Departamento de Ciência da Computação
Arthur Pontes Nader

Modelos Generativos para Recuperação de Informação

Orientador: Rodrygo Luis Teodoro Santos

Belo Horizonte
10 de abril de 2023

1) Introdução

Os modelos de inteligência artificial generativa, como GPT-3 e o recém lançado GPT-4, têm sido amplamente utilizados nas mais diversas tarefas humanas. Desde resumir um tópico de interesse até a criação de códigos em diversas linguagens de programação, essas ferramentas apresentam potencial para alavancar o desenvolvimento de muitos campos da Ciência da Computação.

Uma dessas áreas é a Recuperação de Informação, que, de acordo com (BAEZA-YATES, RIBEIRO-NETO, 2011) no livro “Modern Information Retrieval”, é uma área de pesquisa que lida com o armazenamento de documentos e a recuperação da informação associada a eles a partir de uma necessidade de informação do usuário. A importância desse campo se torna ainda evidente quando se considera a recuperação de informação da Web, que teve um grande avanço desde então. Duas tarefas frequentes dessa área são:

- Extração de informação: consiste em identificar entidades, relacionamentos e eventos em um texto.
- Web Scraping: criação de modelos automatizados que interagem com uma webpage, por exemplo, clicando em botões e preenchendo formulários, com intuito de baixar documentos, recuperar tabelas, etc.

Essas tarefas podem ser facilmente realizadas no cotidiano por humanos, mas demandam muito tempo e se tornam repetitivas para grandes volumes de dados a serem recuperados. Para serem automatizadas, frequentemente é necessário conhecimento sobre algoritmos e linguagens de programação.

O fato é que esses recentes modelos generativos podem estar alterando esse cenário, em que essas tarefas poderão ser feitas por pessoas que não possuem todo o conhecimento técnico necessário, por meio de uma interação direta em linguagem natural com o modelo, tal como: “Identifique quem é o autor da lei mencionada no texto” ou “Clique no botão Download”.

Assim, o principal objetivo desse trabalho é reproduzir essas tarefas utilizando esses modelos generativos e avaliar os resultados obtidos, comparando com outros métodos utilizados até então. Esses conhecimentos serão bastante úteis e podem servir de base para o Projeto Orientado em Computação 2.

2) Referencial Teórico

As redes neurais têm sido o método de inteligência artificial mais popular em modelos de aprendizado de máquina atualmente. Dentre os vários tipos de redes, a arquitetura Transformer, exposta em (VASWANI, et al., 2017), tem sido muito utilizada em modelos para processamento de linguagem natural, sendo inclusive citada em (RADFORD, et al., 2018), um dos artigos iniciais que foram a base para o surgimento dos modelos GPT.

A extração de informação estruturada é um dos desafios da inteligência artificial. Em um texto, podem ter várias entidades relacionadas com um mesmo fato, e, o que é facilmente compreendido por um humano, às vezes se torna uma tarefa árdua para um computador.

Uma das estratégias que usa modelos generativos para resolver isso é exposta em (WEI, et al., 2023) em que os autores apresentam a ferramenta ChatIE, um framework baseado em ChatGPT. Essa ferramenta, por meio de um modo interativo, é capaz de decompor complexas tarefas de extração de informação em várias partes e compor os resultados de cada uma em um resultado final estruturado.

Já para a tarefa de web scraping, há ferramentas bem recentes que usam esses modelos generativos que, apesar de não possuírem artigo científico relacionado, merecem a devida atenção. Um exemplo de ferramenta baseadas em GPT-3 é o ExtractGPT, enquanto em GPT-4 tem-se Scrapeghost e TaxyAI. Todas essas ferramentas possuem uma ampla descrição em seus repositórios/sites.

