Consultor Legal Autónomo

Elaborado por:

Salvador Nicolás Sanchez

ÍNDICE

Preparación del entorno

Librerías y Datos

Para la ejecución del proyecto utilizamos las siguientes librerías:

- os: Interacción con el entorno de python.
- shutil: Mover archivos en el entorno.
- getpass: Ingresar usuarios y contraseñas de forma segura.
- urllib: Descarga de archivos de la web.
- zipfile: Descompresión de archivos zip.
- PyPDF2: Lectura de los archivos PDF.
- csv: Creación y modificación de archivos CSV.
- chromadb: Creación y acceso de base de datos vectoriales.
- pandas: Creación y modificación de DataFrames.
- re: Manejo de expresiones regulares.
- huggingface hub: Descarga de modelos de LLM.
- llama cpp: Creación e interacción con el modelo de LLM “Llama 2”.
- requests: Acceso a APIs de la WEB.
- wiki data integrator: Inicio de sesión de la API de wikidata.
- wikipedia: Consultas de wikipedia.

El modelo de LLM que se escogió fué la versión de 13 mil millones de parámetros de “Llama 2” dado que es un modelo con bastante documentación para su implementación, con buen performance en el lenguaje natural y en su ejecución y con un consumo de VRAM menor al disponible en el entorno de ejecución.

El modelo de embedding que se utilizará será “multilingual-e5-large” por ser entrenado con una gran cantidad de datos en diferentes idiomas. Se crearon las funciones vectorizar_texto y vectorizar_lista_texto para poder realizar las vectorizaciones de una manera más cómoda.

La creación de las base de datos vectoriales se realizaron con la librería “ChromaDB” que permite una interacción sencilla con los datos además de permitir la utilización de embeddings propios para guardar los datos y estar fuertemente optimizada para la adquisición de vectores similares.

Los archivos necesarios para la ejecución se encuentran en el siguiente repositorio “https://github.com/SalvaMrS/TP2_NLP/” que al ejecutar el proyecto se importará al entorno de ejecución para poder trabajar con ellos.

Base de datos CSV

Como base de datos en formato tabular se utilizará un diccionario jurídico en donde se identificará del PDF las frases y su significado para posteriormente ser guardado en un archivo CSV. El archivo posee el siguiente formato en sus definiciones en donde se observa que las frases están separadas de su definición y la definición anterior por dos saltos de líneas con la posibilidad de que se incluya antes de la frase un asterisco (*) por lo que se utilizó esta regularidad para generar la base de datos.

*A non domino

Así se denomina la transferencia de un bien, mueble o inmueble, cuando la efectúa una persona que no es su propietaria. La adquisición a non dominó en los títulos de crédito cambiarios significa que el accipiens de buena fe adquiere la propiedad del título aun cuando su tradens no sea el propietario del mismo. Por su mecanismo se trata de una adquisición y no de una cesión: adquisición originaria y no derivada.

*A puerta cerrada

Expresión empleada cuando, por excepción al principio de publicidad de las contiendas judiciales, sea por mandato de la ley o por decisión del tribunal, se veda al público su permanencia en la sala de audiencias durante los debates, o parte de los mismos. Esta prohibición tiene lugar en los casos en que por la naturaleza de las cuestiones que se suscitan en el litigio, o por la índole de éste, puedan producirse graves inconvenientes o escándalo.

Base de datos Vectorial

Observando la constitución nacional, el código civil y el código penal de Argentina se observan una regularidad en el formato de la definición de cada artículo presente y lo que dicta este artículo; este comienza con un salto de línea seguido de “Art” y termina con un “.- “ por lo que se utilizó esta expresión regular para segmentar los textos y obtener a qué artículo se refiere para posteriormente guardar la información en la base de datos vectorial “Documentacion_de_leyes_Argentinas”.

CONSTITUCIÓN NACIONAL:

Artículo 10.- En el interior de la República es libre de derechos la circulación de los efectos de producción o fabricación nacional, así como la de los géneros y mercancías de todas clases, despachadas en las aduanas exteriores.

CÓDIGO CIVIL:

Art.285.- Los menores no pueden demandar a sus padres sino por sus intereses propios, y previa autorización del juez, aun cuando tengan una industria separada o sean comerciantes.

