

UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA
DEPARTAMENTO INGENIERIA COMERCIAL
SANTIAGO - CHILE



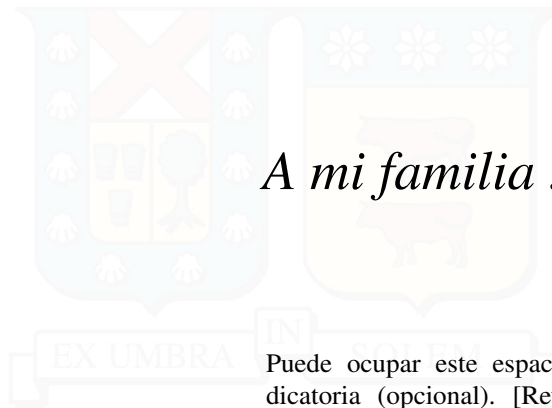
**TÍTULO DE MEMORIA (EL TÍTULO SÓLO PUEDE TENER UN MÁXIMO DE 3
LÍNEAS)**

SANTIAGO JESÚS VASCONCELLO ACUÑA

MEMORIA PARA OPTAR AL TÍTULO DE
INGENIERO COMERCIAL

PROFESOR GUÍA : SR. PABLO ISLA
PROFESOR CORREFERENTE : SR. THIERRY DE SAINT PIERRE.

Diciembre 2023



A mi familia . . .

Puede ocupar este espacio para escribir una dedicatoria (opcional). [Revise el archivo maestro `memoria.tex` para modificar / eliminar esta sección.]

(AGRADECIMIENTOS) [Título es opcional]

Agradezco a quienes contribuyeron para ir mejorando esta plantilla hecha en L^AT_EX.

Los aportes y comentarios de distintas personas en el [Departamento de Industrias](#) fueron muy útiles para que este documento puede ser ocupado para mejorar la presentación de tesis y memorias del Departamento (y la universidad).

Para el impaciente ...

Por favor ocupar `git` :

```
git clone https://github.com/jaimercz/utfsm-thesis.git
```

Para los interesados en `git` revisar [1].

Abra el archivo de configuración `config.tex` para cambiar título, autor, fecha, etc. de la portada y del documento en general.

Abra y compile el documento maestro `memoria.tex` .

```
$ pdflatex memoria.tex
$ biber memoria
$ pdflatex memoria.tex
$ pdflatex memoria.tex
```

Esta version ocupa `biber` en lugar de `natbib` / `bibtex` :

Si hay errores, verifique primero que todos los paquetes L^AT_EX han sido instalados.

Si desea omitir alguna sección (dedicatoria, agradecimientos, etc.), revise el documento maestro `memoria.tex` y agregue o comente (o elimine) las líneas correspondientes.

Por ejemplo, para eliminar esta sección, borre las líneas:

```
_____ memoria.tex (extracto) _____
\section*{Agradecimientos}
\insertFile[plain]{agradecimientos}}
```

RESUMEN EJECUTIVO

Plantilla \LaTeX para las Memorias y Tesis del Departamento de Ingenieria Comercial, UTFSM.

Se incluyen también algunos ejemplos de cómo incorporar tablas y gráficos en distintas presentaciones respetando las Normas de Biblioteca para Memorias y Tesis de la UTFSM.

Palabras Clave. \LaTeX , Plantilla para Memoria, Departamento de Ingenieria Comercial, UTFSM.

¡Importante! [LEAME]

Impresión por un solo lado.

A partir del año 2016, el Departamento de Ingenieria Comercial sólo requiere la entrega digital de los archivos de memorias y tesis. Por este motivo, este documento está preparado para ser impreso por un solo lado de una hoja (“*oneside*”), y facilitar así su lectura en pantallas. Esta configuración es parte de archivo de clase `thesis_utfsm.cls`.

Codificación de caracteres.

Todos los archivos `*.tex` de esta plantilla han sido preparados ocupando la codificación de caracteres por defecto *unicode* (UTF-8). Windows (y algunas versiones de OSX) ocupan otro tipo de codificación (ej. *Windows-1252* o *Mac Roman*).

