ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

NGÔ PHÚ THỊNH

Deep Learning in the Legal System of Vietnam: Opportunities and Challenges

LUẬN VĂN TỐT NGHIỆP

Tp. Hồ Chí Minh - 2023

ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

NGÔ PHÚ THỊNH

Deep Learning in the Legal System of Vietnam: Opportunities and Challenges

LUẬN VĂN TỐT NGHIỆP
CHUYÊN NGÀNH KHOA HỌC DỮ LIỆU

NGƯỜI HƯỚNG DẪN KHOA HỌC
PGS. Nguyễn Thanh Bình

Tp. Hồ Chí Minh - 2023

Lời cảm ơn

Khóa luận tốt nghiệp chuyên ngành Khoa học Dữ liệu với đề tài Deep Learning in the Legal System: Opportunities and Challenges là kết quả cố gắng của bản thân tôi sau 4 năm học tập tại Khoa Toán - Tin học, trường Đại học Khoa học Tự nhiên, ĐHQG-TPHCM và được sự giúp đỡ, động viên từ quý thầy cô, bạn bè và người thân. Qua đây tôi xin gửi lời cảm ơn chân thành đến những người đa giúp đỡ tôi trong quá trình học tập - nghiên cứu khoa học vừa qua.

Lời đầu tiên, tôi xin trân trọng gửi đến PGS. Nguyễn Thanh Bình lời cảm ơn chân thành và sâu sắc nhất. Thầy không chỉ là người tạo cảm hứng cho tôi đến với chuyên ngành Khoa học Dữ liệu, mà còn là người nhiệt tình hướng dẫn cũng như cung cấp cho tôi những kiến thức, tài liệu khoa học cần thiết phục vụ cho đề tài này.

Tôi rất vui mừng và biết ơn khi được công ty King Attorney hỗ trợ trong quá trình nghiên cứu luận văn của tôi. Công ty King Attorney là một đơn vị uy tín và chuyên nghiệp trong lĩnh vực luật. Tôi cũng muốn gửi lời cảm ơn chân thành đến anh Đỗ Hữu Chiến, người đã truyền cho tôi nhiều cảm hứng và ý tưởng sáng tạo cho luận văn của tôi. Anh Đỗ Hữu Chiến là một người thầy tuyệt vời và một người bạn đồng hành đáng tin cậy.

Tôi cũng tri ân đến bạn Lê Huy Hoàng, đã hỗ trợ tôi rất nhiều từ giai đoạn lên ý tưởng đến việc đề xuất những phương pháp hữu ích cho bài luận này.

Nhân dịp này tôi xin gửi lời cảm ơn đến quý thầy cô ở khoa Toán - Tin học đã nhiệt tình truyền đạt cho tôi những kiến thức từ cơ bản đến chuyên sâu trong suốt quá trình học tập tại Khoa. Những kiến thức tích lũy được ở Khoa đã giúp tôi có nền tảng vững vàng cho việc phát triển tương lai sau này.

Cuối cùng, tôi xin cảm ơn gia đình, người thân, bạn bè đã luôn bên cạnh, ủng hộ, động viên.

Tp.HCM, ngày 12 tháng 6 năm 2023

Tác giả

Ngô Phú Thịnh

Mục lục

1. Lời nói đầu

Giới thiệu luận văn…

Nội dung luận văn bao gồm xx chương:

Giới thiệu luận văn…

Nội dung luận văn bao gồm xx chương:

1.1. Động lực

2. Kiến thức chuẩn bị

2.1. Hệ thống văn bản quy phạm pháp luật

Hệ thống những văn bản quy phạm pháp luật là hình thức biểu hiện mối liên hệ bên ngoài của pháp luật thông qua các loại văn bản quy phạm pháp luật có giá trị cao thấp khác nhau được các cơ quan Nhà nước có thẩm quyền ban hành theo một trình tự, thủ tục do pháp luật quy định, nhưng đều tồn tại trong thể thống nhất.

Các văn bản quy phạm pháp luật tạo nên hệ thống pháp luật các văn bản quy phạm pháp luật có những đặc điểm:

- Nội dung của các văn bản quy phạm pháp luật là các quy phạm pháp luật do các cơ quan Nhà nước có thẩm quyền ban hành.
- Các văn bản quy phạm pháp luật đều có tên gọi khác nhau (luật, nghị định, pháp lệnh…) do Hiến pháp quy định. Giá trị pháp lý của chúng cao thấp khác nhau do vị trí của cơ quan Nhà nước trong bộ máy nhà nước có quy định.
- Các văn bản quy phạm pháp luật có hiệu lực trong không gian (hiệu lực trong phạm vi khu vực lãnh thổ) và hiệu lực theo thời gian (bắt đầu có hiệu lực hay hết hiệu lực), hiệu lực theo nhóm người (có hiệu lực đối với nhóm người này và không có hiệu lực đối với nhóm người khác).

Theo Hiến pháp năm 2013 [1], Luật Ban hành văn bản quy phạm pháp luật năm 2015 [2] quy định hệ thống những văn bản quy phạm pháp luật gồm các văn bản có giá trị pháp lý như sau:

1. Hiến pháp.
2. Bộ luật, luật, nghị quyết của Quốc hội.
3. Pháp lệnh, nghị quyết của Ủy ban thường vụ Quốc hội; nghị quyết liên tịch giữa Ủy ban thường vụ Quốc hội với Đoàn Chủ tịch Ủy ban trung ương Mặt trận Tổ quốc Việt Nam; nghị quyết liên tịch giữa Ủy ban thường vụ Quốc hội, Chính phủ, Đoàn Chủ tịch Ủy ban trung ương Mặt trận Tổ quốc Việt Nam.
4. Lệnh, quyết định của Chủ tịch nước.

5. Nghị định của Chính phủ; nghị quyết liên tịch giữa Chính phủ với Đoàn Chủ tịch Ủy ban trung ương Mặt trận Tổ quốc Việt Nam.
6. Quyết định của Thủ tướng Chính phủ.
7. Nghị quyết của Hội đồng Thẩm phán Tòa án nhân dân tối cao.
8. Thông tư của Chánh án Tòa án nhân dân tối cao; thông tư của Viện trưởng Viện kiểm sát nhân dân tối cao; thông tư của Bộ trưởng, Thủ trưởng cơ quan ngang bộ; quyết định của Tổng Kiểm toán nhà nước.
9. Thông tư liên tịch giữa Chánh án Tòa án nhân dân tối cao, Viện trưởng Viện kiểm sát nhân dân tối cao, Tổng Kiểm toán nhà nước, Bộ trưởng, Thủ trưởng cơ quan ngang bộ. Không ban hành thông tư liên tịch giữa Bộ trưởng, Thủ trưởng cơ quan ngang bộ.
10. Nghị quyết của Hội đồng nhân dân tỉnh, thành phố trực thuộc Trung ương (sau đây gọi chung là cấp tỉnh).
11. Quyết định của Ủy ban nhân dân cấp tỉnh.
12. Văn bản quy phạm pháp luật của chính quyền địa phương ở đơn vị hành chính - kinh tế đặc biệt.
13. Nghị quyết của Hội đồng nhân dân huyện, quận, thị xã, thành phố thuộc tỉnh, thành phố thuộc thành phố trực thuộc Trung ương (sau đây gọi chung là cấp huyện).
14. Quyết định của Ủy ban nhân dân cấp huyện.
15. Nghị quyết của Hội đồng nhân dân xã, phường, thị trấn (sau đây gọi chung là cấp xã).
16. Quyết định của Ủy ban nhân dân cấp xã.