CÓDIGO PENAL:

ARTICULO 22.- En cualquier tiempo que se satisfaciera la multa, el reo quedará en libertad.

Del importe se descontará, de acuerdo con las reglas establecidas para el cómputo de la prisión preventiva, la parte proporcional al tiempo de detención que hubiere sufrido.

ARTICULO 22 bis.- Si el hecho ha sido cometido con ánimo de lucro, podrá agregarse a la pena privativa de libertad una multa, aun cuando no esté especialmente prevista o lo esté sólo en forma alternativa con aquélla. Cuando no esté prevista, la multa no podrá exceder de noventa mil pesos.

Selector de Database

Dada una consulta a una estudiante de abogacía sobre qué diría para preguntar sobre alguna consulta legal se observó que las frases seguían patrones según a qué fuente de información se debería considerar para realizar la consulta luego de eso se realizaron pruebas para comprobar la similitud de los embeddings de estas consultas se comprobó su cercanía en el espacio vectorial por lo que esta práctica nos podría servir para identificar qué database deberíamos buscar la información.

Se barajó la posibilidad de comprobar los vectores uno a uno con una función de python pero fué más conveniente aprovechar esta característica de la librería ChromaDB que está optimizada para realizar esta búsqueda de manera eficiente.

La resolución de este problema consistió en cargar a la base de datos vectorial “Clasificador” unas 30 consultas etiquetadas de acuerdo a la fuente de información y a la hora de obtener a que DB debe acceder el modelo para resolver la consulta del usuario obtenemos el objeto más similar. En el caso de que la respuesta del modelo no sea óptima el usuario agregará su consulta a la base de datos optimizando el selector de DB.

Buscadores de fuentes de datos

CSV

Para poder buscar en esta base de datos se debe identificar qué buscar y obtener el índice más acorde a este para luego acceder a la BD. Gracias a la diversidad de palabras que existen en nuestra lengua utilizamos el modelo de LLM para identificar la frase jurídica a la cual se quiere conocer su significado y a este resultado buscamos la frase más similar dentro del diccionario, esta última problemática se soluciona cargando todas las frases a una base de datos vectorial que permite retornar la frase más similar de manera eficiente y una vez obtenido accedemos a la definición almacenada previamente en un dataframe de la librería pandas.

Base de Datos de Grafos

Como base de datos de grafos accederemos con la API de wikidata. La búsqueda se realizará obteniendo de la consulta del usuario el artículo de wikipedia al que hace referencia utilizando el modelo de LLM, con el resultado buscamos en wikidata el artículo y accedemos a la descripción con la API de wikipedia.

Base de Datos Vectorial

Las consultas a la base de datos vectorial se llevarán a cabo mediante la creación del embedding de la consulta completa del usuario, lo que implica la representación numérica de la consulta en un espacio vectorial, y posteriormente se realizará la búsqueda en la base de datos de ChromaDB denominada "Documentacion_de_leyes_Argentinas".