Si deseas ocupar esta plantilla y encuentras problemas con los caracteres acentuados, entonces puedes optar por una de estas tres alternativas:

- i) cambiar tu editor (TexMaker, TexStudio, TexShop, etc.) para que ocupe UTF-8 como codificación de caracteres por defecto; o
- ii) cambiar la codificación de cada documento `*.tex` para que ocupe la codificación nativa de tu sistema operativo; y, modificar el archivo `config.tex` la línea que dice:

OSX, Linux: `\usepackage[utf8x]{inputenc}`

Windows: `\usepackage[latin1]{inputenc}`

Overleaf: `\usepackage[utf8]{inputenc}` <https://overleaf.com>

- iii) escribir todo ocupando caracteres pre-acentuados (ej. `\'a` en lugar de á).

Recuerde:

Mezclar documentos de distintas codificaciones puede generar muchos problemas al momento de compilar.

ABSTRACT

This is a \LaTeX thesis template for the Departamento de Ingenieria Comercial, UTFSM. A few examples about the inclusion of figures and tables are also provided.

(The abstract can be edited by opening the file `includes/abstract.tex` .)

Keywords. \LaTeX , Thesis Template, Departamento de Ingenieria Comercial, UTFSM

Instrucciones para la Plantilla.

Editar el archivo `/includes/abstract.tex` para modificar los contenidos de esta sección.

Si no desea incluir un abstract, editar el archivo `/memoria.tex` , y comentar o borrar la sección que se muestra a continuación.

```
_____ /memoria.tex (extracto) _____  
\section*{ABSTRACT}  
\insertFile[plain]{abstract} % Archivo abstract.tex
```

Índice de Contenidos

1. Introducción	1
1.1. Obejetivos	1
1.1.1. Objetivo General	1
1.1.2. Objetivo Específico	1
1.2. Metodologia	2
1.2.1. Creación del Proyecto	2
1.2.2. Ejemplo de Uso del Proyecto	2
1.2.3. Evaluación de Riesgos	2
2. Creación del Proyecto	3
2.1. ETL	3
2.1.1. Extract	4
2.1.2. Transform	5
2.1.3. Load	6
2.2. Chatbot	7
3. Ejemplo de Uso del Proyecto	8
4. Evaluación de Riesgos	9
4.1. General	9
4.2. ETL	9
4.2.1. Extract	9
4.2.2. Transform	9
4.2.3. Load	9
4.3. Chatbot	9
4.3.1. Entrega de contexto adecuado	9
5. Conclusiones	10

Índice de Tablas



Índice de Figuras

2.1. Estructura basica de la aplicación	3
2.2. Estructura del proceso de ETL para el Buscador Ambiental	4
2.3. Screenshot del Buscador de la pagina del Primer Tribunal Ambiental	4
2.4. Screenshot de una sola reclamación en la pagina web del buscador ambiental	5

1 | Introducción

La inteligencia artificial, también conocida como IA, ha experimentado un notable auge en la industria en los últimos tiempos de la mano de la llamada Industria 4.0 [intro1], especialmente en el ámbito en áreas algo reacias como la administración y finanzas[intro2]. Este incremento no se debe necesariamente a un aumento en la capacidad de cómputo, ya que esta ha ido creciendo gradualmente a lo largo del tiempo (buscar respaldo). Anteriormente, aunque importante, no generaba tanto interés como en la actualidad. No fue sino hasta que OpenAI lanzó ChatGPT el 30 de noviembre de 2022 que el público en general pudo experimentar, probar y comprender de manera más completa la gran revolución llamada inteligencia artificial generativa [intro3].

De acuerdo con Google, "La inteligencia artificial generativa se refiere al uso de la IA para crear contenido, como texto, imágenes, música, audio y videos"[google1]. Gracias a su interfaz amigable, resultó sencillo para personas de diversas industrias descubrir que existía una herramienta capaz de generar texto y responder preguntas de manera comprensible, incluso para aquellos que no eran expertos en tecnología.

