de exceções quando não é possível extrair nenhum artigo do **SciELO** sobre determinado tema ou busca (por exemplo, quando as *queries* de busca não retornam qualquer artigo).
- Incluir na aplicação final um notebook interativo que seja executado apenas a partir do arquivo main.py, para facilitar a execução no Google Collab. No atual estágio é possível executar a proposta aplicando função por função em um notebook. No entanto, a versão interativa que visa receber os parâmetros de maneira mais amigável, só está disponível na execução via terminal.

Referências

Annesley, T. M. (2010). The abstract and the elevator talk: a tale of two summaries. Clinical chemistry, 56(4), 521-524. DOI:10.1373/clinchem.2009.142026

Meneghini, R. (2003). O projeto Scielo (Scientific Electronic Library on Line) e a visibilidade da literatura científica" Periférica". *Química Nova*, *26*, 155-156. 10.1590/S0100-40422003000200001

Zhou, R., Chen, L., & Yu, K. (2024, May). Is LLM a Reliable Reviewer? A Comprehensive Evaluation of LLM on Automatic Paper Reviewing Tasks. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024) (pp. 9340-9351). Disponpivel em https://aclanthology.org/2024.lrec-main.816. Acesso em 11 jul. 2024.