2.2. Large Language Model

Large Language Model (LLM) là một mô hình ngôn ngữ sử dụng deep neural network[1] với số lượng tham số rất lớn (thường là hàng tỷ trọng số hoặc nhiều hơn), được huấn luyện trên lượng lớn văn bản không được gán nhãn bằng cách sử dụng học tự giám sát hoặc học bán giám sát. LLM xuất hiện vào khoảng năm 2018 và thể hiện khả năng xử lý tốt nhiều loại nhiệm vụ khác nhau. Điều này đã thay đổi tâm điểm của nghiên cứu xử lý ngôn ngữ tự nhiên từ mô hình giám sát chuyên biệt cho từng nhiệm vụ sang mô hình đa năng có thể thích ứng với nhiều tình huống. LLM thường được áp dụng trong các ứng dụng xử lý ngôn ngữ tự nhiên (NLP) như hiểu, tóm tắt, dịch, sinh và dự đoán văn bản mới.

Một ví dụ của LLM là GPT, viết tắt của Generative Pre-trained Transformer. GPT là một mô hình biến đổi được tiền huấn luyện trên một tập dữ liệu văn bản rộng lớn, sau đó được tinh chỉnh cho các nhiệm vụ cụ thể như sinh văn bản, trả lời câu hỏi, phân loại văn bản và hơn thế nữa. GPT có khả năng sinh ra các đoạn văn bản có ý nghĩa và trôi chảy từ một đầu vào bất kỳ, chẳng hạn như một câu, một từ khóa hoặc một hình ảnh. Phiên

[1]Deep neural network (DNN) là một mạng nơ-ron nhân tạo (ANN) với nhiều lớp ẩn giữa lớp đầu vào và lớp đầu ra. DNN có thể được huấn luyện với dữ liệu không được gán nhãn và được sử dụng để phân loại, phân cụm và trích xuất đặc trưng. DNN là một phần của họ các mô hình học sâu (deep learning).

bản mới nhất của GPT là GPT-4[3], có khoảng 100 tỷ tham số và được huấn luyện trên khoảng 10 triệu từ.

2.3. Generative Pretrained Transformer

Generative Pre-trained Transformer (GPT), một loại mô hình học sâu có khả năng sinh văn bản tự động dựa trên dữ liệu huấn luyện lớn. GPT được phát triển bởi OpenAI[2]. GPT có nhiều phiên bản khác nhau, từ GPT-1 ra mắt vào năm 2018 đến GPT-3 ra mắt vào năm 2020. Mỗi phiên bản đều có số lượng tham số và khả năng sinh văn bản cao hơn phiên bản trước. Ví dụ, GPT-3 có 175 tỷ tham số và có thể sinh văn bản với độ dài tối đa là 2048 từ. GPT có thể áp dụng cho nhiều ứng dụng khác nhau, như viết tiêu đề, tóm tắt, bài luận, thơ, hội thoại và nhiều thứ khác. Ví dụ, GPT-3 có thể viết một bài luận ngắn về tác dụng của việc đọc sách hoặc một câu chuyện ngắn về một chú mèo tên Tom. GPT là một trong những mô hình học sâu tiên tiến nhất hiện nay trong lĩnh vực xử lý ngôn ngữ tự nhiên.

2.4. Embeddings

Embedding là một kỹ thuật biểu diễn các nội dung số như hình, chữ, âm thanh thành một danh sách các con số (vector). Quá trình này giúp cho các machine learning model có thể "hiểu" được nội dung đó.

Embeddings thường được sử dụng trong các ứng dụng như:
- Search (kết quả được sắp xếp theo mức độ liên quan đến một chuỗi truy vấn)
- Clustering (các chuỗi văn bản được nhóm lại theo độ tương tự)
- Recommendations (các mục có chuỗi văn bản liên quan được đề xuất)
- Anomaly detection (các điểm ngoại lệ có độ tương tự thấp được xác định)
- Diversity measurement (phân tích phân phối độ tương tự)
- Classification (các chuỗi văn bản được phân loại theo nhãn tương tự nhất)

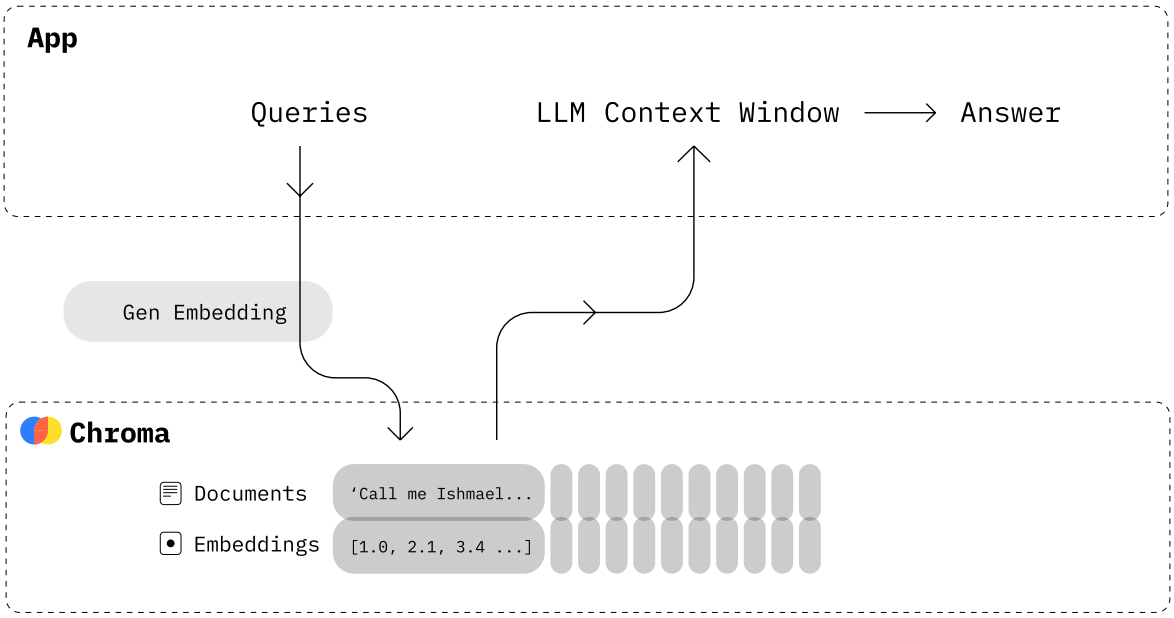
Khoảng cách giữa hai vector đo lường mức độ liên quan của chúng. Khoảng cách nhỏ cho thấy mức độ liên quan cao và khoảng cách lớn cho thấy mức độ liên quan thấp.

2.5. Chroma

Chroma là một cơ sở dữ liệu nhúng mã nguồn mở được thiết kế để lưu trữ các vector nhúng (embeddings) và cho phép tìm kiếm các vector gần nhất thay vì tìm kiếm theo chuỗi con như một cơ sở dữ liệu truyền thống.