La génesis de esta tesis se basa en la experiencia de llevar a cabo un proyecto utilizando estas tecnologías y los riesgos asociados a ellas. En este contexto, entendemos el riesgo como cualquier aspecto que pueda afectar tanto al equipo involucrado en la creación del proyecto como a los resultados obtenidos. El proyecto se centró en el uso de un modelo de lenguaje de gran envergadura, conocido como LLM por sus siglas en inglés, limitándose a la generación de texto. Por lo tanto, no profundizaremos en otros tipos de inteligencia artificial generativa, como la generación de imágenes o audio. El enfoque principal de este trabajo se concentra solo en el área del procesamiento del lenguaje natural aplicados a LLM.

1.1. Objetivos

1.1.1. Objetivo General

Determinar los factores de riesgo que pueden llegar a influir, tanto de la creación como del uso de aplicaciones que utilicen modelos grandes de lenguaje (LLM) aplicados a la industria, usando de base el proyecto de búsqueda de jurisprudencia de los tribunales ambientales.

1.1.2. Objetivo Específico

1. Determinar una posible estructura de una aplicación usando LLM
2. Desarrollar los estados del Arte del uso de LLM y de los modelos generativo en si
3. Desarrollar los problemas que conlleva el uso de información para alimentar dichos problemas
4. Analizar un proceso de ETL de principio a fin para observar sus posibles riesgos

1.2. Metodología

La metodología empleada en esta tesis se estructura en torno a tres componentes esenciales: la creación del proyecto, un ejemplo de uso concreto y la evaluación de los riesgos asociados a cada etapa del proceso, tanto en la fase de desarrollo del proyecto como en su aplicación práctica.

1.2.1. Creación del Proyecto

Esta fase inicial comprende el desarrollo del proyecto basado en IA generativa. Incluye los siguientes pasos:

- **Definición de Objetivos y Alcance:** Establecimiento claro de los propósitos y límites del proyecto, identificando las metas a alcanzar.
- **Selección de Tecnologías y Herramientas:** Evaluación y elección de las tecnologías y herramientas apropiadas para la implementación del proyecto.
- **Diseño de la Arquitectura:** Desarrollo de la estructura y componentes del proyecto, considerando aspectos de escalabilidad y rendimiento.
- **Implementación y Desarrollo:** Construcción efectiva del proyecto, incluyendo la programación y configuración de la inteligencia artificial generativa.

1.2.2. Ejemplo de Uso del Proyecto

Esta fase implica la aplicación práctica del proyecto en un contexto específico, demostrando su funcionalidad y utilidad. En este caso nuestro interés mas que en el output que genere la aplicación, es como funciona internamente el proceso cosa que para la siguiente etapa sea más fácil

1.2.3. Evaluación de Riesgos

Se trabajará la identificación y análisis de los riesgos potenciales en cada etapa del proceso, así como los riesgos derivados del caso de uso. Incluye:

- **Riesgos en la Creación del Proyecto:** Identificación de posibles obstáculos y contratiempos durante la etapa de concepción y desarrollo.
- **Riesgos en el Ejemplo de Uso:** Consideración de los riesgos asociados a la implementación práctica del proyecto en el contexto definido.

Esta metodología proporciona un enfoque integral para la creación y aplicación de un proyecto utilizando LLM, permitiendo una evaluación de los riesgos en cada etapa del proceso y en el caso de uso específico. Esto facilita la toma de decisiones informadas y la formulación de estrategias para mitigar posibles contratiempos.

2 | Creación del Proyecto

Los Tribunales Ambientales son órganos jurisdiccionales especiales, sujetos a la superintendencia directiva, correccional y económica de la Corte Suprema, cuya función es resolver las controversias medioambientales de su competencia y ocuparse de los demás asuntos que la ley somete a su conocimiento [3]. Estos tribunales generan una cantidad de jurisprudencia que puede ser encontrada en su portal de consulta llamado buscador ambiental [2].

El proyecto consiste en la generación de un chatbot en donde se pueda preguntar sobre la jurisprudencia de estos tribunales, aunque por razones de capacidad el chatbot se vea acotado solamente a las reclamaciones recibidas por el tribunal.