LLM

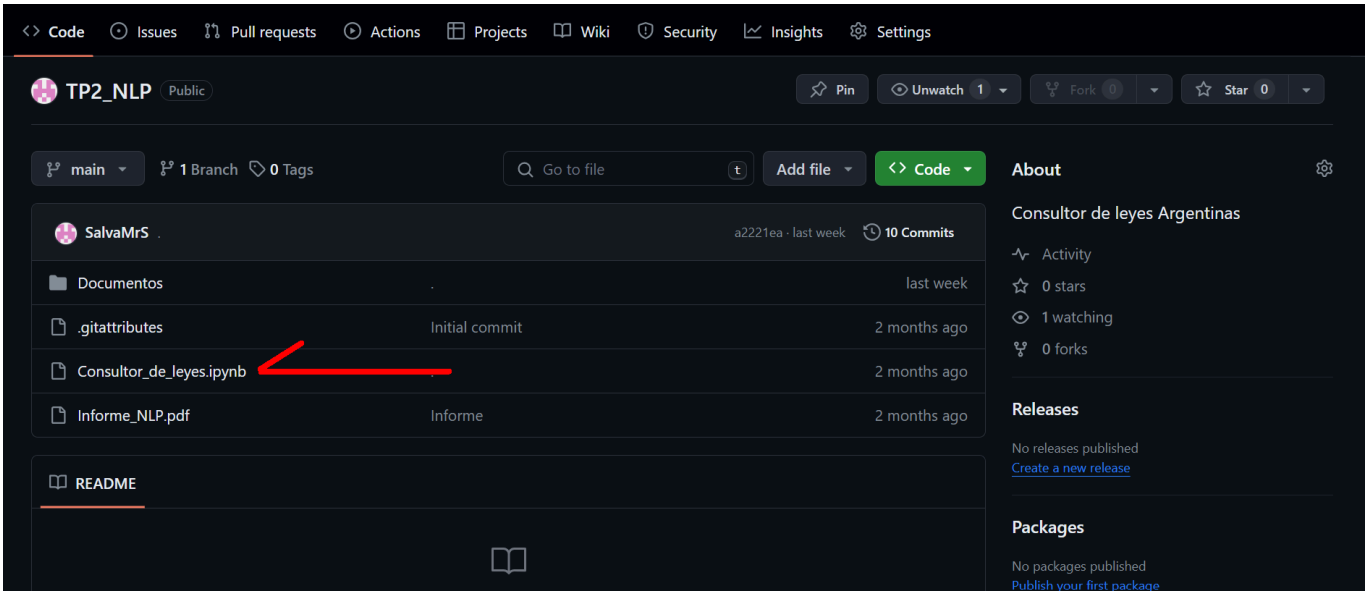
Para facilitar la interacción con el usuario, es necesario que el modelo de lenguaje comprenda claramente el papel que debe desempeñar. Esto implica proporcionarle una descripción detallada de su función, así como la información relevante relacionada con la consulta en cuestión. Además, se le suministra al modelo la pregunta formulada por el usuario y un punto de partida para la respuesta que debe generar.

Chatbot

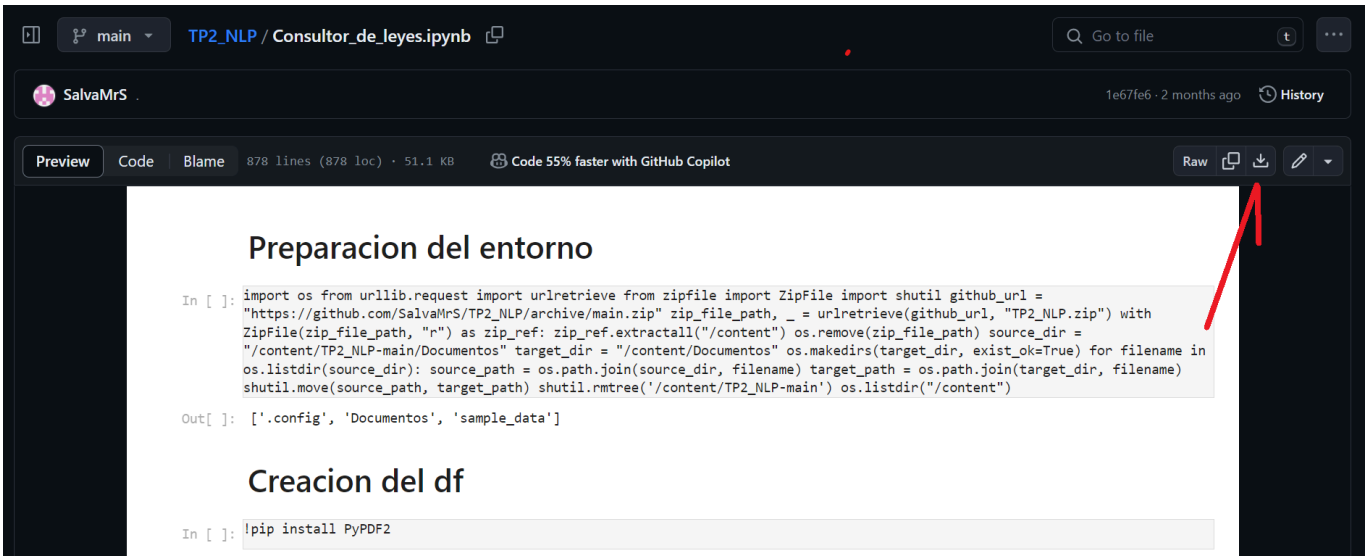
En primer lugar, el chatbot recibe la consulta del usuario y analiza en qué fuente de información se encuentra la información requerida por el modelo de lenguaje. Dependiendo de la fuente de datos, se busca la información y se ingresa al LLM, que devolverá una respuesta adecuada que se mostrará al usuario. En caso de que la respuesta no sea satisfactoria para el usuario, se le preguntará dónde buscaría la información para adaptar la fuente de información.