3) Metodologia

Para facilitar a obtenção dos resultados, primeiramente é necessário integrar o ChatGPT a um ambiente Python, linguagem de programação que será usada no desenvolvimento das tarefas. Isso pode ser feito utilizando a biblioteca requests e enviados solicitações HTTP à API. A partir daí, faz-se requisições de extração de informação à ferramenta tal qual feito pelos pesquisadores criadores do ChatIE. Pode-se comparar os resultados obtidos com outras bibliotecas, como a que utiliza redes neurais tradicionais para extração de informação, spaCy.

Já para as tarefas de Web Scrape, pode-se utilizar a própria interface anterior para solicitar à ferramenta criação de códigos em BeautifulSoup e Selenium. Apesar disso, GPT-3 e GPT-4 são mais poderosas que ChatGPT. Por isso, é interessante explorar o uso de ferramentas como ExtractGPT, Scrapeghost e TaxyAI para essas tarefas específicas. Essas ferramentas já possuem bibliotecas em Python e podem ser integradas ao ambiente de desenvolvimento.

Para avaliar a eficácia das ferramentas de Web Scrape, pode-se comparar os resultados obtidos com aqueles realizados usando diretamente um código escrito por mim. É importante lembrar que a qualidade dos dados extraídos pode ser afetada pela qualidade do código gerado pelas ferramentas, sendo necessário realizar testes e ajustes na configuração dos parâmetros de cada ferramenta para obter os melhores resultados possíveis.

É de grande interesse também avaliar a usabilidade de alternativas gratuitas ao GPT-4, tal como a ferramenta gpt4all. Isso é necessário pois, caso as estratégias de Scrape citadas se mostrem uma boa solução para facilitar a recuperação de informação, a ampliação do seu uso proporcionaria gastos consideráveis a longo prazo.

4) Resultados Esperados

Espera-se que ao final das atividades, tenha-se obtido um conhecimento mais profundo da arquitetura Transformer e dos modelos generativos que estão sendo muito usados atualmente.

Além disso, busca-se atingir resultados semelhantes aos obtidos em (WEI, et al., 2023) para avaliar a usabilidade para automatização de extração de informação para usuários que não possuem conhecimento específico em computação.

Espera-se também que as ferramentas usadas para Scrape sejam bastante intuitivas, servindo, como por exemplo, para uso na automatização da coleta de dados na Internet por diferentes grupos de interesse. Cada uma dessas ferramentas possui especificidades e são direcionadas a diferentes casos de uso. O uso combinado delas pode ser um grande impulso para facilitar a coleta de informação, frente à crescente complexidade e volume de dados da Web.

Esses conhecimentos e resultados obtidos possibilitarão uma melhor definição de ações a serem tomadas para continuidade da exploração de modelos generativos no Projeto Orientado em Computação 2.

5) Etapas e Cronogramas

As etapas e cronogramas para desenvolvimento das atividades foram divididas semanalmente da seguinte forma:

Semana	Atividades
1	Revisão de bibliografia
2	Estudo detalhado sobre as arquiteturas
3	Reprodução do ChatIE
4	Avaliação dos resultados obtidos Aprimorar o modelo, dependendo dos resultados
5	Preparação para apresentação parcial
6	Início da escrita do relatório final ChatGPT para geração de código de Web Scrape
7	Uso e avaliação do ExtractGPT
8	Uso e avaliação do Srapeghost
9	Uso e avaliação do TaxyAI
10	Análise de ferramenta alternativa: gpt4all
11	Comparação entre os modelos de Scrape
12	Confecção do Pôster

6) Referências Bibliográficas

R. Baeza-Yates, B. Ribeiro-Neto. Modern Information Retrieval. 2011. 2nd ed. New York. Addison-Wesley.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In Advances in Neural Information Processing Systems, pages 6000–6010, 2017.

A. Radford, K. Narasimhan, T. Salimans, I. Sutskever. Improving Language Understanding by Generative Pre-Training. OpenAI, 2019. Disponível em: https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf. Acesso em: 07 abr. 2023.

X. Wei, et al. Zero-Shot Information Extraction via Chatting with ChatGPT. Disponível em: https://arxiv.org/pdf/2302.10205.pdf. Acesso em: 07 abr. 2023.