Por lo que, este proyecto consiste en un proceso de extracción de datos desde el buscador ambiental, transformación de estos datos para su utilización, generación de vectores de estos datos para que puedan interactuar con la aplicación, carga de estos en una base de datos, para que después la aplicación pueda interactuar con ellos y mandando esa información a el LLM, siendo en este caso gpt-4 perteneciente a OpenAI.

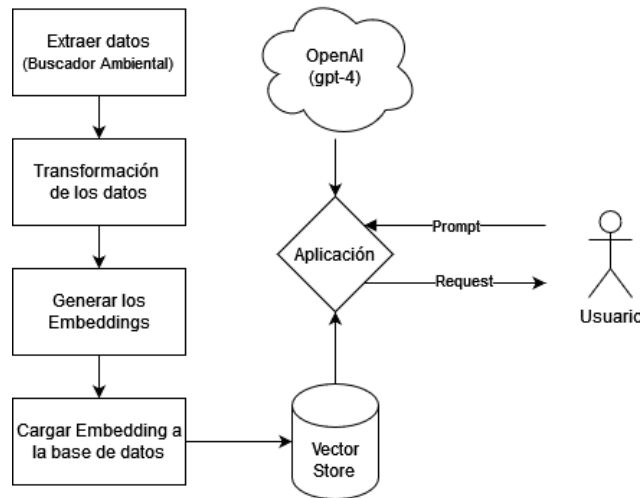


Figura 2.1: Estructura básica de la aplicación

(Fuente: Elaboración propia)

A partir de esta estructura mientras se avance en el desarrollo, se explicará parte por parte el proceso y con ello los riesgos de cada uno de ellos.

2.1. ETL

Para realizar el proyecto fue necesario realizar un proceso de ETL. El término ETL se refiere a las técnicas de "Extracción, Transformación y Carga" (Extract, Transform, Load), que constituyen un proceso clave para los datos necesarios para el proyecto. Este proceso implica la extracción de datos de fuentes heterogéneas, su transformación para

ajustarse a las necesidades del negocio y su posterior carga en un destino que, por lo general, es un almacén de datos diseñado para el análisis y la generación de informes[ETL1].

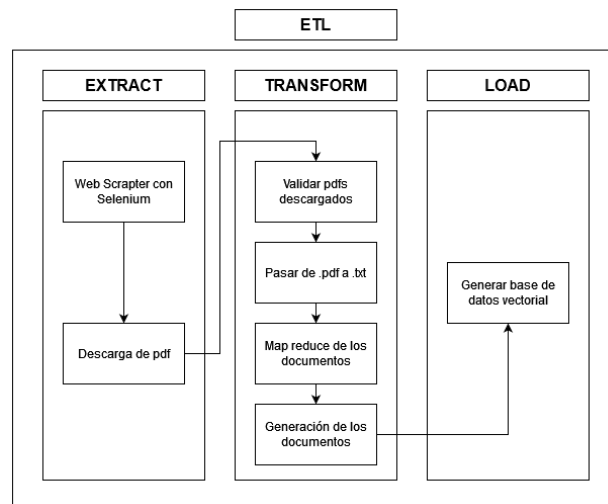


Figura 2.2: Estructura del proceso de ETL para el Buscador Ambiental

(Fuente: Elaboración propia)

La fase de extracción implica la recolección de datos de múltiples fuentes, que pueden variar desde bases de datos estructuradas hasta información no estructurada en la web. La transformación se refiere al proceso de limpieza, conversión, y consolidación de estos datos en un formato adecuado para el análisis. Finalmente, la carga es el proceso de transferir los datos transformados al sistema de destino, donde se pueden almacenar y utilizar para la toma de decisiones estratégicas [ETL1].

2.1.1. Extract

La información requerida para el desarrollo del Chatbot se obtuvo del "Buscador ambiental" del Tribunal de Protección Ambiental de Chile a través de su sitio web[2]. Este portal aloja todos los documentos públicos disponibles para su consulta en cualquiera de los tres tribunales ambientales. Para acceder a la base de datos necesaria, se llevó a cabo la creación de un bot capaz de recopilar las entradas de este buscador de manera análoga a un usuario convencional.