Ejecución del Programa

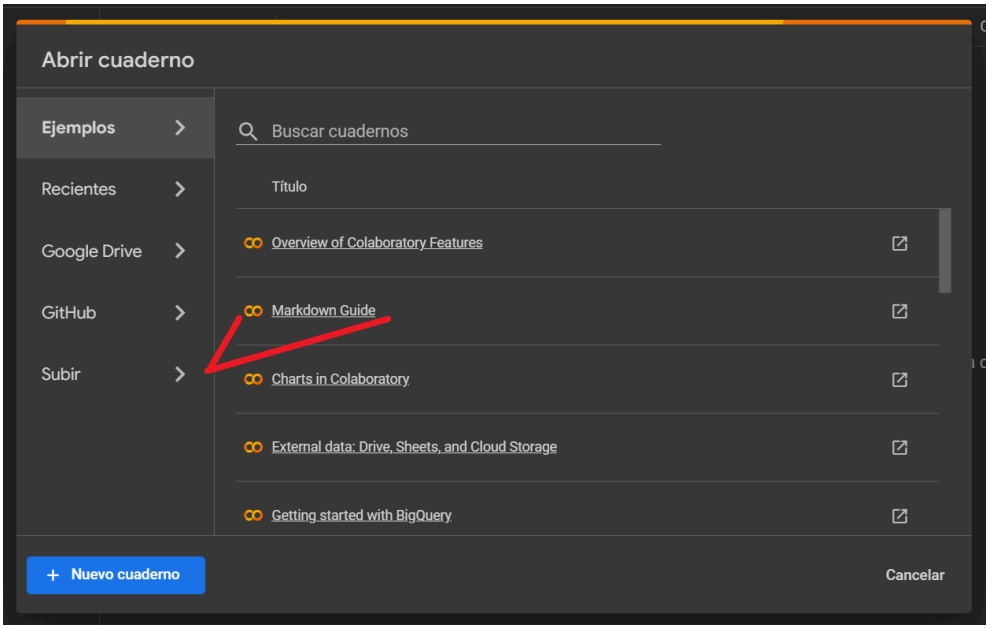
Ingresa a https://github.com/SalvaMrS/TP2_NLP/ y selecciona el archivo “Consultor_de_leyes.ipynb”.



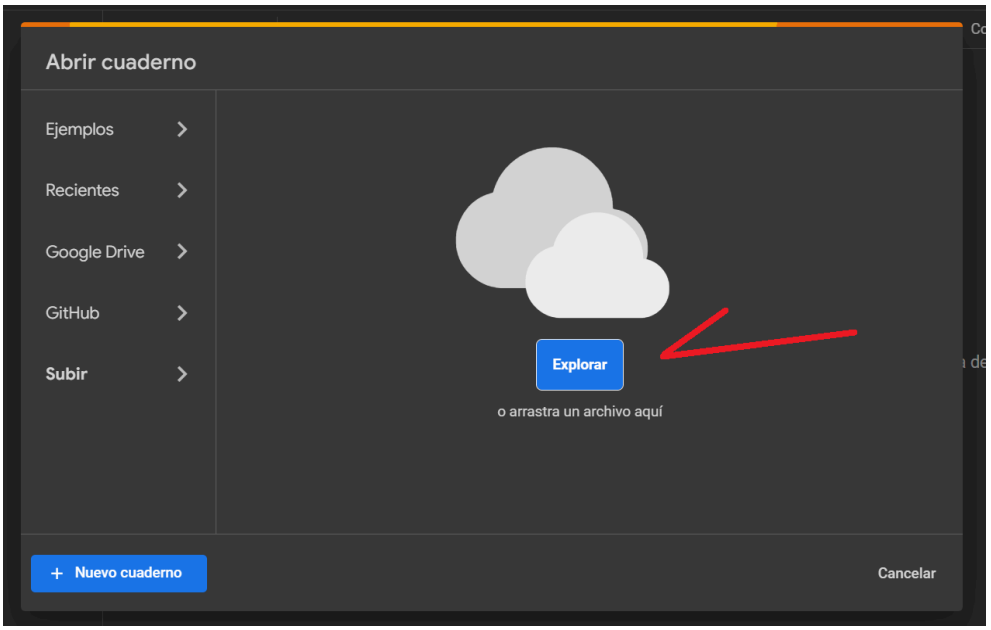
Selecciona el botón de descarga y espera a que se complete.