Chroma cung cấp các công cụ để:
- Lưu trữ embeddings và metadata (dữ liệu mô tả) của chúng
- Nhúng tài liệu và truy vấn
- Tìm kiếm embeddings

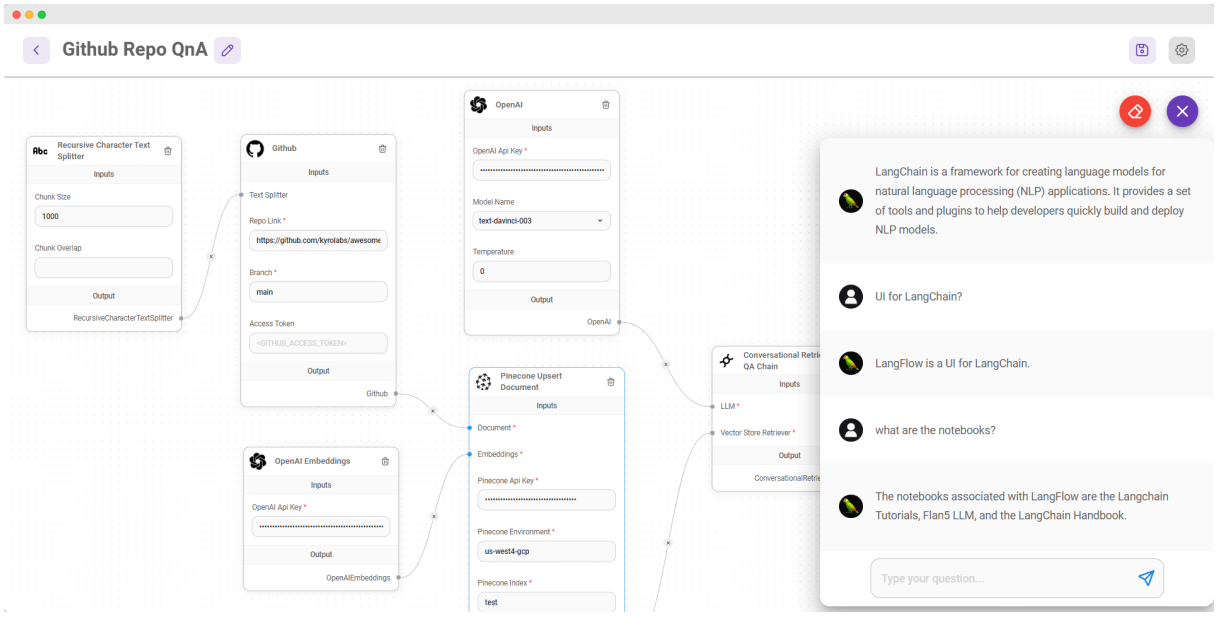
[2]OpenAI là một tổ chức nghiên cứu trí tuệ nhân tạo phi lợi nhuận được thành lập vào tháng 12 năm 2015, có trụ sở tại San Francisco, California. OpenAI được thành lập bởi Elon Musk, Sam Altman và các nhà nghiên cứu khác, với mục tiêu "điều tra và thúc đẩy một trí tuệ nhân tạo thân thiện với con người



Hình 1: Cơ sở dữ liệu Chroma

2.6. Langchain

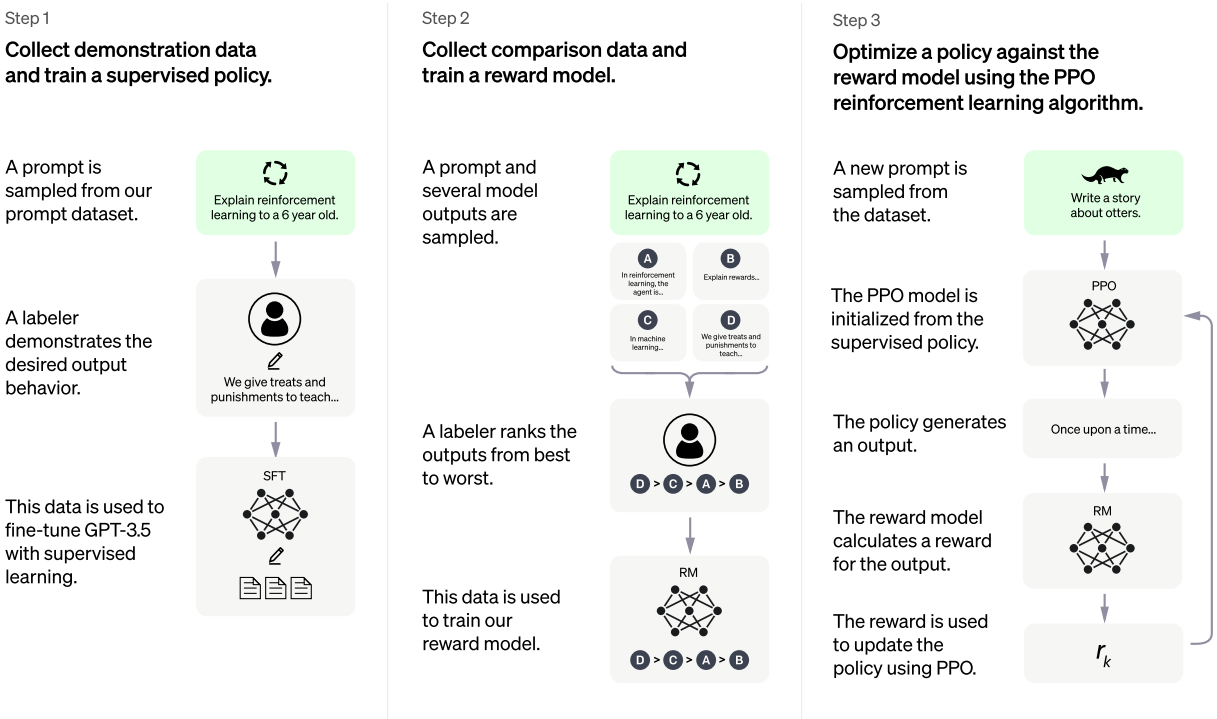
Langchain là một framework được sinh ra để tận dụng sức mạnh của các mô hình ngôn ngữ lớn LLM như ChatPGT, LLaMA… để tạo ra các ứng dụng trong thực tế. Nó giúp cho việc tương tác với các mô hình ngôn ngữ lớn trở nên dễ dàng hơn và cho phép các ứng dụng tận dụng thêm các thông tin từ nhiều nguồn data khác của bên thứ 3 như Google, Notion, Facebook… cũng như cung cấp các component cho phép sử dụng các language model trong nhiều tình huống khác nhau trên thực tế.