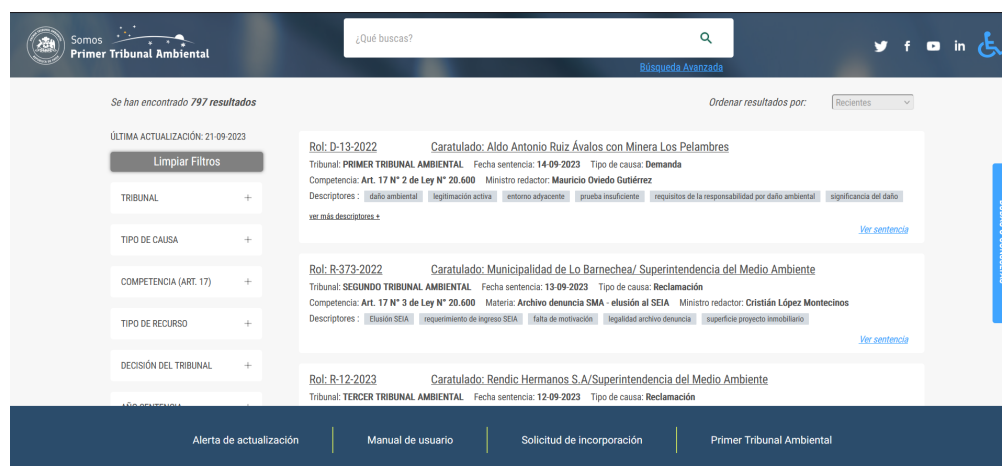


Figura 2.3: Buscador de la página del Primer Tribunal Ambiental

(Fuente: Página del Primer Tribunal Ambiental)

Para esta tarea, se empleó Selenium, una herramienta originalmente diseñada para generar pruebas, pero que, debido a la naturaleza reactiva y dinámica de los sitios web, así como a la detección de bots por parte de algunas páginas, resultó ser la elección más apropiada. Este bot, después de explorar todas las páginas del buscador ambiental, como se ilustra en la Figura 2.3, logró recuperar cada uno de los enlaces individuales que conducen a las páginas específicas de cada caso, tal como se muestra en la Figura 2.4.

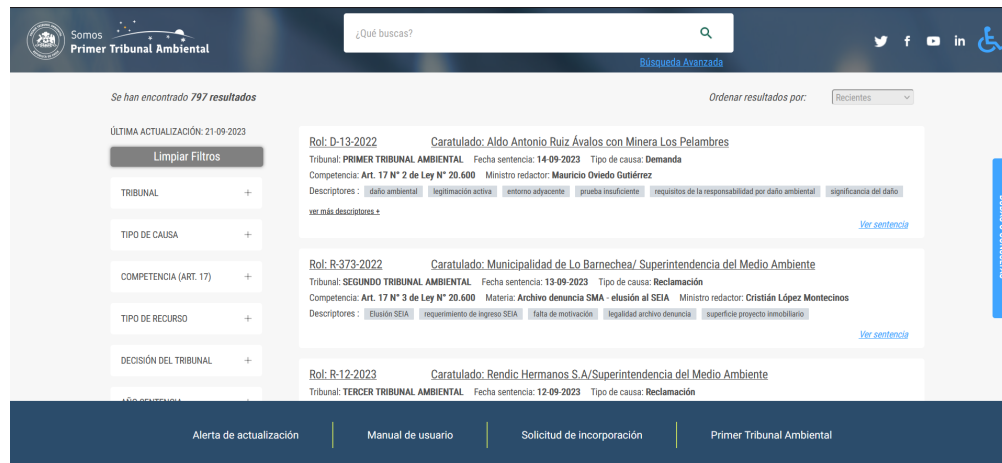


Figura 2.4: Screenshot de una sola reclamación en la página web del buscador ambiental
(Fuente: Página del Primer Tribunal Ambiental)

Posteriormente, se contemplaba la posibilidad de obtener tanto los enlaces a cada documento en formato PDF como la información detallada de cada uno de estos documentos mediante la creación de un nuevo bot. Sin embargo, durante el proceso de desarrollo de este bot, se logró acceder a la API que permitía obtener directamente todos los datos mencionados anteriormente. Esto suprimió la necesidad de crear otro tipo de bot utilizando Selenium, ya que bastaba con realizar una solicitud a la mencionada API.