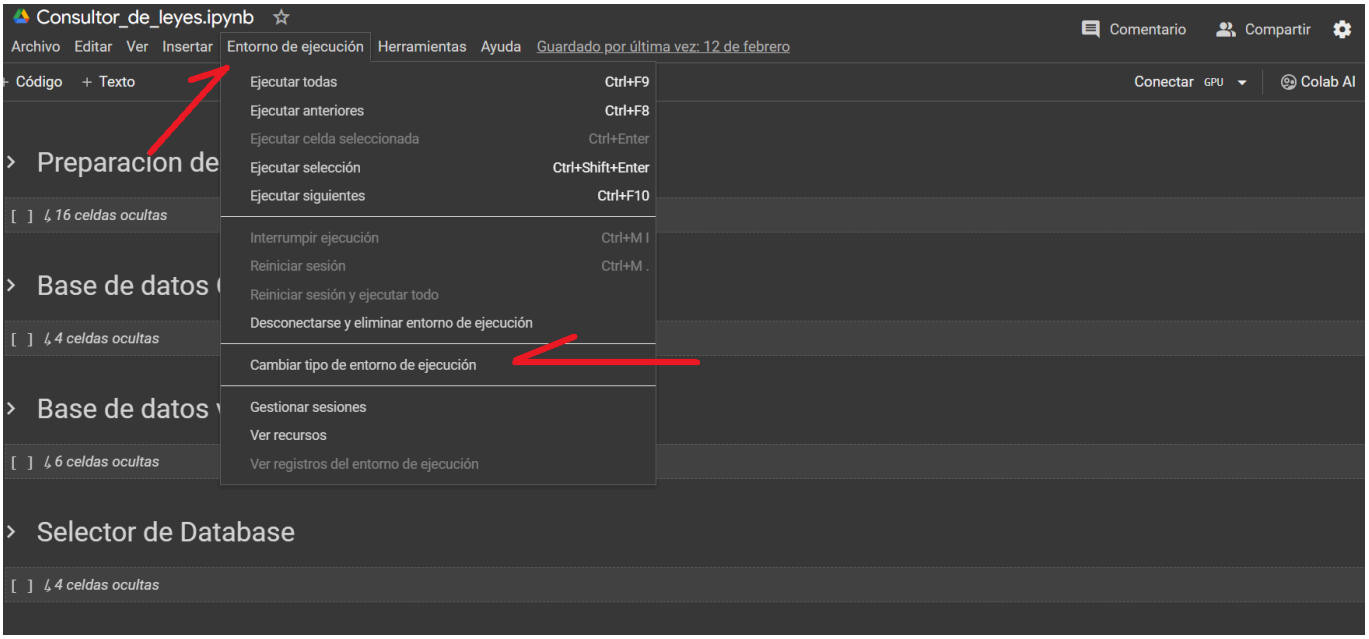
Una vez tengas el archivo ingresa a https://colab.research.google.com/?hl=es y selecciona subir.