Hình 2: Flowise, visual tool để xây dựng các ứng dụng LLM, được xây dựng trên nền tảng Langchain

2.7. ChatGPT

ChatGPT là một chatbot AI hoạt động dựa trên mô hình GPT-3.5 được phát triển bởi OpenAI. ChatGPT có khả năng tương tác với người dùng thông qua việc trả lời các câu hỏi và hoàn thành các tác vụ liên quan đến ngôn ngữ như viết kịch bản, lời thoại, dịch thuật, tìm kiếm thông tin,… mà không giới hạn về chủ đề. ChatGPT được đào tạo bằng phương pháp Học tăng cường từ phản hồi của con người (RLHF – Reinforcement Learning from Human Feedback)[4], nên có thể hiểu ngữ cảnh, ghi nhớ thông tin người dùng nói, dự đoán nhu cầu của họ để đưa ra các phản hồi chính xác nhất. ChatGPT là một ứng dụng nổi bật của GPT-3, một trong những mô hình xử lý ngôn ngữ tự nhiên (Natural Language Processing) tiên tiến nhất hiện nay. ChatGPT có thể được áp dụng cho nhiều lĩnh vực khác nhau như chăm sóc khách hàng, sáng tạo nội dung, giáo dục, … ChatGPT là một bước tiến quan trọng trong lĩnh vực trí tuệ nhân tạo và có tiềm năng thay đổi cách con người giao tiếp và học tập trong tương lai.



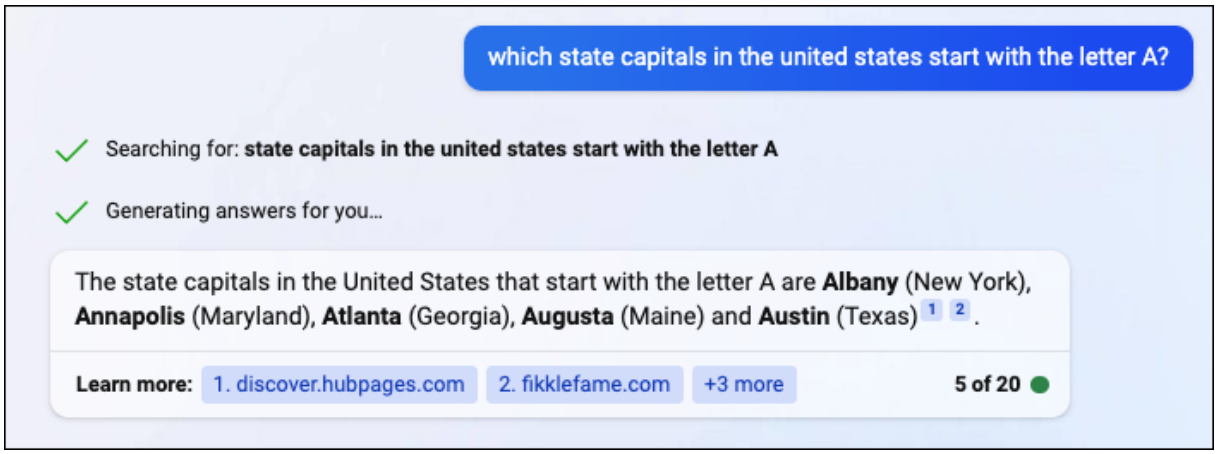
Hình 3: Sơ đồ hoạt động của ChatGPT

2.8. Bing AI

Bing AI[5] là một chatbot trí tuệ nhân tạo (AI) được phát triển bởi Microsoft và ra mắt vào năm 2023. Nó được xây dựng trên nền tảng của mô hình ngôn ngữ lớn (LLM) GPT-4 của OpenAI và đã được tinh chỉnh sử dụng cả các kỹ thuật học có giám sát và học tăng cường.

Bing AI không chỉ sinh văn bản dựa theo xác suất như ChatGPT của OpenAI, mà còn có thể dẫn được nguồn của văn bản mà nó tham chiếu tới do đó nội dung có tính xác thực cao hơn. Ngoài ra, Bing AI còn có thể trả lời các câu hỏi phức tạp, tương tác với người

dùng qua chat, và tạo ra nội dung sáng tạo như thơ, truyện, mã nguồn, bài viết, bài hát và nhiều thứ khác.



Hình 4: Giao diện của Bing AI

2.9. Multimodal Model

Multimodal Model là một hệ thống trí tuệ nhân tạo xử lý nhiều dạng dữ liệu cảm quan cùng lúc. Học trong Multimodal Model kết hợp các dữ liệu từ các cảm biến và nguồn khác vào một mô hình, tạo ra các dự đoán linh hoạt hơn.

Multimodal Model gồm nhiều mạng nơ-ron unimodal, xử lý từng dạng dữ liệu riêng biệt. Sau đó, các đặc trưng được mã hóa từ các mạng unimodal được kết hợp lại để tạo ra một đại diện chung cho tất cả các dạng dữ liệu. Cuối cùng, đại diện chung này được sử dụng để thực hiện các nhiệm vụ mong muốn.

Multimodal Model là đề tài nóng của trí tuệ nhân tạo. Ví dụ nổi bật là GPT-4 của OpenAI, một mô hình lớn xử lý văn bản và hình ảnh và tạo ra văn bản. GPT-4 đã đạt được hiệu suất ở mức con người trên nhiều tiêu chuẩn chuyên nghiệp và học thuật. Multimodal Model có tiềm năng ứng dụng trong nhiều lĩnh vực khác nhau.

GPT-4 visual input example, École Polytechnique Exam Question:

User Answer question I.1.a. Think step-by-step.

I. Principe de la détection de rayonnement avec un bolomètre

Comme illustré sur la figure 1 un bolomètre est constitué d'un absorbeur qui reçoit le rayonnement que l'on désire détecter. Sa température T, supposée uniforme, est mesurée à l'aide d'un thermomètre incorporé, constitué d'un matériau conducteur dont la résistance $R(T)$ varie avec la température T ; cette variation est caractérisée par le coefficient $\alpha = \frac{1}{R}\frac{dR}{dT}$. L'ensemble possède la capacité thermique C_{th}.

Un barreau, conducteur thermique, homogène, de longueur L, de section S et de conductivité thermique λ et sans échanges thermiques latéraux, relie le bolomètre à un thermostat de température T_b fixe.

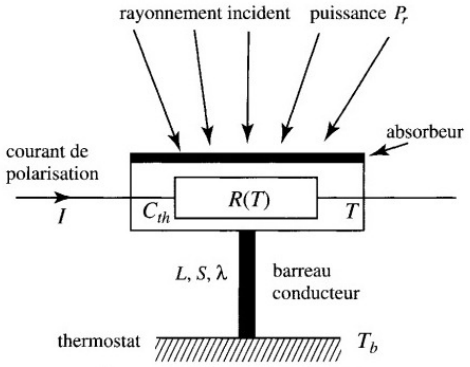


Figure 1 - Schéma d'un bolomètre

I.1. Dans cette première partie, on étudie le comportement du bolomètre en régime stationnaire, à la température de fonctionnement $T = T_0$.

I.1.a) Exprimer la température $T(x)$ en tout point du barreau conducteur repéré par son abscisse x ; on prendra $T(0) = T_0$ et $T(L) = T_b$.

GPT-4 To answer question I.1.a, we need to express the $T(x)$ in every point of the conductive bar, which is characterized by the abscissa x. We are given that $T(0) = T_0$ and $T(L) = T_b$.