Para completar la fase de extracción de datos (Extract), una vez que se había obtenido toda la información mediante las solicitudes a la API, el último paso consistió en generar nuevas solicitudes con el objetivo de descargar todos los archivos PDF de cada una de las entradas. Estos archivos ahora están descargados y listos para la próxima etapa del proyecto, que implica la transformación de los datos con el fin de obtener la información necesaria para construir la base de datos a partir de los documentos.

2.1.2. Transform

En la continuación del proceso de ETL (Extracción, Transformación y Carga), los PDFs que previamente han sido descargados requieren ser sometidos a modificaciones con el objetivo de convertir la información que inicialmente se presenta en un estado "sucio," en datos "limpios" que puedan ser adecuadamente utilizados en el proyecto. Este proceso se denomina "transformación," o "transform," en inglés.

Entre los datos descargados, nos encontramos con un extenso número de PDFs que presentan dificultades significativas para su manipulación. Esto se debe a que el Tribunal Ambiental no sigue un formato estándar en la estructura de las reclamaciones presentadas. En consecuencia, cada uno de los textos posee un formato propio, lo que complica en gran medida la extracción eficiente de las diversas secciones contenidas en dichos textos. Sin embargo, gracias al funcionamiento del proceso de semejanza semántica, esta diversidad de formatos no representa un problema insuperable para el proyecto.

No obstante, surgen dificultades adicionales cuando se trata de las reclamaciones que son presentadas a los tribunales ambientales en formato digital o, en su defecto, en forma de fotocopias. Esto implica que no todos los documentos están habilitados para su procesamiento. En consecuencia, el primer paso en el proceso de transformación involucra la discriminación de qué PDFs son susceptibles de ser procesados y cuáles no. Para llevar a cabo esta tarea, se

ha desarrollado un script capaz de detectar texto dentro de un archivo PDF. Si el texto es legible, se almacena; de lo contrario, se elimina.

Una vez separados los PDFs legibles y adecuados para el trabajo posterior, se procede con la transformación de estos documentos al formato TXT (texto plano). Esta etapa se lleva a cabo considerando la conveniencia de trabajar con archivos en formato de texto en comparación con los archivos en formato PDF puro, dado que el próximo método de transformación, que implica el uso de map-reduce en Langchain, requiere que los datos estén en formato de texto.

Dentro del contexto del uso de Langchain, el proceso de map-reduce implica, en primer lugar, la aplicación de una cadena de LLM (Modelo de Lenguaje de Gran Tamaño) a cada documento individual docs[i], considerándolos de manera independiente (la fase de "Map"). Esto implica tratar la salida de la cadena como un nuevo documento, resumido. Posteriormente, todos los nuevos documentos resumidos se envían a una cadena de combinación de documentos para obtener una única salida por archivo (la fase de Reduce"). En este contexto, la implementación de map-reduce en los archivos de texto (TXT) conlleva la aplicación de la cadena de LLM a cada documento de manera individual, generando así una nueva representación del mismo.

Sin embargo, es importante destacar que un archivo .txt puede contener un número de tokens demasiado elevado como para ser reducido de manera inmediata. En situaciones de este tipo, es necesario recurrir a un proceso de subdivisión que fragmente los textos en segmentos con un número de tokens inferior al límite impuesto por la API de OpenAI. Cada archivo .txt puede ser dividido, resumido y exportado a un nuevo archivo .txt una vez que ha sido fragmentado previamente en segmentos.