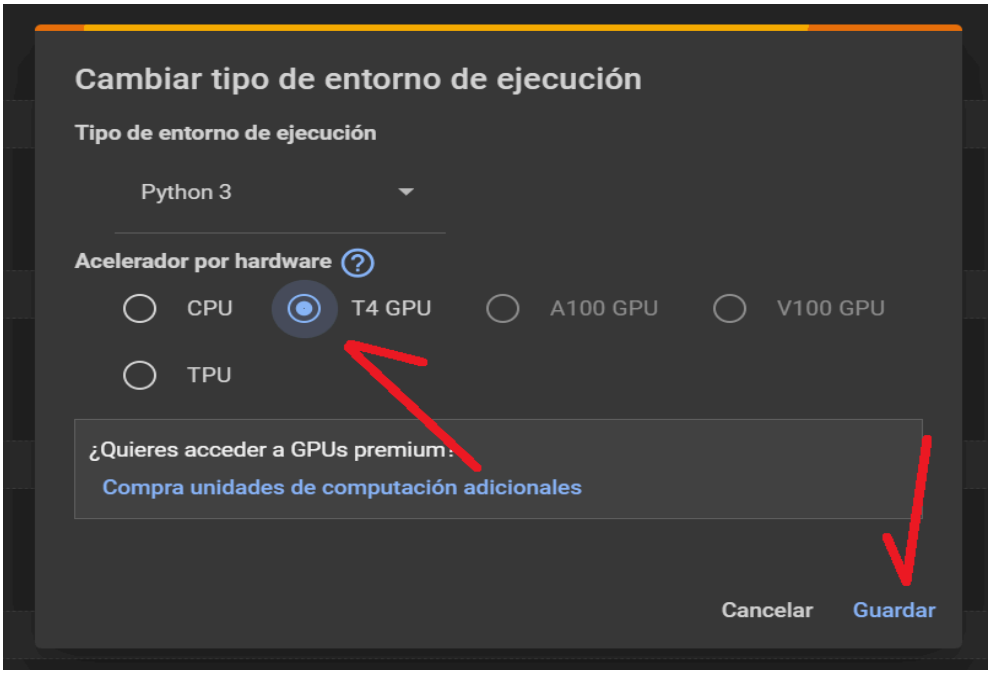
Apreta el botón explorar y selecciona el archivo que descargamos en la carpeta de descargas.



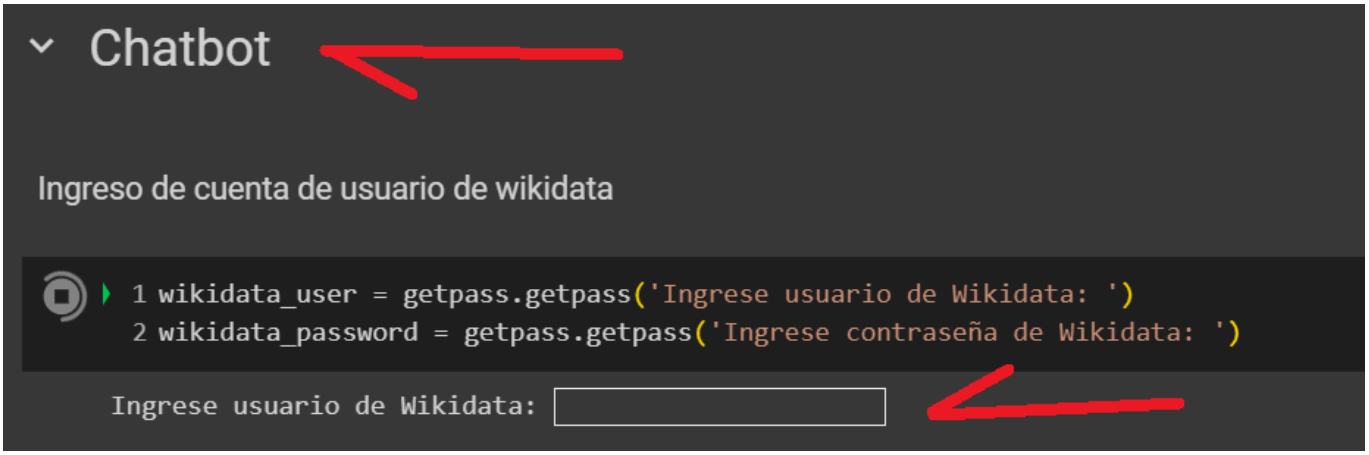
Una vez se haya cargado el archivo vamos a la casilla “Entorno de ejecución” y seleccionamos “cambiar tipo de entorno de ejecución”.



Seleccionamos la opción “T4 GPU” y guardar.



A continuación apretamos las teclas “Ctrl” y “F9” al mismo tiempo y nos dirigimos a la sección de “Chatbot” al final del archivo en donde debemos ingresar nuestro usuario y contraseña de “Wikidata”, si no tienes una cuenta puedes crearla en el siguiente link https://www.wikidata.org/w/index.php?title=Special:CreateAccount&returnto=Wikidata:Main+Page&campaign=loginCTA.



Una vez realizados estos pasos podemos ingresar las consultas que queramos, en el caso de salir del chatbot podemos volver a ingresarlo ejecutando la última celda.