For a stationary regime, the heat equation for the one-dimensional heat conduction in the bar is:
$d^2T \,/\, dx^2 = 0$

Integrating once, we get:
$dT \,/\, dx = C_1$

Integrating a second time:
$T(x) = C_1 x + C_2$

To determine the constants C_1 and C_2, we use the boundary conditions:
$T(0) = T_0$, which gives $C_2 = T_0$
$T(L) = T_b$, which gives $C_1 = (T_b - T_0) \,/\, L$

Finally, the expression for $T(x)$ is:
$T(x) = (T_b - T_0) * (x \,/\, L) + T_0$

This represents a linear temperature profile along the conductive bar, with a slope of $(T_b - T_0) \,/\, L$.

,

Hình 5: Ví dụ về Multimodal Model, người dùng yêu cầu GPT4 trả lời câu hỏi về vấn đề vật lý được viết bằng tiếng Pháp

3. Ứng dụng và thách thức

3.1. AI trong tra cứu văn bản

3.2. AI trong soạn thảo

3.3. Robot luật sư

3.4. Robot luật sư cá nhân

4. Thử nghiệm

Trong phạm vi của bài luận này, tôi chỉ sử dụng các văn bản liên quan tới lĩnh vực bảo hiểm xã hội và việc làm để thử nghiệm và đánh giá:

- Luật Bảo hiểm xã hội 2014
- Văn bản hợp nhất 2089/VBHN-BHXH năm 2020 hợp nhất Quyết định về Quy trình thu bảo hiểm xã hội, bảo hiểm y tế, bảo hiểm thất nghiệp, bảo hiểm tai nạn lao động, bệnh nghề nghiệp; quản lý sổ bảo hiểm xã hội, thẻ bảo hiểm y tế do Bảo hiểm xã hội Việt Nam ban hành
- Thông tư 59/2015/TT-BLĐTBXH quy định chi tiết và hướng dẫn thi hành một số điều của Luật bảo hiểm xã hội về bảo hiểm xã hội bắt buộc do Bộ trưởng Bộ Lao động - Thương binh và Xã hội ban hành
- Nghị định 115/2015/NĐ-CP hướng dẫn Luật bảo hiểm xã hội về bảo hiểm xã hội bắt buộc
- Nghị định 146/2018/NĐ-CP hướng dẫn Luật bảo hiểm y tế
- Nghị định 28/2015/NĐ-CP hướng dẫn Luật Việc làm về bảo hiểm thất nghiệp
- Luật việc làm 2013
- Bộ luật Lao động 2019
- Thông tư 56/2017/TT-BYT về hướng dẫn Luật bảo hiểm xã hội và Luật an toàn vệ sinh lao động thuộc lĩnh vực y tế do Bộ trưởng Bộ Y tế ban hành
- Nghị định 134/2015/NĐ-CP hướng dẫn Luật Bảo hiểm xã hội về bảo hiểm xã hội tự nguyện
- Quyết định 28/2021/QĐ-TTg quy định về thực hiện chính sách hỗ trợ người lao động và người sử dụng lao động bị ảnh hưởng bởi đại dịch COVID-19 từ Quỹ bảo hiểm thất nghiệp do Thủ tướng Chính phủ ban hành
- Luật bảo hiểm y tế 2008
- Luật Bảo hiểm y tế sửa đổi 2014
- Quyết định 166/QĐ-BHXH năm 2019 về Quy trình giải quyết hưởng chế độ bảo hiểm xã hội, chi trả chế độ bảo hiểm xã hội, bảo hiểm thất nghiệp do Bảo hiểm xã hội Việt Nam ban hành
- Nghị định 61/2020/NĐ-CP về sửa đổi Nghị định 28/2015/NĐ-CP hướng dẫn Luật Việc làm về bảo hiểm thất nghiệp

4.1. Xây dựng bộ dữ liệu văn bản vi phạm pháp luật

4.1.1. Sơ lược về dữ liệu

Theo dữ liệu từ Thư viện pháp luật[3], hiện nay Việt Nam có khoảng 303936 văn bản vi phạm pháp luật. Bao gồm 20 loại văn bản và 27 lĩnh vực khác nhau:

Loại văn bản	Số lượng	Loại văn bản	Số lượng
Quyết định	188360	Hướng dẫn	1772
Nghị quyết	30709	Báo cáo	1494
Kế hoạch	23301	Điều ước quốc tế	1331
Thông tư	15067	Công điện	1244
Thông báo	13588	Sắc lệnh	997
Chỉ thị	13438	Lệnh	526
Nghị định	5191	Luật	486
Văn bản khác	2608	Pháp lệnh	228
Thông tư liên tịch	2605	Văn bản WTO	68
Văn bản hợp nhất	2162	Hiến pháp	5

Bảng 1: Số lượng văn bản vi phạm pháp luật theo loại văn bản

[3]thuvienphapluat.vn là trang chuyên cung cấp cơ sở dữ liệu, tra cứu và thảo luận pháp luật

Lĩnh vực	Số lượng
Bộ máy hành chính	105445
Tài chính nhà nước	42216
Văn hóa - Xã hội	39014
Tài nguyên - Môi trường	25490
Thương mại	22388
Xây dựng - Đô thị	21410
Bất động sản	21149
Thể thao - Y tế	19734
Thuế - Phí - Lệ Phí	17592
Giáo dục	16278
Giao thông - Vận tải	14825
Lao động - Tiền lương	14374
Doanh nghiệp	12744
Đầu tư	12718

Lĩnh vực	Số lượng
Công nghệ thông tin	12217
Xuất nhập khẩu	11535
Lĩnh vực khác	8607
Quyền dân sự	5505
Tiền tệ - Ngân hàng	4954
Bảo hiểm	2697
Dịch vụ pháp lý	2619
Thủ tục Tố tụng	2350
Vi phạm hành chính	2225
Kế toán - Kiểm toán	1752
Trách nhiệm hình sự	1515
Sở hữu trí tuệ	965
Chứng khoán	771

Bảng 2: Số lượng văn bản vi phạm pháp luật theo lĩnh vực

Các thuộc tính của một văn bản quy phạm pháp luật gồm: tên văn bản, số hiệu văn bản, loại văn bản, nơi ban hành, người ký, ngày ban hành, ngày hiệu lực, ngày công báo, số công báo.

Ngoài ra các thuộc tính trên, còn có lược đồ thể hiện mối quan hệ giữa các văn bản quy phạm pháp luật dựa trên *văn bản đang tham chiếu*:

- Văn bản được hướng dẫn: là văn bản ban hành trước, có hiệu lực pháp lý cao hơn *Văn bản tham chiếu* và được *Văn bản tham chiếu* hướng dẫn hoặc quy định chi tiết nội dung của nó.
- Văn bản được hợp nhất: Là văn bản ban hành trước, bao gồm các văn bản được sửa đổi, bổ sung và văn bản sửa đổi, bổ sung, được *Văn bản tham chiếu* hợp nhất nội dung lại với nhau.
- Văn bản bị sửa đổi bổ sung: Là văn bản ban hành trước, bị *Văn bản tham chiếu* sửa đổi, bổ sung một số nội dung.
- Văn bản bị đính chính: Là văn bản ban hành trước, bị *Văn bản tham chiếu* đính chính các sai sót như căn cứ ban hành, thể thức, kỹ thuật trình bày,…
- Văn bản bị thay thế: Là văn bản ban hành trước, bị *Văn bản tham chiếu* quy định thay thế, bãi bỏ toàn bộ nội dung.
- Văn bản được dẫn chiếu: Là văn bản ban hành trước, trong nội dung của *Văn bản tham chiếu* có quy định dẫn chiếu trực tiếp đến điều khoản hoặc nhắc đến nó.