Los documentos procesados son combinados utilizando otra cadena de procesamiento para obtener un resultado final consolidado. Para concluir el proceso de transformación, los resúmenes generados después de haber pasado por el procedimiento de map-reduce se someten a un último paso antes de ser incorporados en la base de datos. Este paso implica la fusión de los resúmenes con la información obtenida a través de las solicitudes a la API del Tribunal Ambiental, presentada en formato de texto. Este proceso resulta en la creación de un único documento que engloba toda la información, al cual nos referiremos como "documentos finales". Con esto, se concluye la fase de transformación y se procede al último procedimiento, conocido como carga (Load), que consiste en cargar estos documentos finales en la base de datos.

2.1.3. Load

Al culminar el proceso de Extracción, Transformación y Carga (ETL), resulta fundamental llevar a cabo la fase de carga, también conocida como "load.^{en} inglés, en la cual se incorporan todos los descriptores previamente descargados y transformados en una base de datos. Para este proyecto, en el cual se utiliza LangChain, resulta de vital importancia fragmentar los documentos en secciones más pequeñas.

Esta necesidad surge debido a que los documentos deben ser sometidos a un proceso de incrustación (embedding) antes de ser introducidos en la base de datos. Esto se debe principalmente a que las funciones de incrustación tienen un límite en la extensión de grupos de caracteres, conocidos como "tokens," que pueden ser procesados. En el contexto del modelo de incrustación "text-embedding-ada-002," este límite se establece en 8191 tokens [1], lo que constituye la longitud máxima de los fragmentos.

Por lo tanto, cuando se trabaja con documentos extensos, es imperativo dividirlos en fragmentos más pequeños antes de proceder con su incorporación. Según la información proporcionada en el Blog de OpenAI, los embeddings son representaciones numéricas de conceptos convertidos en secuencias numéricas, lo que facilita que las computadoras comprendan las relaciones entre los conceptos.^{En} términos sencillos, los embeddings son representaciones vectoriales de texto que permiten su comprensión por parte del Modelo de Lenguaje de Gran Tamaño (LLM). Dado que los LLM son redes neuronales, el proceso de incrustación resulta esencial para traducir el texto en números, que es el formato comprensible para esta red neuronal basada en Transformers.

Los embeddings resultantes se almacenan posteriormente en una base de datos vectorial denominada ChromaDB. Esta base de datos ha sido diseñada para ser compacta, escalable y eficiente, con el propósito de almacenar y recuperar vectores de manera efectiva. ChromaDB genera índices que permiten una recuperación rápida y eficiente de los embeddings en función de las consultas realizadas por los usuarios.

2.2. Chatbot



3 | Ejemplo de Uso del Proyecto



4 | Evaluación de Riesgos

4.1. General

4.2. ETL

4.2.1. Extract

4.2.2. Transform

4.2.3. Load

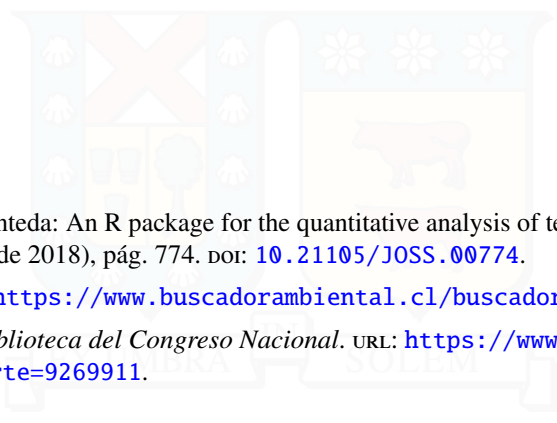
4.3. Chatbot

4.3.1. Entrega de contexto adecuado

5 | Conclusiones



Bibliografía

- 
- [1] Kenneth Benoit y col. “quanteda: An R package for the quantitative analysis of textual data”. En: *Journal of Open Source Software* 3 (30 oct. de 2018), pág. 774. doi: [10.21105/JOSS.00774](https://doi.org/10.21105/JOSS.00774).
- [2] *Buscador Ambiental*. URL: <https://www.buscadorambiental.cl/buscador/#/>.
- [3] *Ley Chile - Ley 20600 - Biblioteca del Congreso Nacional*. URL: <https://www.bcn.cl/leychile/navegar?idNorma=1041361&idParte=9269911>.