- Văn bản được căn cứ: Là văn bản ban hành trước *Văn bản tham chiếu*, bao gồm các văn bản quy định thẩm quyền, chức năng của cơ quan ban hành *Văn bản tham chiếu* văn bản có hiệu lực pháp lý cao hơn quy định nội dung, cơ sở để ban hành *Văn bản tham chiếu*.
- Văn bản liên quan ngôn ngữ: Là bản dịch Tiếng Anh của *Văn bản tham chiếu*.
- Văn bản hướng dẫn: Là bản tiếng Việt của *Văn bản tham chiếu*.
- Văn bản hợp nhất: Là văn bản ban hành sau, hợp nhất lại nội dung của *Văn bản tham chiếu* và văn bản sửa đổi, bổ sung của *Văn bản tham chiếu*.
- Văn bản sửa đổi bổ sung: Là văn bản ban hành sau, sửa đổi, bổ sung một số nội dung của *Văn bản tham chiếu*.
- Văn bản đính chính: Là văn bản ban hành sau, nhằm đính chính các sai sót như căn cứ ban hành, thể thức, kỹ thuật trình bày,…của *Văn bản tham chiếu*.
- Văn bản thay thế: Là văn bản ban hành sau, có quy định đến việc thay thế, bãi bỏ toàn bộ nội dung của *Văn bản tham chiếu*
- Văn bản liên quan cùng nội dung: Là văn bản có nội dung tương đối giống, hoặc có phạm vi điều chỉnh, đối tượng điều chỉnh tương tự *Văn bản tham chiếu*.

Mục lục của văn bản là phần quan trọng không thể thiếu. Tuy nhiên không phải văn bản nào cũng có mục lục, và cũng không có một định dạng chuẩn cho mục lục. Các chỉ mục thường thấy là: phần, chương, mục, điều, khoản, điểm.

4.1.2. Xây dựng cơ sở dữ liệu

Cấu trúc dữ liệu của datasets gồm 3 bảng chính: `VanBan`, `LuocDo`, `ChiMuc` được mô tả như sau:

Tên trường	**Kiểu dữ liệu**	**Mô tả**
id	integer (PK)	ID của văn bản
ten_van_ban	string	Tên văn bản
so_hieu_van_ban	string	Số hiệu văn bản
loai_van_ban	string	Loại văn bản
noi_ban_hanh	string	Nơi ban hành
nguoi_ky	string	Người ký
ngay_ban_hanh	date	Ngày ban hành
ngay_hieu_luc	date	Ngày hiệu lực
ngay_cong_bao	date	Ngày công báo
so_cong_bao	string	Số công báo
noi_dung_van_bang	string	Nội dung văn bản dạng text
linh_vuc	string	Lĩnh vực của văn bản

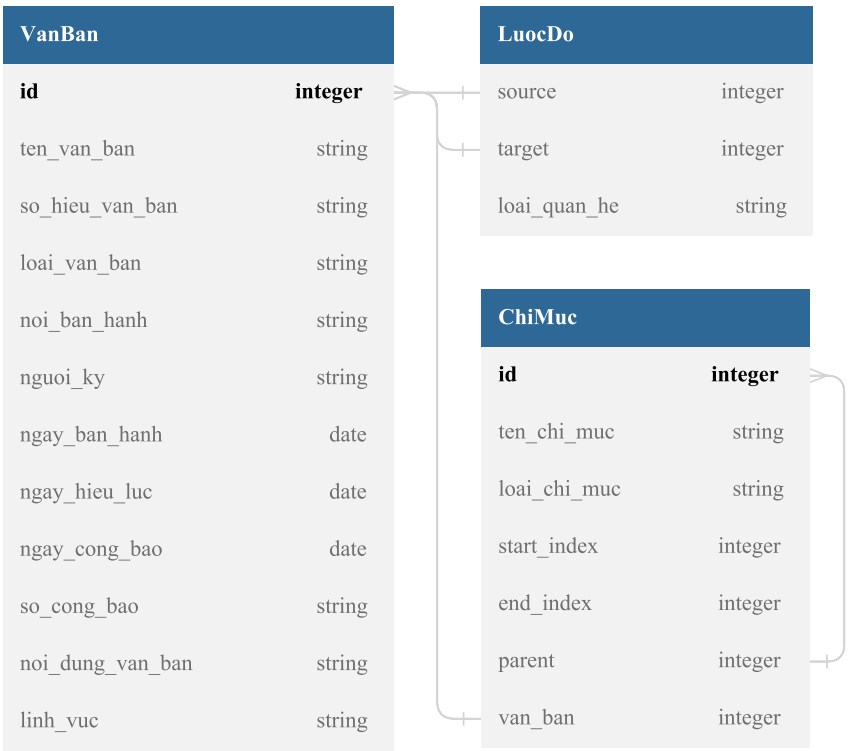
Bảng 3: Bảng `VanBan` chứa thông tin về văn bản vi phạm pháp luật

Tên trường	Kiểu dữ liệu	Mô tả
source	integer (FK)	ID của văn bản nguồn
target	integer (FK)	ID của văn bản đích
loai_quan_he	string	Loại quan hệ giữa văn bản nguồn và văn bản đích. VD: thay thế, hướng dẫn, sửa đổi bổ sung…

Bảng 4: Bảng `LuocDo` chứa thông tin về mối quan hệ giữa các văn bản vi phạm pháp luật

Tên trường	Kiểu dữ liệu	Mô tả
ten_chi_muc	string	Tên của chỉ mục
loai_chi_muc	string	Loại của mục lục. VD: phần, chương, mục, điều, khoản, điểm…
start_index	integer	Vị trí bắt đầu của nội dung của chỉ mục trong văn bản
end_index	integer	Vị trí kết thúc của nội dung của chỉ mục trong văn bản
parent_id	integer (FK)	ID của chỉ mục cha (nếu có), thể hiển tree structure[4].
vanban_id	integer (FK)	ID của văn bản

Bảng 5: Bảng `ChiMuc`: chứa thông tin về mục lục của văn bản vi phạm pháp luật.



Hình 6: Cấu trúc dữ liệu của cơ sở dữ liệu văn bản vi phạm pháp luật

[4]Tree structure hay cây là một cấu trúc dữ liệu được sử dụng rộng rãi gồm một tập hợp các nút (node) được liên kết với nhau theo quan hệ cha-con

Xử lý văn bản: Văn bản sau khi tải xuống có định dạng HTML[5], do đó cần phải xử lý để lấy được nội dung văn bản dạng text. Để làm được điều này, tôi sử dụng thư viện BeautifulSoup[6] để lấy nội dung dạng text của văn bản.

QUỐC HỘI
CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM Độc lập - Tự do - Hạnh phúc
Luật số: 58/2014/QH13
Hà Nội, ngày 20 tháng 11 năm 2014

LUẬT BẢO HIỂM XÃ HỘI
Căn cứ Hiến pháp nước Cộng hòa xã hội chủ nghĩa Việt Nam;
Quốc hội ban hành Luật bảo hiểm xã hội.
Chương I:NHỮNG QUY ĐỊNH CHUNG
Điều 1. Phạm vi điều chỉnh
Luật này quy định chế độ, chính sách bảo hiểm xã hội; quyền và trách nhiệm của người lao động, người sử dụng lao động; cơ quan, tổ chức, cá nhân có liên quan đến bảo hiểm xã hội, tổ chức đại diện tập thể lao động, tổ chức đại diện người sử dụng lao động; cơ quan bảo hiểm xã hội; quỹ bảo hiểm xã hội; thủ tục thực hiện bảo hiểm xã hội và quản lý nhà nước về bảo hiểm xã hội.
Điều 2. Đối tượng áp dụng
1. Người lao động là công dân Việt Nam thuộc đối tượng tham gia bảo hiểm xã hội bắt buộc, bao gồm:
a) Người làm việc theo hợp đồng lao động không xác định thời hạn, hợp đồng lao động xác định thời hạn, hợp đồng lao động theo mùa vụ hoặc theo một công việc nhất định có thời hạn từ đủ 03 tháng đến dưới 12 tháng, kể cả hợp đồng lao động được ký kết giữa người sử dụng lao động với người đại diện theo pháp luật của người dưới 15 tuổi theo quy định của pháp luật về lao động;
b) Người làm việc theo hợp đồng lao động có thời hạn từ đủ 01 tháng đến dưới 03 tháng;
c) Cán bộ, công chức, viên chức;
d) Công nhân quốc phòng, công nhân công an, người làm công tác khác trong tổ chức cơ yếu;
đ) Sĩ quan, quân nhân chuyên nghiệp quân đội nhân dân; sĩ quan, hạ sĩ quan nghiệp vụ, sĩ quan, hạ sĩ quan chuyên môn kỹ thuật công an nhân dân; người làm công tác cơ yếu hưởng lương như đối với quân nhân;……

Hình 7: Văn bản sau khi xử lý

[5]HTML là viết tắt của cụm từ Hypertext Markup Language (tạm dịch là Ngôn ngữ đánh dấu siêu văn bản). HTML được sử dụng để tạo và cấu trúc các thành phần trong trang web hoặc ứng dụng, phân chia các đoạn văn, heading, titles, blockquotes…

Tạo mục lục: để tạo chỉ mục cho văn bản, tôi sử dụng regex[6] để tìm kiếm các chỉ mục trong văn bản. Như đã nêu trong Phần 4.1.1 các regex để tìm chỉ mục là:

Regex	**Loại chỉ mục**	**Chú giải**
^(Phần thứ [\d\w]+.*)$	Phần	Tìm các chỉ mục có dạng "Phần thứ <số|chữ> <nội dung>". Ví dụ: "Phần thứ nhất: Những quy định chung"
^(Chương [\d\w]+.*)$	Chương	Tìm các chỉ mục có dạng "Chương <số|chữ> <nội dung>". Ví dụ: "Chương I: ĐIỀU KHOẢN CƠ BẢN"
^(Mục [\d|I|V|X|L|C|D|M]+.*)$	Mục	Tìm các chỉ mục có dạng "Mục <số| chữ số la mã> <nội dung>". Ví dụ: "Mục 1. QUY ĐỊNH CHUNG VỀ QUYẾT ĐỊNH HÌNH PHẠT"
^(Điều \d+.*)$	Điều	Tìm các chỉ mục có dạng "Điều <số> <nội dung>". Ví dụ: "Điều 51. Các tình tiết giảm nhẹ trách nhiệm hình sự"
^(\d+\. .*)$	Khoản	Tìm các chỉ mục có dạng "<số>. <nội dung>". Ví dụ: 1. Người phạm tội phải trả lại tài sản đã…
^(\w\).*)$	Điểm	tìm các chỉ mục có dạng "<chữ>. <nội dung>". Ví dụ: a) Người phạm tội đã ngăn chặn h…

Bảng 6: Các regex để tìm chỉ mục trong văn bản

Phương pháp sử dụng regex tuy tốt nhưng vẫn chỉ là bán tự động, vì có một số trường hợp đặc biết vẫn cần sự can thiệp từ con người để có được kết quả tốt nhất.

Để đơn giản khi lập trình, tôi lưu kết quả sau khi xử lý thành định dạng JSON[7]:

[6]Regex là một chuỗi các ký tự đặc biệt được định nghĩa để tạo nên các mẫu (pattern) và được sử dụng để tìm kiếm và thay thế các chuỗi trong một văn bản

[7]JSON là viết tắt của Javascript Object Notation, là một bộ quy tắc về cách trình bày và mô tả dữ liệu trong một chuỗi lớn thống nhất được gọi chung là chuỗi JSON. Chuỗi JSON được bắt đầu bằng ký tự { và kết thúc bởi ký tự }

```
{
  "id":1,
  "ten_van_ban": "Luật Bảo hiểm xã hội 2014",
  "so_hieu_van_ban": "58/2014/QH13",
  "loai_van_ban": "Luật",
  "linh_vuc": "Bảo hiểm, Lao động - Tiền lương",
  "noi_ban_hanh": "Quốc hội",
  "nguoi_ky": "Nguyễn Sinh Hùng",
  "ngay_ban_hanh": "20/11/2014",
  "ngay_hieu_luc": "01/01/2016",
  "ngay_cong_bao": "29/12/2014",
  "so_cong_bao": "Từ số´1163 đến số´1164",
  "noi_dung_van_ban":"...",
  "tree": {
    "ten_chi_muc": "Luật Bảo hiểm xã hội 2014",
    "loai_chi_muc": "root",
    "start_index": 0,
    "end_index": 110927,
    "children": [
      {
        "ten_chi_muc": "Chương I:NHỮNG QUY ĐỊNH CHUNG",
        "loai_chi_muc": "chương",
        "start_index": 280,
        "end_index": 14742,
        "children": [
          {
            "ten_chi_muc": "Điều 1. Phạm vi điều chỉnh",
            "loai_chi_muc": "điều",
            "start_index": 307,
            "end_index": 681,
            "children": []
          },
          {
            "ten_chi_muc": "Điều 2. Đối tượng áp dụng",
            "loai_chi_muc": "điều",
            "start_index": 708,
            "end_index": 3086,
            "children": [...]
          },
          ...
        ]
      }
      ...
    ]
  }
}
```


Hình 8: Kết quả sau khi xử lý văn bản ở định dạng JSON

Sau khi xây dựng được datasets về luật, tôi có tạo thêm một python package dùng để truy vấn dữ liệu một cách dễ dàng [7]:

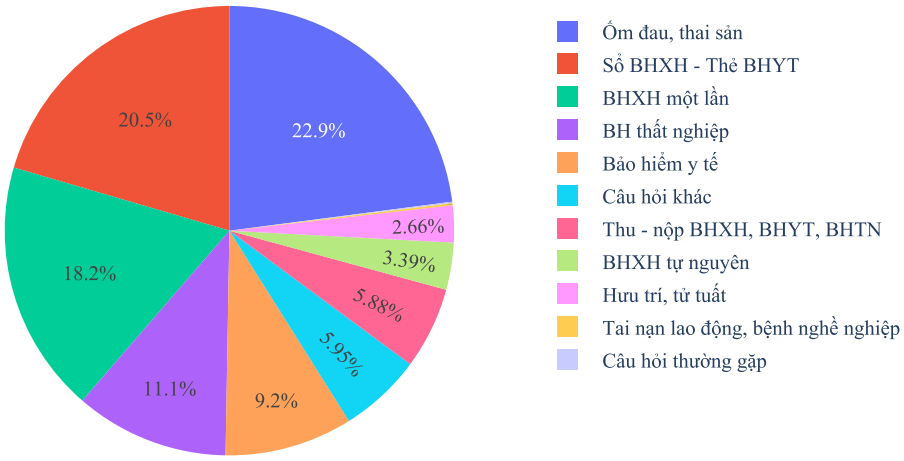
```
from lawquery import Engine

# create engine, law_id là số hiệu của văn bản luật
engine = Engine(law_id='58/2014/QH13')
# query single node
engine.query(node_type='điều', node_id='1')
# => [Điều 1. Phạm vi điều chỉnh]
engine.query(node_type='phần')
# => [Phần thứ nhất..., Phần thứ hai...]
engine.query(name='hôn nhân')
# => [Điều 67. Các trường hợp hưởng trợ cấp tuất hằng tháng]
# query by path: from parent to child
node = engine.query_by_path([
    {
        'node_type': 'phần',
        'node_id': 'hai'
    },
    {
        'node_type': 'chương',
        'node_id': 'I'
    },
    {
        'node_type': 'mục',
        'node_id': '1'
    },
    {
        'node_type': 'điều',
        'node_id': '50'
    }])
# => [Điều 50. Trợ cấp phục vụ]
node.content
# => Nội dung của điều luật
```

Chương trình 1: Sử dụng package `lawquery` để truy vấn dữ liệu

4.2. Xây dựng bộ dữ liệu hỏi đáp luật

Từ những gì đã làm được ở Intelligent Retrieval System on Legal Information[8], tôi tiếp tục phát triển bộ dữ liệu hỏi đáp về luật với chủ đề là bảo hiểm xã hội.

Bộ câu hỏi được lấy từ website baohiemxahoi.gov.vn, cổng thông tin điện tử bảo hiểm xã hội Việt Nam. Bao gồm 19330 bộ câu hỏi-trả lời được phân vào nhiều lĩnh vực khác nhau, xem Hình 9. Tôi chỉ lấy các data point mà câu trả lời có trích dẫn đến các văn bản luật, xem Hình 10. Do đó, số lượng thật sự của bộ dữ liệu là 4368 câu hỏi.



Hình 9: Số lượng câu hỏi theo lĩnh vực

Nội dung câu hỏi:
Sổ BHXH của tôi đã được chốt tại BHXH Ba Đình - Hà Nội. Hiện tại tôi bị mất 2 tờ rời của sổ , tôi đang sinh sống ở tỉnh Long An thì có thể ra cơ quan BHXH của tỉnh để xin cấp lại tờ rời BHXH hay không? hay phải ra cơ quan BHXH đã chốt sỗ thì mới có thể xin cấp lại được? Xin cảm ơn

Câu trả lời:
Theo quy định tại **Tiết a Điểm 2.1** và **Tiết a Điểm 2.2 Khoản 2 Điều 3 Văn bản hợp nhất số 2089/VBHN-BHXH** thì:

- BHXH huyện được cấp lại sổ BHXH cho người đang bảo lưu thời gian đóng BHXH, BHTN, BHTNLĐ, BNN ở huyện, tỉnh khác.
- BHXH tỉnh được cấp lại sổ BHXH cho người đã hưởng BHXH hoặc đang bảo lưu thời gian đóng BHXH, BHTN, BHTNLĐ, BNN ở huyện, tỉnh khác.

Đồng thời, theo quy định tại **Tiết a Điểm 1.1 Khoản 1 Điều 27 Văn bản hợp nhất số 2089/VBHN-BHXH** ngày 26/6/2020 của BHXH Việt Nam ban hành Quy trình thu BHXH, BHYT, BHTN, BHTNLĐ, BNN; quản lý sổ BHXH, thẻ BHYT thì hồ sơ để cấp lại sổ BHXH gồm Tờ khai tham gia, điều chỉnh thông tin BHXH, BHYT (Mẫu TK1-TS). Vì vậy, nếu Bạn thuộc các trường hợp nêu trên thì có thể nộp hồ sơ xin cấp lại sổ BHXH tại cơ quan BHXH ở Long An nơi Bạn đang sinh sống.

Hình 10: Ví dụ câu hỏi và câu trả lời

Sau khi có được bộ dữ liệu hỏi đáp luật, tôi tiếp tục sử dụng Label Studio[9] để gán nhãn cho câu trả lời.

4.3. Information Retrieval

name	top5	top10	top20	top50	top100
tdidf	0.1	0.1869	0.3205	0.5022	0.661
bm25l	0.0099	0.0291	0.0686	0.1939	0.4022
bm25plus	0.066	0.1273	0.2381	0.4144	0.6077
bm25okapi	0.0653	0.1348	0.2348	0.4243	0.6114
tdidf_ws	0.1021	0.1878	0.3207	0.5088	0.6779
bm25l_ws	0.0108	0.0301	0.074	0.2036	0.4186
bm25plus_ws	0.0852	0.1545	0.2703	0.4466	0.6311
bm25okapi_ws	0.088	0.1613	0.2759	0.4536	0.6344
instructor-base	0.0146	0.0258	0.0458	0.0949	0.1714
instructor-large	0.0082	0.0171	0.0373	0.1052	0.2024
instructor-xl	0.0204	0.0364	0.0669	0.1397	0.2458

4.4. Tra cứu văn bản luật bằng ChatGPT API

5. Kết luận

Phụ lục

Tài liệu tham khảo

[1] "Hiến pháp năm 2013," Quốc hội, 2013.

[2] "Luật ban hành văn bản quy phạm pháp luật 2015," Quốc hội, 2015.

[3] OpenAI, "Gpt-4 technical report," 2023.

[4] N. Lambert, L. Castricato, L. von Werra, and A. Havrilla, "Illustrating reinforcement learning from human feedback (rlhf)," *Hugging Face Blog*, 2022.

[5] Y. Mehdi, "Reinventing search with a new ai-powered microsoft bing and edge, your copilot for the web," 2023. [Online]. Available: https://blogs.microsoft.com/blog/2023/02/07/reinventing-search-with-a-new-ai-powered-microsoft-bing-and-edge-your-copilot-for-the-web/

[6] L. Richardson, "Beautiful soup documentation," *April*, 2007.

[7] T. Ngo, "LawQuery." [Online]. Available: https://github.com/Th1nhNg0/law-query

[8] "Intelligent retrieval system on legal information," *Aciids 2023*, 2023.

[9] "Open source data labeling platform." [Online]. Available: https://labelstud.io/